Bootstrapping with Multi-Level Data

Lacey Etzkorn

May 8, 2019

Multi-Level Models

Outline

1. Review sampling distributions and CLT
2. Review of bootstrapping in the simple case:
    a. What is the bootstrap?
    b. Why should we use it?
    c. A example from Biostat 621.
3. Extending the bootstrap to multi-level data.
    a. How to do the non-parametric bootstrap for multi-level data.
    b. Data analysis example:
        i. CI for an intraclass correlation coefficient

Review: Sampling Distribution of a Statistic

Sampling distributions allow us to calculate confidence intervals and perform statistical inference.

Central Limit Theorem:

$$X_1, X_2, X_3 \ldots X_n \text{ are iid } ; E(X_i) = \mu, Var(X_i) = \sigma^2$$
$$\text{Then } \bar{X}_n \sim N(\mu, \sigma^2/n) \text{ for large } n$$

CLT can be extended to derive the sampling distribution of other statistics like model coefficients.

- Requires n to be large enough for the statistic to be “approximately normal”.
- Need to be able to calculate the standard error.

Why Bootstrap?

Bootstrap when we don't know the sampling distribution of a statistic of interest.

a. Not sure if n is large enough for the CLT to hold for a normal-based confidence interval.
b. When the statistic is "weird" and doesn't follow a normal distribution, or maybe not even any other known distribution.

Bootstrap allows empirical non-parametric approximation of the sampling distribution of a statistic.

Obtain a confidence interval for a "tough" problem.

What is the bootstrap procedure?

Estimate the sampling distribution of a statistic by leveraging the observed distribution of the data.

1. Resample n observations from the data many (m = 10,000) times.
2. Calculate the desired statistic for each data set.
3. Use 2.5 and 97.5 percentiles of the m (10,000) statistics as a confidence interval.

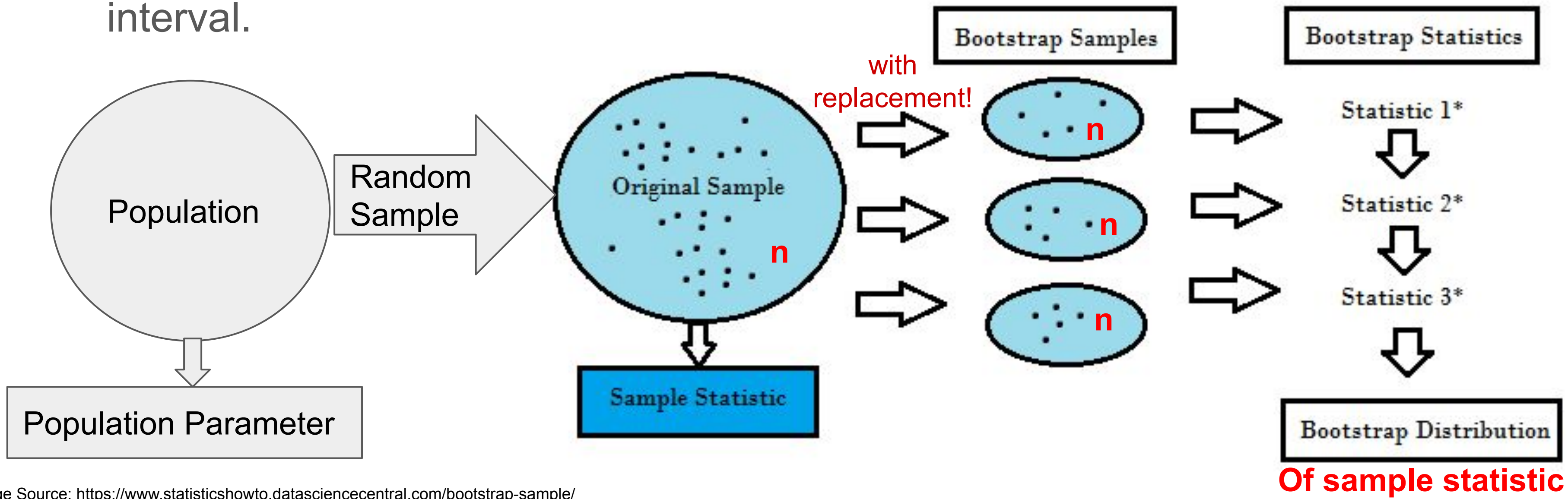


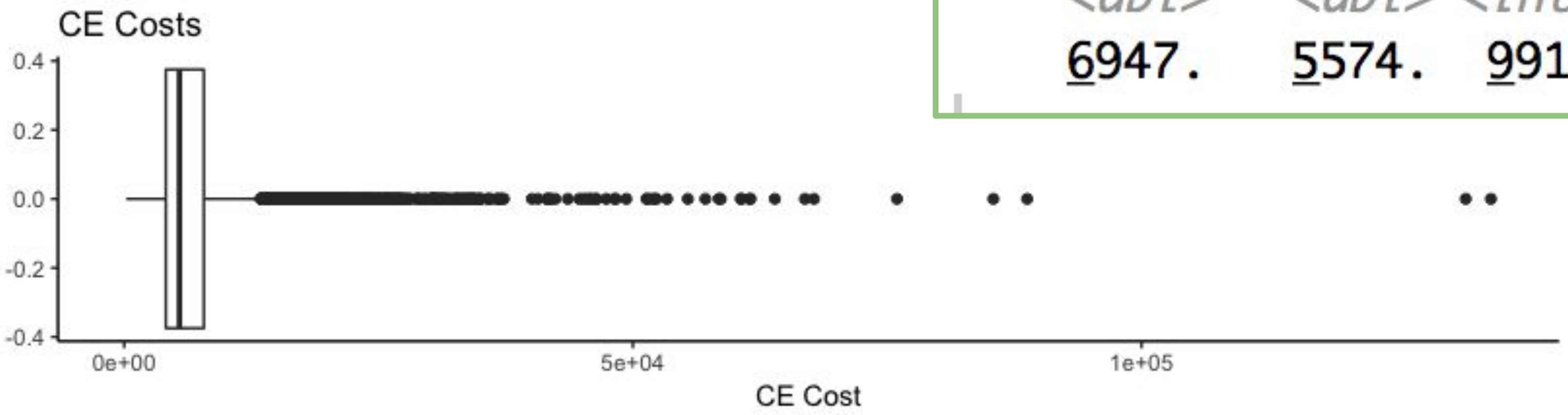
Image Source: https://www.statisticshowto.datasciencecentral.com/bootstrap-sample/

What is the (basic) Bootstrap? An example.

Data: Costs of carotid endarterectomy (CE) in Maryland.

Question: What is the average cost for CE?

mean_cost	sd_cost	n
<dbl>	<dbl>	<int>
6947.	5574.	9918



What is the Bootstrap? An example.

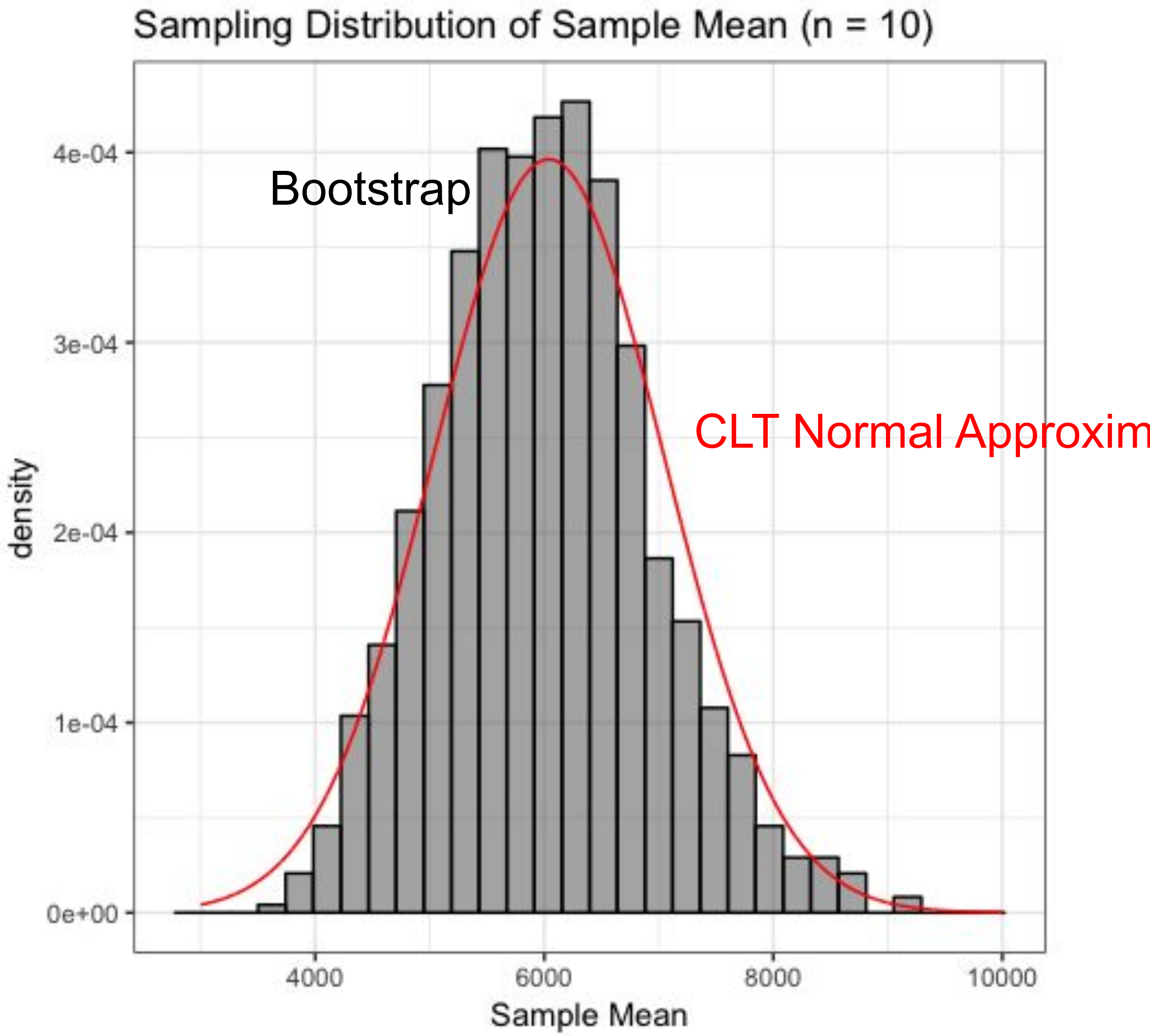
Question: What is the average cost for CE?

Method: Sample n = 10. Point estimate and t-based confidence interval.

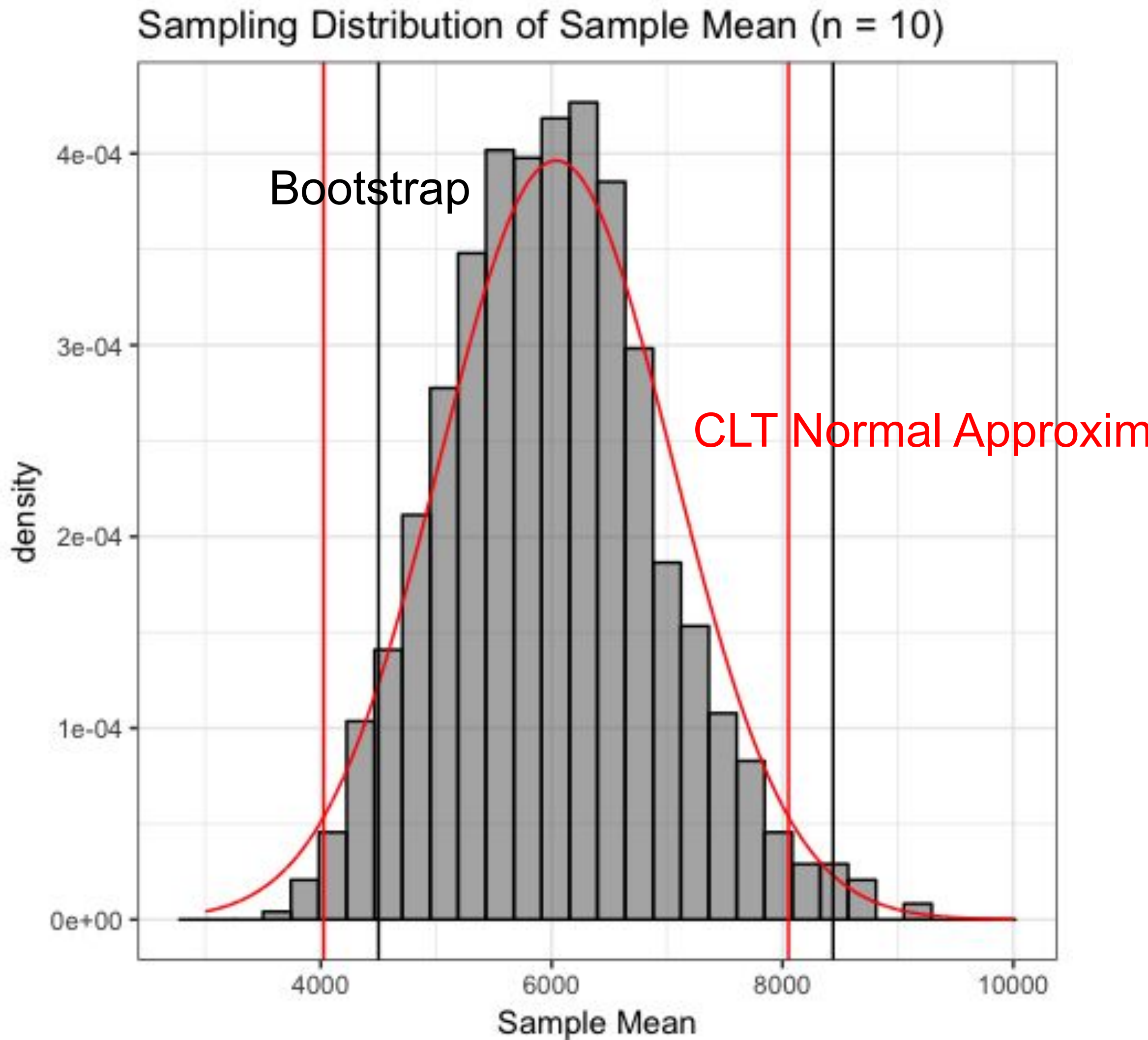
Assumptions for t-interval: (what will make our inferences correct?)

1. Observations independent.
2. Random sample from population.
3. Data normally-distributed.

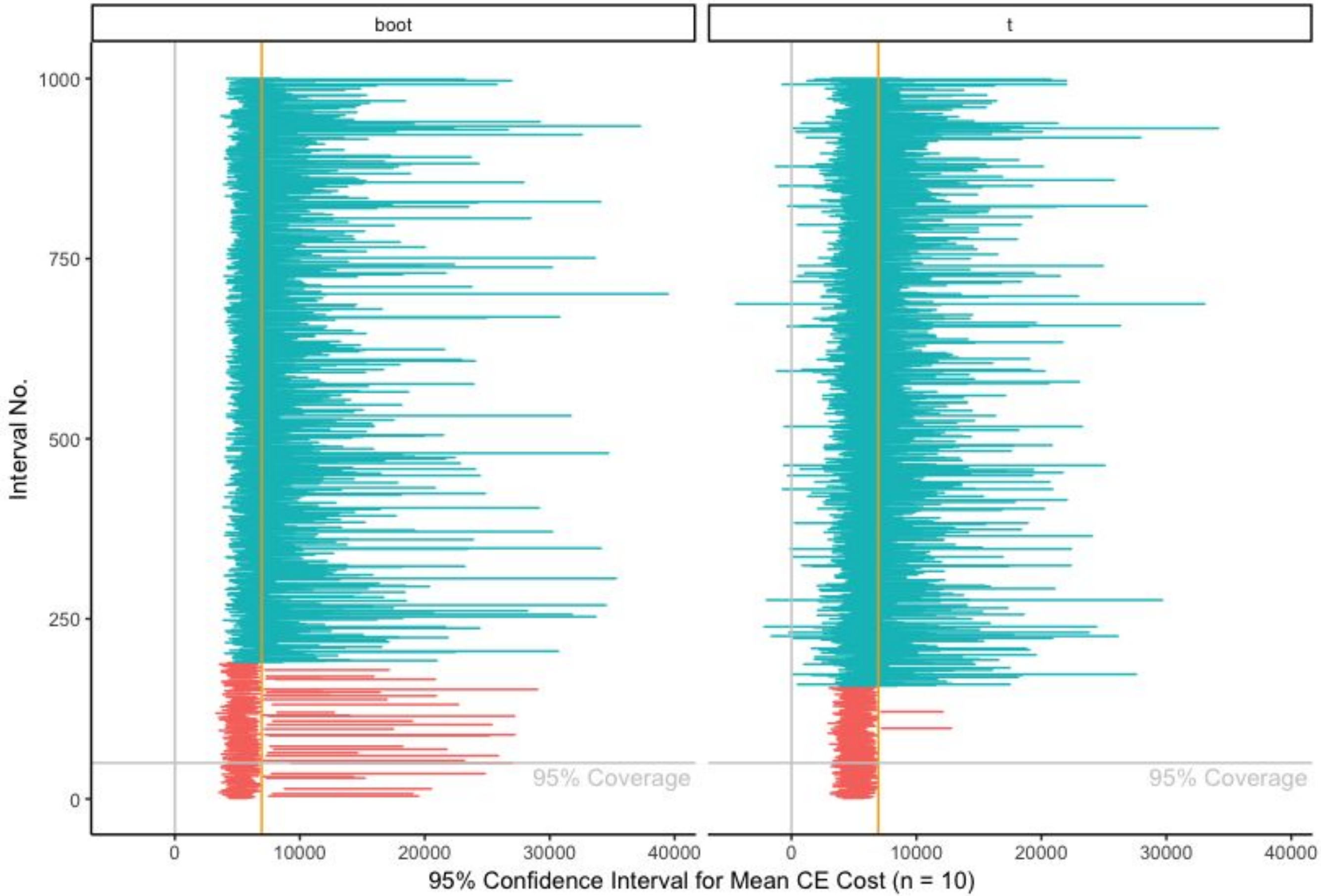
What is the Bootstrap? An example.



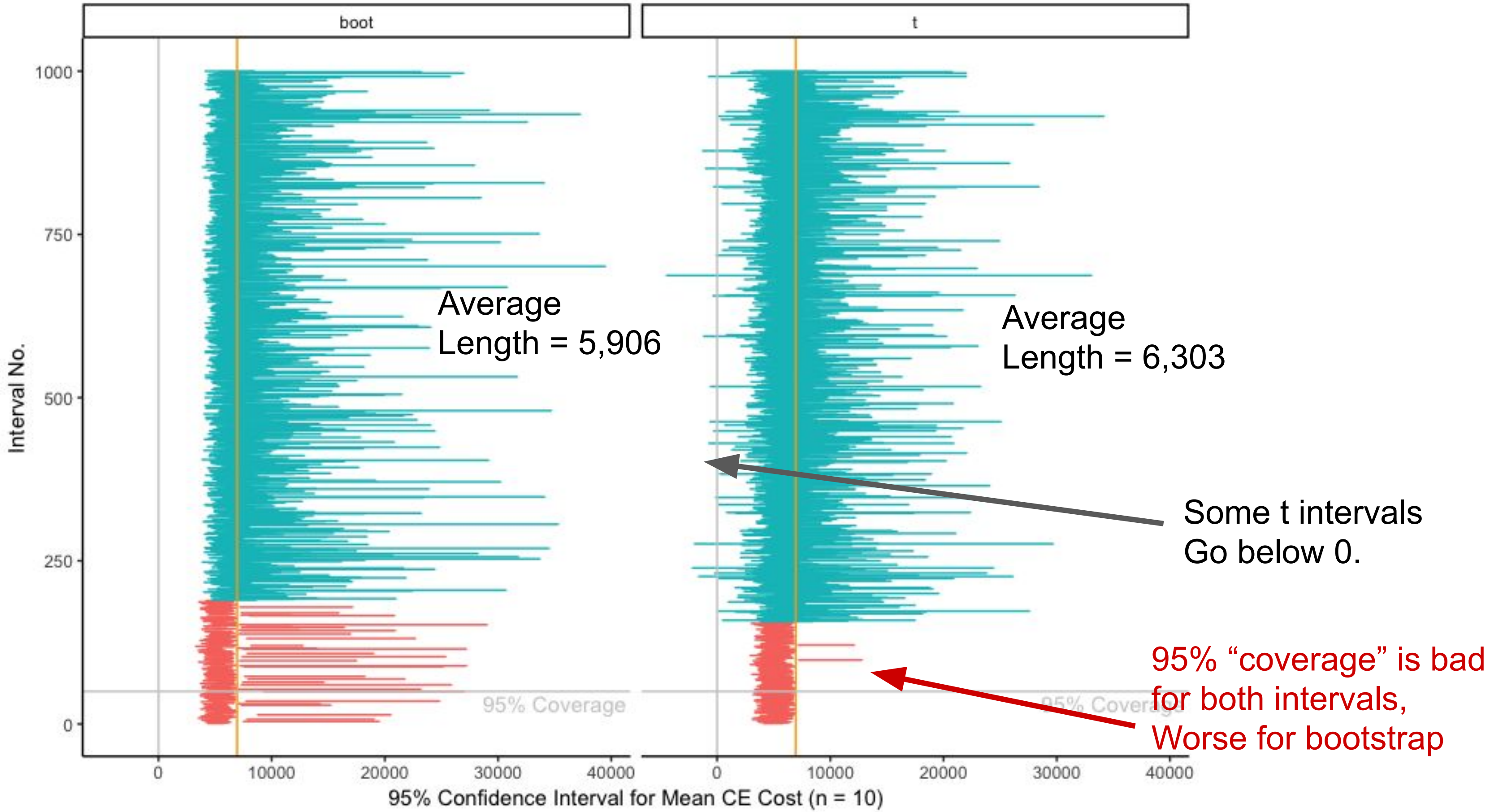
What is the Bootstrap? An example.



Intervals from resamples of n = 10



Intervals from resamples of n = 10



Summary of Simple Bootstrap

- Requires independent observations.
- Makes fewer assumptions about the distribution of the data.
- Easier to implement when the math becomes hard.
- NOT a panacea for statistical inference.

Extension to the Multilevel Case

- Simple Random Sample: Simple Bootstrap
  - SRS from Population to generate **Data**
  - SRS from **Data** (with replacement) to generate many bootstrap data sets
- Multilevel Sample: ?
  - Data is no longer a simple random sample from population.
  - Sampling is correlated

Should we still use SRS from the lowest level of the data for our bootstrap?

How do we account for correlated sampling in a multilevel bootstrap?

Extension to the Multilevel Case

Unaccepted Approach: Simple Random Resampling

Perform simple-random-resampling on the lowest level (students).

Accepted Approach: Cluster Resampling

Ren et al. (2010): *Nonparametric bootstrapping for hierarchical data*

3-Level Data: Sample the highest level with replacement. Keep data within the highest level as is. (schools)

Case Study: Math Achievement ICC (Homework 1)

Data: High School and Beyond (HS&B) Study,
Outcome of Interest: Mathematical Achievement
Estimand of Interest: Intraclass Correlation Coefficient

	newid	student	minority	female	ses	mathach	size	sector	pracad	disclim	himinty
	<int>	<int>	<int>	<int>	<dbl>	<dbl>	<int>	<int>	<dbl>	<dbl>	<int>
1	1	1	1	0	-1.66	0.122	842	0	0.35	1.60	0
2	1	2	1	0	-1.53	4.14	842	0	0.35	1.60	0
3	1	3	0	1	-1.53	5.88	842	0	0.35	1.60	0
4	1	4	0	1	-1.45	9.48	842	0	0.35	1.60	0
5	1	5	0	1	-1.40	6.13	842	0	0.35	1.60	0
6	1	6	1	1	-1.25	4.75	842	0	0.35	1.60	0

Case Study: Math Achievement ICC (Homework 1)

Outcome of Interest: Mathematical Achievement
Estimand of Interest: Intraclass Correlation Coefficient

$$MA_{ij} = \mu + \alpha_i + \epsilon_{ij}; \ \ \alpha_i \sim N(0, \sigma^2_{school}); \ \ \epsilon_{ij} \sim N(0, \sigma^2_{student})$$

$$ICC = \frac{\sigma^2_{school}}{\sigma^2_{school} + \sigma^2_{student}}$$

Case Study: Math Achievement ICC (Homework 1)

H0: Equal student-to-student and school-to-school variability in MA scores.

- ICC = 0.5

H1: More student-to-student variability in MA scores.

- ICC < 0.5

Approach: Use non-parametric bootstrapping to create a confidence interval. See whether null value (0.5) is in the confidence interval.

Case Study: Math Achievement ICC (Homework 1)

School Resampling

```
> boot.ci(boot.icc, type = c("perc","bca"), conf = 0.95)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.icc, conf = 0.95, type = c("perc", "bca"))

Intervals :
Level      Percentile            BCa
95%   ( 0.1297,  0.2131 )   ( 0.1348,  0.2159 )
Calculations and Intervals on Original Scale
```

Conclusion: **We estimate that only 17% of the variation in MA scores is due to school-to-school variation (95% CI = [13%, 21%]). Hence the majority of variation in MA scores can be attributed to student-level differences.**

Case Study: Math Achievement ICC (Homework 1)

But wait: The clustered bootstrap didn't produce very different results from the simple bootstrap.

Clustered School Resampling

```
> boot.ci(boot.icc, type = c("perc","bca"), conf = 0.95)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.icc, conf = 0.95, type = c("perc", "bca"))

Intervals :
Level      Percentile            BCa
95%   ( 0.1297,  0.2131 )   ( 0.1348,  0.2159 )
Calculations and Intervals on Original Scale
```

Independent Student Resampling

```
> boot.ci(boot.icc.incorrect, type = c("perc"), conf = 0.95)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.icc.incorrect, conf = 0.95, type = c("perc"))

Intervals :
Level      Percentile
95%   ( 0.1743,  0.2082 )
Calculations and Intervals on Original Scale
```

Conclusions

- The bootstrap is ONE way of empirically/non-parametrically estimating the sampling distribution of a statistic.
- Tends to be more trusted when parametric assumptions are likely not to hold.
- Resampling independently with replacement at the lowest level can cause problematic inferences in hierarchical data, especially when outcomes are highly correlated within clusters.
- Resampling the highest level and maintaining all the data within these highest level clusters is the most accepted strategy for bootstrapping hierarchical data.

References

Nonparametric bootstrapping for hierarchical data (2013): https://doi.org/10.1080/02664760903046102

Using Cluster Bootstrapping to Analyze Nested Data With a Few Clusters (2013):
https://doi-org.proxy1.library.jhu.edu/10.1177/0013164416678980