

Comparison of Rounding Methods

By, S.E. Van Bramer
2/13/97 revised 11/8/97

The motivation for this worksheet came after a lengthy discussion of rounding techniques with a colleague in physics. During which I realized that neither one of us had a way to "prove" our point. We were both using select examples to show our points. This completely misses the point of rounding, which is to minimize the accumulation of uncertainty when manipulating experimental data. I wrote this Mathcad document in an attempt to find an unbiased solution to this question.

I start out with a data set consisting of two measurements. Since this is a simulation, I have the luxury of defining the "true" value and the "true" uncertainty. The "Population" is defined below:

Measurement A

Population mean $\mu_a := 1.508$

Population standard deviation $\sigma_a := 0.001$

Measurement B

Population mean $\mu_b := \frac{356.4856}{\mu_a} \quad \mu_b = 236.396286472$

Population standard deviation $\sigma_b := 1$

Next, I simply multiply the two measurements. This is a step frequently used with experimental data. Because I have defined the population it is possible to use error propagation to calculate the standard deviation of the product population.

Product: $x_{true} := \mu_a \cdot \mu_b \quad x_{true} = 356.4856$

Standard deviation of product population. This is the uncertainty in the final results.

$$\sigma_x := x_{true} \cdot \sqrt{\left(\frac{\sigma_a}{\mu_a}\right)^2 + \left(\frac{\sigma_b}{\mu_b}\right)^2} \quad \sigma_x = 1.526416458$$

Now that I have rigorously defined this experimental population I can use it to compare different methods of rounding. Keeping in mind that the purpose of rounding is to provide an easy way to approximate the propagation of error when using experimental data. With the goal of providing a realistic estimate of the uncertainty in the "answer".

First I will generate a matrix of experimental data from the population defined above. This data is a random normal distribution from this population, where the measurement is repeated N times and averaged. This would be typical of a situation where it is important to avoid having error "accumulate" because of a bias in the data processing.

The real "story" here comes from repeating this exercise J times. When we typically discuss which method of rounding is "best" someone comes up with a couple of examples to "prove" their method. In this simulation Mathcad will repeat the exercise a VERY large number of times so that we can compare the results. The number used for J is limited by the RAM and processing speed of your computer.

The Matrix:

Number of pts averaged N := 5 i := 0, 1.. N - 1

Number of experiments J := 500 j := 0, 1.. J - 1

Generate data sets (N measurements of a and b repeated J times)

$$\text{NORM}(\mu_n, \sigma_n) := \mu_n + \sigma_n \cdot \sqrt{-2 \cdot \ln(\text{rnd}(1))} \cdot \cos(2 \cdot \pi \cdot \text{rnd}(1))$$

$$a_{i,j} := \text{NORM}(\mu_a, \sigma_a)$$

$$b_{i,j} := \text{NORM}(\mu_b, \sigma_b)$$

$$x_{i,j} := a_{i,j} \cdot b_{i,j}$$

Now Mathcad will calculate the "results" several different ways.

-short: This method rounds all values <5 down and all values > or = 5 up.

-long: This method rounds all values < 5 down; all values > 5 up; and all values = 5 up if odd, down if even.

-truncate: This method simply truncates all values

-true: This method carries through calculations using all precision available in Mathcad

The Rounding Methods:

$$\text{short}(x) := \text{if}(x - \text{floor}(x) > 0.5, \text{ceil}(x), \text{floor}(x))$$

$$\text{long}(x) := \text{if}\left(\frac{\text{floor}(x \cdot 10)}{10} - \text{floor}(x) \leq 0.4, \text{floor}(x), \text{if}\left(\frac{\text{floor}(x \cdot 10)}{10} - \text{floor}(x) \geq 0.6, \text{ceil}(x), \text{if}\left(\frac{\text{floor}(x)}{2} - \text{floor}\left(\frac{x}{2}\right) > 0.2, \text{ceil}(x), \text{floor}(x)\right)\right)\right)$$

$$\text{truncate}(x) := \text{floor}(x)$$

Create the rounded data sets:

$$\text{data short}_{i,j} := \text{short}(x_{i,j})$$

$$\text{data long}_{i,j} := \text{long}(x_{i,j})$$

$$\text{data truncate}_{i,j} := \text{truncate}(x_{i,j})$$

$$\text{data true}_{i,j} := x_{i,j}$$

Now from the data set, calculate the J "average" results for each data set.

$$\text{mean short}_j := \text{mean}(\text{data short}^{<j>})$$

$$\text{mean long}_j := \text{mean}(\text{data long}^{<j>})$$

$$\text{mean truncate}_j := \text{mean}(\text{data truncate}^{<j>})$$

$$\text{mean true}_j := \text{mean}(\text{data true}^{<j>})$$

Now we can take a look at the "results". The first thing we can do is "average" the J results for each method. If everything works out "right" we will get the "true" value $x_{\text{true}} = 356.4856$.

Results from the "long" rounding method.

$$\text{mean}(\text{mean long}) = 356.492$$

Results from the "short" rounding method.

$$\text{mean}(\text{mean short}) = 356.5372$$

Results from truncating.

$$\text{mean}(\text{mean truncate}) = 356.0552$$

Results without any rounding.

$$\text{mean}(\text{mean true}) = 356.547051913$$

However, this does not give a very complete comparison. Recall that the data set (like all real measurements) has a population distribution. Unless we take an infinite number of samples, the results will vary some from $x_{\text{true}} = 356.4856$.

One statistical technique for determining if there is a "significant" difference between two averages is the t-test. For more information on the t-test, see any statistics textbook (there is a section on this included in most analytical, instrumental, and p-chem textbooks).

In this step I calculate the "t-score" for each experimental result ($J = 500$ results for each technique). The distribution for this t-score is well characterized and can be used for comparison. The larger the value of the t-score, the further the experimental result is from the "true" result. The average t-score for each technique is shown below.

$$t_{\text{true}_j} := \frac{|(x_{\text{true}} - \text{mean}_{\text{true}_j})|}{\sigma_x} \cdot \sqrt{N} \qquad \text{mean}(t_{\text{true}}) = 0.785228961$$

$$t_{\text{short}_j} := \frac{|x_{\text{true}} - \text{mean}_{\text{short}_j}|}{\sigma_x} \cdot \sqrt{N} \qquad \text{mean}(t_{\text{short}}) = 0.804375764$$

$$t_{\text{long}_j} := \frac{|x_{\text{true}} - \text{mean}_{\text{long}_j}|}{\sigma_x} \cdot \sqrt{N} \qquad \text{mean}(t_{\text{long}}) = 0.804614838$$

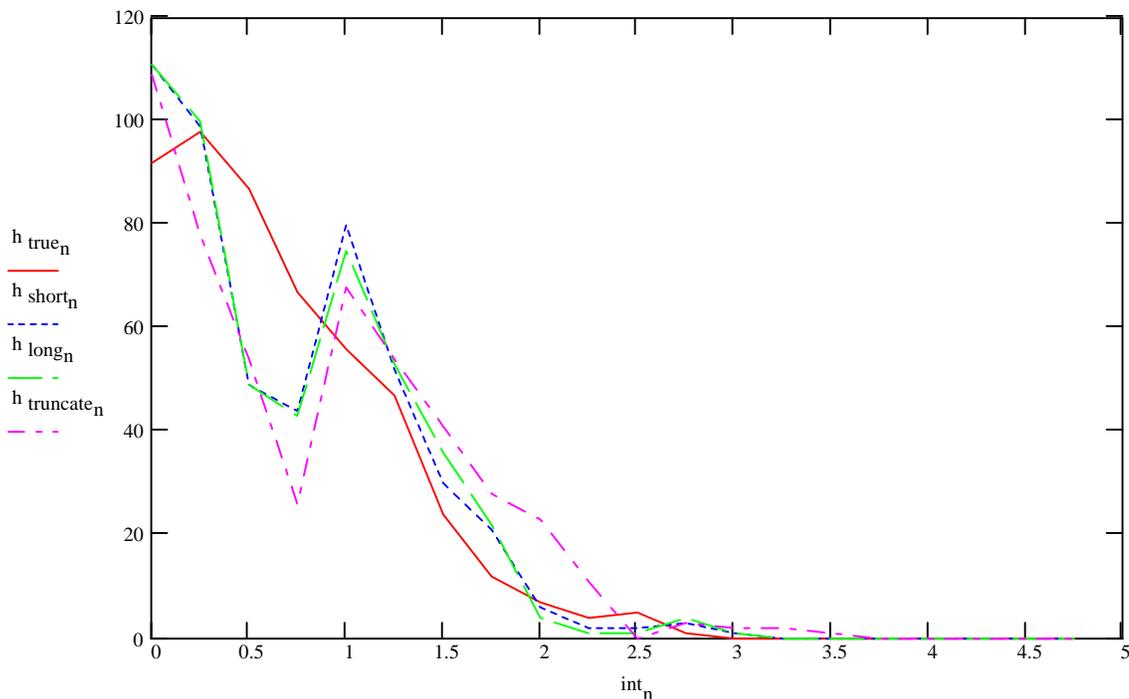
$$t_{\text{truncate}_j} := \frac{|x_{\text{true}} - \text{mean}_{\text{truncate}_j}|}{\sigma_x} \cdot \sqrt{N} \qquad \text{mean}(t_{\text{truncate}}) = 0.934455389$$

These results may be displayed graphically as a histogram, showing the number of experiments with each value for the t-score.

$$k := 0, 1..20 \quad n := 0, 1..19 \quad \text{int}_k := k \cdot 25$$

$$h_{\text{true}} := \text{hist}(\text{int}, t_{\text{true}}) \quad h_{\text{short}} := \text{hist}(\text{int}, t_{\text{short}}) \quad h_{\text{long}} := \text{hist}(\text{int}, t_{\text{long}}) \quad h_{\text{truncate}} := \text{hist}(\text{int}, t_{\text{truncate}})$$

Graph Histograms



The t-test is typically used by selecting a "confidence" interval. This corresponds to the area under the curves shown above. This is expressed as a percentage of the area that is less than a given value.

In this experiment, 90% of the data will have a t-score less than $-qt(0.1, N) = 1.475884049$ and 99% of the data will have a t-score less than $-qt(0.01, N) = 3.364929999$

Next, the area under the curve for each technique is calculated:

At the 90 percent confidence interval:

$$g := 0, 1 .. 1 \quad \text{lim}_g := g - qt(0.1, N)$$

$$\frac{\text{hist}(\text{lim}, t_{\text{true}})}{J} = (0.88)$$

$$\frac{\text{hist}(\text{lim}, t_{\text{short}})}{J} = (0.87)$$

$$\frac{\text{hist}(\text{lim}, t_{\text{long}})}{J} = (0.862)$$

$$\frac{\text{hist}(\text{lim}, t_{\text{truncate}})}{J} = (0.778)$$

At the 95 percent confidence interval:

$$g := 0, 1 .. 1 \quad \text{lim}_g := g - qt(0.05, N)$$

$$\frac{\text{hist}(\text{lim}, t_{\text{true}})}{J} = (0.966)$$

$$\frac{\text{hist}(\text{lim}, t_{\text{short}})}{J} = (0.972)$$

$$\frac{\text{hist}(\text{lim}, t_{\text{long}})}{J} = (0.978)$$

$$\frac{\text{hist}(\text{lim}, t_{\text{truncate}})}{J} = (0.916)$$