Classroom Simulation: Distributions of Ties Induced by Rounding Continuous Data

Bruce E. Trumbo and Eric A. Suess

Dept of Statistics and Biostatistics
California State University, East Bay
Hayward, CA 94542

bruce.trumbo@csueastbay.edu
eric.suess@csueastbay.edu

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 nonnormality of rounded normal data.
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.

INTRODUCTION

Foundational paper is Churchill Eisenhart: "Effects of rounding or grouping data." In ***Selected Techniques of Statistical Analysis for Scientific and Industrial Research, and Production and Management Engg***, 187–223. McGraw-Hill, New York (1947), & online.

The IEEE rounding, implemented in R and other statistical software packages "goes to the even digit" when rounding a 5. Thus, "on average" rounding of numbers ending in 5 will tend not to bias the data either upwards or downwards.

Occasional exceptions in R, 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. Exceptions are rarely of practical importance because rounding is usually from many decimal places to a few.

```
> x = c(1.0,-1.1,1.5,
        2.5,-2.6,2.8)
x.r = round(x)
> cbind(x, x.r)
        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 SD are changed only slightly as a result of a modest degree of IEEE rounding, but severe rounding can have consequences of practical importance in particular cases.

Here are results of rounding 1000 simulated standard normal observations to 4 and 0 decimal places.

Rounding to integers yields many ties & 8 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
```

HOW MANY TIES?

Truncate $n = 6$ obs. from UNIF(0, 1) to $d = 1$ place.

Combinatorics: Similar to birthday matching problem.

$$\text{P}\{\text{No ties}\} = \text{P}\{T = 0\} = \text{P}(10, 6)/10^6 = 0.1512.$$

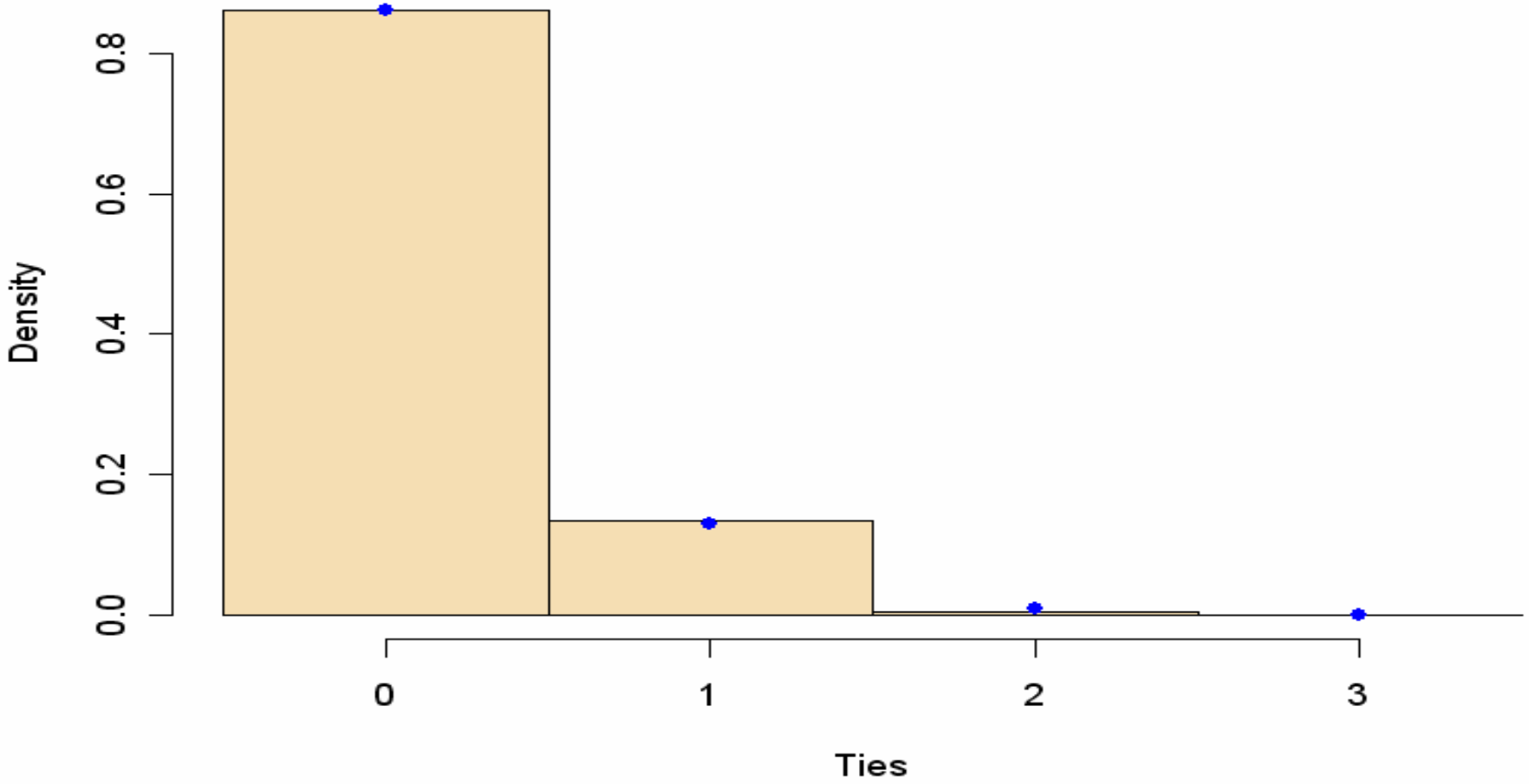
Simulation:

```
set.seed(1066)
m = 100000;  n = 6 # number of iterations, sample size
t = numeric(m)     # vector of 0s, 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  # Sim. Dist'n
```

So $\text{P}\{T = 0\} \approx 0.15083$, $\text{E}(T) \approx 1.317$, $\text{SD}(T) \approx 0.816$.

Truncate $n = 6$ obs. from UNIF(0, 1) to $d = 2$ places.

Simulation: $\mathrm{P}\{T=0\} \approx 0.86$, $\mathrm{E}(T) \approx 0.15$, $\mathrm{V}(T) \approx 0.14$.



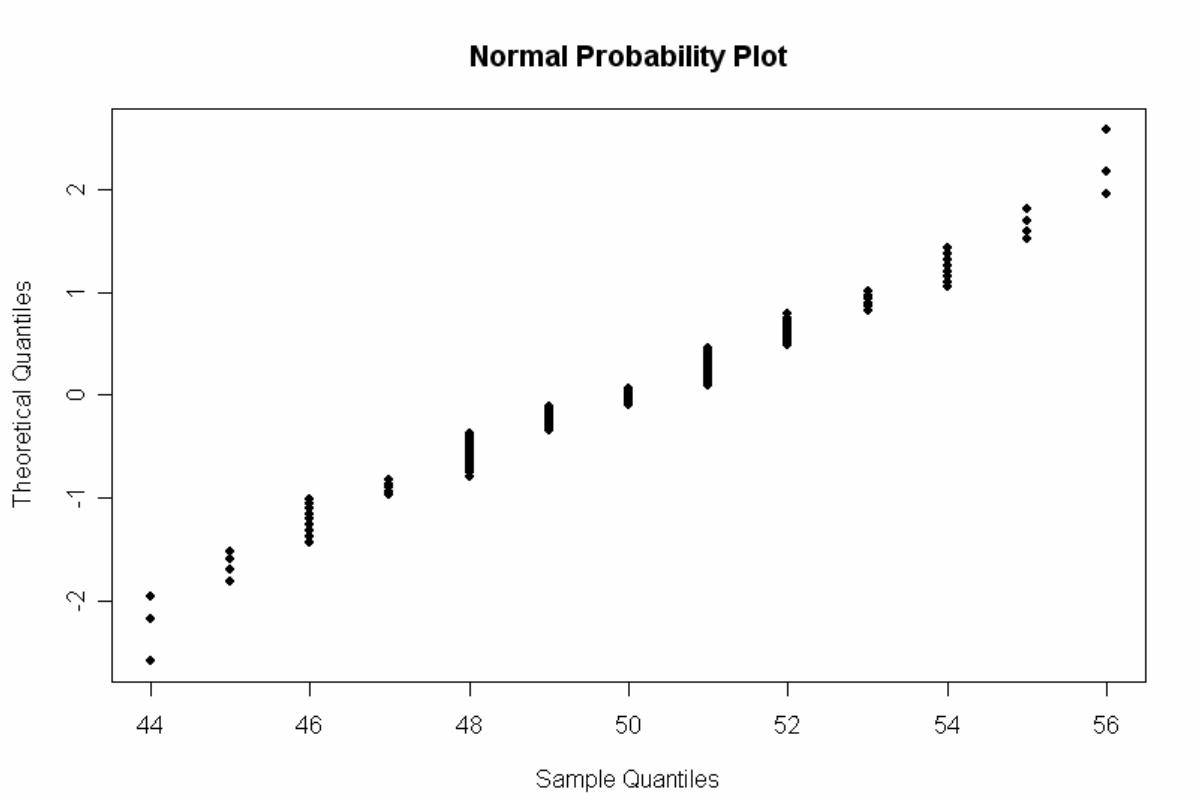
Histogram bars from simulation. Dots from POIS(0.15).

Poisson approx. has mean $\lambda = \mathrm{C}(n, 2)/10^d = 0.15$.

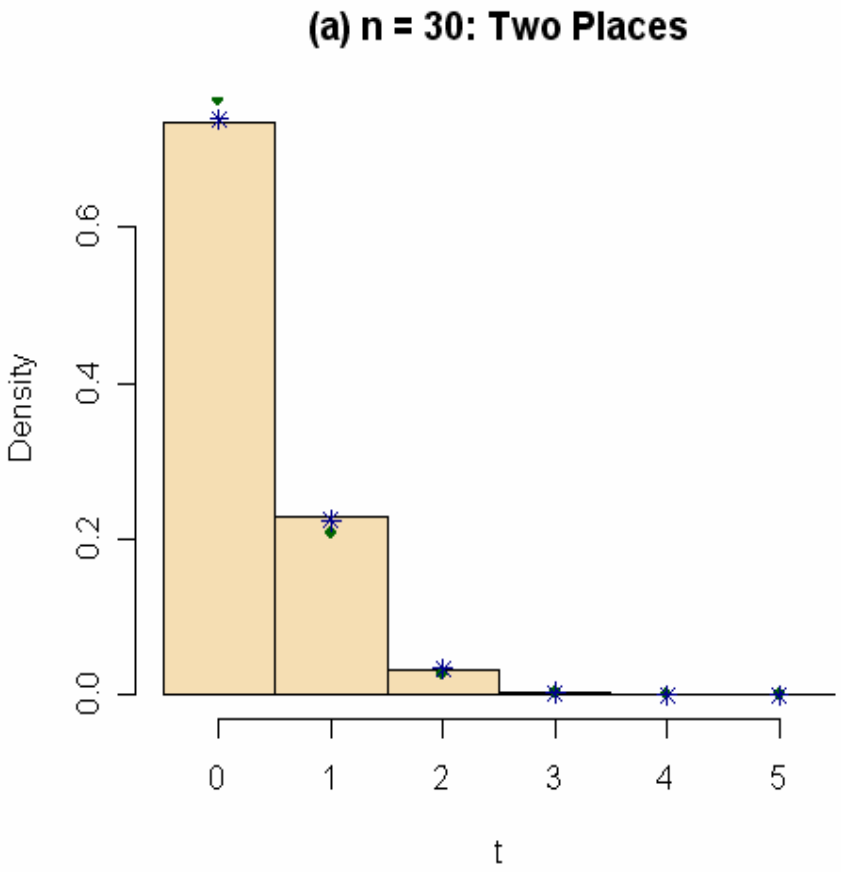
Rounding NORM(50, 3) to integers: n = 100.

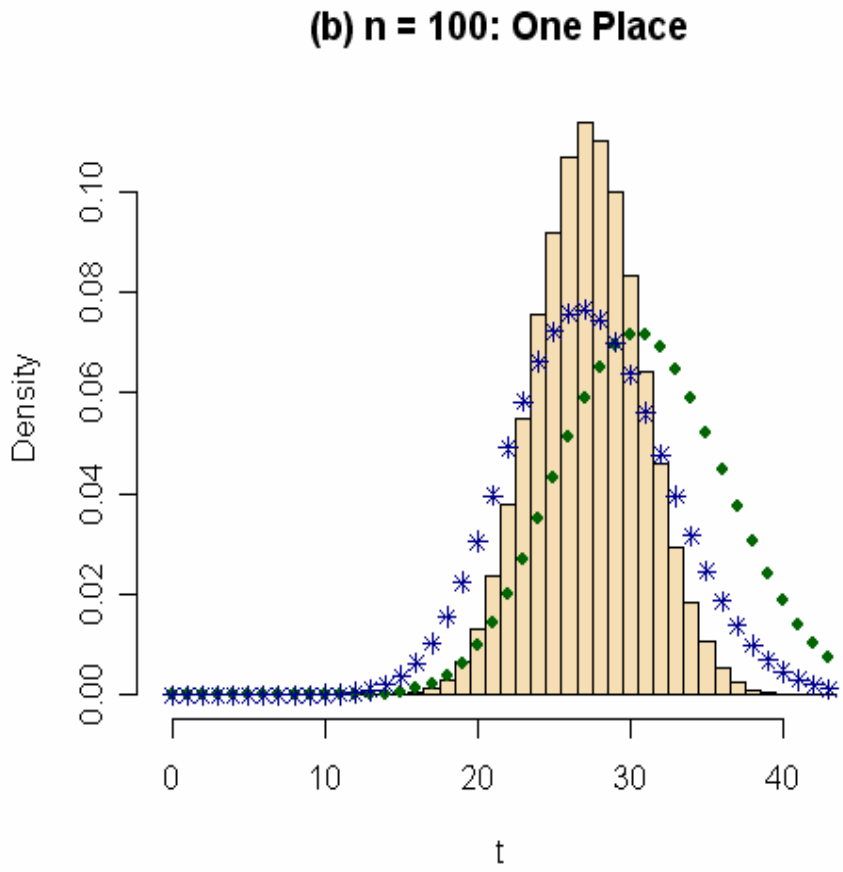
Now nonnormal. Power of Shapiro-Wilk test is ≈ 26%.

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

        Shapiro-Wilk normality test

data:  round(rnorm(100, 50, 3))
W = 0.9738, p-value = 0.04341  # Normality rejected here
```



Rounding NORM(50, 4): Distributions of ties.





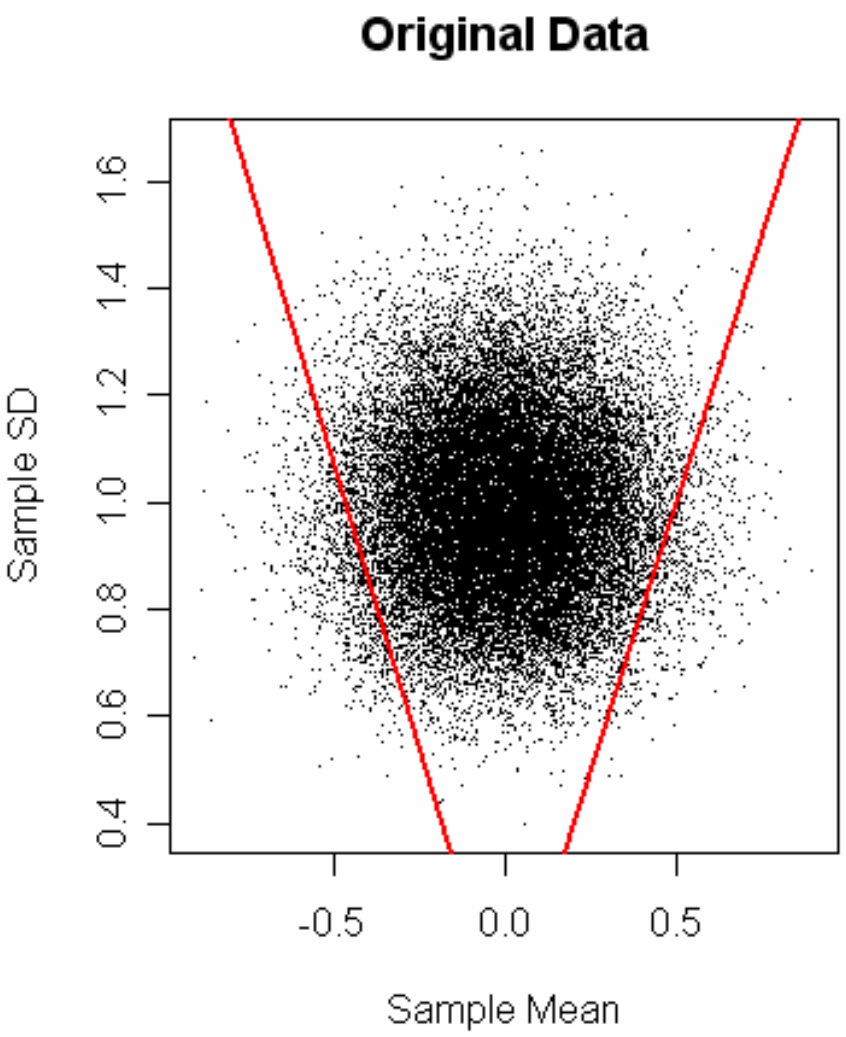
Left: Relatively few ties,100,000 simulated samples. Dots from POIS($\lambda = 0.272$), where λ is guessed from C(30, 2) pairs and about $v = 4\sigma\times10^2$ values; *s use $\lambda = 0.302$ from simulation.

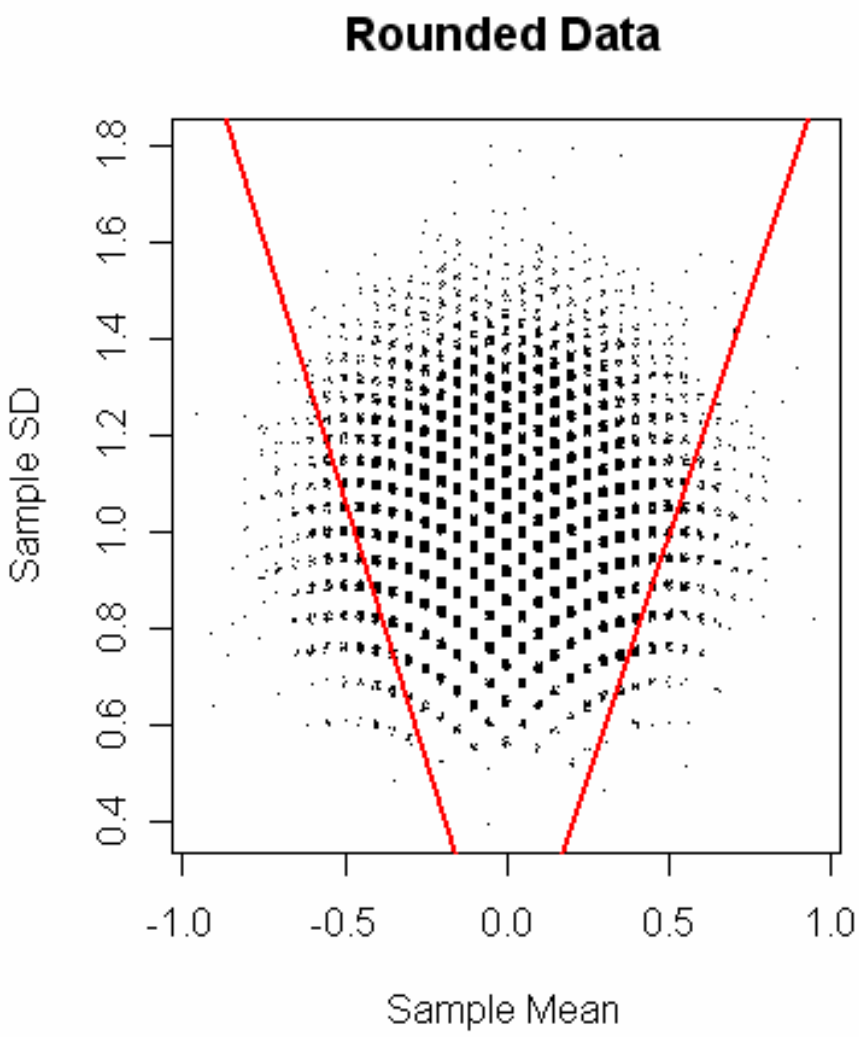
Right: Too many ties, so ***no*** Poisson approximation is useful.

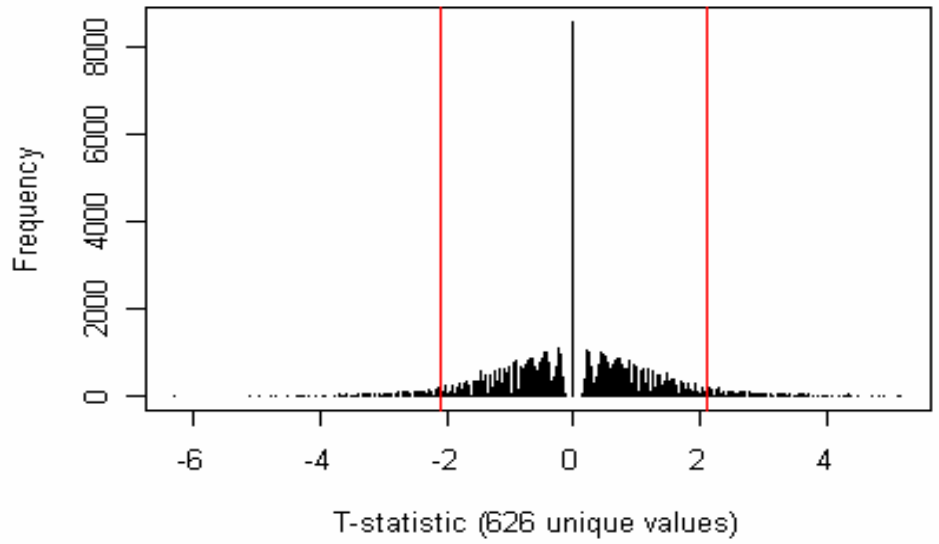
CONSEQUENCES FOR 1-SAMPLE INFERENCE

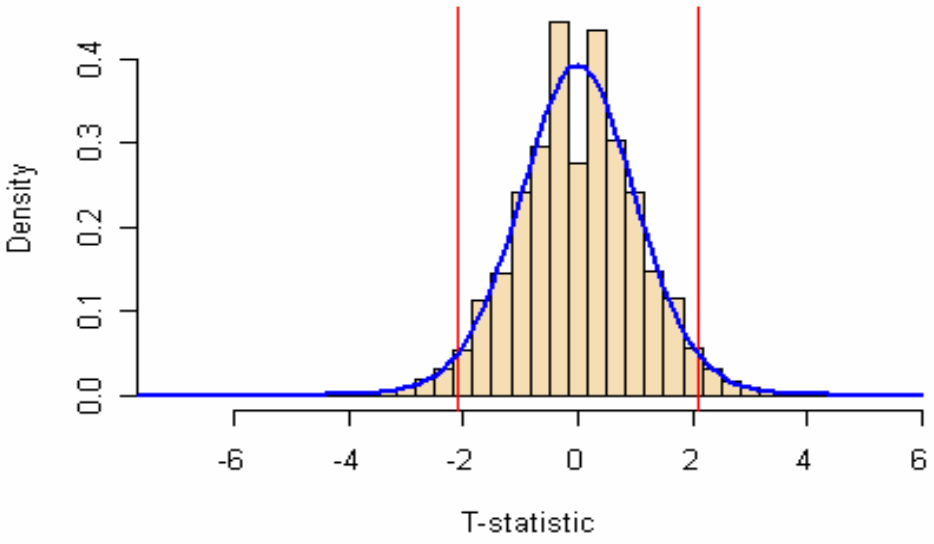
One-sample t test: Normal, $n = 20$, $\sigma = 1$. $d = 1$ place

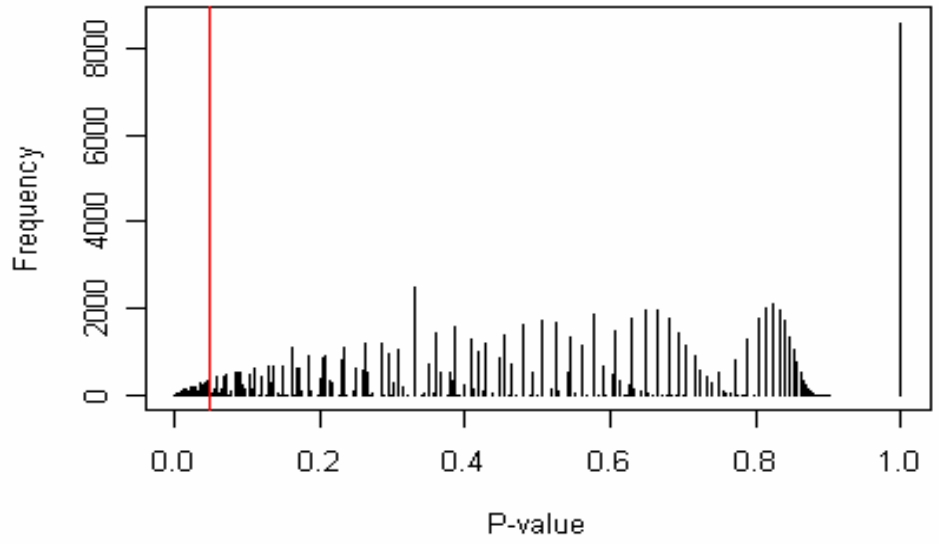
Null hypothesis H_0: $\mu = 0$. (Rejected outside boundaries)

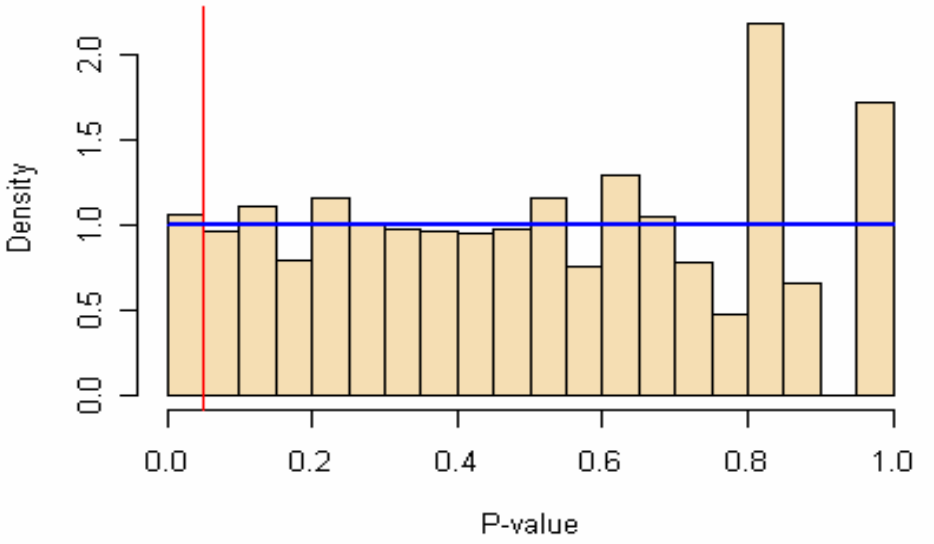








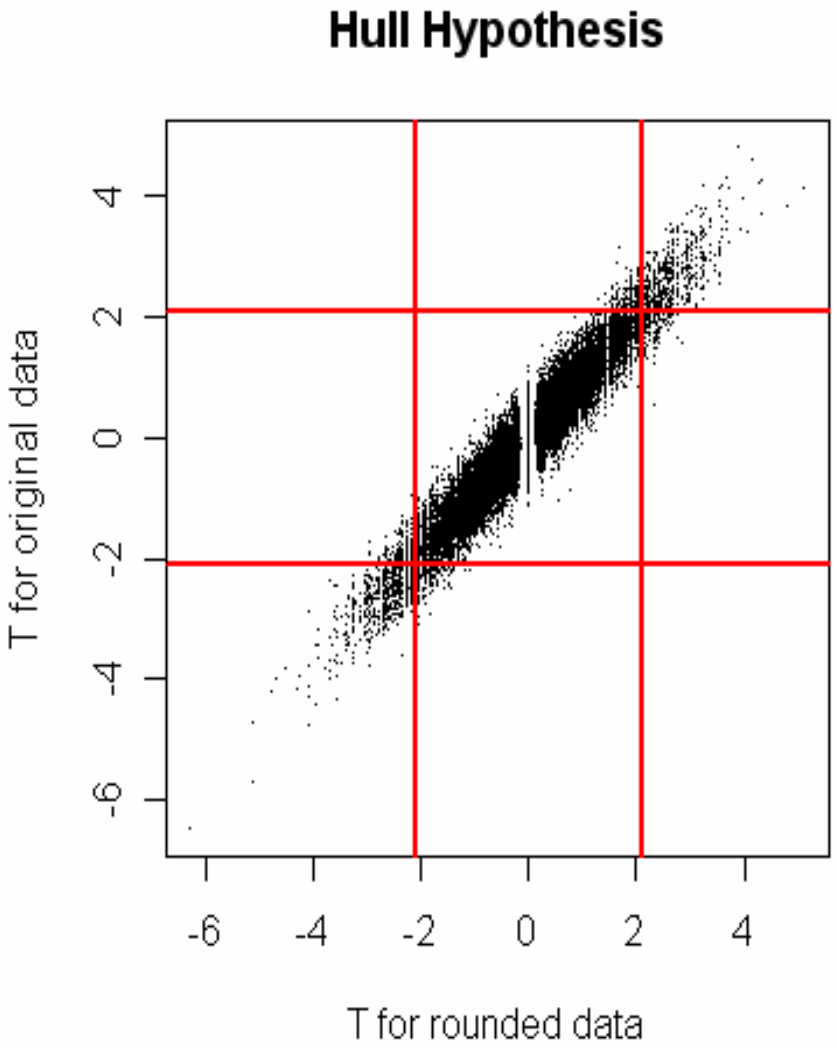


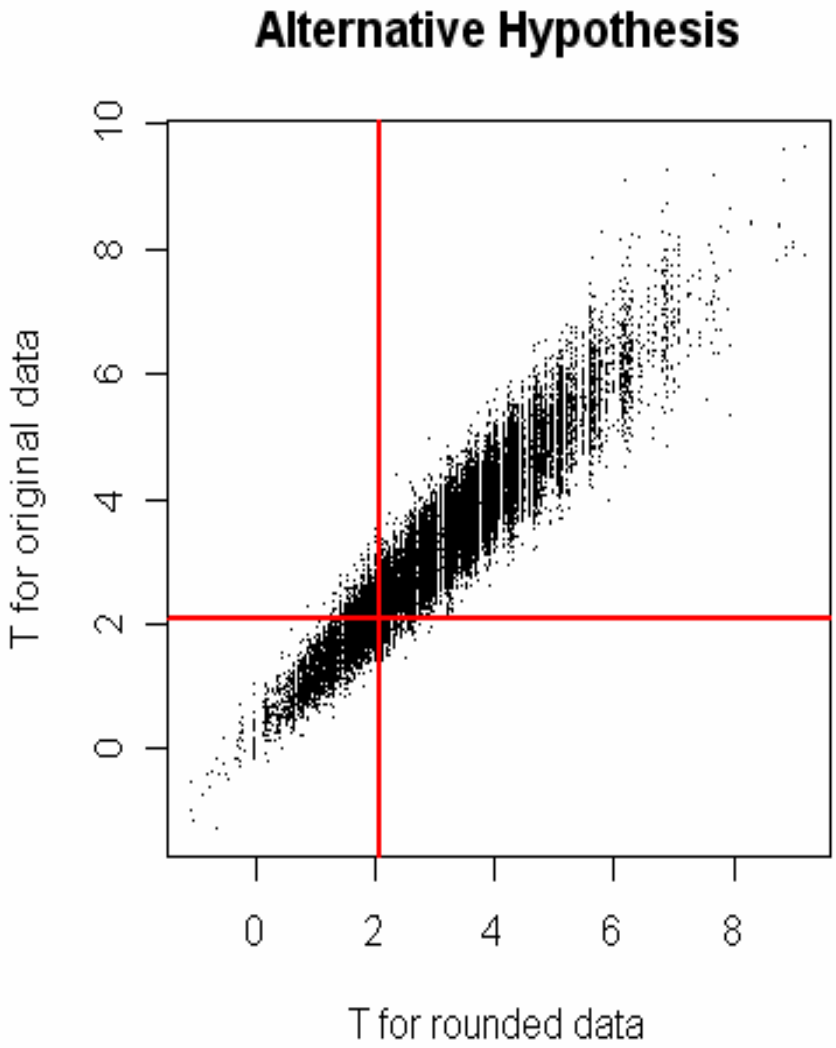


Rounded. ***Top:*** *T* values not T(19). Only 626 unique values of *T* among 100,000 simulated. ***Bottom:*** P-values not UNIF(0, 1). (Heavy lines at right show distributions for unrounded data.)

Rejection probabilities under H_0: $\mu = 0$ and H_a: $\mu = 0.7$ ($\sigma = 1$)

Reject if $|T| \geq 2.093$. Graphs show 30,000 samples of size 20.





Tabulations for original and rounded data based on 100,000 samples of size $n = 20$.

Significance level under H_0: $\mu = 0$ (higher when rounded)

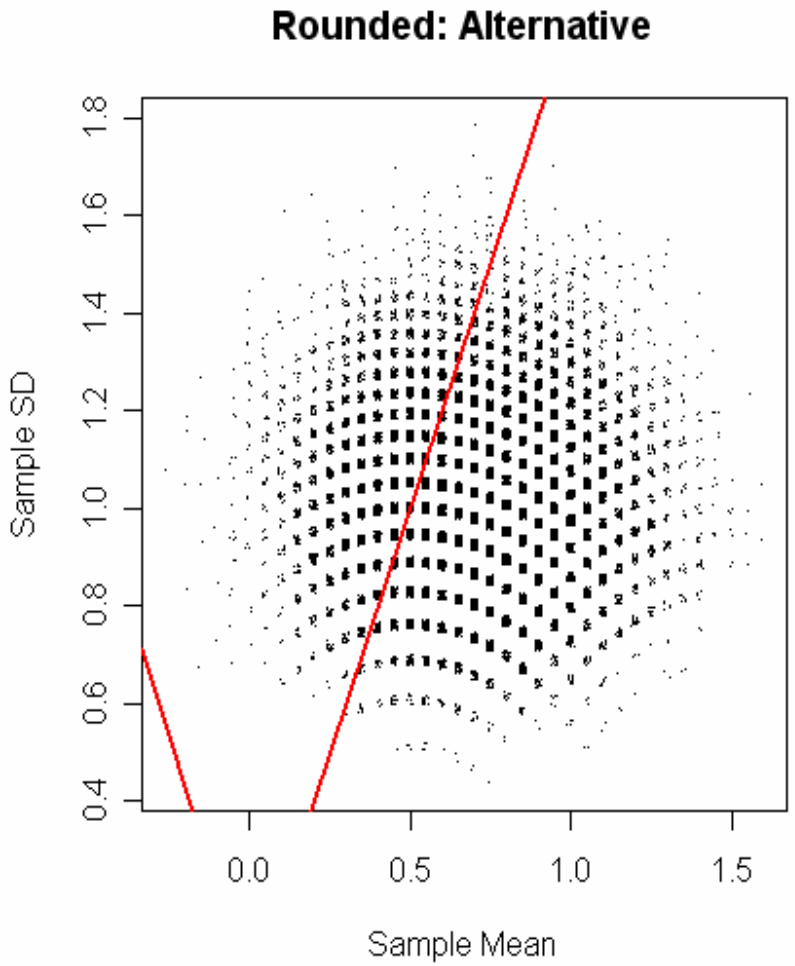
	Rounded		
Orig	Acc	Rej	Tot
Acc	0.934	0.015	0.949
Rej	0.013	0.038	0.051
Tot	0.947	0.053	1

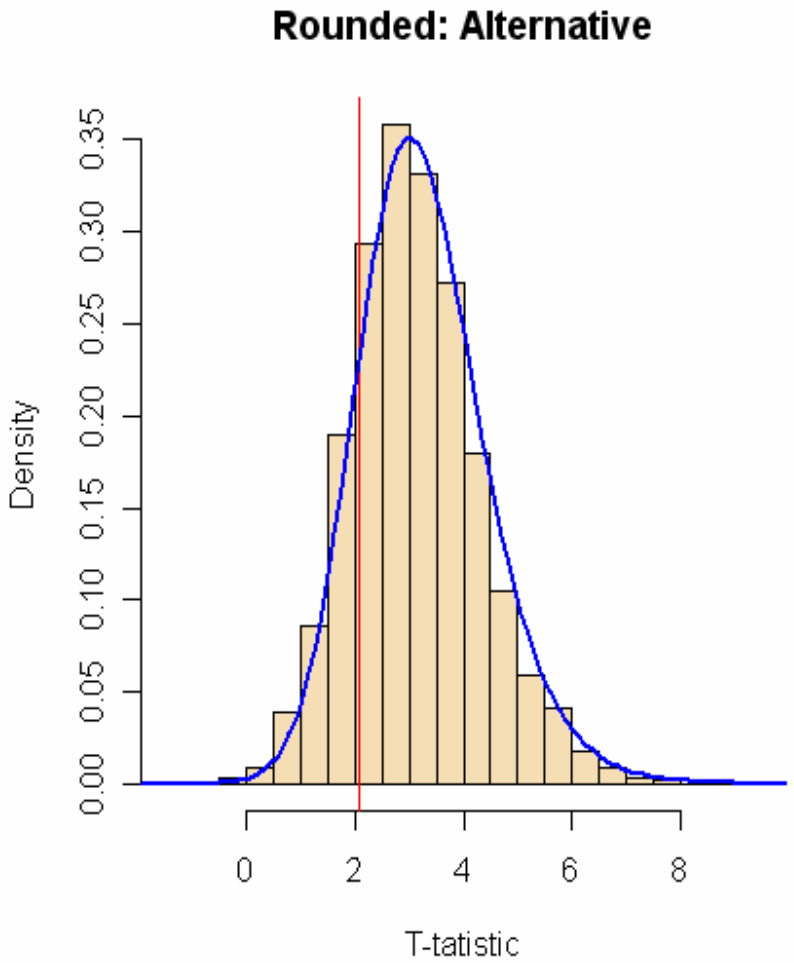
Power against alternative H_a: $\mu = 0.7$ (lower)

	Rounded		
Orig	Acc	Rej	Tot
Acc	0.138	0.018	0.165
Rej	0.046	0.798	0.843
Tot	0.183	0.817	1

Cheating possible: If original data accept H_0, then try rounding for "bogus power" = 0.018 + 0.843 = 0.862.

Rounded data: Power against the alternative H_a: $\mu = 0.7$.
At left: sample SD vs. mean. ***At right:*** simulated distribution of *T* statistics (histogram) does not quite match the slightly skewed noncentral t density for unrounded data (curve).





Consequences of rounding: Real data, paired t

Investigators A and B measuring heights of 20 students in mm. [Data from Majumdar & Rao: Sankhya (1958)]

```
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)
```

Unrounded mm:

```
t.test(A, B, pair=T)$p.value
```

returns P-value = 0.023 ⇒ Signif. difference

Rounded to cm:

```
t.test(round(A/10), round(B/10), pair=T)$p.value
```

returns P-value = 0.102 ⇒ NO signif. difference

Modeling student height data (incl. multiple measurements of 41 students), *and simulating* with 100,000 iterations.

```
set.seed(1234);  m = 100000;  n = 41
A.sim = round(rnorm(m*n, 1700, 80))     # millimeters
B.sim = round(A.sim - rnorm(m*n, 2, 3)) # millimeters
DTAD = matrix(round(A.sim/10)-round(B.sim/10), nrow=m)#cm
D.bar = rowMeans(DTAD);  D.sd = apply(DTAD, 1, sd);
T.obs = D.bar*sqrt(n)/D.sd
mean(abs(T.obs[D.sd > 0]) > qt(.975, n-1)) # Prob of rej
```

Original data

H_0 true, significance level 5.0% (agrees with theory)

Ha, power (δ = 2mm) 98.6% (agrees with theory)

Rounded to cm

H_0 true, significance level 4.4% (slightly decreased)

Ha, power (δ = 2mm) 69.6% (from code given above)

Rounding markedly decreases the power of this paired t test.

RELATED CLASSROOM DISCUSSIONS

Do not round data before computing test statistics. How is this general advice supported by the previous slides?

Birthday Matching Problem. Use combinatorics to obtain P{No Match} for 25 people, then simulation to approximate the distribution and mean of the number of such matches. Discuss similarities (differences) to the number of ties in UNIF(0, 1) data truncated to one place.

Poisson approximations. When can the number of ties in rounded data (or the number of birthday matches) be well approximated by a Poisson distribution with $\lambda = C(n,2)/v$?

Independence of sample mean and SD for normal data. For simulated samples of size 5, a scatterplot as on slide 11 suggests independence. Similar plots show lack of indep. for data from exponential and BETA(.5, .5) distributions.

The P-value of a test is UNIF(0, 1) under H_0 (for a *continuous* test statistic). A plot on slide 12 illustrates this. Why is this true? Explore the distribution of the P-value under H_a. In each case, what is P{P-value < 5%}?

Noncentral t distribution. Use R functions **`pt`** and **`dt`** with noncentrality parameter $\Delta\sqrt{n}/\sigma$ to discuss power of a t test.

Wilcoxon signed-rank test for paired data. When rounding produces 0s or ties among the differences, R approximates the P-value (using a normal approximation). Use simulation to assess the accuracy of this approximation.

Exponential and Laplace data. Use simulation to explore the effect of rounding on data from these distributions. Also, to illustrate that the difference between indep. exponentials with rate λ is Laplace with $\mu = 0$, $\sigma^2 = 2/\lambda^2$, and that the mean of n indep. exponential obs. with rate λ is GAMMA(n, $n\lambda$).

REFERENCES

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

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.

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.

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 the two books listed below.

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

Trumbo, Bruce E.: *Learning Statistics with Real Data,* Chapter 4, Duxbury, Pacific Grove CA (2002). Chapter with data from 2e draft available from the author.

