It’s time to retire
“n >= 30”

Tim Hesterberg, Google Inc.
July 2008
Common Statistical Practice

Is $n \geq 30$, and data not too skewed?

- Yes: use $t$ intervals and tests
- No: use them anyway 😞
Dustbin

Old rule: use normal-based $t$ inferences if $n \geq 30$ (and data not too skewed)

• Ambiguous
  ▪ “not too skewed”?  
  ▪ Statistics other than the mean?

• Inaccurate

Relic of pre-computer era
  ▪ Toss into the dustbin of history
Instead

Better diagnostics for normality of sampling distribution

- Bootstrap
- Answers are surprising

Alternative confidence intervals and tests

- More accurate
Bootstrap Diagnostics

Sampling distribution of estimator
- Bootstrap histogram & Normal quantile plot
- Variety of examples

Sampling distribution of $t$-statistic
<table>
<thead>
<tr>
<th>Basic</th>
<th>Extended</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.95</td>
<td>3.383</td>
</tr>
<tr>
<td>10.013</td>
<td>7.8</td>
</tr>
<tr>
<td>10.62</td>
<td>9.416</td>
</tr>
<tr>
<td>10.15</td>
<td>4.66</td>
</tr>
<tr>
<td>8.583</td>
<td>5.36</td>
</tr>
<tr>
<td>7.62</td>
<td>7.63</td>
</tr>
<tr>
<td>8.233</td>
<td>4.95</td>
</tr>
<tr>
<td>10.35</td>
<td>8.013</td>
</tr>
<tr>
<td>11.016</td>
<td>7.8</td>
</tr>
<tr>
<td>8.516</td>
<td>9.58</td>
</tr>
</tbody>
</table>
Bootstrap Distns, Basic & Pay

![Density plots for Commercial Time (Basic) and Commercial Time (Extended)]
Bootstrap Distn for Difference

bootstrap : tv$commercialTime : mean : basic - extended

Density

mean

Observed
Mean
### Example - Verizon

<table>
<thead>
<tr>
<th></th>
<th>Number of Observations</th>
<th>Average Repair Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILEC (Verizon)</td>
<td>1664</td>
<td>8.4</td>
</tr>
<tr>
<td>CLEC (other carrier)</td>
<td>23</td>
<td>16.5</td>
</tr>
</tbody>
</table>

Is the difference statistically significant?
Bootstrap Distn for CLEC mean

observed

mean

Density

bootstrap : CLEC$Time : mean

mean

quantiles of standard normal

mean

quantiles of standard normal
Bootstrap Distn for ILEC mean

![Bootstrap Distn for ILEC mean](image)
Take another look

Take another look at the previous two figures.

Is the amount of non-normality/asymmetry there a cause for concern?

Note – we’re looking at a sampling distribution, not the underlying distribution. This is after the CLT effect!
Verizon bootstrap diff in means

bootstrap : Verizon
$Time : mean : ILEC - CLEC
...difference in trimmed means

bootstrap : Verizon : mean(Time, trim =... : ILEC - CLEC
Bushmeat

Brashares et al., *Science* 2004

• “Bushmeat hunting, wildlife declines, and fish supply in West Africa”

• 1970-1999, biomass of 30 species in national parks, and fish supply (kg/person/year)

• Relate change in biomass to fish supply
Bushmeat Data

Per capita fish supply (kg)

Wildlife biomass

year

20 25 30 35 40


200 400 600 800 1000
Bushmeat, ΔBiomass vs Fish

[Graph showing the relationship between per capita fish supply (kg) and percent change in wildlife biomass. Two panels: original data and one bootstrap sample.]
Bootstrap Bushmeat
Kyphosis vs. Start

![Graph showing relationship between Kyphosis and Start]
Graphical bootstrap of predictions
Bootstrap Coefficients

\texttt{bootstrap : glm(formula = Kyp... : coef(glm(data))}
Bootstrap Scatterplots

(Intercept)  Age  Start  Number
Example: Relative Risk

<table>
<thead>
<tr>
<th>Blood Pressure</th>
<th>Cardiovascular Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>55/3338 = 0.0165</td>
</tr>
<tr>
<td>Low</td>
<td>21/2676 = 0.0078</td>
</tr>
</tbody>
</table>

Estimated Relative Risk = 2.12
...bootstrap Relative Risk

![Image of bootstrap Relative Risk graph]

Observed Mean

Density

mean

0.0
0.2
0.4
0.6
0.8

0.2
0.4
0.6
0.8

0.0
2
3
4
5

1
2
3
4
5

BCa
percentile
t

tilt
Sampling Distn of $t$ statistic

Previous examples: sampling distributions of estimators

Now sampling distn of $t$ statistic
Sampling distribution of \( t \)-statistic

If population Normal,

- \( \bar{x} \) and \( s \) are independent
- \( t \) statistic has a \( t \) distribution

Bootstrap \( t \)

- \( t^* = (\bar{x}^* - \bar{x}) / (s^* / \sqrt{n}) \)
- CI: \( \bar{x} - t^*_{(1-\alpha)} (s / \sqrt{n}) \)
Bootstrap Distn for ILEC mean

![Bootstrap Distribution for ILEC Mean](image)
ILEC bootstrap xbar & s

Not independent!
ILEC bootstrap $t$ distribution

![Graph showing the distribution of t.mean with observed probabilities 0.0394 and 0.0212.](image-url)
Visual inspection not enough

Numerical summary of difference from Normal or $t$ distribution

In the previous example, 3.94% of bootstrap $t$ statistics are below $t_{0.025}$; it should be 2.5%.
What now?

\( n = 1664 \) is too small for accurate \( t \) inferences for ILEC data.

- One-sided noncoverage 3.94%
- (58\% too high)

Is this a general problem?

- Mean for skewed populations
- Statistics other than the mean
  - *Inherent skewness, also bias*
What $n$ does $t$ need?

Goal “reasonable accuracy”, 10% error

• 95% interval: nominal 2.5% each side
• Allow (2.25%, 2.75%)

Example:

• Exponential distribution

What $n$ gives reasonable accuracy?
Simulation Results

Johnson 1978 JASA
Need $n > 5000$ for $t$ accuracy!
Skewness of Exponential

Exponential distribution

• Moderate skewness (looks deceive)
• Many real distributions are more skewed
• This $F$ distn has the same skewness:
• For comparison, Gamma($a=10$)
  ▪ Need $n \geq 500$
Asymptotic Accuracy

$t$ interval, bootstrap percentile CI:
- First-order correct
- Consistent, coverage error $O(1/\sqrt{n})$

Johnson $t$, ABC, bootstrap $t$, BCa and Tilting:
- Second-order correct
- Coverage error $O(1/n)$
- Skewness, bias, transformations
Need $n > 220$ for Johnson $t$
Other Intervals
Tests

Use permutation tests

Or skewness-adjusted procedures
  • (and bias, transformation invariance)
Verizon permutation test

permutation: Verizon
Time: mean: ILEC - CLEC
Verizon test results

Pooled-variance t-test
  \( t = -2.6125, \) \( \text{df} = 1685, \) \textbf{p-value} = 0.0045

\[
\text{permutationTestMeans}(\text{data} = \text{Verizon}\$\text{Time}, \\
\text{treatment} = \text{Verizon}\$\text{Group}, \text{B} = 499999, \\
\text{alternative} = \"less\", \text{seed} = 99)
\]

Number of Replications: 499999

Summary Statistics:
<table>
<thead>
<tr>
<th>Observed</th>
<th>Mean</th>
<th>SE</th>
<th>alternative</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var</td>
<td>-8.098</td>
<td>-0.001288</td>
<td>3.105</td>
<td>less</td>
</tr>
</tbody>
</table>
Why aren’t these issues known?

Need bootstrap diagnostics

Need 64,000 bootstrap samples

Need fast computers
  • 20,000 hours for simulations in 1981

Sour Grapes
  • Accept low accuracy
  • Adjust criteria to fit practicality
5% total / 2.5% each side?

What matters?

• Non-coverage on each side? (3%, 2%)
• Total non-coverage (5%)

If you believe

• All statistical applications are really two-sided
  ▪ Missing too high/too low are equivalent
• People never interpret a 95% interval as meaning 2.5% on each side
• Your students understand CIs well

Then only total non-coverage matters
Caveats

Need care with really small $n$

- Second-order correct requires skewness correction
- Skewness hard to estimate with small $n$
- Shrink the skewness correction

Need 64000 bootstrap samples

- For accurate bootstrap diagnostics – MOE \(0.25\%/2\)
Summary

• Don’t use “n >= 30” rule
  ▪ Pre-computer rule
  ▪ Not accurate
  ▪ Use bootstrap diagnostics instead

• Use $t$ intervals and tests?
  ▪ Not if you want reasonable accuracy (10% error)
  ▪ Require $n > 5000$ for reasonable accuracy
  ▪ Use a second-order accurate procedure
Resources for Teaching

www.insightful.com/bootstrap

- Downloadable versions of BMPT for
  - Moore, McCabe, & Craig
  - Moore McCabe Duckworth & Alwan.
- Data packages for those books
- Free student version of S-PLUS