# Classroom Simulation: Distributions of Ties Induced by Rounding Continuous Data

Bruce E. Trumbo and Eric A. Suess
Department of Statistics and Biostatistics, California State University, East Bay,
Hayward, CA 94542

**Abstract**
Graphical and numerical illustrations of the effects of moderate to severe rounding of normal data on the values of test statistics, actual significance level, and power for one-sample t tests and paired t tests. Goodness-of-fit tests confirm that rounded normal data are not normal. Combinatorial computations, simulations, and graphs are made using R. The level of programming, exposition, and suggested extensions is suitable for upper division and MS level statistics students.

**Key Words:** IEEE rounding, one-sample tests, R software, simulation, statistics education.

## 1. Introduction

Distributions such as the normal, exponential, and uniform are viewed as "continuous" from a theoretical point of view. However, in practice, there is no such thing as a continuous distribution because observations must be rounded to some number of decimal places. Methods and consequences of rounding have been widely studied. A foundational technical paper on these topics is a paper by Churchill Eisenhart [1], and a good way to find subsequent papers and books on rounding data is to search for papers that reference this paper. (An earlier reference is [2].) Of the thousands of articles on rounding rules at various technical levels and degrees of authenticity, the Wikipedia article (as it appeared in mid-2014) seemed to be one reasonable starting point. Our purpose here is to use simulation (with R) to illustrate some of the consequences of rounding for statistical inference at a level accessible to beginning statistics students.

*The IEEE method of rounding,* implemented in R (among many other statistical software packages) "goes to the even digit" when rounding a 5. That is, round down if the next least significant digit is 1, 2, 3, or 4; to round up if the next least digit is 6, 7, 8, or 9; and, in case the next least digit is 5, to round so that the new final digit is even. The idea is that "on average" rounding of numbers ending in 5 will not systematically bias the data either upwards or downwards. Examples below show simple rounding in R—essentially according to these rules. Occasional exceptions, such as the ones noted below for a *specially chosen* set of numbers, are due to the way decimals are stored as binary numbers in R. Such exceptions are rarely of practical importance, especially when rounding from many decimal places to only a few.

```
> x = c(1.0,-1.1,1.5,2.5,-2.6,2.8)
> cbind(x, x.r = round(x))
        x x.r
[1,]  1.0   1
[2,] -1.1  -1
[3,]  1.5   2
[4,]  2.5   2
[5,] -2.6  -3
[6,]  2.8   3
```

```
> y = x + .05
> cbind(y, y.r = round(y,1))
          y  y.r
[1,]  1.05  1.1  # Exception
[2,] -1.05 -1.1  # Exception
[3,]  1.55  1.6
[4,]  2.55  2.5  # Exception
[5,] -2.55 -2.6
[6,]  2.85  2.8
```

The sample mean and standard deviation are changed only slightly as a result of a modest degree of IEEE rounding as implemented in R, but severe rounding can have consequences of practical importance in particular cases. Below we show results of rounding 1000 simulated standard normal observations to four decimal places and to integers, respectively. In the simulation run shown below, rounding to integers yielded many ties and only eight unique values.

```
> set.seed(1066);  x = rnorm(1000);  x4 = round(x, 4);  x0 = round(x)
> c(mean(x), mean(x4), mean(x0))
[1] -0.005800276 -0.005799100 -0.021000000
> c(sd(x), sd(x4), sd(x0))
[1] 1.009047 1.009047 1.046735
> c(length(unique(x)), length(unique(x4)), length(unique(x0)))
[1] 1000  988    8
```

## 2. Ties Induced by Rounding Uniform and Normal Data

In theory, there are no ties in continuous data. However, as soon as data are rounded to some number of decimal places, it is possible to see ties. In this section, we investigate the proportion of ties induced by rounding in various situations.

### 2.1. Uniform Data

*Combinatorial approach.* Suppose an experiment consists of choosing six observations at random from the uniform distribution on the interval [0, 1) and truncating each observation to show only the first decimal place. We use truncation rather than rounding for simplicity. [An equivalent experiment would be one-place *rounding* of six values sampled from UNIF(–0.05, 0.95).] Only ten truncated values are possible: 0.0, 0.1, ..., 0.9. By a simple combinatorial argument, the probability of getting no ties (duplicated results) is $P(10, 6)/10^6 = 0.1512$, where $P(10, 6) = 10!/(10 – 6)!$. This can be computed in R as `prod(10:5)/10^6`. (The argument parallels that of the famous birthday matching problem [3, 4].) Below we show how to simulate this experiment. The simulation shown resulted in 1 tie; two additional iterations resulted in 2 and 0 ties, respectively.

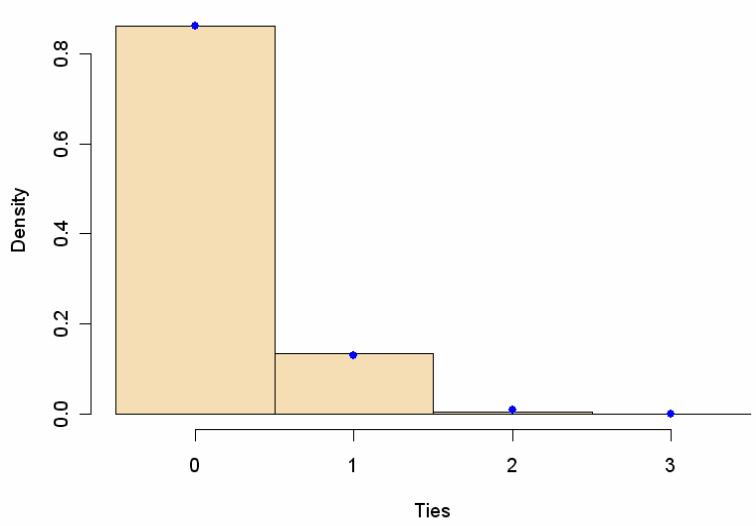
```
> uf = floor(10*runif(6))/10   # 'floor' truncates to the integer part
> uf;  length(unique(uf))
[1] 0.0 0.9 0.6 0.1 0.0 0.7    # 0.0 appears twice
[1] 5                          # five unique values; one tie
```

*Simulating the distribution of ties.* From above, we know that the probability of getting no ties is 0.1512 when six randomly chosen observations from UNIF(0, 1) are truncated to one decimal place. However, it is not so easy to use combinatorial methods to find the entire distribution of the number $T$ of such ties. By simulating the experiment 100,000 times, as shown below, we can get a useful approximation of this distribution, which has $E(T) \approx 1.317$ and $SD(T) \approx 0.816$. With 100,000 iterations, simulated probabilities are accurate to two (maybe three) decimal places. Although very unlikely, one may also see five ties. Our simulation is not adequate to approximate the probability of this event, but a simple combinatorial argument gives $P\{T = 5\} = 6 \times 10^{-6}$.

```
set.seed(1066)
m = 100000;  n = 6   # number of iterations, sample size
t = numeric(m)       # vector of 0s, each element changed in loop
for (i in 1:m)  {  uf = floor(10*runif(n))/10
  t[i] = n - length(unique(uf)) }

>  summary(as.factor(t))/m
      0       1       2       3       4  # Values of t
0.15083 0.45283 0.32812 0.06521 0.00301  # Simulated P{T = t}
```

By contrast, consider the number of ties if we truncate the six observations to the next lower hundredth. Then ties are much less likely, and we have $P(T = 0) = P(100, 6)/100^6 = 0.8583$ (to four places). If we were to assume that ties occur only when some of $C(6, 2) = 15$ independent pairs of data are equal, each with probability 1/100, then we would have $T \sim$ BINOM(15, .01) with $E(T) = 0.15$ This distribution is very similar to Poisson with mean $\lambda = C(6, 2)/100 = 0.15$. The argument above is not quite right, but it can give a useful approximation when ties are the relatively rare. (Of course, ties need not occur as independent pairs; for example, three or more observations may share the same value.) For our specific problem, if we assume $T \sim$ POIS(0.15), then $P(T = 0) = \exp(-0.15) = 0.8607$, which agrees with the combinatorial result to two places.




| | P{T=t} | |
|---|---|---|
| t | Simu-lated | POIS Approx |
| 0 | 0.858 | 0.861 |
| 1 | 0.136 | 0.129 |
| 2 | 0.006 | 0.010 |
| 3 | 0.000 | 0.000 |
| ... | | |
| E(T) | 0.147 | 0.150 |
| V(T) | 0.138 | 0.150 |

**Figure 1:** Distribution of the number of ties when six random observations from UNIF(0, 1) are truncated to the next lower 0.01. The histogram shows simulated probabilities based on 100,000 iterations of this experiment. The heavy dots atop histogram bars show probabilities from the approximating distribution POIS(0.15).

### 2.2. Normal Data

*A rounded normal sample is no longer normal.* The null hypothesis of a Shapiro-Wilk test is that data come from some member of the normal family (mean and standard deviation unspecified). When normal data are rounded, this test is sometimes sufficiently powerful to detect that the data are no longer normal. The following R program illustrates rounding in which samples of size $n = 100$ from NORM(50, 3) are rounded to the nearest integer.

```
set.seed(12)
m = 10000;  pv = pvr = numeric(m); mu = 50; sg = 3; n = 100
for (i in 1:m) {  x = rnorm(n, mu, sg)
  pv[i] = shapiro.test(x)$p.value;  pvr[i] = shapiro.test(round(x))$p.value  }
mean(pv < .05);  mean(pvr < .05)

> mean(pv < .05);  mean(pvr < .05)
[1] 0.0483       # Rejection for about 5% of unrounded samples
[1] 0.2696       # Rejection for over 25% of rounded samples
```

In this situation, the simulation results show that the Shapiro-Wilk test detects the nonnormality of the rounded data in over a quarter of the instances. (Rejection occurs even more often for rounded samples with larger sample sizes $n$.) For the particular seed used in the simulation above, the first rounded sample out of ten thousand is one of the 2696 detected as nonnormal. The results of the Shapiro-Wilk test for that sample are shown below for original (left) and rounded data, followed by a normal probability plot (QQ-plot) of the rounded data in Figure 2.

```
> set.seed(12)
> x = rnorm(100, 50, 3)
> shapiro.test(x)

  Shapiro-Wilk normality test

data:  x
W = 0.9894, p-value = 0.6201
```

```
> set.seed(12)
> xr = round(rnorm(100, 50, 3))
> shapiro.test(xr)

  Shapiro-Wilk normality test

data:  xr
W = 0.9738, p-value = 0.04341
```

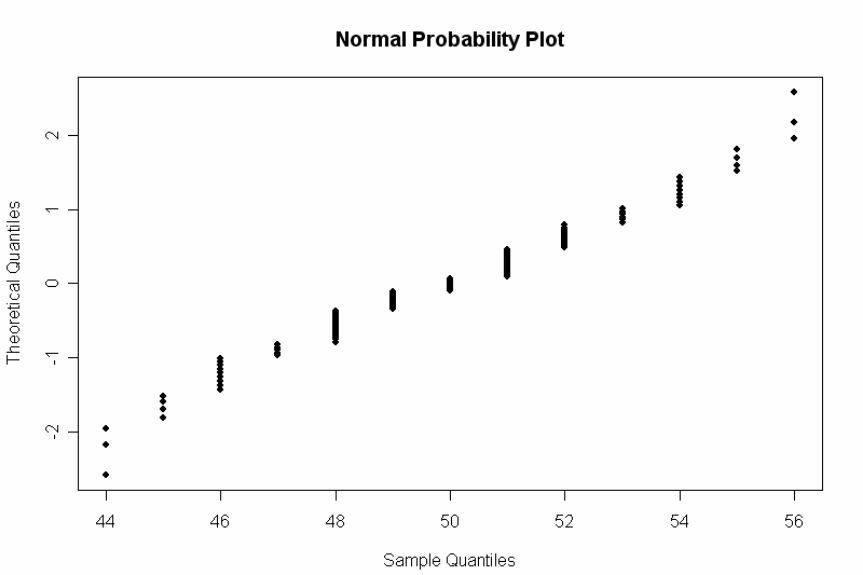



**Figure 2:** Normal probability plot of 100 observations from NORM(50, 3), rounded to integers. Among the 100, there are only 13 different values in the rounded sample.

*Simulation of the distribution of ties.* With smaller samples or less severe rounding than in Figure 2, there are fewer opportunities for ties. Then it may be possible to model the number of ties as having approximately a Poisson distribution—and with a mean that we can roughly guess in advance of the simulation. We need to recognize that such an approximation cannot be as good as for uniform data because rounded values have a wide variety of probabilities of occurrence. Assuming that the "most likely" rounded values lie within about two standard deviations of the mean, there will be about $10^d 4\sigma$ of these values, where $\sigma$ is the population standard deviation and $d$ is the number of decimal places after rounding. Here we assume incorrectly that these values occur equally often and hope for the best, using a Poisson distribution with approximate mean $\lambda = C(n, 2)/(10^d 4\sigma)$.

```
set.seed(1776)
m = 100000;  n = 30;  mu  = 50;  sd = 4;  d = 2
la = choose(n,2)/(4*10^d*sd)
t = numeric(m)
for (i in 1:m)  {
  x = round(rnorm(n, mu, sd), d);  t[i] = sum(duplicated(x))  }
mx = max(t);  cutp = -1:mx + .5
hist(t, prob=T, breaks = cutp, col="wheat")
points(0:mx, dpois(0:mx, lam), pch=19, col="darkgreen")
points(0:mx, dpois(0:mx, mean(t)), col="purple")
mean(t);  t.avg;  sd(t);  var(t);  la
r = rle(sort(t));   t.obs=r$values;  p.sim = round(r$lengths/m, 3)
cbind(t.obs,p.sim,round(dpois(0:mx,la),3), round(dpois(0:mx,mean(t)),3))
```

```
> mean(t);  sd(t);  var(t);  la
[1] 0.30186
[1] 0.5434367
[1] 0.2953235
[1] 0.271875
```

| t | Simulated | POIS(.272) | POIS(.302) |
|---|---|---|---|
| 0 | 0.737 | 0.762 | 0.739 |
| 1 | 0.227 | 0.207 | 0.223 |
| 2 | 0.033 | 0.028 | 0.034 |
| 3 | 0.003 | 0.003 | 0.003 |

The near agreement of E($T$) and V($T$) suggests that the distribution of integer-valued $T$ may be approximately POIS(0.302). Also, it seems that our initial approximate value of $\lambda = 0.272$ is not far off in this particular case. In Figure 3(a) the histogram bars show the simulated probabilities, solid dots show POIS(0.272) and asterisks show POIS(0.302). The first of these models, with a mean guessed in advanced, provides a surprisingly good fit, but the second model is better because it uses a value closer to the true mean. The last two lines of code above provide the probabilities shown at the end of the program. Simulated values should be accurate to two (perhaps three) places. Among the 100,000 simulated samples, there were a few instances with 4 and 5 ties, but they are not apparent when proportions are expressed to three places.

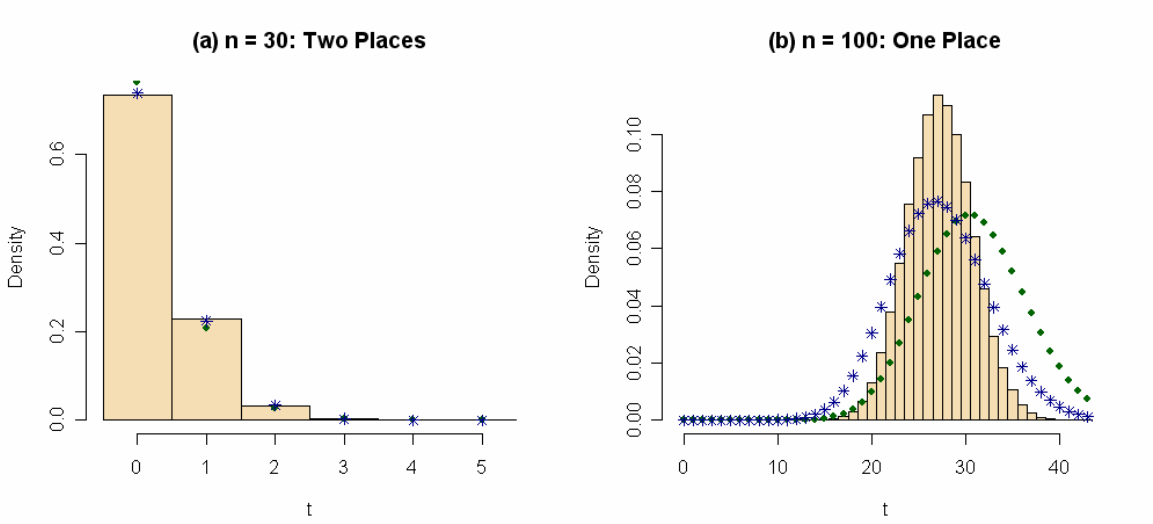



**Figure 3:** (a) Histogram of the simulated distribution of ties when $n$ = 30 observations from NORM(50, 4) are rounded to two decimal places. Heavy dots show POIS(.272), where Poisson the mean was guessed in advance of the simulation, and asterisks show POIS(.302), where the mean is the simulated value of E($T$). (b) When $n$ = 100 observations are rounded to one place, there are many more ties, and no Poisson distribution comes close to fitting the distribution of $T$.

By contrast, when the sample size is relatively large and the number of likely rounded values is relatively small, the average number of ties can no longer be considered as rare and independent events. Thus, the distribution of $T$ is no longer approximated by any Poisson distribution. In particular, Figure 3(b) shows the simulated distribution of the number of ties when samples of $n$ = 100 observations from NORM(50, 4) are rounded to only one decimal place.

### 3. Some Consequences for One-Sample Inference

The distribution theory for one-sample t tests is based on normal data. We have seen that rounded observations from a normal population are no longer normal. The distribution theory for the Wilcoxon signed-rank test is based on continuous data, and adjustments are necessary if rounding induces ties. In this section we illustrate effects of aggressive rounding on the significance level and power of t tests.

**3.1. One-sample t tests.**
We commented in Section 1 that moderate rounding has relatively little effect on the sample mean and standard deviation. So we might suppose rounding has almost no effect on t procedures, which are based on the sample mean and standard deviation. But that is not quite the end of the story. Here we use simulated samples of size $n = 20$ taken from normal populations with $\sigma = 1$, rounded to integer values. Specifically, we look at rejection probabilities of a two-sided t test of $H_0$: $\mu = 0$, assuming first $\mu = 0$ against $H_a$: $\mu = 0.7$, using $\sigma = 1$ for both simulations.

*Significance level.* Under $H_0$, the P-value of a t test on unrounded data is a random variable distributed UNIF(0, 1). Not surprisingly, the P-values of t tests on $m = 100{,}000$ samples nearly UNIF(0, 1), the probability of Type I error (probability of rejection) is $\alpha = 5\%$, and the $T$ statistic has Student's t distribution with $\nu = 19$ degrees of freedom.

The $T$-statistic is computed from the sample mean and standard deviation. For unrounded normal data these statistics are independent, as illustrated in the left panel of Figure 4. Rounding introduces discreteness, and also some dependence as shown in the right panel, so that the $T$ statistic does not have exactly a t distribution. (But by symmetry, the sample mean and standard deviation are uncorrelated: $|r| < 0.005$.)

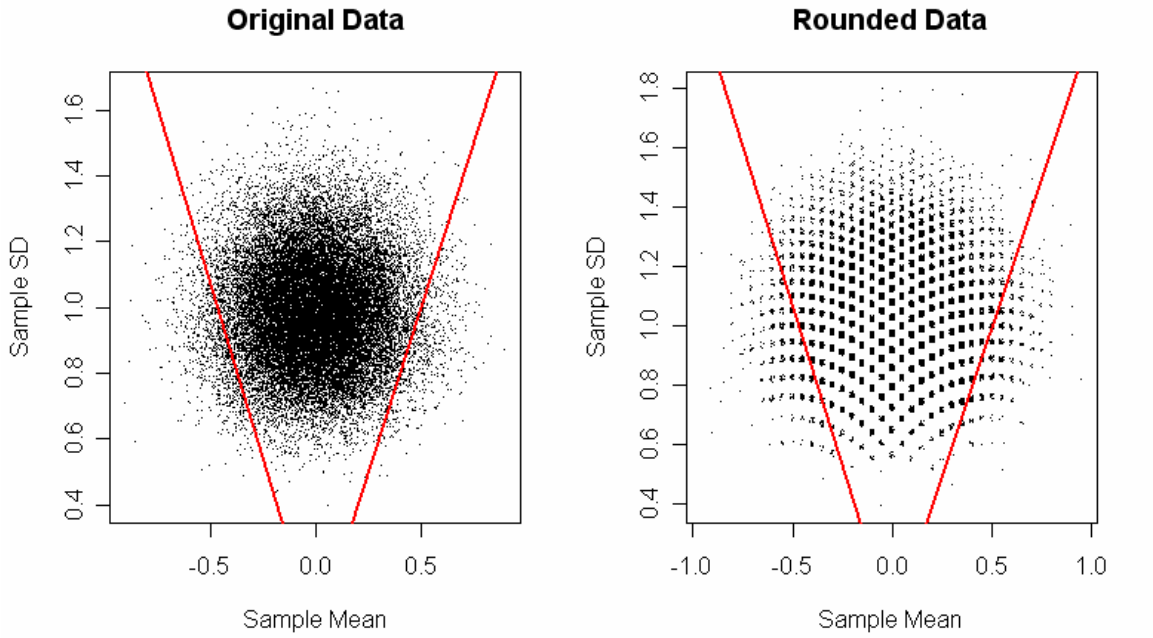



**Figure 4:** Plots of the sample standard deviation against the sample mean. The plot at the right shows the effects of rounding. Points outside the slanted lines represent samples for which $H_0$ is wrongly rejected at a nominal 5% level (based on critical values ±2.093). To avoid overplotting many coincident points, the plot at right is slightly jittered. Each plot shows results for the first 30,000 of 100,000 samples of size 20 from standard normal (before rounding).

When simulated data were rounded to integers, we found only 626 distinct values of the $T$ statistic among $m = 100{,}000$ cases; see Figure 5. The most obvious effect on $T$ of rounding is the peculiar behavior near 0. (If the mean is exactly 0, then the standard deviation is irrelevant.) But it mainly the tails of the distribution that matter in inference. The discreteness evident in this figure gives rise to an actual rejection probability of 5.3 ± 0.14 percent—noticeably above 5%.

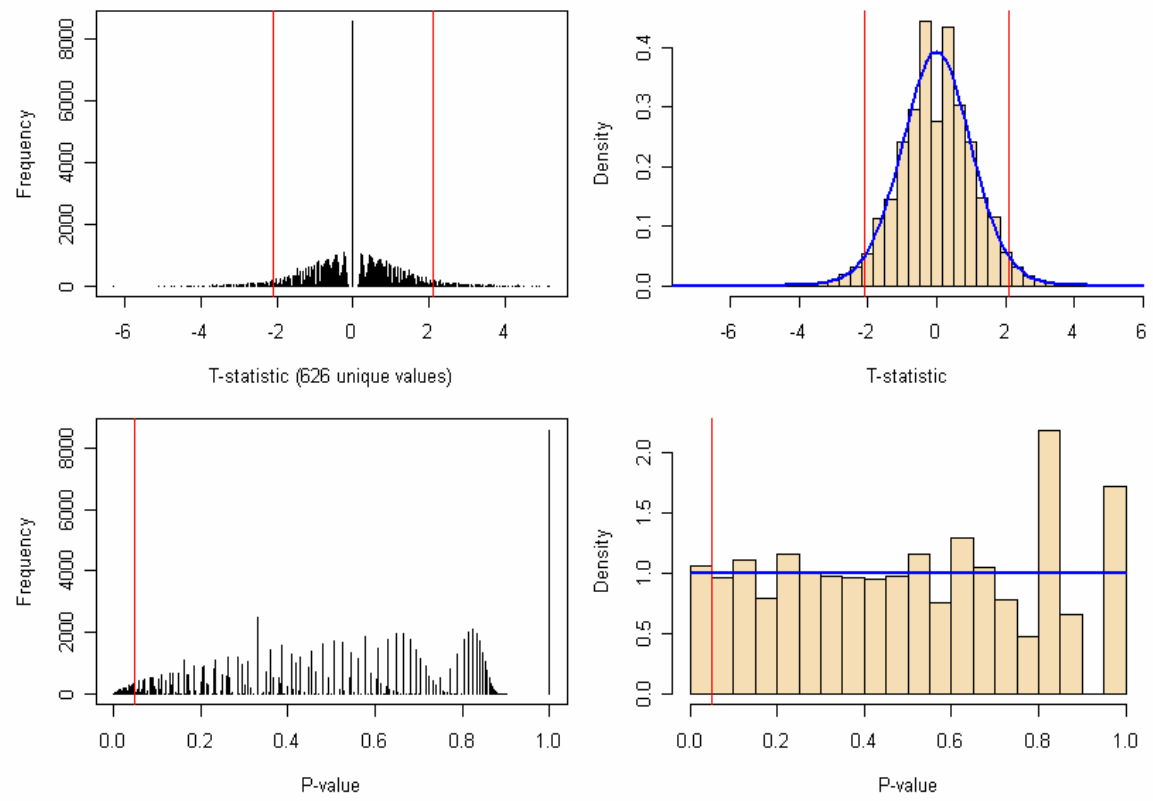
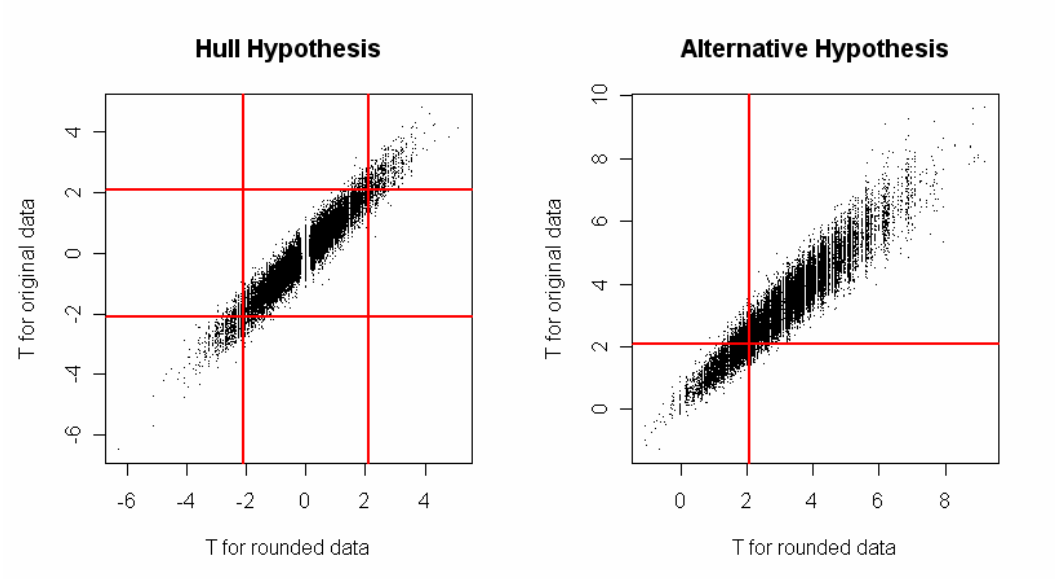



**Figure 5:** Under $H_0$: 100,000 simulated samples of size 20 from NORM(0, 1) rounded to integers. Top row: Graphs show that *T* values are not distributed as T(19). Bottom row: P-values are not UNIF(0, 1). Densities through both histograms show exact distributions for unrounded data.

Perhaps more important, there were many samples in which the decision whether to reject depended on whether data where rounded. This is illustrated in the left side of Figure 6, where each point represents one of the first 30,000 iterations and the accompanying tabulation (accurate to almost three places) is for all 100,000 iterations. Thus at the borderline, an unscrupulous investigator could sometimes "achieve rejection" for a particular experiment by arbitrarily rounding the data or not, to obtain a bogus level of about 6.6% for a nominal "5% level" test (0.013 + 0.038 + 0.015 = 0.066). It is important to realize that any such tampering would not be an adjustment in the *critical value* of the test, but a change in the *data* being analyzed.

*Power.* Now suppose $\mu_a = 0.7$ (still with $n = 20$ and $\sigma = 1$ as above). Then a computation using the noncentral t distribution shows the power (probability of rejection) of the one-sample t test of the hypothesis $H_0$: $\mu = 0$ against the alternative $H_a$: $\mu = 0.7$ to be $\pi(0.7) = 0.8435$.

For data rounded to integers, the *T* statistic is only approximately distributed according to a noncentral t distribution. A simulation with 100,000 iterations shows that the power against $H_a$ is closer to 81.7% than to 84.4%. Thus, rounding gives both larger $\alpha$ and smaller power. Again here, there are many datasets for which the decision whether to reject depends on the choice whether the data were rounded. Improperly making this choice in favor of rejection one could achieve a bogus "power" of about 79.8 + 1.8 + 4.6 percent, or about 86.2%. Results from 30,000 simulated datasets are plotted in the right side of Figure 6, and the tabulation there is based on 100,000 iterations and rounded to three places (approximately the precision of the simulation). Figure 7 shows graphs analogous to two of those in Figures 4 and 5, but illustrating results for rounded data under the alternative hypothesis $H_a$: $\mu = 0.7$.




| Original | Rounded: Accept | Rounded: Reject | Total |
|---|---|---|---|
| Accept | 0.934 | 0.015 | 0.949 |
| Reject | 0.013 | 0.038 | 0.051 |
| Total | 0.947 | 0.053 | 1 |

| Original | Rounded: Accept | Rounded: Reject | Total |
|---|---|---|---|
| Accept | 0.138 | 0.018 | 0.165 |
| Reject | 0.046 | 0.798 | 0.843 |
| Total | 0.183 | 0.817 | 1 |

**Figure 6:** At left: If the null hypothesis $H_0$: $\mu = 0$ is true ($n = 20$, $\sigma = 1$), a test with critical value $t^* = 2.093$ of rounded data has actual significance level about 5.3%. The central square of the plot contains points representing about 93.4% of simulated datasets, for which $H_0$ is accepted whether or not data are rounded. At Right: For the alternative $H_a$: $\mu = 0.7$, the power is 84.4% for original data and only 81.7% for rounded data.

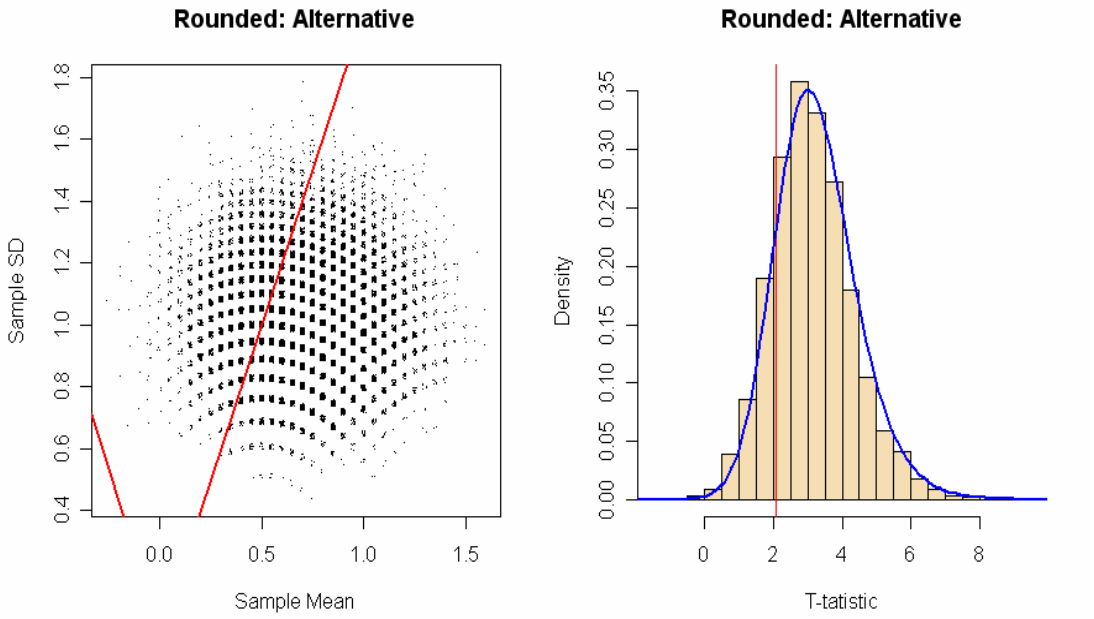



**Figure 7:** Plots for rounded data under $H_a$: $\mu = 0.7$. At left: A plot of the sample SD against the sample mean, analogous to the right-hand side of Figure 4. At right: The simulated distribution of $T$ statistics for rounded data under the alternative (histogram) only roughly matches the skewed noncentral t density for unrounded data (curve). Compare with the upper-right panel of Figure 5.

### 3.2 An Example Based on Real Data: Rounding Paired Data

In a study of the heights of young men [5], two anthropologists, call them A and B, made careful measurements of the heights of 41 students at a boarding school. Each of the investigators made two measurements in the evening and two again the following morning. They attempted to measure heights to the nearest millimeter.

Apparently measuring heights to the nearest millimeter is a challenging task. For each investigator, the median absolute difference of the two evening measurements was about 2mm. Also, across the entire study, measurements by B tended to be 2 or 3mm larger than those by A. This systematic discrepancy between measurements by A and B turned out to be too small to spoil the main finding of the study—that students shrink in height by about a centimeter during their daily activities (with full height restored after a night's sleep). However, to pursue optimal precision, it would have been desirable to know about the difference in measurement techniques between A and B early on, perhaps to agree on more consistent measurement procedures. Accordingly, we take a speculative look at a few of their earliest measurements—specifically the first measurement each investigator made on each of the first 20 subjects in the evening.

*Significance of a paired t test can be spoiled by rounding.* A two-sided paired t test shows that the initial measurements (in millimeters) made by B were significantly larger than those made by A by about 1.75mm. Data and some results from R are shown below. Thus the two investigators could have discovered early on that they were not taking measurements in exactly the same way. For the data below, `t.test(A, B, pair=T)$p.value` returns 0.023.

```
A = c(1717, 1528, 1454, 1778, 1664, 1573, 1662, 1709, 1633, 1782,
      1815, 1788, 1729, 1711, 1714, 1743, 1715, 1593, 1747, 1663)
B = c(1724, 1526, 1454, 1775, 1668, 1572, 1667, 1709, 1639, 1782,
      1814, 1788, 1732, 1713, 1723, 1744, 1718, 1595, 1748, 1662)
```

By contrast, suppose the investigators had decided that measuring heights to the nearest millimeter is futile, and recorded measurements rounded to the nearest centimeter. Then the mean of the differences between the measurements made by the two investigators is no longer significantly different from 0: `t.test(round(A/10), round(B/10), pair=T)$p.value` returns 0.1036. Apparently, rounding these height measurements to centimeters would have been a bad idea.

Moreover, inference from data rounded to centimeters is even more problematic than may be seen just from the results of the paired t test above. Differences of observations rounded to the nearest centimeter are highly discrete, putting the validity of the t test in question. A Shapiro-Wilk test on these differences very strongly rejects normality (P-value $3.4 \times 10^{-5}$). Also, the Wilcoxon signed-rank test is inappropriate here because of the large number of ties and zeros among these differences. A permutation test (for these particular data, equivalent to a sign test) has P-value 0.21875.

*Examples such as the previous one are not rare.* Based on information in the initial 20 pairs of measurements, it seems reasonable to model data vector $A \sim \text{NORM}(1700, 90)$, rounded to integer millimeters, and $B \sim A - \text{NORM}(1.75, 3)$, also rounded to integers. Among $m = 100{,}000$ simulated samples using this model there were only about 9200 unique values of the paired $T$ statistic, but one can see that the power of a 5% level test is about 69.7%. This is essentially the same power as obtained from the noncentral t distribution, assuming continuous data:

```
phi = 1.75*sqrt(20)/3  # noncentrality parameter
1-diff(pt(qt(c(.025,.975), 19), 19, phi)) # Power: n=20, δ=1.75, σ=3, unrounded
[1] 0.6968201
```

However, when the simulated $A$ and $B$ were rounded to *centimeters* before taking differences, there were about 170 iterations out of $m$ in which it was not possible to compute the $T$ statistic because of a null standard deviation. And among the computable $T$ statistics, the proportion leading to rejection was only about 31%. Based on all of the data in [5], one might model the $n = 41$ pairs as $A \sim$ NORM(1700, 80) and $B \sim A -$ NORM(2, 3). Then the power of a paired t test at level 5% is 98% or 99% for data rounded to millimeters before taking differences and only 69% or 70% for data rounded to centimeters (all $T$ statistics computable). Here is the simulation for centimeters.

```
set.seed(1234);  m = 100000;  n = 41
A.sim = round(rnorm(m*n, 1700, 80));  B.sim = round(A.sim - rnorm(m*n, 2, 3))
DTAD = matrix(round(A.sim/10) - round(B.sim/10), nrow=m)
D.bar = rowMeans(DTAD);  D.sd = apply(DTAD, 1, sd)
T.obs = D.bar*sqrt(n)/D.sd
mean(abs(T.obs) > qt(.975, n-1)) # Approx. power: n=40, δ=.2, σ=.3, rounded
[1] 0.69574
```

Based on our simulations, it seems to be particularly counterproductive to round paired data before computing differences. Generally speaking, data should rarely be rounded before doing statistical inference. This is true even when the precision of their final digits may be in question. Rounding should be done at the final step, when results of statistical procedures are reported. Reporting the key finding of [5], for example, one might say that a 95% confidence interval for the difference between morning and evening heights is $9.56 \pm 0.86$ mm, (8.70 mm, 10.42 mm) or (0.87 cm, 1.04 cm). In the end, reporting the margin of error as $\pm 0.86327$ mm seems pointless.

## 4. Activities and Extensions

Finally, we suggest a few activities based on topics discussed above and some extensions to additional types of tests and population distributions. These might be homework problems, classroom activities or individual student projects. Some would require writing simple R programs; some of the programs we used for this paper are available from the authors.

- *Birthday matching problem.* If one assumes equally likely birthdays among $n = 23$ randomly chosen people, the number $Y$ of birthday matches has $P\{Y = 0\} = 0.43$ and $E(Y) = 0.81$. Discuss whether $Y$ is "approximately" Poisson with mean $\lambda = C(23, 2)/365 = 0.6932$? In what other ways is $Y$ different from, and similar to, the number of ties $T$ in our §2.1? What facts about the distribution of the number of birthday matches (or ties in rounded data) can approximated by simulation, but not easily found by combinatorics?

- *Plots of sample mean against standard deviation.* For a t test $n$-dimensional data have been reduced to two dimensions in Figure 4. Find the equations of the boundary lines in Figure 4. For unrounded data, which points represent samples whose confidence intervals include $\mu = 0$? Make similar plots for the alternative where $\mu = 0.7$. Why does the plot of $\bar{X}$ against $S$ for original normal data suggest that these two statistics are independent? Also, make similar plots for samples of size five from exponential data to illustrate dependence of $\bar{X}$ and $S$, compute their correlation. Repeat for data from BETA(.5, .5), where the correlation is zero, but $\bar{X}$ and $S$ are obviously dependent.

- *P-value considered as a random variable.* Consider a two-sided t test of unrounded normal samples of size $n$ where the null hypothesis is true. Explain why the P-value is a random variable distributed as UNIF(0, 1). For unrounded data, make histograms similar to those in Figure 5. Repeat for the alternative where $\mu = 0.7$, but without the theoretical lines.

- *Noncentral t distribution.* Use R functions **pt** and **dt** with noncentrality parameter $\psi = \Delta\sqrt{n}/\sigma$ to discuss power of a t test. Here, $n$ is sample size, $\Delta = |\mu_0 - \mu_a|$, and $\sigma$ is the population standard deviation; use $\psi$ as the third parameter of each of the R functions.

- *Wilcoxon signed-rank test.* The null distribution is assumed to be symmetrical and continuous with differences centered at 0. Rounding may produce ties or zeros or both, in which case R uses a normal distribution to give an approximate P-value (along with a warning message that it is not exact). Explore the effects of various degrees of rounding on the these approximations.

- *Nonnormal data.* In well-specified scenarios, consider the effect of rounding exponential and Laplace data. (Laplace data may be simulated by randomly changing the sign of exponential variates.) Explore the effect of rounding on data from these distributions. Also, use simulation (suggested R code below) to illustrate that the difference between independent exponentials with rate $\lambda$ is Laplace with $\mu = 0$, $\sigma^2 = 2/\lambda^2$, and that the mean of $n$ independent exponential observations with rate $\lambda$ is GAMMA($n$, $n\lambda$).

```
m = 100000;  lam = 3;  y = rexp(m, lam) - rexp(m, lam)
mean(x);  sd(x);  sqrt(2)/lam;  plot.ecdf(y)
curve(.5*exp(lam*x), min(y), 0, col="green", add=T)
curve(1-.5*exp(-lam*x), 0, max(y), col="green", add=T)
```

```
m = 100000;  n = 5;  lam = 3;  DTA = matrix(rexp(m*n, lam), nrow=m)
s = rowMeans(DTA);  mean(s);  sd(s);  1/(sqrt(n)*lam)
plot.ecdf(s);  curve(pgamma(x, n, n*lam), col="green", add=T)
```

**References**


[1] Eisenhart, Churchill: Effects of rounding or grouping data. In Eisenhart, Churchill; Hastay, Millard W.; Wallis, W. Allen: *Selected Techniques of Statistical Analysis for Scientific and Industrial Research, and Production and Management Engineering*, 187–223. McGraw-Hill, New York (1947). Available online: eBook and Texts > American Libraries > University of Florida George A. Smathers Libraries > University of Florida Duplicates.

[2] Kelley, Truman: How many figures are significant? *Science,* Vol. 60, No.1562, 524 (1924).

[3] Feller, William: *An Introduction to Probability Theory and Its Applications,* Vol.1, 2nd ed, Wiley, New York (1957). An early mention of the birthday matching problem in a probability text book.

[4] Suess, Eric A.; Trumbo, Bruce E.: *Introduction to Probability Simulation and Gibbs Sampling with R.* Chapter 1. Springer, New York (2002). Treatments of the birthday matching problem using combinatorial and simulation methods.

[5] Majumdar, D. N; Rao, C. R.: Bengal Anthropometric Survey, 1945, *Sankhya*, Vol.19, Parts 3 and 4 (1958). Data and analyses of the student height study are also available in [6] and [7].

[6] Rao, C. R.: *Statistics and Truth: Putting Chance to Work,* International Cooperative Publishing House, Fairland. MD (1989).

[7] Trumbo, Bruce E.: *Learning Statistics with Real Data,* Chapter 4, Duxbury, Pacific Grove CA (2002).