7-2001

Interval Estimation of a Normal Process Mean from Rounded Data

Chiang-Sheng Lee
Iowa State University

Stephen B. Vardeman
Iowa State University, vardeman@iastate.edu

Interval Estimation of a Normal Process Mean from Rounded Data

Abstract
Standard statistical methods are based on an implicit assumption that numerical data are exact. But in truth, all real data are rounded to some smallest unit of measure related to the precision of the device used to produce them. When the degree of rounding is severe, ignoring the rounding produces statistical methods with operating characteristics far from nominal. We discuss the interval estimation of the parameter μ when rounded data come from the N(μ,σ2) distribution.

Keywords
Coverage Probability, Gaging, Likelihood

Disciplines
Statistics and Probability

Comments
This article is published as Lee, Chiang-Sheng, and Stephen B. Vardeman. "Interval estimation of a normal process mean from rounded data." *Journal of Quality Technology* 33, no. 3 (2001): 335-348. DOI: 10.1080/00224065.2001.11980083. Posted with permission.

Interval Estimation of a Normal Process Mean from Rounded Data

CHIANG-SHENG LEE and STEPHEN B. VARDEMAN

Iowa State University, Ames, IA 50011-1210

Standard statistical methods are based on an implicit assumption that numerical data are exact. But in truth, all real data are rounded to some smallest unit of measure related to the precision of the device used to produce them. When the degree of rounding is severe, ignoring the rounding produces statistical methods with operating characteristics far from nominal. We discuss the interval estimation of the parameter μ when rounded data come from the $N(\mu,\sigma^2)$ distribution.

Introduction

IT is an important practical problem that the collection of measurement data is sometimes done using relatively crude gaging. This is especially obvious and common where digital gages are used to measure precisely machined part dimensions and very few distinct values are typically recorded in a sample. Elementary methods of point estimation of distribution parameters and the construction of confidence intervals are based on an implicit assumption that observed data are essentially "exact." It is of interest to know what happens to the statistical properties of these methods when, in fact, the available data are produced by relatively crude gaging. Do nominal (or exact data) statistical properties carry over to the case of crudely gaged data? And if they do not, are there reasonable replacements for these standard methods?

Our main purpose in this paper is to investigate the properties of interval estimators of the parameter μ based on rounded normal data. Two methods will be compared. One is the traditional t interval (appropriate for exact normal data), and the other is obtained from inversion of (rounded data) likelihood ratio tests for μ. Our end goal is to find reliable confidence intervals for μ. We first discuss the *likelihood function* for rounded (interval-censored) normal data. The construction of the rounded data confidence intervals for μ is provided next. We then compare the t and likelihood intervals in terms of coverage probability and average length for various sample sizes. Next we provide a discussion of computational issues related to implementation of our intervals. Finally, we provide some summary comments.

The Model for Rounded Normal Data

Without loss of generality, it is convenient to assume that all observations available for data analysis take on integer values. (Measurements can be expressed in an integer number of smallest possible increments above a nominal value.) Table 1, for example, shows 5 (real) samples of size $n = 3$ of journal diameter measurements taken in the routine process monitoring of a grinder in an engine remanufacturing plant. The units are 0.0001 inch over the nominal diameter and measurements of this type were the best available.

As a second example, Stein (2000) considers a sample of $n = 10$ readings from a digital gage, 4 of which are 1.2 and 6 of which are 1.3, and says that "by taking the mean of 1.26 you can add another digit of resolution to your process." In units of .1 above 1.0, Stein's sample consists of the $n = 10$ crudely gaged values

$$3, 2, 3, 3, 3, 2, 2, 3, 2, 3.$$

In rough terms, his verbiage can be taken to imply that he believes some true value is close to 2.6 with fair certainty. In this article we investigate what, in fact, can really be learned about a mean from data

Mr. Lee is a Ph. D. Candidate in Industrial Engineering in the Industrial and Manufacturing Systems Engineering Department. His email address is chiang@iastate.edu.

Dr. Vardeman is a Professor in the Statistics and Industrial and Manufacturing Systems Engineering Departments. He is a Senior Member of ASQ.

TABLE 1. 5 Samples of $n = 3$ Journal Diameter Measurements

Sample	Rounded Diameters		
1	−1	−1	−1
2	0	−1	0
3	0	0	0
4	0	−1	0
5	0	0	0

like those in one of the samples in Table 1 or from Stein's data.

One plausible model for a single sample of such data is that unrounded observations (reflecting measurement variation, and where appropriate, part to part variation) are a random sample from a normal distribution with mean μ and standard deviation σ, and because of crude gaging, what is observed are the corresponding (integer) rounded values. Under this model, the probability that n observations $X_1, X_2, \ldots, X_n$ take the integer values $x_1, x_2, \ldots, x_n$ is

$$\begin{aligned} f(X;\mu,\sigma) &= \Pr(X_1 = x_1,\ X_2 = x_2, \ldots,\ X_n = x_n) \\ &= \prod_{i=1}^{n} \left\{ \Phi\left(\frac{x_i + 0.5 - \mu}{\sigma}\right) \right. \\ &\qquad \left. - \Phi\left(\frac{x_i - 0.5 - \mu}{\sigma}\right) \right\} \qquad (1) \\ &= \prod_{i} \left\{ \Phi\left(\frac{i + 0.5 - \mu}{\sigma}\right) \right. \\ &\qquad \left. - \Phi\left(\frac{i - 0.5 - \mu}{\sigma}\right) \right\}^{n_i}, \qquad (2) \end{aligned}$$

where $\Phi(x)$ is the standard normal cumulative probability function, the product in Equation (2) is over integer values i, and n_i is the number of observed values which equal the integer i. (The appropriateness of these model assumptions can be investigated on the basis of a large sample using a standard χ^2 goodness of fit test with bins (cells) of width 1.0 centered at the integers.)

Treating the data in hand as fixed and plugged into Equation (2), this function of μ and σ can be termed the *likelihood function.* It will be convenient to work with the natural logarithm of the likelihood and thus define the *log likelihood function* by

$$L(\mu,\sigma) = \sum_i n_i \ell n \left\{ \Phi\left(\frac{i + 0.5 - \mu}{\sigma}\right) - \Phi\left(\frac{i - 0.5 - \mu}{\sigma}\right) \right\}. \qquad (3)$$

Finally, define

$$L^*(\mu) = \sup_{\sigma > 0} L(\mu,\sigma). \qquad (4)$$

Then, $L^*(\mu) \leq 0$ is often called the *profile loglikelihood function* for μ, and for fixed μ can be explained as the "maximum" or supremum value of $L(\mu,\sigma)$ over $\sigma > 0$. "Maximum likelihood" point estimation of (μ,σ) requires maximization of $L(\mu,\sigma)$. We will use $L^*(\mu)$ to define interval estimators of μ.

The notion of using a "rounded data likelihood" like Equation (1) is far from novel. The papers of Shapiro and Gulati (1998), Schader and Schmid (1984), and Swan (1969), for example, treat estimation based on such a likelihood. However, the published work on the subject concentrates on point estimation (and algorithms for the same) and the large n asymptotics for the problem. Here our focus is confidence interval estimation for samples of practical size.

We remark too that the general issue of rounding/grouping of continuous data has been one of continuing practical interest. Some recent references are Tricker, Coates, and Okell (1998) and Vardeman and Jensen (1989). Early discussions and bibliographies can be found in Haitovsky (1982), Heitjan (1989), and Sheppard (1898).

Construction of the Intervals for μ

Two methods of making confidence intervals for μ are discussed in this section. First, if we ignore rounding and treat the rounded data as "exact" normal data, then the usual nominal $(1-\alpha)$ level confidence interval for μ is

$$\left[\bar{x} - \left(\frac{s}{\sqrt{n}}\right) t_{(n-1,\ 1-\frac{\alpha}{2})}, \bar{x} + \left(\frac{s}{\sqrt{n}}\right) t_{(n-1,1-\frac{\alpha}{2})}\right], \qquad (5)$$

where $s = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}$, and $t_{(n-1,1-\frac{\alpha}{2})}$ is the $1-\alpha/2$ quantile of the t distribution with $n-1$ degrees of freedom.

The formula in Equation (5) works for "exact" normal data. It is thus sensible to expect that it will work well in those circumstances where σ is (*a priori* unbeknownst to an investigator) large. It is clear that it can not have a real confidence level close to the nominal one if σ is small. For an extreme example, if $\mu = 0.25$ and $\sigma = 0.001$, for any reasonable sample

size one is essentially certain to observe a sample of all 0's, find $\bar{x} = 0$ and $s = 0$, and fail to cover $\mu = 0.25$. The confidence level of a nominally 90% interval is near 0.

Furthermore, large sample size does not cure this problem. That is, while it is true that even for a small σ case like $\mu = 0.25$ and $\sigma = 0.001$, as n increases one will in theory observe all integer values, $\bar{x}$ will converge to the mean of the rounded data distribution, and s will be positive and converge to the standard deviation of the rounded data distribution, this is not good news. The problem is that the mean of the rounded data distribution is not, in general, μ. So for large n, in a case like $\mu = 0.25$ and $\sigma = 0.001$, the real confidence level associated with applying the formula in Equation (5) to rounded normal data is essentially 0.

We seek a confidence interval method that doesn't suffer from the deficiencies of Equation (5) applied to rounded data. To accomplish this, we use the likelihood ideas introduced in the previous section.

A method of explicitly using the rounded data joint distribution in display (1) to construct the confidence intervals for μ is to invert likelihood ratio tests of $H_0 : \mu = \mu_0$ and apply the (asymptotic) chi-square null distribution associated with the likelihood ratio test statistic (see, for example, Bickel and Doksum (1977) page 229). That is, if $\mu = \mu_0$, then

$$-2\,\ell\mathrm{n}\left(\frac{\sup_{\sigma>0} f(X;\mu_0,\sigma)}{\sup_{\sigma>0}\sup_{\mu\in R} f(X;\mu,\sigma)}\right) \dot{\sim} \chi^2_{(1)}.$$

Or using the notation in this paper, if $\mu = \mu_0$

$$-2(L^*(\mu_0) - \sup_{\mu\in R} L^*(\mu)) \dot{\sim} \chi^2_{(1)}. \tag{6}$$

(Note that $\sup_{\mu\in R} L^*(\mu)$ is also the supremum log-likelihood value.)

Given a desired small α, we construct an interval of means μ satisfying the inequality

$$\sup_{\mu\in R} L^*(\mu) - L^*(\mu) \le \frac{1}{2}\chi^2_{(1,1-\alpha)}, \tag{7}$$

and we conclude from the approximation in Equation (6) that the resulting interval has (large n, or) approximate coverage probability $(1-\alpha)$.

We ran simulations to investigate the performance of our initial attempt to use the likelihood ideas represented by display (6). We found that while intervals defined by Equation (7) are conservative when σ is small and have coverage probability $(1-\alpha)$ when n is large, they don't hold their nominal confidence level for small to moderate n and large σ. So some adjustment to our initial likelihood-based method is required.

In particular, one might seek $c(n,\alpha) > \chi^2_{(1,1-\alpha)}$, so that as n gets large $c(n,\alpha)$ approximates $\chi^2_{(1,1-\alpha)}$ and so that

$$\Pr[-2(L^*(\mu) - \sup_{\mu\in R} L^*(\mu)) \le c(n,\alpha)] \ge (1-\alpha),$$

for most (μ,σ) pairs. If this can be done, we can then use the likelihood-based intervals defined by

$$\sup_{\mu\in R} L^*(\mu) - L^*(\mu) \le \frac{1}{2}c(n,\alpha). \tag{8}$$

Our reasoning to produce such $c(n,\alpha)$ is as follows. "Large σ" is the situation where rounding is perhaps negligible. So it might be expected that for large σ, the likelihood ratio test based on "rounded" data is equivalent to the likelihood ratio test of the same hypothesis, $H_0 : \mu = \mu_0$, based on "exact" data. The (standard) development of this exact data test (see Bickel and Doksum (1977, pages 209-212)) shows that the exact data version of $-2(L^*(\mu) - \sup_{\mu\in R} L^*(\mu))$ is

$$n\ell\mathrm{n}\left[1 + \frac{n}{n-1}\left(\frac{\bar{x}-\mu}{s}\right)^2\right].$$

Now, with exact normal data, $T = \sqrt{n}(\bar{x}-\mu)/s$ is well known to have a t_{n-1} distribution. This suggests that a choice of $c(n,\alpha)$ likely to produce correct large σ coverage probabilities for our likelihood-based intervals is

$$c(n,\alpha) = n\ell\mathrm{n}\left(\frac{t^2_{(n-1,1-\frac{\alpha}{2})}}{n-1} + 1\right). \tag{9}$$

Table 2 gives $c(n,\alpha)$ values for several combinations of n and α.

The rounded data likelihood-based intervals that we ultimately offer for use are those defined by Equations (8) and (9), namely

$$\left\{\mu \mid L^*(\mu) \ge \sup_{\mu\in R} L^*(\mu) - \frac{1}{2}c(n,\alpha)\right\} \tag{10}$$

for $c(n,\alpha)$ of form in Equation (9).

Properties of the Intervals

We used the two methods discussed in previous section to find intervals for the parameter μ from

TABLE 2. $c(n, \alpha)$ Values

	α		
n	0.05	0.10	0.20
2	10.18	7.42	4.70
3	6.98	4.98	3.06
4	5.90	4.18	2.55
5	5.37	3.80	2.31
6	5.05	3.57	2.17
7	4.84	3.42	2.08
8	4.70	3.31	2.01
9	4.59	3.23	1.96
10	4.50	3.17	1.93
15	4.26	3.00	1.82
∞	3.84	2.71	1.64

simulated normal samples rounded to the nearest integer. A variety of values of μ, σ, n, and α were used in the simulations to provide a thorough comparison of the two methods. We considered $\mu \in \{0, 0.1, 0.2, \cdots, 0.9, 1.0\}$, $\sigma \in \{0.01, 0.25, 0.5, 1, 5, 10\}$, $n \in \{2, 5, 10, 15\}$, and $\alpha \in \{0.05, 0.10, 0.20\}$. Figures 1-4 show graphs of the estimated coverage probabilities for the t-intervals and likelihood-based intervals. In those graphs, the solid lines indicate the estimated coverage probabilities for the t-intervals, and the dashed lines indicate the estimated coverage probabilities for the likelihood-based intervals (based on 1000 samples for each (μ, σ, n, α) combination). The actual coverage probabilities are symmetric about $\mu = 0.5$, so for a given σ and α we have averaged estimated coverage probabilities for μ and $(1 - \mu)$ before plotting.

After analyzing these graphs, we can make two conclusions:

(1) When σ is small, say $\sigma = 0.01$ or $\sigma = 0.25$, the graphs display basically the same pattern for all combinations of n and α. We can also see that the coverage probability for the likelihood method (10) is almost always bigger than that for the t method (5), except for the special points $\mu = 0$, 0.5, and 1.0. These points deserve explanation.

First, we focus on the coverage probabilities for the likelihood-based intervals (indicated by the dashed lines). If $0.0 \leq \mu < 0.5$ and σ is "small" with

$$\Phi\left(\frac{0.5 - \mu}{\sigma}\right) - \Phi\left(\frac{-0.5 - \mu}{\sigma}\right) \doteq 1,$$

then all of the "exact" sample will typically fall below 0.5 and the rounded values will all be 0. Similar reasoning applies to the interval $0.5 < \mu \leq 1.0$, but this time all rounded data will typically have the value 1. Because the likelihood interval for μ for samples with only one distinct value i_0 for these σ always contains $(i_0 - 0.5, i_0 + 0.5)$, the true parameter μ is essentially always contained in the interval. This is why the estimated coverage probabilities for the likelihood method (10) always have the value 1. But when $\mu = 0.5$ and σ is small, the values in the rounded sample will typically be a (binomial) mixture of 0's and 1's, so the coverage probability will be smaller than 1.

Second, we check the solid lines on the pictures and consider the t interval coverage probabilities. The solid lines indicate coverage probabilities larger than 0 at $\mu = 0$, 0.5, and 1.0, but 0 probabilities for other μ. Since, when σ is small, the rounded samples all tend to contain the single value 0 if $\mu \in [0, 0.5)$ or the value 1 if $\mu \in (0.5, 1.0]$, the t method tends to produce intervals degenerate at $\bar{x} = 0$ or $\bar{x} = 1$. So the method brackets μ with probability near 1 only when $\mu = 0$ or $\mu = 1$. This explains why the coverage probability is always 1 at the two points $\mu = 0$ and $\mu = 1$, but is 0 for $\mu \in (0, 0.5)$ and $\mu \in (0.5, 1)$. As to the situation when $\mu = 0.5$, the same kind of reasoning applies here as was applied to the method (10).

(2) For σ not small, we see that our replacement of the χ^2 percentile in (7) with $c(n, \alpha)$ has solved the problem of sub-nominal coverage probability for small n and moderate to large σ. The method meets our goal of providing reliable coverage of μ regardless of the true values of μ and σ and for all n. The same can not be said of the t intervals. As argued before, for small σ, their coverage probability will remain close to 0 for many μ's, *even for large* n.

In addition to estimating coverage probabilities, we also ran simulations to compare average interval lengths for the t-method and likelihood-based method. Table 3 presents average lengths for 1,000 t intervals (from(5)) and 1,000 likelihood-based intervals (from(10)) for various μ, σ, and n.

The general character of the results in these tables is as follows:

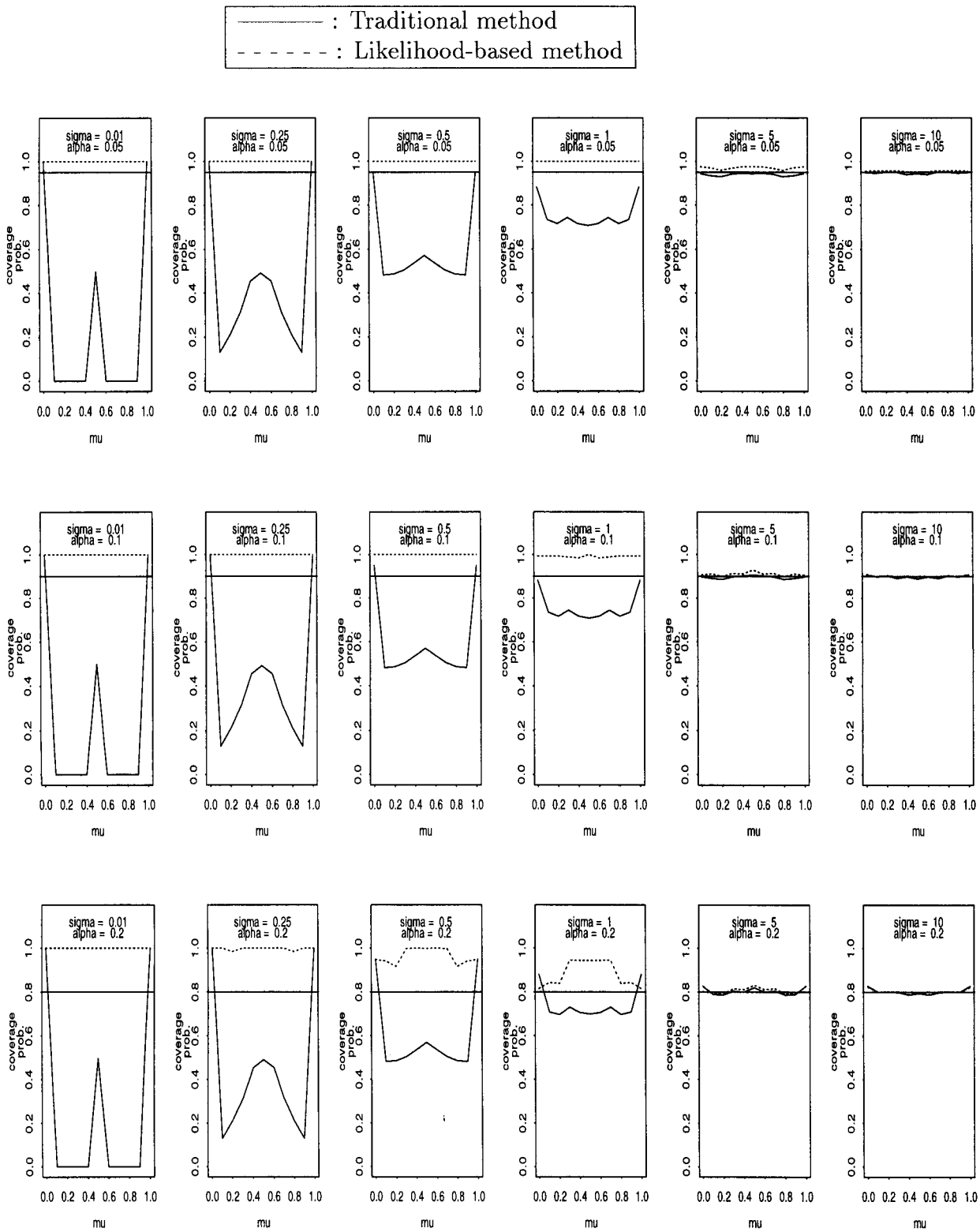

FIGURE 1. Estimated Coverage Probability for Sample Size $n = 2$.

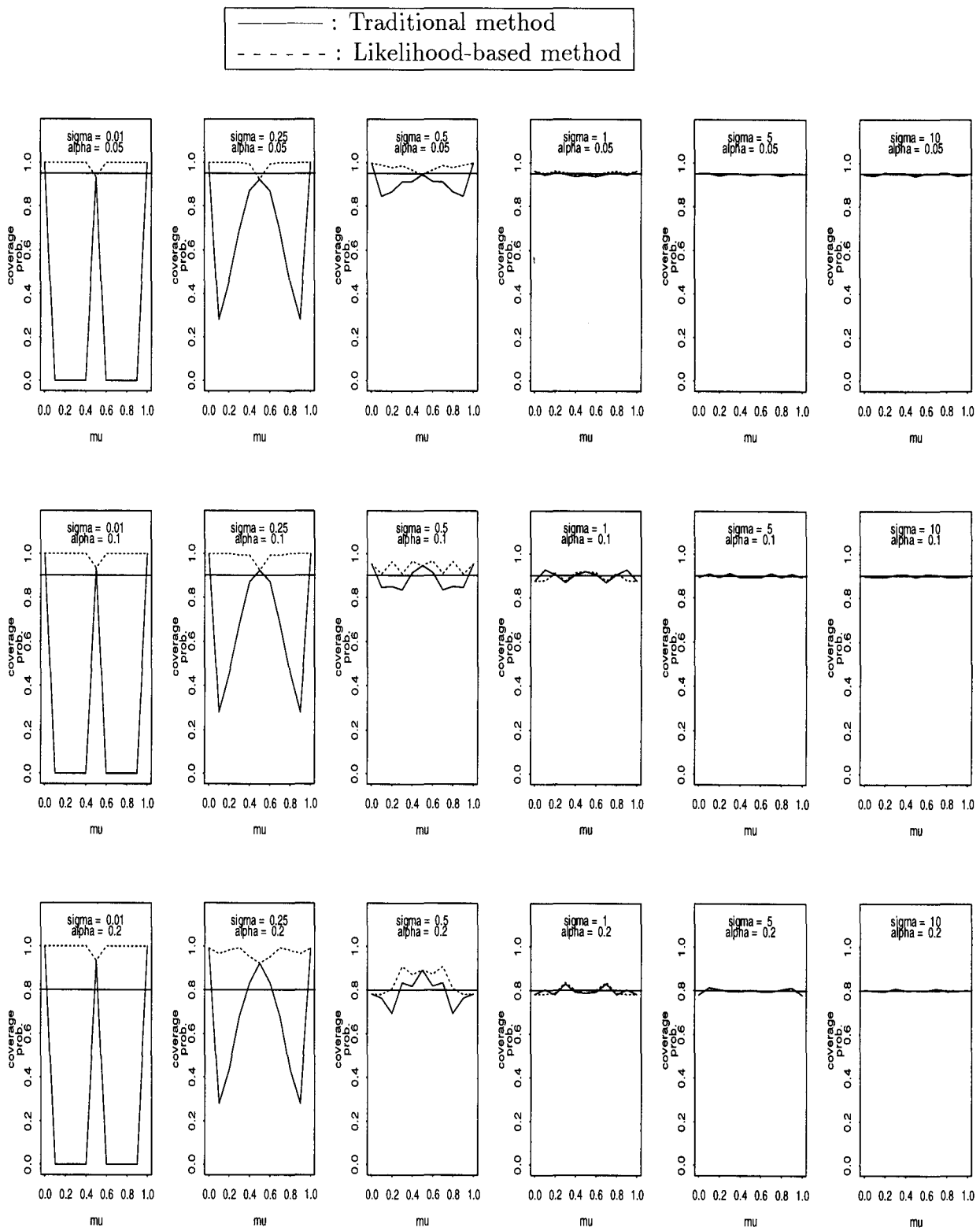

FIGURE 2. Estimated Coverage Probability for Sample Size $n = 5$.

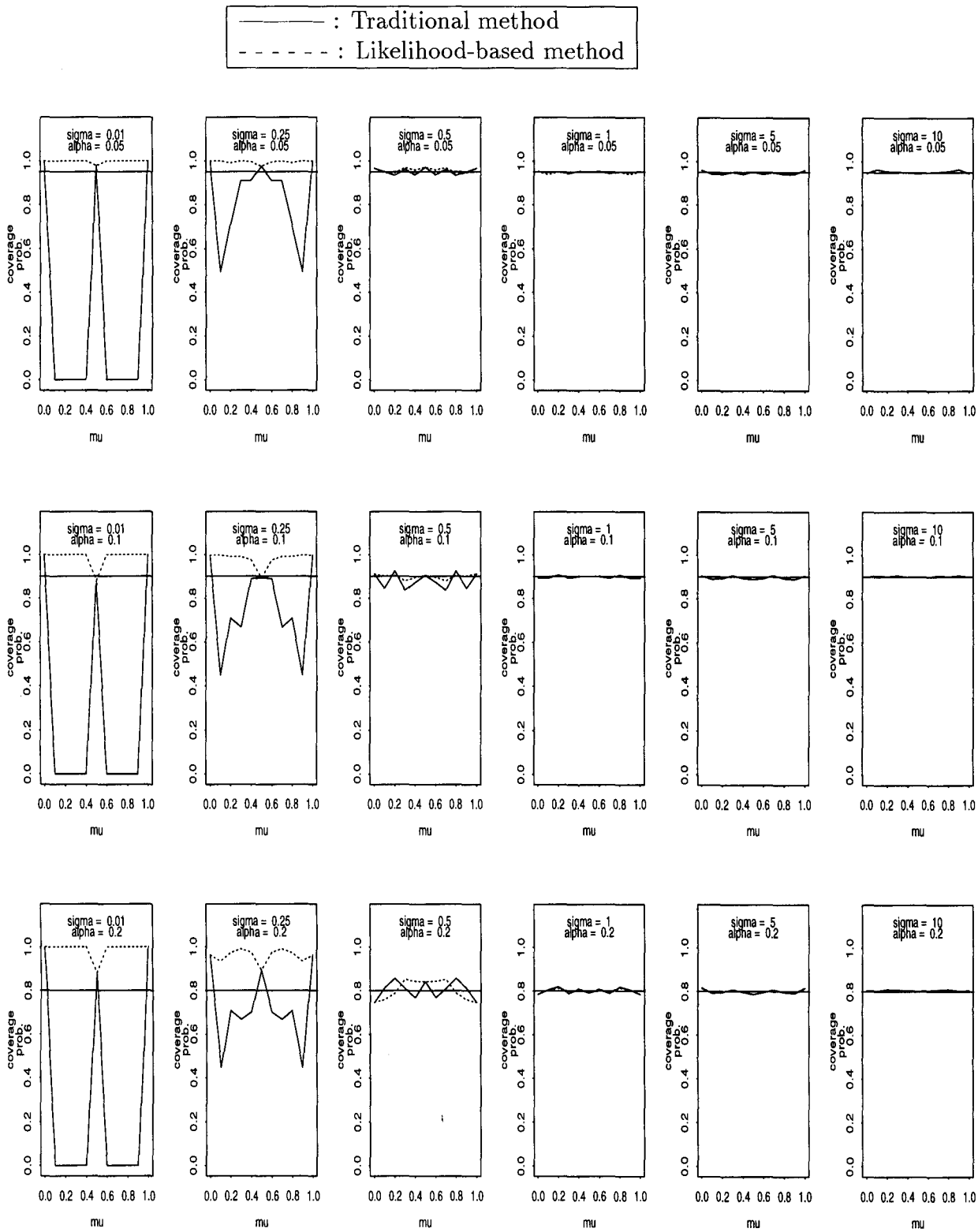

FIGURE 3. Estimated Coverage Probability for Sample Size $n = 10$.

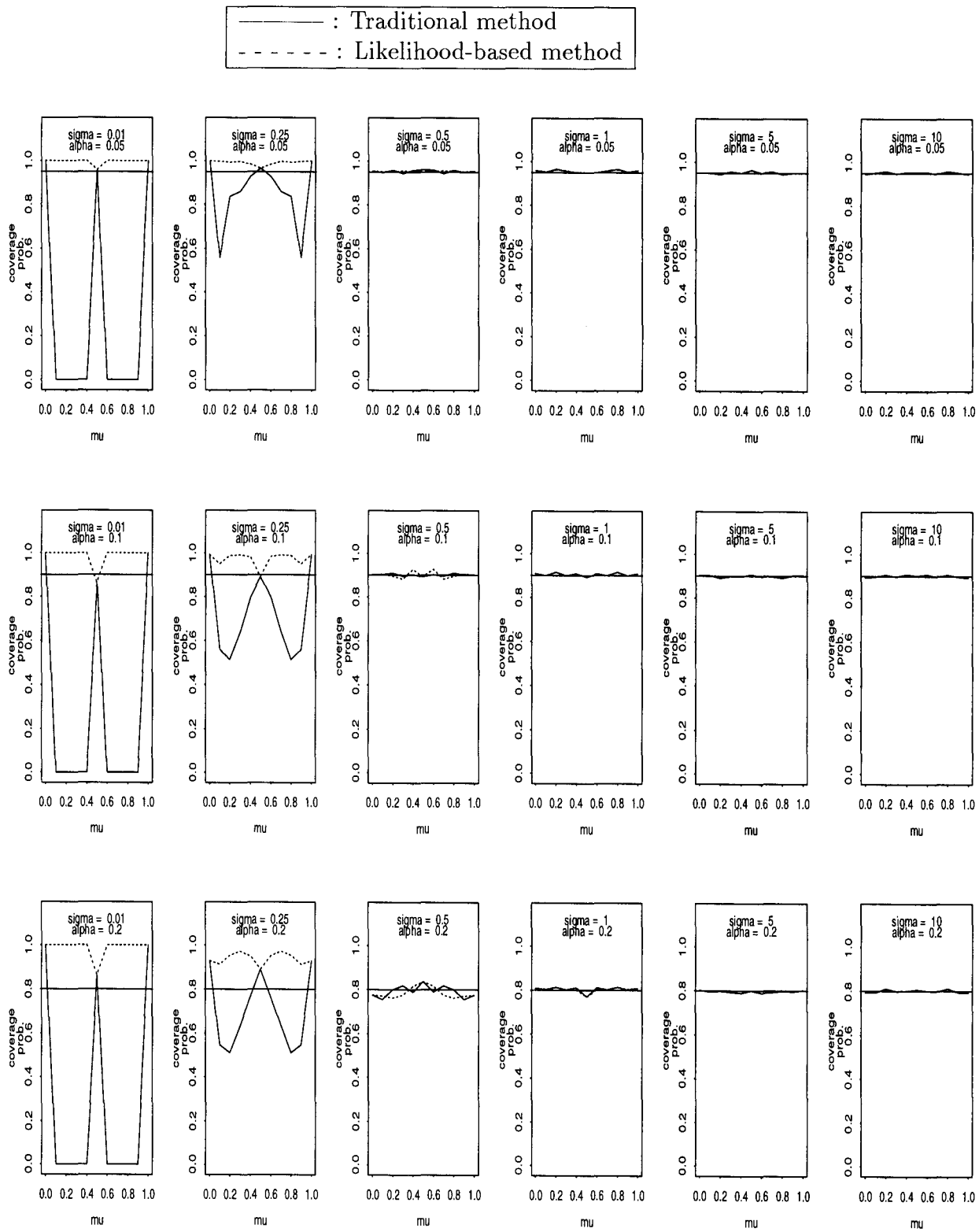

FIGURE 4. Estimated Coverage Probability for Sample Size $n = 15$.

TABLE 3. The Average Simulated Lengths for the t and Likelihood (l) Methods for Different Values of μ

$\mu = 0.0$											
σ		0.01		0.25		0.50		1.00		5.00	
n	α	t	l	t	l	t	l	t	l	t	l
2	0.05	0.000	6.167	0.889	6.596	6.760	9.726	14.574	16.050	69.795	70.098
	0.10	0.000	3.094	0.442	3.304	3.359	4.846	7.242	7.982	34.681	34.835
	0.20	0.000	1.571	0.215	1.669	1.637	2.392	3.530	3.907	16.906	16.997
3	0.05	0.000	1.553	0.391	1.727	2.241	2.710	4.508	4.601	21.663	21.651
	0.10	0.000	1.124	0.266	1.232	1.521	1.859	3.060	3.126	14.702	14.697
	0.20	0.000	1.000	0.171	1.032	0.982	1.286	1.976	2.036	9.494	9.483
4	0.05	0.000	1.035	0.258	1.122	1.553	1.753	2.998	3.013	14.639	14.628
	0.10	0.000	1.000	0.191	1.026	1.148	1.345	2.217	2.234	10.826	10.817
	0.20	0.000	1.000	0.133	0.969	0.799	0.998	1.543	1.560	7.534	7.523
5	0.05	0.000	1.000	0.236	1.033	1.276	1.402	2.465	2.459	11.897	11.897
	0.10	0.000	1.000	0.181	0.979	0.980	1.108	1.893	1.889	9.135	9.143
	0.20	0.000	1.000	0.130	0.928	0.704	0.833	1.361	1.356	6.570	6.567
$\mu = 0.25$											
2	0.05	0.000	6.167	3.138	7.680	7.446	10.140	14.485	16.027	72.959	73.282
	0.10	0.000	3.094	1.559	3.837	3.700	5.051	7.198	7.971	36.254	36.418
	0.20	0.000	1.571	0.760	1.917	1.804	2.489	3.509	3.902	17.672	17.770
3	0.05	0.000	1.553	1.185	2.056	2.277	2.681	4.507	4.608	21.742	21.727
	0.10	0.000	1.124	0.804	1.434	1.545	1.837	3.059	3.131	14.755	14.748
	0.20	0.000	1.000	0.519	1.082	0.998	1.260	1.975	2.041	9.528	9.516
4	0.05	0.000	1.035	0.822	1.318	1.639	1.757	2.955	2.968	14.805	14.794
	0.10	0.000	1.000	0.608	1.090	1.212	1.330	2.185	2.200	10.948	10.940
	0.20	0.000	1.000	0.423	0.906	0.844	0.964	1.521	1.536	7.619	7.609
5	0.05	0.000	1.000	0.676	1.098	1.314	1.384	2.358	2.362	11.418	11.418
	0.10	0.000	1.000	0.519	0.943	1.009	1.081	1.810	1.817	8.767	8.775
	0.20	0.000	1.000	0.374	0.798	0.726	0.798	1.302	1.307	6.305	6.303
$\mu = 0.5$											
2	0.05	6.391	9.249	6.035	9.077	7.459	10.077	14.320	15.658	68.588	68.904
	0.10	3.176	4.608	2.999	4.529	3.706	5.019	7.116	7.786	34.082	34.242
	0.20	1.548	2.275	1.462	2.236	1.807	2.473	3.469	3.810	16.613	16.708
3	0.05	2.169	2.474	2.166	2.473	2.493	2.765	4.590	4.666	22.463	22.446
	0.10	1.472	1.691	1.470	1.691	1.692	1.887	3.115	3.170	15.244	15.237
	0.20	0.950	1.150	0.949	1.150	1.093	1.270	2.011	2.061	9.844	9.831
4	0.05	1.517	1.575	1.464	1.557	1.715	1.779	2.974	2.983	14.725	14.713
	0.10	1.122	1.183	1.083	1.177	1.268	1.334	2.199	2.210	10.889	10.880
	0.20	0.781	0.845	0.753	0.851	0.882	0.951	1.530	1.542	7.577	7.567
5	0.05	1.199	1.221	1.201	1.219	1.382	1.398	2.455	2.453	11.680	11.680
	0.10	0.921	0.947	0.922	0.944	1.061	1.081	1.885	1.885	8.969	8.977
	0.20	0.662	0.689	0.663	0.686	0.763	0.783	1.356	1.354	6.450	6.448

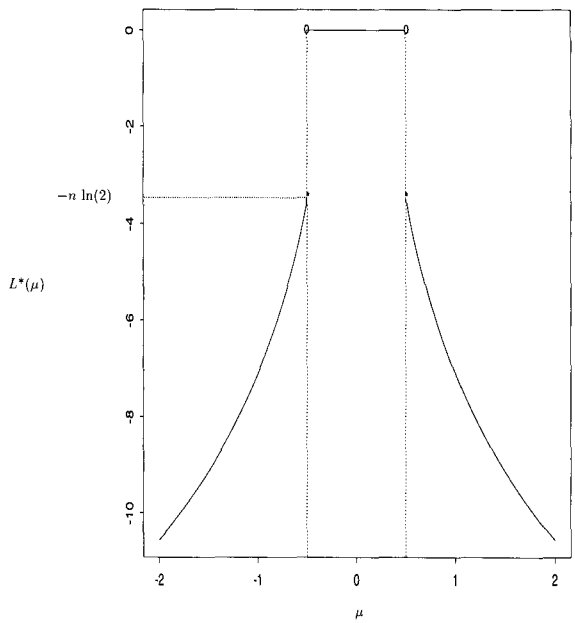

FIGURE 5. Representative Graph of $L^*(\mu)$ When a Sample Contains Only One Distinct Value. This Particular Graph is for a Case Where a Sample of Size $n = 5$ Contains Only the Value 0.

(1) At $\mu = 0.0$ and 0.25, the t method average length is much smaller than the likelihood method average length for $\sigma = 0.01$ and 0.25. And as σ grows, the mean lengths become quite similar. The difference in lengths when σ is small derives from the fact that many of the samples have range 0, and the poor coverage probabilities for the t method evident in Figures 1-4 show the impact of the small mean lengths for the t method.

(2) At $\mu = 0.5$ the average lengths for two methods are comparable.

Computational Considerations

Here we discuss what is needed in terms of computations in order to implement the likelihood-based interval (10). There are 3 cases to be considered, corresponding to the observed range of the rounded sample. The nature of the profile loglikelihood (4) is different depending upon whether the sample range, R, is 0, 1, or at least 2.

Consider first the $R = 0$ case, and suppose all n rounded observations have the value i_0. Figure 5 is representative of this situation. The profile loglikelihood is maximum (and 0) on the interval $(i_0 - 0.5, i_0 + 0.5)$. There are discontinuities at $i_0 - 0.5$ and $i_0 + 0.5$, where the profile loglikelihood drops to $-n\ell\mathrm{n}(2)$. So, unless both n and α are small (the sample size is small and the nominal confidence is large), the interval prescribed by (10) will be

$$(i_0 - 0.5, i_0 + 0.5). \tag{11}$$

TABLE 4. Δ for Small n, $R = 0$ Intervals in Equation (13)

	α		
n	0.05	0.10	0.20
2	3.084	1.547	0.785
3	0.776	0.562	
4	0.517		

For cases with small n and α, finding limits prescribed by (10) requires numerical search to the left of $\mu = i_0 - 0.5$ and to the right of $\mu = i_0 + 0.5$ to find roots of

$$L^*(\mu) = -\frac{1}{2}c(n, \alpha). \tag{12}$$

For the values $\alpha = 0.05, 0.10$, and 0.20 considered in Table 2, numerical search is required only when n is, respectively, 4 or less, 3 or less, or 2. Table 4 gives values Δ so that $i_0 - \Delta$ and $i_0 + \Delta$ solve (12), and the likelihood interval is thus

$$(i_0 - \Delta, i_0 + \Delta). \tag{13}$$

Where Table 4 is not needed, the interval is as in (11).

Now consider the situation where the range of the rounded data is $R = 1$. For definiteness, say the sample contains distinct values i_0 and $i_0 + 1$. Figures 6 and 7 are representative of the profile loglikelihood when $R = 1$. Unless $n_{i_0} = n_{i_0+1}$ the profile loglikelihood is discontinuous at $i_0 + 0.5$. The profile loglikelihood is maximum near $i_0 + 0.5$, and the maximum value is

$$n_{i_0}\ell\mathrm{n}\left(\frac{n_{i_0}}{n}\right) + n_{i_0+1}\ell\mathrm{n}\left(\frac{n_{i_0+1}}{n}\right). \tag{14}$$

To find the interval prescribed by (10) in this case, one must always search to one side of $i_0 + 0.5$ (and usually on both sides of $i_0 + 0.5$) for a root of

$$L^*(\mu) = n_{i_0}\ell\mathrm{n}\left(\frac{n_{i_0}}{n}\right) + n_{i_0+1}\ell\mathrm{n}\left(\frac{n_{i_0+1}}{n}\right) - \frac{1}{2}c(n, \alpha). \tag{15}$$

When $n_{i_0} \geq n_{i_0+1}$ one must certainly search to the

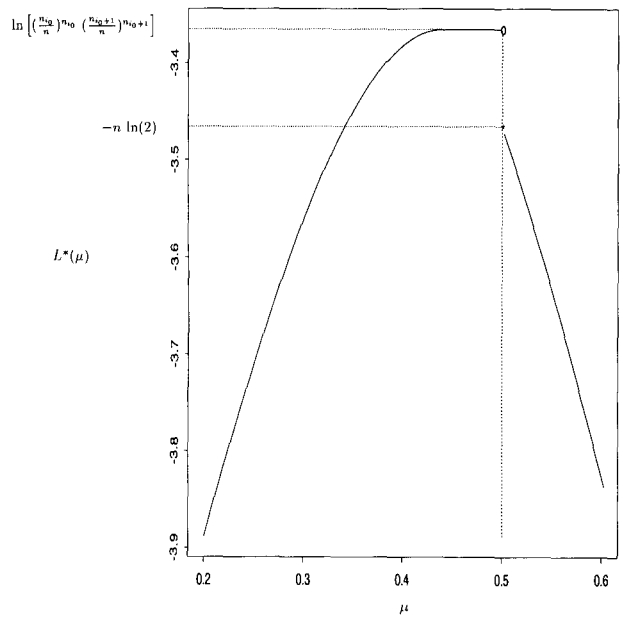

FIGURE 6. Representative Graph of $L^*(\mu)$ When a Sample Contains Only Two Distinct Values With Range 1. This Particular Graph is for a Case Where a Sample of Size $n = 5$ Gives $n_0 = 3$ and $n_1 = 2$.

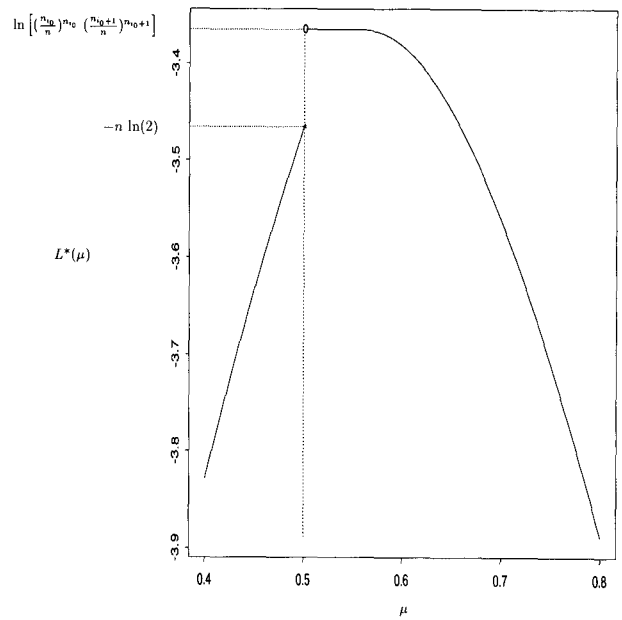

FIGURE 7. Representative Graph of $L^*(\mu)$ When a Sample Contains Only Two Distinct Values With Range 1. This Particular Graph is for a Case Where a Sample of Size $n = 5$ Gives $n_0 = 2$ and $n_1 = 3$.

left of $i_0 + 0.5$ and when $n_{i_0} \leq n_{i_0+1}$ one must certainly search to the right of $i_0 + 0.5$. Exactly when $n_{i_0} \neq n_{i_0+1}$ and

$$\frac{1}{2}c(n,\alpha) \leq n_{i_0}\ell\mathrm{n}\left(\frac{n_{i_0}}{n}\right) + n_{i_0+1}\ell\mathrm{n}\left(\frac{n_{i_0+1}}{n}\right) + n\ell\mathrm{n}(2)$$

only a single search is required, because in these cases $i_0 + 0.5$ is one of the interval end points.

Table 5 collects some pairs (Δ_1, Δ_2) with $\Delta_1 \geq \Delta_2$ for making intervals (10) in $R = 1$ cases. These Δ's were obtained from solving equation (15), except where $\Delta_2 = 0$. Where $n_{i_0} \geq n_{i_0+1}$ the interval prescribed by (10) is

$$(i_0 + 0.5 - \Delta_1, i_0 + 0.5 + \Delta_2), \tag{16}$$

while if $n_{i_0} \leq n_{i_0+1}$ the interval is

$$(i_0 + 0.5 - \Delta_2, i_0 + 0.5 + \Delta_1). \tag{17}$$

Finally, consider the case where the sample range is $R \geq 2$. Here the profile loglikelihood is concave (and therefore continuous) and generally well-behaved. There is no simple formula for the maximum (profile) loglikelihood, so numerical search over μ and σ is required to find the maximum

$$M = \sup_{\mu \in R} L^*(\mu) = \sup_{\mu \in R} \sup_{\sigma > 0} L(\mu, \sigma).$$

Subsequently, numerical search is required to find two roots of the equation

$$L^*(\mu) = M - \frac{1}{2}c(n,\alpha) \tag{18}$$

in order to implement formula (10). It has been our experience that effective starting values for a two-dimensional numerical search for M are

$$\mu = \bar{x} \quad \text{and} \quad \sigma = \sqrt{\frac{(n-1)s^2}{n} - \frac{1}{12}} \tag{19}$$

and that effective starting values for roots of (18) in $R \geq 2$ cases are the end points of the t intervals (5). The σ starting point in (19) is the famous "Sheppard's corrected" sample standard deviation for integer grouped data.

Discussion and Conclusions

We began by considering the rounded journal diameters in Table 1 and the $n = 10$ digital gage readings of Stein. Our work with the likelihood-based intervals defined in (10) shows, for example, that 95% confidence limits for μ based on the first $n = 3$ sample in Table 1 are

$$-1 - 0.776 = -1.776 \quad \text{and} \quad -1 + 0.776 = -0.224$$

using (13) and Table 4. Further, 95% confidence limits for μ based on the second $n = 3$ sample in Table

TABLE 5. (Δ_1, Δ_2) for $R = 1$ Intervals (16) or (17) with $m = \max\{n_{i_0}, n_{i_0+1}\}$

		α		
n	m	0.05	0.10	0.20
2	1	(6.147, 6.147)	(3.053, 3.053)	(1.485, 1.485)
3	2	(1.552, 1.219)	(1.104, 0.771)	(0.765, 0.433)
4	3	(1.025, 0.526)	(0.820, 0.323)	(0.639, 0.149)
	2	(0.880, 0.880)	(0.646, 0.646)	(0.441, 0.441)
5	4	(0.853, 0.257)	(0.721, 0.132)	(0.592, 0.024)
	3	(0.748, 0.548)	(0.592, 0.393)	(0.443, 0.248)
6	5	(0.772, 0.116)	(0.673, 0.032)	(0.569, 0.000)
	4	(0.680, 0.349)	(0.562, 0.235)	(0.444, 0.126)
	3	(0.543, 0.543)	(0.420, 0.420)	(0.299, 0.299)
7	6	(0.726, 0.035)	(0.645, 0.000)	(0.556, 0.000)
	5	(0.640, 0.218)	(0.545, 0.130)	(0.446, 0.046)
	4	(0.534, 0.393)	(0.432, 0.293)	(0.329, 0.193)
8	7	(0.698, 0.000)	(0.626, 0.000)	(0.547, 0.000)
	6	(0.616, 0.129)	(0.534, 0.058)	(0.446, 0.000)
	5	(0.527, 0.281)	(0.439, 0.197)	(0.347, 0.113)
	4	(0.416, 0.416)	(0.327, 0.327)	(0.236, 0.236)
9	8	(0.677, 0.000)	(0.613, 0.000)	(0.541, 0.000)
	7	(0.599, 0.065)	(0.526, 0.010)	(0.448, 0.000)
	6	(0.521, 0.196)	(0.443, 0.124)	(0.361, 0.054)
	5	(0.429, 0.321)	(0.350, 0.242)	(0.267, 0.163)
10	9	(0.662, 0.000)	(0.604, 0.000)	(0.537, 0.000)
	8	(0.587, 0.020)	(0.521, 0.000)	(0.450, 0.000)
	7	(0.515, 0.129)	(0.446, 0.069)	(0.371, 0.012)
	6	(0.437, 0.242)	(0.365, 0.174)	(0.289, 0.105)
	5	(0.346, 0.346)	(0.275, 0.275)	(0.200, 0.200)

1 are

$$-0.5 - 1.219 = -1.719 \quad \text{and} \quad -0.5 + 1.552 = 1.052$$

using (17) and Table 5.

After similarly using (17) and Table 5 with the integer coded version of Stein's data set and translating back to his original scale we find 95% confidence limits for μ

$$1.226 \qquad \text{and} \qquad 1.294.$$

Our analysis shows that Stein's belief that the true value is likely very close to 1.26 is overly optimistic. When σ is small enough to make small sample ranges of rounded values likely, one simply doesn't learn much about μ.

More generally, from our analyses we reach the following conclusions about crudely gaged data in the interval estimation of μ.

(1) When it is *a priori* clear that σ could be small in comparison to "rounding precision," and one obtains a rounded sample with all values equal to i_0, there is really no way to estimate μ reliably beyond saying that $\mu \in (i_0 - 0.5, i_0 + 0.5)$. (Of course, in such cases, the best option in terms of quality of estimation is to find another gage that is not so crude.)

If it is *a priori* clear that it is possible that $\sigma < 0.5$ and obtaining better gaging is not an option, then it is best to use the likelihood-

based method (10), since the actual confidence level of the standard t intervals can be significantly below the nominal one (and essentially 0).

(2) When one is *a priori* sure that $\sigma \geq 0.5$, both methods (10) and (5) can be used except for $n = 2$. The simulations show that for $n = 2$ the likelihood-based method is much better than the t-method when $\sigma = 0.5$ and 1 (See Figure 1).

Crudely gaged data are, of course, not ideal. They are, however, very common in industrial applications. The analysis in this paper: (1) shows clearly the peril of ignoring the rounding issue; (2) helps identify for various n the degree of rounding that can be tolerated by standard statistical methodology; and (3) provides a method of interval estimation for μ that is reliable even when rounding proves to be important.

Appendix

It can be helpful to have approximations for the end points of the likelihood-based intervals. (For one thing, these can be starting values for numerical searches for more exact values.) We provide such approximations in this Appendix for $R = 1$ cases.

Continue to let n_{i_0} be the number of values i_0 observed and n_{i_0+1} be the number of values $(i_0 + 1)$ observed. To find approximations for the intervals prescribed by display (10) when $R = 1$, we plug $\widehat{\sigma}_\mu = \sqrt{\sum_{i=1}^{n}(x_i - \mu)^2/n}$ into the exact data loglikelihood modified by an empirically derived "correction factor" k to produce the approximation

$$L^*(\mu) \doteq k \sum_{i=i_0}^{i_0+1} n_i \, \ell n \left(\frac{1}{\widehat{\sigma}_\mu} \phi \left(\frac{i - \mu}{\widehat{\sigma}_\mu} \right) \right),$$

for

$$k = \begin{cases} 1 & \text{if } n_{i_0+1} < n_{i_0} \text{ when computing a lower bound for } \mu \\ 0.975 & \text{if } n_{i_0+1} < n_{i_0} \text{ when computing an upper bound for } \mu \\ 0.975 & \text{if } n_{i_0+1} > n_{i_0} \text{ when computing a lower bound for } \mu \\ \frac{1}{0.975} & \text{if } n_{i_0+1} > n_{i_0} \text{ when computing an upper bound for } \mu. \end{cases}$$

Substituting this approximation into display (15) and solving the quadratic equation in μ that results when there is equality, we get two solutions for μ. For convenience in what follows, let $\widehat{\sigma}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$ and $w = (2M - c(n, \alpha))/n$, for M the supremum of the loglikelihood given in display (14).

Case A : When $n_{i_0+1} < n_{i_0}$ is observed.

If $(n\ell n(2) + M) > 1/2c(n, \alpha)$, then the interval for μ prescribed by display (10) is approximately

$$\left(\bar{x} - \sqrt{\frac{e^{-1-w}}{2\pi} - \widehat{\sigma}^2}, \quad i_0 + 0.5 \right).$$

Otherwise, the interval for μ is approximately

$$\left(\bar{x} - \sqrt{\frac{e^{-1-w}}{2\pi} - \widehat{\sigma}^2}, \bar{x} + \sqrt{\frac{e^{-1-(w/0.975)}}{2\pi} - \widehat{\sigma}^2} \right).$$

Case B : When $n_{i_0+1} > n_{i_0}$ is observed.

If $(n\ell n(2) + M) > 1/2c(n, \alpha)$, then the interval for μ prescribed by display (10) is approximately

$$\left(i_0 + 0.5, \quad \bar{x} + \sqrt{\frac{e^{-1-(0.975w)}}{2\pi} - \widehat{\sigma}^2} \right).$$

Otherwise, the interval for μ is approximately

$$\left(\bar{x} - \sqrt{\frac{e^{-1-\frac{w}{0.975}}}{2\pi} - \widehat{\sigma}^2}, \bar{x} + \sqrt{\frac{e^{-1-(0.975w)}}{2\pi} - \widehat{\sigma}^2} \right).$$

References

BICKEL, P. J. and DOKSUM, K. A. (1977). *Mathematical Statistics : Basic Ideas and Selected Topics*. Holden-Day, San Francisco, CA.

HAITOVSKY, Y. (1982). "Grouped Data" in *Encyclopedia of Statistical Sciences* 3, pp. 527–536. Wiley, New York, NY.

HEITJAN, D. (1989). "Inference from Grouped Continuous Data: A Review". *Statistical Science* 4, pp. 164–183.

SCHADER, M. and SCHMID, F. (1984). "Computation of Maximum Likelihood Estimates for μ and σ from a Grouped Sample of a Normal Population - A Comparison of Algorithms". *Statistical Papers* 25, pp. 245–258.

SHAPIRO, S. S. and GULATI, S. (1998). "Estimating the Mean of an Exponential Distribution from Grouped Observations". *Journal of Quality Technology* 30, pp. 107–118.

SHEPPARD, W. (1898). "On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equidistant Divisions of a Scale". *Proceedings of the London Mathematical Society* 29, pp. 353–380.

STEIN, P. (2000). "Careful Interpolation Yields Useful Information". *Quality Progress* January, pp. 67–69.

SWAN, A. V. (1969). "Maximum Likelihood Estimation from Grouped and Censored Normal Data". *Applied Statistics* 18, pp. 110–114.

TRICKER, A. ; COATES, E. ; and OKELL, E. (1998). "The Effect on the R Chart of Precision of Measurement". *Journal of Quality Technology* 30, pp. 232–239.

VARDEMAN, S. B. and JENSEN, K. L. (1989). "$\overline{X}$ and R Control Charts for Rounded Data". Preprint Number 89-33, Statistical Laboratory, Iowa State University, Ames, IA.

Key Words: *Coverage Probability, Gaging, Likelihood.*