Using Randomization Methods to Build Conceptual Understanding in Statistical Inference: Day 1

Lock, Lock, Lock, Lock, Lock, and Lock
MAA Minicourse – Joint Mathematics Meetings
San Diego, CA
January 2013
The Lock⁵ Team

Dennis
Iowa State

Kari
Duke

Robin
St. Lawrence

Patti
St. Lawrence

Eric
Duke
Introductions:

Name

Institution
Schedule: Day 1
Wednesday, 1/9, 9:00 – 11:00 am

1. Introductions and Overview

2. Bootstrap Confidence Intervals
   • What is a bootstrap distribution?
   • How do we use bootstrap distributions to build understanding of confidence intervals?
   • How do we assess student understanding when using this approach?

3. Getting Started on Randomization Tests
   • What is a randomization distribution?
   • How do we use randomization distributions to build understanding of p-values?

4. Minute Papers
Schedule: Day 2
Friday, 1/11, 9:00 – 11:00 am

5. More on Randomization Tests
   • How do we generate randomization distributions for various statistical tests?
   • How do we assess student understanding when using this approach?

6. Connecting Intervals and Tests

7. Connecting Simulation Methods to Traditional

8. Technology Options
   • Brief software demonstration (Minitab, Fathom, R, Excel, ...)
     – pick one!

9. Wrap-up
   • How has this worked in the classroom?
   • Participant comments and questions

10. Evaluations
Why use Randomization Methods?
These methods are great for teaching statistics...

(the methods tie directly to the key ideas of statistical inference so help build conceptual understanding)
And these methods are becoming increasingly important for *doing* statistics.
"Actually, the statistician does not carry out this very simple and very tedious process [the randomization test], but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method."

-- Sir R. A. Fisher, 1936
... and the way of the future

“... the consensus curriculum is still an unwitting prisoner of history. What we teach is largely the technical machinery of numerical approximations based on the normal distribution and its many subsidiary cogs. This machinery was once necessary, because the conceptually simpler alternative based on permutations was computationally beyond our reach. Before computers statisticians had no choice. These days we have no excuse. Randomization-based inference makes a direct connection between data production and the logic of inference that deserves to be at the core of every introductory course.”

-- Professor George Cobb, 2007

(see full TISE article by Cobb in your binder)
Question

Do you teach Intro Stat?

A. Very regularly (most semesters)
B. Regularly (most years)
C. Occasionally
D. Rarely (every few years)
E. Never (or not yet)
Question

How familiar are you with simulation methods such as bootstrap confidence intervals and randomization tests?

A. Very
B. Somewhat
C. A little
D. Not at all
E. Never heard of them before!
Question

Have you used randomization methods in Intro Stat?

A. Yes, as a significant part of the course
B. Yes, as a minor part of the course
C. No
D. What are randomization methods?
### Question

Have you used randomization methods in any statistics class that you teach?

A. Yes, as a significant part of the course  
B. Yes, as a minor part of the course  
C. No  
D. What are randomization methods?
Intro Stat – Revise the Topics

• Descriptive Statistics – one and two samples
• Normal distributions
• **Bootstrap confidence intervals**
• Data production (samples/experiments)
• **Randomization-based hypothesis tests**
• Sampling distributions (mean/proportion)
• Normal distributions
• Confidence intervals (means/proportions)
• Hypothesis tests (means/proportions)
• ANOVA for several means, Inference for regression, Chi-square tests
We need a snack!
What proportion of Reese’s Pieces are Orange?

Find the proportion that are orange for your “sample”.
BUT – In practice, can we really take lots of samples from the same population?
Bootstrap Distributions

Or: How do we get a sense of a sampling distribution when we only have ONE sample?
Suppose we have a random sample of 6 people:
Create a “sampling distribution” using this as our simulated population
**Bootstrap Sample:** Sample with replacement from the original sample, using the same sample size.
Simulated Reese’s Population

Sample from this “population”

Original Sample
Create a bootstrap sample by sampling *with replacement* from the original sample.

Compute the relevant statistic for the bootstrap sample.

Do this many times!! Gather the bootstrap statistics all together to form a bootstrap distribution.
Example: What is the average price of a used Mustang car?

Select a random sample of n=25 Mustangs from a website (autotrader.com) and record the price (in $1,000’s) for each car.
Sample of Mustangs:

Our best estimate for the average price of used Mustangs is $15,980, but how accurate is that estimate?

\[ n = 25 \quad \bar{x} = 15.98 \quad s = 11.11 \]

Our best estimate for the average price of used Mustangs is $15,980, but how accurate is that estimate?
We need technology!

Introducing StatKey.

www.lock5stat.com/statkey
StatKey

Confidence Interval for a Mean, Median, Std. Dev.

Bootstrap Dotplot of Mean

Original Sample
- $n = 25$, mean $= 15.98$
- median $= 11.9$, stdev $= 11.14$

Bootstrap Sample
- $n = 25$, mean $= 13.712$
- median $= 11.9$, stdev $= 9.723$

Std. dev of $\bar{x}'s = 2.18$
Using the Bootstrap Distribution to Get a Confidence Interval – Method #1

The standard deviation of the bootstrap statistics estimates the **standard error** of the sample statistic.

Quick interval estimate:

\[ \text{Original Statistic } \pm 2 \cdot SE \]

For the mean Mustang prices:

\[
15.98 \pm 2 \cdot 2.18 = 15.98 \pm 4.36 \\
= (11.62, 20.34)
\]
Using the Bootstrap Distribution to Get a Confidence Interval – Method #2

We are 95% sure that the mean price for Mustangs is between $11,930 and $20,238
Bootstrap Confidence Intervals

Version 1 (Statistic $\pm 2$ SE):
Great preparation for moving to traditional methods

Version 2 (Percentiles):
Great at building understanding of confidence intervals
Playing with StatKey!

See the purple pages in the folder.
Traditional Inference

1. Which formula?
\[ \bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}} \] or \[ \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}} \]

2. Calculate summary stats
\[ n = 25, \ \bar{x} = 15.98, s = 11.11 \]

3. Find t*
95% CI \[ \Rightarrow \alpha/2 = \frac{1-0.95}{2} = 0.025 \]
\[ df = 25 - 1 = 24 \]
\[ t^* = 2.064 \]

4. df?
\[ df = 25 - 1 = 24 \]
\[ t^* = 2.064 \]

5. Plug and chug
\[ 15.98 \pm 2.064 \cdot \frac{11.11}{\sqrt{25}} \]
\[ 15.98 \pm 4.59 = (11.39, 20.57) \]

6. Interpret in context

7. Check conditions
We want to collect some data from you. What should we ask you for our one quantitative question and our one categorical question?
What quantitative data should we collect from you?

A. What was the class size of the Intro Stat course you taught most recently?
B. How many years have you been teaching Intro Stat?
C. What was the travel time, in hours, for your trip to Boston for JMM?
D. Including this one, how many times have you attended the January JMM?
E. ???
What categorical data should we collect from you?

A. Did you fly or drive to these meetings?
B. Have you attended any previous JMM meetings?
C. Have you ever attended a JSM meeting?
D. ???
E. ???
Why does the bootstrap work?
BUT, in practice we don’t see the “tree” or all of the “seeds” – we only have ONE seed
Bootstrap Distribution

What can we do with just one seed?

Grow a NEW tree!

Bootstrap “Population”

Estimate the distribution and variability (SE) of x’s from the bootstraps
Golden Rule of Bootstraps

The *bootstrap statistics* are to the *original statistic* as the *original statistic* is to the *population parameter*.
How do we assess student understanding of these methods (even on in-class exams without computers)?

See the green pages in the folder.
Paul the Octopus

http://www.youtube.com/watch?v=3ESGpRUMj9E
Paul the Octopus

• Paul the Octopus predicted 8 World Cup games, and predicted them all correctly

• Is this evidence that Paul actually has psychic powers?

• How unusual would this be if he were just randomly guessing (with a 50% chance of guessing correctly)?

• How could we figure this out?
Simulate!

• Each coin flip = a guess between two teams
• Heads = correct, Tails = incorrect
• Flip a coin 8 times and count the number of heads. Remember this number!

Did you get all 8 heads?

(a) Yes
(b) No
Hypotheses

Let $p$ denote the proportion of games that Paul guesses correctly (of all games he may have predicted)

\[ H_0 : p = 1/2 \]
\[ H_a : p > 1/2 \]
Randomization Distribution

• A *randomization distribution* is the distribution of sample statistics we would observe, just by random chance, if the null hypothesis were true.

• A randomization distribution is created by simulating many samples, assuming $H_0$ is true, and calculating the sample statistic each time.
Randomization Distribution

• Let’s create a randomization distribution for Paul the Octopus!
• On a piece of paper, set up an axis for a dotplot, going from 0 to 8
• Create a randomization distribution using each other’s simulated statistics
• For more simulations, we use StatKey
p-value

• The \textit{p-value} is the probability of getting a statistic as extreme (or more extreme) as that observed, just by random chance, if the null hypothesis is true.

• This can be calculated directly from the randomization distribution!
\[
\left( \frac{1}{2} \right)^8 = 0.0039
\]
Randomization Test

• Create a randomization distribution by simulating assuming the null hypothesis is true

• The p-value is the proportion of simulated statistics as extreme as the original sample statistic
Coming Attractions - Friday

• How do we create randomization distributions for other parameters?
• How do we assess student understanding?
• Connecting intervals and tests
• Connecting simulations to traditional methods
• Technology for using simulation methods
• Experiences in the classroom
Using Randomization Methods to Build Conceptual Understanding of Statistical Inference: Day 2

Lock, Lock, Lock, Lock, and Lock
MAA Minicourse- Joint Mathematics Meetings
San Diego, CA
January 2013
Schedule: Day 2
Friday, 1/11, 9:00 – 11:00 am

5. More on Randomization Tests
   • How do we generate randomization distributions for various statistical tests?
   • How do we assess student understanding when using this approach?

6. Connecting Intervals and Tests

7. Connecting Simulation Methods to Traditional

8. Technology Options
   • Brief software demonstration (Minitab, R, Excel, more StatKey...)

9. Wrap-up
   • How has this worked in the classroom?
   • Participant comments and questions

10. Evaluations
Cocaine Addiction

• In a randomized experiment on treating cocaine addiction, 48 people were randomly assigned to take either Desipramine (a new drug), or Lithium (an existing drug)

• The outcome variable is whether or not a patient relapsed

• Is Desipramine significantly better than Lithium at treating cocaine addiction?
1. Randomly assign units to treatment groups
2. Conduct experiment

3. Observe relapse counts in each group

\[ \hat{P}_{\text{new}} - \hat{P}_{\text{old}} = \frac{10}{24} - \frac{18}{24} = -0.333 \]

10 relapse, 14 no relapse

18 relapse, 6 no relapse

R = Relapse
N = No Relapse

1. Randomly assign units to treatment groups

Desipramine

Lithium
Randomization Test

• Assume the null hypothesis is true
• Simulate new randomizations
• For each, calculate the statistic of interest
• Find the proportion of these simulated statistics that are as extreme as your observed statistic
10 relapse, 14 no relapse

18 relapse, 6 no relapse
Simulate another randomization

Desipramine

\[ \hat{p}_N - \hat{p}_O = \frac{16 - 12}{24 - 24} = \frac{0.167}{24 - 24} = 0.167 \]

16 relapse, 8 no relapse

Lithium

12 relapse, 12 no relapse
Simulate another randomization

Desipramine

\begin{align*}
\hat{p}_N - \hat{p}_O &= \frac{17}{24} - \frac{11}{24} \\
&= 0.250
\end{align*}

Lithium

17 relapse, 7 no relapse

11 relapse, 13 no relapse
Physical Simulation

• Start with 48 cards (Relapse/No relapse) to match the original sample.

• Shuffle all 48 cards, and rerandomize them into two groups of 24 (new drug and old drug)

• Count “Relapse” in each group and find the difference in proportions, $\hat{p}_N - \hat{p}_O$.

• Repeat (and collect results) to form the randomization distribution.

• How extreme is the observed statistic of $-0.33$?
A randomization sample must:

• Use the data that we have  
  (That’s why we didn’t change any of the results on the cards)  
  AND

• Match the null hypothesis  
  (That’s why we assumed the drug didn’t matter and combined the cards)
The probability of getting results as extreme or more extreme than those observed *if the null hypothesis is true*, is about .02.
How can we do a randomization test for a correlation?
Is the number of penalties given to an NFL team positively correlated with the “malevolence” of the team’s uniforms?
Ex: NFL uniform “malevolence” vs. Penalty yards

\[ r = 0.430 \]
\[ n = 28 \]

Is there evidence that the population correlation is positive?
Key idea: Generate samples that are (a) consistent with the null hypothesis (b) based on the sample data.

\[ H_0 : \rho = 0 \]

\[ r = 0.43, \ n = 28 \]

How can we use the sample data, but ensure that the correlation is zero?
Randomize one of the variables!

Let’s look at StatKey.
Traditional Inference

1. Which formula?

$$t = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

2. Calculate numbers and plug into formula

$$0.43\sqrt{28 - 2}$$

$$\frac{\sqrt{1 - 0.43^2}}{\sqrt{1 - 0.43^2}}$$

3. Plug into calculator

$$= 2.43$$

4. Which theoretical distribution?

5. df?

6. find p-value

$$0.01 < p-value < 0.02$$
How can we do a randomization test for a mean?
Example: Mean Body Temperature

Is the average body temperature really 98.6°F?

\( H_0 : \mu = 98.6 \)

\( H_a : \mu \neq 98.6 \)

Data: A random sample of \( n=50 \) body temperatures.

Data from Allen Shoemaker, 1996 JSE data set article

\[ n = 50 \quad \bar{x} = 98.26 \quad s = 0.765 \]
**Key idea:** Generate samples that are 
(a) consistent with the null hypothesis 
(b) based on the sample data.

$H_0: \mu = 98.6$

How to simulate samples of body temperatures to be consistent with $H_0: \mu = 98.6$?

**Sample:**

$n = 50, \bar{x} = 98.26, s = 0.765$
Randomization Samples

How to simulate samples of body temperatures to be consistent with $H_0: \mu=98.6$?

1. Add 0.34 to each temperature in the sample (to get the mean up to 98.6).
2. Sample (with replacement) from the new data.
3. Find the mean for each sample ($H_0$ is true).
4. See how many of the sample means are as extreme as the observed $\bar{x} = 98.26$. 
Let’s try it on StatKey.
Playing with StatKey!

*See the orange pages in the folder.*
Choosing a Randomization Method

Example: Word recall

<table>
<thead>
<tr>
<th>A=Sleep</th>
<th>14</th>
<th>18</th>
<th>11</th>
<th>13</th>
<th>18</th>
<th>17</th>
<th>21</th>
<th>9</th>
<th>16</th>
<th>17</th>
<th>14</th>
<th>15</th>
<th>mean=15.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>B=Caffeine</td>
<td>12</td>
<td>12</td>
<td>14</td>
<td>13</td>
<td>6</td>
<td>18</td>
<td>14</td>
<td>16</td>
<td>10</td>
<td>7</td>
<td>15</td>
<td>10</td>
<td>mean=12.25</td>
</tr>
</tbody>
</table>

$H_0: \mu_A = \mu_B$  vs. $H_a: \mu_A \neq \mu_B$

Option 1: Randomly scramble the A and B labels and assign to the 24 word recalls.

Option 2: Combine the 24 values, then sample (with replacement) 12 values for Group A and 12 values for Group B.

Reallocate

Resample
Question

In Intro Stat, how critical is it for the method of randomization to reflect the way data were collected?

A. Essential
B. Relatively important
C. Desirable, but not imperative
D. Minimal importance
E. Ignore the issue completely
How do we assess student understanding of these methods (even on in-class exams without computers)?

*See the blue pages in the folder.*
Connecting CI’s and Tests

Randomization body temp means when $\mu=98.6$

Bootstrap body temp means from the original sample

What’s the difference?
Sample mean is in the “rejection region” ⇔ Null mean is outside the confidence interval
What about Traditional Methods?
Transitioning to Traditional Inference

AFTER students have seen lots of bootstrap distributions and randomization distributions...

Students should be able to

• Find, interpret, and understand a confidence interval
• Find, interpret, and understand a p-value
Bootstrap and Randomization Distributions

Slope: Restaurant tips

Correlation: Malevolent uniforms

Mean: Body Temperatures

Means: Finger taps

Proportion: Owners/dogs

Mean: Atlanta commutes

All bell-shaped distributions!
The students are primed and ready to learn about the normal distribution!
Transitioning to Traditional Inference

• Introduce the normal distribution (and later t)
• Introduce “shortcuts” for estimating SE for proportions, means, differences, slope...

Confidence Interval:

\[ \text{Sample Statistic} \pm z^* \cdot SE \]

Hypothesis Test:

\[ \frac{\text{Sample Statistic} - \text{Null Parameter}}{SE} \]
Confidence Intervals

![Graph showing a normal distribution with 95% confidence interval. The critical Z-scores are labeled as $-Z^*$ and $Z^*$, with corresponding probabilities of 0.025 and 0.975.]
Hypothesis Tests

Test statistic

95% Area is p-value

T Distribution

Reset Plot

T Distribution

2.430

0.011

df

26

Edit Parameters
Yes! Students see the general pattern and not just individual formulas!

Confidence Interval:

Sample Statistic $\pm z^* \cdot SE$

Hypothesis Test:

Sample Statistic $-$ Null Parameter $\over SE$
Brief Technology Session
Choose One!

R (Kari) Excel (Eric)
Minitab (Robin) TI (Patti)
More StatKey (Dennis)

(Your binder includes information on using Minitab, R, Excel, Fathom, Matlab, and SAS.)
## Student Preferences

Which way of doing inference gave you a better conceptual understanding of confidence intervals and hypothesis tests?

<table>
<thead>
<tr>
<th>Bootstrapping and Randomization</th>
<th>Formulas and Theoretical Distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>113</td>
<td>51</td>
</tr>
<tr>
<td>69%</td>
<td>31%</td>
</tr>
</tbody>
</table>
Student Preferences

Which way did you prefer to learn inference (confidence intervals and hypothesis tests)?

<table>
<thead>
<tr>
<th>Bootstrapping and Randomization</th>
<th>Formulas and Theoretical Distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>105</td>
<td>60</td>
</tr>
<tr>
<td>64%</td>
<td>36%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Simulation</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP Stat</td>
<td>31</td>
<td>36</td>
</tr>
<tr>
<td>No AP Stat</td>
<td>74</td>
<td>24</td>
</tr>
</tbody>
</table>
Student Behavior

• Students were given data on the second midterm and asked to compute a confidence interval for the mean

• How they created the interval:

<table>
<thead>
<tr>
<th>Bootstrapping</th>
<th>t.test in R</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>94</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>84%</td>
<td>8%</td>
<td>8%</td>
</tr>
</tbody>
</table>
A Student Comment

"I took AP Stat in high school and I got a 5. It was mainly all equations, and I had no idea of the theory behind any of what I was doing.

Statkey and bootstrapping really made me understand the concepts I was learning, as opposed to just being able to just spit them out on an exam."

- one of Kari’s students
Thank you for joining us!

More information is available on www.lock5stat.com

Feel free to contact any of us with any comments or questions.
Please fill out the Minicourse evaluation form at

www.surveymonkey.com/s/JMM2013MinicourseSurvey