

Module 7

Accounting for quantization/digitalization effects and "off-scale" values
in measurement

Prof. Stephen B. Vardeman
Statistics and IMSE
Iowa State University

March 4, 2008

Two Seemingly Unrelated Problems in Measurement With a Common Solution

Two apparently unrelated practical problems of measurement are

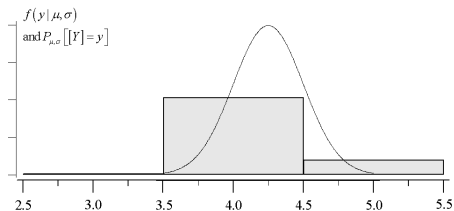
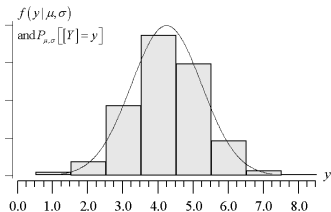
- digitalization/quantizations effects arising because real-world instruments can only report values "to the nearest something," and
- the possibility of a measurement being out of the range of values over which an instrument can be reliably read.

Common naive "fixes" here are in the first case to ignore the effects and in the second case to either "throw away" off-scale values or to substitute the min or max value the instrument will read for any off-scale observation. These "fixes" often grossly misrepresent the truth about a distribution of measurements.

A *real* fix for both these problems is to use the statistical machinery of "censored" data analysis.

Digitalization/Quantization Effects

Suppose that we admit that instead of observing a real number (infinite number of decimal places) value y , we can only observe/record it "to the nearest unit." A distribution for what is observed need look nothing like the underlying distribution for y unless the spread of the underlying distribution is large in comparison to the degree of quantization.



Ignoring this effect when it matters will produce foolish answers.

Digitalization/Quantization Effects (cont.)

Notice that

- this problem is *on top of* the issues raised in Modules 2A and 2B regarding the practical meaning of parameters μ and σ of the distribution of the real number y , and
- ordinary statistical inferences (e.g. t and χ^2 confidence intervals) can at best be expected to capture the properties of the *observed distribution* and *not* the properties of the underlying distribution of real number y .

For the two cases (of integer quantization) pictured on the previous slide

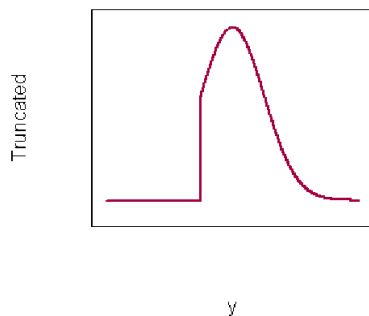
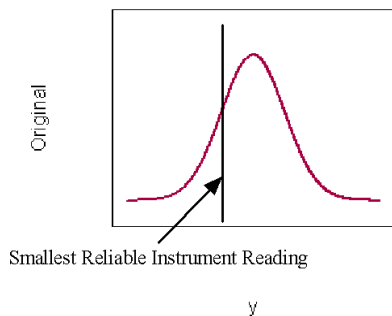
μ	σ	mean of quantized values	std. dev. of quantized values
4.25	1.0	4.2500	1.0809
4.25	.25	4.1573	.3678

- large sample sizes *do not address this problem* (they only help zero-in on the properties of the quantized distribution).

Off-Scale Observations

Treating off-scale observations in naive ways also produces distributions for what is observed that may be little like an underlying distribution for y . (And elementary statistical methods can at best be expected to capture the properties of what is observed.)

- Ignoring off-scale values will produce a "truncated" version of a distribution for y (that can be quite unlike the original distribution).



Off-Scale Observations (cont.)

- Replacing off-scale observations by a corresponding end (or ends) of the scale over which an instrument can be trusted has the effect of replacing a tail (or tails) of a distribution for y with a "spike" (or spikes) at the end point(s) of the scale. Again, these distributions (that can't be pictured conveniently with smooth probability density curves) can be quite unlike the original distribution.

As for the effects of digitalization, large samples do not help address this problem. (They can only show us definitively what the distribution of the *observable* y looks like.)

The "Right" Fix for These Problems: A Censored Data Analysis

A way to "fix" both problems caused by important quantization effects and by off-scale measurements is to properly encode what is really known about observations y from their digital or "off-scale" values into a statistical "censored data likelihood function" and use it to guide the making of inferences. (Vardeman and Lee discuss this in the digitalization context in "Likelihood-based statistical estimation from quantized data" *IEEE Transactions on Instrumentation and Measurement*, 2005, Vol. 54, No. 1, pp. 409-414.)

What follows is an overview of what can be done.

The "Right" Fix (cont.)

For $[y]$ a quantized/digital version of y where an instrument reads "to the nearest Δ " what one knows from the observation is that

$$[y] - .5\Delta < y < [y] + .5\Delta$$

For Φ the standard normal cumulative probability function, if Y is normal with mean μ and standard deviation σ , the probability associated with such a quantized observation is then

$$\begin{aligned} P_{\mu,\sigma} [[y] - .5\Delta < Y < [y] + .5\Delta] &= P \left[Z < \frac{[y] + .5\Delta - \mu}{\sigma} \right] \\ &\quad - P \left[Z < \frac{[y] - .5\Delta - \mu}{\sigma} \right] \\ &= \Phi \left(\frac{[y] + .5\Delta - \mu}{\sigma} \right) \\ &\quad - \Phi \left(\frac{[y] - .5\Delta - \mu}{\sigma} \right) \end{aligned}$$

The "Right" Fix (cont.)

In a similar fashion, if the instrument produces only values in the finite range from L to U , what one knows from an out-of range measurement is

either that $y < L$ or that $y > U$.

(Again assuming Y to be normal) corresponding probabilities are respectively

$$P_{\mu,\sigma} [Y < L] = P \left[Z < \frac{L - \mu}{\sigma} \right] = \Phi \left(\frac{L - \mu}{\sigma} \right) \quad \text{and}$$

$$P_{\mu,\sigma} [U < Y] = 1 - P \left[Z < \frac{U - \mu}{\sigma} \right] = 1 - \Phi \left(\frac{U - \mu}{\sigma} \right)$$

(If it is important to do so, L and/or U could be adjusted by $.5\Delta$ to account for quantization in the stating of the range of the instrument.)

A Censored Data "Likelihood Function"

If, for all observations available one multiplies together terms of one of the three types above, a function of μ and σ that could be called a "censored data likelihood function" is created. That is, the likelihood function is

$$L(\mu, \sigma) = \prod_{\substack{\text{digital} \\ \text{observations } [y]}} \left(\Phi \left(\frac{[y] + .5\Delta - \mu}{\sigma} \right) - \Phi \left(\frac{[y] - .5\Delta - \mu}{\sigma} \right) \right) \\ \cdot \left(\Phi \left(\frac{L - \mu}{\sigma} \right) \right)^{\text{number of below-scale observations}} \\ \cdot \left(1 - \Phi \left(\frac{U - \mu}{\sigma} \right) \right)^{\text{number of above-scale observations}}$$

(L and U may need to be ".5 Δ adjustments" of nominal values.)

Example 7-1

Suppose that y has a Normal distribution with $\mu = 10$ and $\sigma = .5$, but that values of y can be read only to the nearest .1 and that only read values of 9.8 through 10.4 are possible, anything more extreme than that being "off-scale." Under this model, normal probability calculations show that in the long run the following distribution of values would be seen.

Value	r.f	Value	r.f
" < 9.8"	.3085	10.2	.0736
9.8	.0736	10.3	.0665
9.9	.0781	10.4	.0579
10.0	.0796	" > 10.4"	.1841
10.1	.0781		

Example 7-1 (cont.)

Below are $n = 10$ simulated values (to 3 decimals) from the normal distribution for y (with $\mu = 10$ and $\sigma = .5$).

10.283, 10.002, 10.328, 10.343, 9.708, 9.538, 10.465, 9.669, 10.069, 9.390

Under the circumstances described in this example however, what would have been observed was

10.3, 10.0, 10.3, 10.3, " < 9.8", " < 9.8", " > 10.4", " < 9.8", 10.1, " < 9.8"

Example 7-1 (cont.)

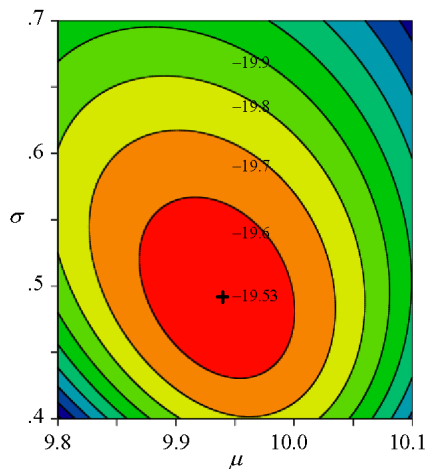
In this context, an appropriate likelihood function would be

$$\begin{aligned} L(\mu, \sigma) = & \left(\Phi\left(\frac{10.35 - \mu}{\sigma}\right) - \Phi\left(\frac{10.25 - \mu}{\sigma}\right) \right)^3 \\ & \cdot \left(\Phi\left(\frac{10.05 - \mu}{\sigma}\right) - \Phi\left(\frac{9.95 - \mu}{\sigma}\right) \right) \\ & \cdot \left(\Phi\left(\frac{10.15 - \mu}{\sigma}\right) - \Phi\left(\frac{10.05 - \mu}{\sigma}\right) \right) \\ & \cdot \left(\Phi\left(\frac{9.75 - \mu}{\sigma}\right) \right)^4 \cdot \left(1 - \Phi\left(\frac{10.45 - \mu}{\sigma}\right) \right) \end{aligned}$$

and this can guide inference about μ and σ .

Example 7-1 (cont.)

Here's a contour plot of the (log) likelihood function for these data.



Using A Likelihood Function

Uses that can be made of a likelihood function include the following.

- Finding the location of "the top" of a likelihood function is a sensible way to estimate of μ and σ .
- How fast the likelihood falls off away from its summit (or the curvature of the likelihood at its top) can be used to set confidence limits on μ and σ . (Details of exactly how this is done are beyond what is sensible to present here.)

Implementing an Analysis Based on a Censored Data Likelihood

Standard statistical software can be used to get confidence intervals from a censored data likelihood function. However, in some cases a slightly indirect route must be taken. This is because the kind of censoring described here has historically been most commonly associated with reliability/life data analyses, not measurement analyses. And where intrinsically positive life-lengths have been studied, it has been standard fare to consider models that treat the *logarithms* of life lengths as normal. So it may be a convenient expedient to 1) exponentiate one's original data (to make "pseudo-lifetime data") and then 2) use life data analysis programs and so-called "lognormal" models to make inferences for μ and σ .

Example 7-1 (cont.)

An exponentiated version of the original data set used earlier (treating the "off-scale" observations as less than 9.75 or larger than 10.45) can be made

- treating values off-scale on the low side as satisfying $\exp(y) < \exp(9.75) = 17154.23$,
- treating the value recorded as 10.0 as satisfying $20952.22 = \exp(9.95) < \exp(y) < \exp(10.05) = 23155.78$,
- treating the value recorded as 10.1 as satisfying $23155.78 = \exp(10.05) < \exp(y) < \exp(10.15) = 25591.10$,
- treating the values recorded as 10.3 as satisfying $28282.54 = \exp(10.25) < \exp(y) < \exp(10.35) = 31257.04$, and
- treating values off-scale on the high side as satisfying $34544.37 = \exp(10.45) < \exp(y)$.

Example 7-1 (cont.)

Here is a TMJMP data sheet and report from use of the TMJMP Survival/Reliability censored data analysis routine. It provides point estimates and 95% confidence limits for μ and σ that take account of the interval censoring and off-scale nature of 5 of the $n = 10$ observations in this example.

The screenshot displays two windows from the JMP software. The left window, titled 'example7-1', shows a data table with the following data:

	Lower	Upper	Count
1		17154.23	4
2	20952.22	23155.78	1
3	23155.78	25591.1	1
4	28282.54	31257.04	3
5	34544.37		1

The right window, titled 'example7-1 - Survival/Reliability of Lower, Upper', shows the 'Product-Limit Survival Fit' report. The 'LogNormal Parameter Estimates' section is expanded, showing the following results:

Parameter	Estimate	Lower 95%	Upper 95%	N Failed
μ	9.9374168	9.3941636	10.298277	5
σ	0.4914818	0.2699302	1.2020566	5

Notice, for example, that the estimated mean and standard deviation correspond to the location of the summit of the (log) likelihood function.