Likelihood-Based Statistical Estimation From Quantized Data

by

Stephen B. Vardeman*
Statistics and IMSE Departments
Iowa State University
Ames, Iowa
vardeman@iastate.edu

Chiang-Sheng Lee
Department of Industrial Management
National Taiwan University of Science and Technology
Taipei, Taiwan, R.O.C.
chiang@im.ntust.edu.tw

Abstract

Most standard statistical methods treat numerical data as if they were real (infinite-number-of-decimal-places) observations. The issue of quantization or digital resolution is recognized by engineers and metrologists, but is largely ignored by statisticians and can render standard statistical methods inappropriate and misleading. This article discusses some of the difficulties of interpretation and corresponding difficulties of inference arising in even very simple measurement contexts, once the presence of quantization is admitted. It then argues (using the simple case of confidence interval estimation based on a quantized random sample from a normal distribution as a vehicle) for the use of statistical methods based on "rounded data likelihood functions" as an effective way of dealing with the issue.

I. Introduction

"Quantization" (see [1]-[3]) or "digital resolution" (see [4],[5]) of measurement is well-recognized as a source of measurement error by engineers and metrologists. But it is typically ignored by statisticians as they develop methods of statistical inference, whose inputs in any real application are potentially subject to quantization effects. The matter is never even considered in the exposition of basic or intermediate statistical methods, and one is then perhaps left to wonder whether quantization is irrelevant as far as simple statistical analysis is concerned.

Take, for example, the case in described in [5] where 10 readings taken with a digital gauge are

* The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of Complexity in Multivariate Data Structures") through the University of Dortmund is gratefully acknowledged.

$$1.3, 1.2, 1.3, 1.3, 1.3, 1.2, 1.2, 1.3, 1.2 \text{ and } 1.3 \ .$$

The author notes that these numbers average to 1.26 and says "By taking the mean of 1.26, you can add another digit of resolution to your process." He seems to imply that 1) his measurement are "only good to the nearest .1" and 2) standard elementary statistical operations are appropriate with such values. (Indeed, he says that a simple average of 10 values provides insight into the underlying phenomenon that is an order of magnitude more revealing than the individual raw data themselves.)

Our purpose here is to examine the question of when the issue of digital resolution may safely be ignored for purposes of elementary statistical inference, and to identify reliable means of dealing with it when it can not be ignored. In the end, we will conclude that the author in [5] was wise to explicitly recognize that his values were only good to the nearest .1, but naïve in assuming that standard elementary statistical calculations are necessarily appropriate under such conditions, and wildly optimistic in expecting that his average of 10 observations was in any sense an order of magnitude better than a single observation.

II. Continuous Distributions and "Rounding"

Most standard small sample statistical methods are built on models that say the mechanism generating observations can be described by a continuous probability distribution like, for example, the normal distribution with mean μ and standard deviation σ that has the probability density

$$f\left(x \mid \mu,\sigma\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right) \tag{1}$$

pictured in Fig. 1. Under such a model, the long run fraction of values falling in any interval $\left(a,b\right)$ is

$$P_{\mu,\sigma}\left(a < X < b\right) = \int_a^b f\left(x \mid \mu,\sigma\right)dx = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{2}$$

where $\Phi\left(z\right)$ is the standard normal cumulative distribution function,

$$\Phi\left(z\right) = \int_{-\infty}^{z} f\left(x \mid 0,1\right)dx \tag{3}$$

In this framework, the model parameters μ and σ become the objects of interest and the implicit assumption is then that one actually observes and works with real numbers from the continuous distribution.

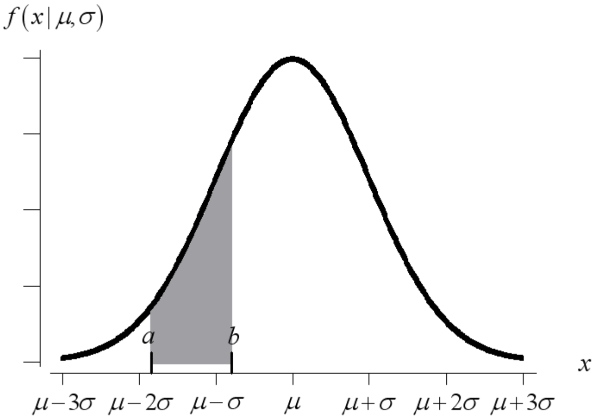

Fig. 1. A normal distribution model.

In this kind of elementary statistical modeling, the number "4" is typically interpreted as "4.0000000000…" (just as is the number "4.0"). But even when a continuous model is a good description of a physical phenomenon, it need not adequately describe what can be observed. There is the matter of quantization of measurement.

Take for example a "4.0 mm" reading from the digital gauge pictured in Fig. 2. How should that value be interpreted and then used in statistical analysis? After all, the gauge can read out only the numbers

$$\ldots, 3.8, 3.9, 4.0, 4.1, \ldots \; .$$

It can not read out a number like 4.1111111111… . It would seem that a better interpretation of the "4.0 mm" reading than

$$4.0000000000\ldots \text{ mm}$$

is the interpretation

$$\text{between } 3.950000000000\ldots \text{ mm and } 4.050000000000\ldots \text{ mm} \; .$$

While conceptually there might be a "real number" measurement corresponding to a recorded "4.0 mm" value, all we know about that number from the gauge is that it is within .05 mm of what is read/recorded. Whether this distinction is important to a statistical analysis depends upon how variable are the real numbers that stand behind what is recorded. Let us elaborate.

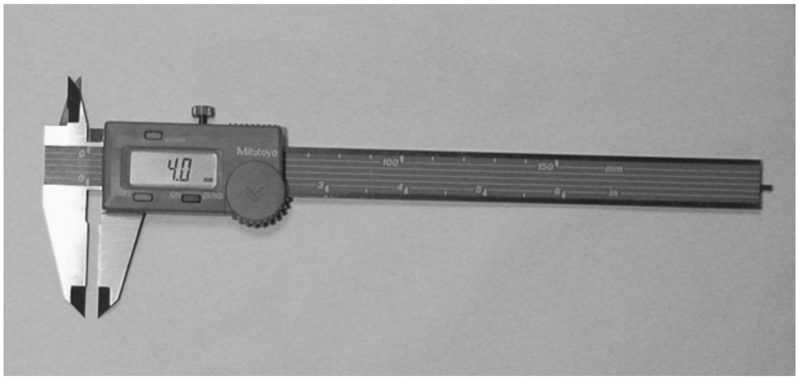

Fig. 2. 4.0 mm on a digital gauge.

A way to describe the kind of "to the nearest .1 mm" interpretation we're suggesting for the observation "4.0 mm" is through the notion of "interval censoring" or (more simply) "rounding." (See [6] and Chapters 2 and 3 of [7] for discussions of the notion of censoring in the statistical literature.) That is, suppose that conceptual real number measurements are read only after rounding to the nearest full unit of observation (to the nearest .1 mm in the case of the digital gauge pictured in Fig. 2). Then the kind of continuous distribution model pictured in Fig. 1 ought to be replaced by a discrete distribution for what is read/recorded. The probabilities for the discrete distribution should be related to continuous model as follows: If real number measurement X with normal distribution with mean μ and standard deviation σ mean "rounds" to Y, where the finest unit of observation is Δ,

$$P_{\mu,\sigma}\left[Y=y\right]=P_{\mu,\sigma}\left[y-\frac{\Delta}{2}<X<y+\frac{\Delta}{2}\right]=\Phi\left(\frac{y+\frac{\Delta}{2}-\mu}{\sigma}\right)-\Phi\left(\frac{y-\frac{\Delta}{2}-\mu}{\sigma}\right) \qquad (4)$$

Fig. 3 illustrates this correspondence.

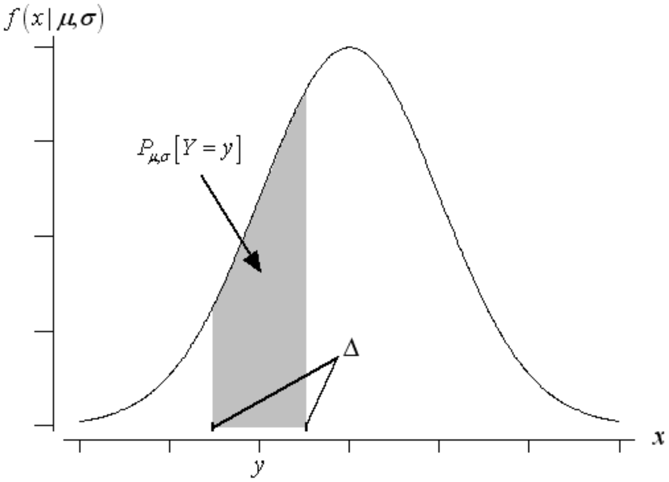

Fig. 3. Relationship between probabilities for X and those for Y.

Notice that this is quantization in the sense of [1]-[3], and that much of the related engineering literature concerns the nature of the quantization error

$$Q = Y - X \tag{5}$$

(particularly in signal processing contexts).

Fig.s 4 and 5 show two different normal distributions and corresponding "probability histograms" (representing respectively the distributions of *X* and of *Y*). In the first σ is "not small" compared to Δ, while in the second σ is small compared to Δ. In the first case the probability histogram looks roughly like the normal curve and in the second it does not. Table I records the parameters used to make the two pairs of graphs (Δ, μ and σ) and the corresponding means (μ_Y) and standard deviations (σ_Y) for the rounded observation *Y*. Notice that not only are the two graphs in Fig. 5 quite different, but *the mean and/or standard deviation of Y (μ_Y and σ_Y) can be substantially different from those of X* (μ and σ).

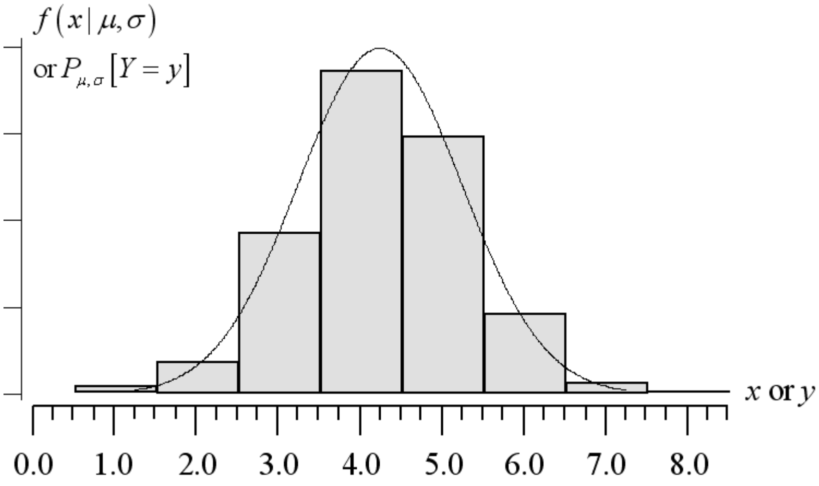

Fig. 4. $\Delta = 1.0, \mu = 4.25$ and $\sigma = 1.0$ distributions of X and Y.

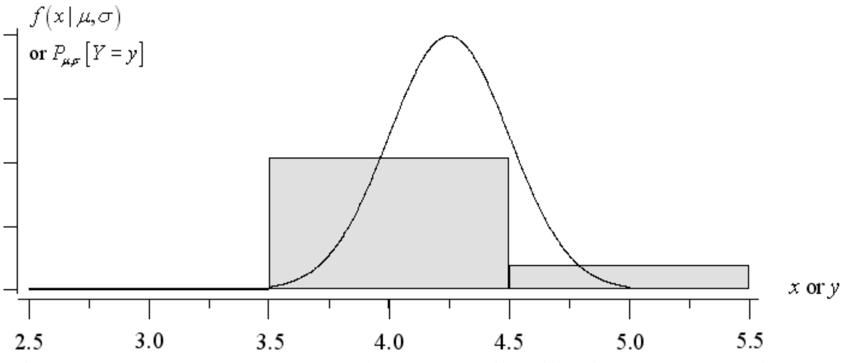

Fig. 5. $\Delta = 1.0, \mu = 4.25$ and $\sigma = .25$ distributions of X and Y.

Table I
Two Sets of X Distribution Parameters and Corresponding Means and Standard Deviations of Y

Figure	Δ	μ	σ	μ_Y, Mean of Y	σ_Y, Standard Deviation of Y
4	1.00	4.25	1.00	4.2500	1.0809
5	1.00	4.25	.25	4.1573	.3678

We should remark that although the probability histograms used in Fig.s 4 and 5 are common in statistical circles, it could be argued that a better representation of the discrete distributions of Y might be in terms of line or spike graphs that more forcefully indicate that the distributions are concentrated on the integers. We note also that a referee has pointed out that applying Sheppard's correction (see [8]) to the values of σ_Y produces $\sqrt{(1.0809)^2 - 1/12} = 1.0416$ for the case of Fig. 4 and $\sqrt{(.3678)^2 - 1/12} = .2279$ for the case of Fig. 5. In these cases, the correction is effective and these values provide better matches to σ than do the values of σ_Y in Table I.

Numerical calculation with the discrete distribution of Y establishes that as long as σ is at least $\Delta/2$, there is good agreement between μ and μ_Y. (In fact for $\sigma > \Delta/2$, μ_Y is within $\Delta/200$ of μ.) On the other hand, for σ small (compared to Δ), μ_Y can differ from μ by nearly $\Delta/2$. (Take, for example, a case where μ is almost but not quite exactly half way between two successive possible rounded values, and σ is tiny.) And the situation as regards standard deviations is similarly complex. Provided $\sigma > .15\Delta$, σ_Y exceeds σ, and for $\sigma > \Delta/2$ the fractional increase going from σ to σ_Y is no more than $.141$ (and this decreases as σ increases). But when σ is small (compared to Δ), σ_Y can be many times σ (for example in a case where μ is exactly half way between two successive possible rounded values) and it can negligible in comparison to σ (for example in a case where μ is exactly equal to a possible rounded value). (Note, by the way, that in this latter circumstance Sheppard's correction will be of no help.)

A referee has commented that (for a given lower bound on σ) if it is important enough that the distribution of Y (and, for example, its moments) approximate that of X, engineering resources can almost always be brought to bear to improve measurement by appropriately reducing Δ. That is, the quality of the match between X and Y is subject to engineering cost/benefit considerations. We don't disagree, but our emphasis here is really a different one. For a reliable statistical analysis, it is not necessary that Y match X. But it *is* necessary that 1) the issue of quantization and the kind of effects it produces be recognized and 2) that relevant allowance be made for its presence in the statistical methodology employed.

III. Statistical Inference From “Rounded” Data

Fig.s 4 and 5, Table I, and the forgoing discussion represent a serious and very real problem for elementary data analysis. What are typically of most interest are the characteristics of the “real number” (*X*) distribution, like μ and σ. But if treated as itself a “real number,” what is observed (*Y*) can have characteristics quite unlike those of interest when σ is small (compared to Δ). And this possibility can not simply be ignored as if it never matters.

If one knew exactly the *Y* (rounded observation) distribution or even μ_Y and σ_Y, it would be possible to determine μ and σ (that is, the distribution of *X*) from that information. But the further problem of statistics is that one has only empirical observations $y_1, y_2, \ldots, y_n$ from the *Y* distribution, and these give only a noisy or approximate view of the rounded data distribution.

Elementary statistical summaries (made treating the rounded data as real numbers) like the sample mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6}$$

and the sample standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{7}$$

are at best approximations for μ_Y and σ_Y, not for μ and σ. And contrary to naïve intuition (that perhaps assumes that all problems are solved by large samples), this phenomenon doesn't “go away” as *n* gets larger. (Indeed it shouldn’t, as large samples will only let one see clearly μ_Y and σ_Y !) For example, the standard elementary confidence limits for a mean applied to the rounded data

$$\bar{y} \pm t\frac{s}{\sqrt{n}} \tag{8}$$

will for large samples “zero in” on μ_Y, not on μ, giving real coverage probability approaching 0, not the nominal confidence level.

So there is a real question as to how one might develop reasonably elementary statistical methods that take account of the fact that *Y* is not *X* and of the fact that in any case one has only a noisy view of the *Y* distribution. One kind of answer to this question has been developed in [9] and [10] using the notion of a rounded data “likelihood function.”

If one models what is observed as independent realizations from a normal distribution with mean μ and standard deviation σ rounded to the nearest Δ, the probability associated with a possible sample $y_1, y_2, \ldots, y_n$ is (from equation (4))

$$L(\mu,\sigma)=\prod_{i=1}^{n}P_{\mu,\sigma}\left[y_i-\frac{\Delta}{2}<X<y_i+\frac{\Delta}{2}\right]=\prod_{i=1}^{n}\left(\Phi\left(\frac{y_i+\frac{\Delta}{2}-\mu}{\sigma}\right)-\Phi\left(\frac{y_i-\frac{\Delta}{2}-\mu}{\sigma}\right)\right) \quad (9)$$

The (data-dependent) function of μ and σ in display (9) is called the "likelihood function" and its logarithm

$$l(\mu,\sigma)=\log L(\mu,\sigma) \quad (10)$$

is (not surprisingly) called the "log-likelihood function." These can be used to reliably guide inference about the parameters of the "real number" (*X*) distribution based on rounded observations $y_1, y_2, \ldots, y_n$. These are large for values of the parameters that are in some sense compatible with the data in hand, and small for values that are incompatible with what has been observed. (Incompatible here means that observations in hand could essentially never be generated by a model with such parameters.) Basing inference for μ and σ on the rounded data likelihood function (or its logarithm) is a way of explicitly accounting for the fact that we know that what is observed are not real numbers and that they do not even definitively identify the *Y* distribution.

Fig. 6 is a contour plot (a topographic map) of a function very closely related to the log-likelihood function for the $n=10$ data points of [5] used as an example in the introduction, namely

$$l^*(t,\sigma)=l\left((1.25+.25\sigma)+t\sigma,\sigma\right) \quad . \quad (11)$$

(In making the plot we've used base 10 logarithms. We would actually have preferred to plot $l(\mu,\sigma)$ directly, but matters of scaling make it far easier to obtain and interpret the present plot.) Fig. 6 indicates that what these 4 values 1.2 and 6 values 1.3 really suggest is 1) σ is small (compared to $\Delta=.1$) and 2) $t\approx 0$.

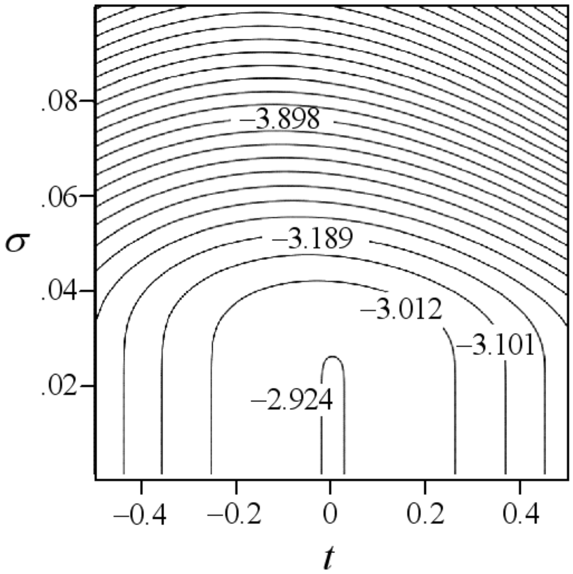

Fig. 6. Contour plot of $l^*(t,\sigma)=l\left((1.25+.25\sigma)+t\sigma,\sigma\right)$ for the data of [5].

This is in complete agreement with informed intuition. The condition $t \approx 0$ can be written as

$$\mu \approx 1.25 + .25\sigma \qquad (12)$$

and under this condition parameters μ and σ are such that about 40% of the X distribution is to the left of $x = 1.25$ and about 60% is to the right, just as is the case with the rounded values. (That is, 1.25 appears in the expressions above because it is half way between the rounded values that are observed, and .25 appears because $.25 = \Phi^{-1}(.6)$.) Notice also that the result (12) is not necessarily compatible with Stein's suggestion that to two decimal places the data indicate that some true value is 1.26.

One important quantitative use of a likelihood function is in producing single number (point) estimates of parameters. It is common to adopt parameter values that make it as large as possible as estimates of the unknown parameters. These are "maximum likelihood estimates." There are no simple formulas for these estimates, but finding them numerically is not hard, at least after one knows what to expect in terms of the behavior of $l(\mu,\sigma)$.

Strictly speaking, $l(\mu,\sigma)$ has no maximum unless the range of rounded values

$$R = \max_{i=1,\ldots,n} y_i - \min_{i=1,\ldots,n} y_i \qquad (13)$$

is larger than Δ. But when $R = 0$ (all observed rounded values are the same), for any μ within $\Delta/2$ of the common observed value, provided σ is small the limiting value $l(\mu,\sigma) = 0$ ($L(\mu,\sigma) = 1$) is very nearly achieved. (All one has learned from the data in hand is that the standard deviation is small and the mean is within a half unit of the recorded value.) When $R = \Delta$ (there are only two different observed rounded values, separated by one unit of observation), the limiting value is very nearly achieved for σ small and μ nearly half way between the two rounded values and linearly related to σ so that the underlying normal distribution of X puts appropriate fractions of its probability to the left and right of $\frac{1}{2}\left(\max_{i=1,\ldots,n} y_i + \min_{i=1,\ldots,n} y_i\right)$. (This is the case illustrated by the example of [5].) Finally, when $R > \Delta$ the likelihood (or log-likelihood) is mound-shaped and simple numerical analysis will easily find maximum likelihood estimates.

For the sake of illustrating the discussion of the forgoing paragraph, Fig.s 7 and 8 are contour plots complementing Fig. 6. Fig. 7 is a plot of $l(\mu,\sigma)$ for a $R = 0$ case where $n = 10$ rounded values $y_1, y_2, \ldots, y_{10}$ are all 1.2. Fig. 8 is a plot of $l(\mu,\sigma)$ for a $R > \Delta$ case where among 10 rounded values $y_1, y_2, \ldots, y_{10}$ there is a single value 1.1, seven values 1.2 and two values 1.3. Notice that in this last case $\overline{y} = 1.2100$ and $s = .0568$, while maximum likelihood estimates are $\hat{\mu} = 1.2102$ and $\hat{\sigma} = .0465$ respectively.

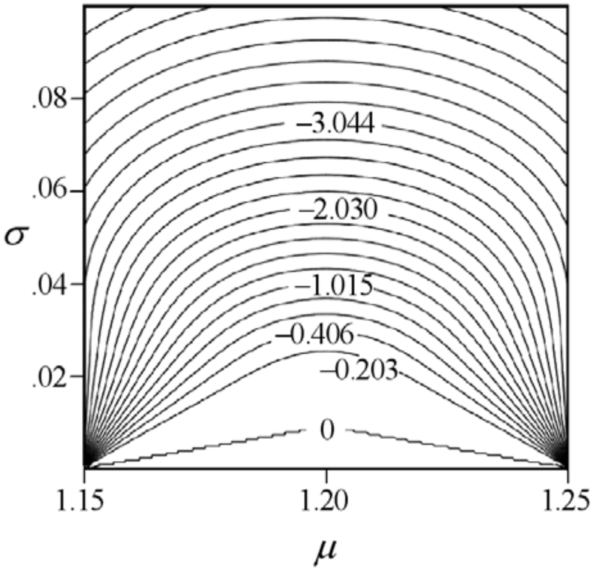

Fig. 7. Contour plot of $l(\mu,\sigma)$ when $n=10$ rounded values $y_1, y_2, \ldots, y_{10}$ are all 1.2.

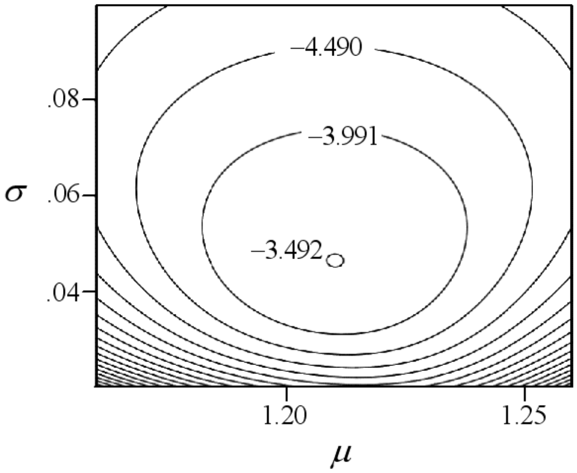

Fig. 8. Contour plot of $l(\mu,\sigma)$ where among 10 rounded values $y_1, y_2, \ldots, y_{10}$ there is a single value 1.1, seven values 1.2 and two values 1.3.

IV. Confidence Intervals Based on the Likelihood Function

It is not only possible to use the likelihood function (9) to guide qualitative statements about the parameters μ and σ and to find maximum likelihood estimates, but it can be used to decide how much it is appropriate to "hedge" the estimates in light of sampling variability. That is, it can be the basis of confidence interval estimation of the parameters.

Let *M* stand for the maximum value of the log-likelihood function (or its limiting value in the $R = 0$ and $R = \Delta$ cases). An intuitively reasonable way to identify values of μ consistent with data in hand is to look for ones which when paired with some appropriate standard deviation produce a log-likelihood value not too much smaller than *M*. That is, one might look for means μ with

$$M - \max_{\sigma>0} l(\mu,\sigma) < c \tag{14}$$

for some appropriate value *c*. Standard large sample theory implies that for large *n*, the set of means satisfying (14) for *c* an upper percentage point of a χ_1^2 distribution can serve as a confidence interval for μ. It is the main technical contribution of [9] to identify positive constants *c* (that depend upon *n* and a desired confidence level) so that the set of means satisfying relationship (14) can serve as a confidence interval for μ for any *n*, small or large. When this idea is applied to the data set in the introduction, a 95% confidence interval for μ is seen to be

$$(1.226,1.294) \quad .$$

In light of this interval, if one were to interpret the author's statement about 1.26 as a statement about μ, he is clearly overly optimistic about the precision of his empirical information.

Note, by the way, that the potentially quite inappropriate "*t*" confidence limits for μ_Y here are (1.223,1.297), which in this case are not radically different from the limits for μ prescribed in [9]. However, it is not at all hard to find cases where the limits are radically different (for example, when $R = 0$) and it is equally easy to give examples where nominally 95% "*t*" limits have actual confidence level for estimating μ near 0 (for example, when σ is very small and μ is about $\Delta / 4$ from a possible rounded value). The virtue of using (14) is that the confidence intervals from [9] hold their nominal confidence level for estimating μ no matter what be μ and σ. We have further found empirically, that when *R* is many times Δ (the data suggest that σ is not small compared to Δ and that the rounding doesn't seem important), the intervals produced reasoning from (14) agree numerically with the "*t*" confidence limits. The likelihood approach thus protects one from the blunder of ignoring rounding when it is important, while reducing to a standard analysis when it is not.

A similar story can be told for estimating σ. "Plausible" values of σ are those with

$$M - \max_{\mu} l(\mu,\sigma) < c \tag{15}$$

for some appropriate *c*. It is the main technical contribution of [10] to identify positive constants *c* (that depend upon *n*, a desired confidence level and whether $R = 0$, $R = \Delta$ or $R > \Delta$) so that the set of standard deviations satisfying relationship (15) serve as a confidence interval for σ. When this idea is applied to the data of [5], a 95% confidence interval for σ is seen to be

$$(0,0.0851) \tag{16}$$

As it turns out, $M - \max_{\mu} l(\mu,\sigma)$ appearing in (15) is decreasing in σ for $R = 0$ and $R = \Delta$ cases. So the one-sided nature of the interval in (16) is completely typical and in agreement with intuition. In contrast, the usual "χ^2" 95% confidence limits for σ_Y can be made to produce two-sided intervals in such cases. For $R > \Delta$ the intervals of [10] are two-sided, and for large *R* they empirically seem to agree numerically with "shortest-length" two-sided "χ^2" confidence intervals for σ. Thus, as in the case of the mean, the likelihood approach protects one from the blunder of ignoring rounding when it is important, while reducing to a standard analysis when it is not.

V. Conclusion

We have hopefully demonstrated to the reader's satisfaction that for purposes of statistical analysis, data read to some nearest unit of observation can not always be treated as if they were real numbers. As a theoretical matter, σ must be several times Δ before there is no important difference between the properties of *X* and those of *Y* and elementary inference methods applied treating observations $y_1, y_2, \ldots, y_n$ as real numbers are reliable guides to the properties of *X*. As a practical matter, one should be comfortable applying those methods to $y_1, y_2, \ldots, y_n$ only when *R* is an order of magnitude larger than Δ. For data less variable than this, the situation is potentially subtle, and use of data analysis methods that explicitly treat observations as the rounded values that they really are is one's only insurance against falling unaware into errors of logic and inference.

REFERENCES

[1] R.M. Gray and D.L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, pp. 2325-2383, October 1998.

[2] B. Widrow, I. Kollár, and M-C Liu, "Statistical theory of quantization," *IEEE Transactions on Instrumentation and Measurement*, vol. 45, pp. 353-361, April 1996.

[3] I. Kollár, "Bias of mean value and mean square value measurements based on quantized data," *IEEE Transactions on Instrumentation and Measurement*, vol. 43, pp. 733-739, October 1994.

[4] K-D Sommer and M. Kochsiek, "Role of measurement uncertainty in deciding conformance in legal metrology," *Organisation Internationale de Métrologie Légale Bulletin*, vol. 43, pp. 19-24, April 2002.

[5] P. Stein, "Careful interpolation yields useful information," *Quality Progress*, vol. 33, pp. 67-69, January 2000.

[6] W.Q. Meeker and L.A. Escobar, “Maximum likelihood methods for fitting parametric statistical models to censored and truncated data,” Chapter 8 in *Statistical Methods for Physical Science*, edited by J. Stanford and S.B. Vardeman, New York: Academic Press, 1994.

[7] W.Q. Meeker and L.A. Escobar, *Statistical Methods for Reliability Data*, New York: John Wiley & Sons, 1998.

[8] A. Stuart and J.K. Ord, *Kendall’s Advanced Theory of Statistics, Vol. 1, Distribution Theory*, 6th Edition, London: Edward Arnold--New York: John Wiley & Sons, 1994.

[9] C-S Lee and S.B. Vardeman, “Interval estimation of a normal process mean from rounded data,” *Journal of Quality Technology*, vol. 33, pp. 335-348, July 2001.

[10] C-S Lee and S.B. Vardeman, “Interval Estimation of a normal process standard deviation from rounded data,” *Communications in Statistics--Simulation and Computation*, vol. 31, pp. 13-34, March 2002.