Bootstrapping and Permutation Tests

Keegan Flanagan, Jeffrey LeDue, Will Casazza, Katlyn Richardson, Megan Pawluk

Pulling Ourselves Up By Our Own Bootstraps.

- A phrase that refers to the completely absurd.
- Despite its namesake, many empirical tests have found that bootstrapping works incredibly well.

What is bootstrapping?

- A process where you estimate the sampling distribution of a statistic by resampling from your data with replacement.
  1. You sample a population and get data
  2. You randomly choose points from your data (resample) and allow yourself to take the same points multiple times (replacement).
  3. You repeat step two many many times and get thousands of resamples.
  4. You calculate a sample statistic from each resample (eg the mean) and then build a distribution of your sample statistics.
- This distribution is called the bootstrap distribution, and it is used as an estimate for the sampling distribution.

The Idea Behind Bootstrapping

"The original sample represents the population from which it was drawn. So, resamples from this sample represent what we would get if we took many samples from the population. The bootstrap distribution of a statistic, based on many resamples, represents the sampling distribution of the statistic, based on many samples"

Companion Chapter 18: Bootstrap Methods And Permutation Tests for the Practice of Business Statistics. David S. Moore, George P. McCabe, William M. Duckworth II, Stanley L. Sclove

Let's Give It A Try.

- Go into the Bootstrap_work subfolder and open up the t-Bootstrap_script_working rmd file with Rstudio.

Why Use Bootstrapping?

- A great way to explore your data.
- Does not require assumption of normality.
- Less sensitive to sample sizes (but not insensitive).
- Can be generalized to explore statistics other than the mean.
- Determine SEs and CIs.
- Perform permutation tests.

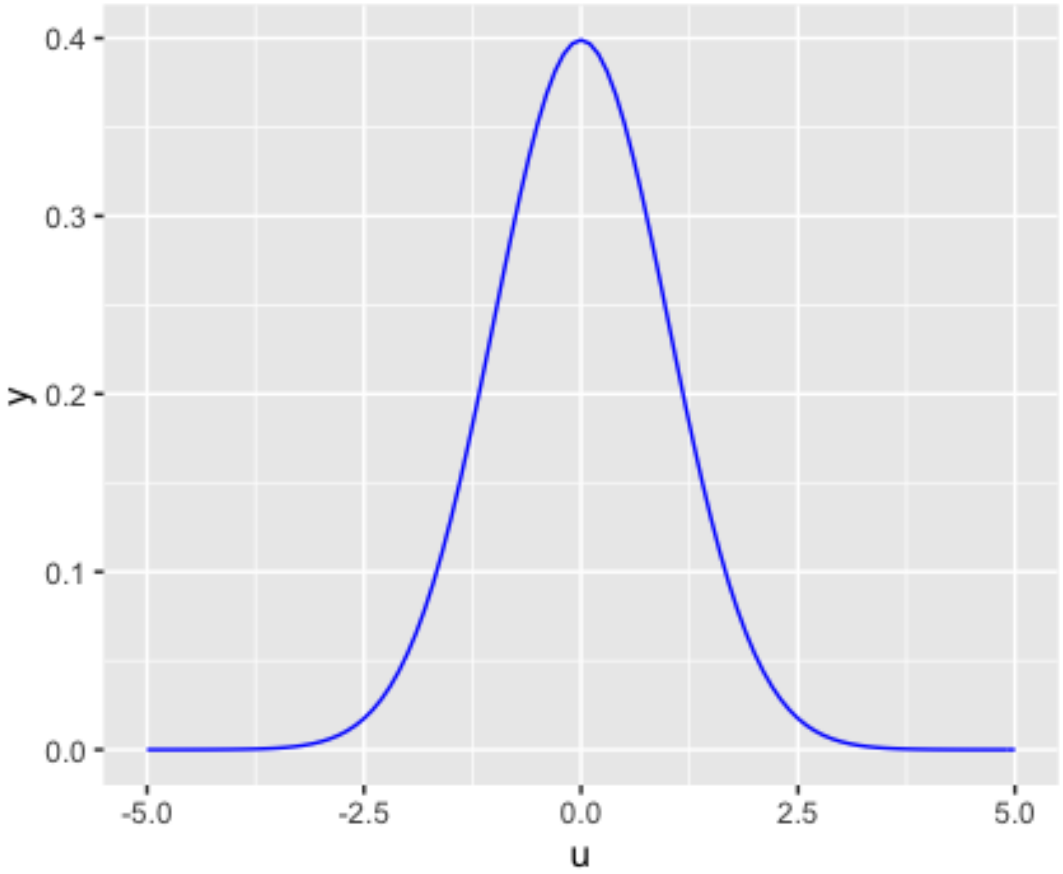
Standard Error.

- The standard error of a variable can be defined as the standard deviation of that variable's sampling distribution.
- The bootstrap distribution is an estimate of the sampling distribution
- Take the standard deviation of the bootstrap distribution.
- You're done.

Confidence Intervals.

Calculating confidence intervals when the bootstrap distribution is normal and not skewed (symmetric).

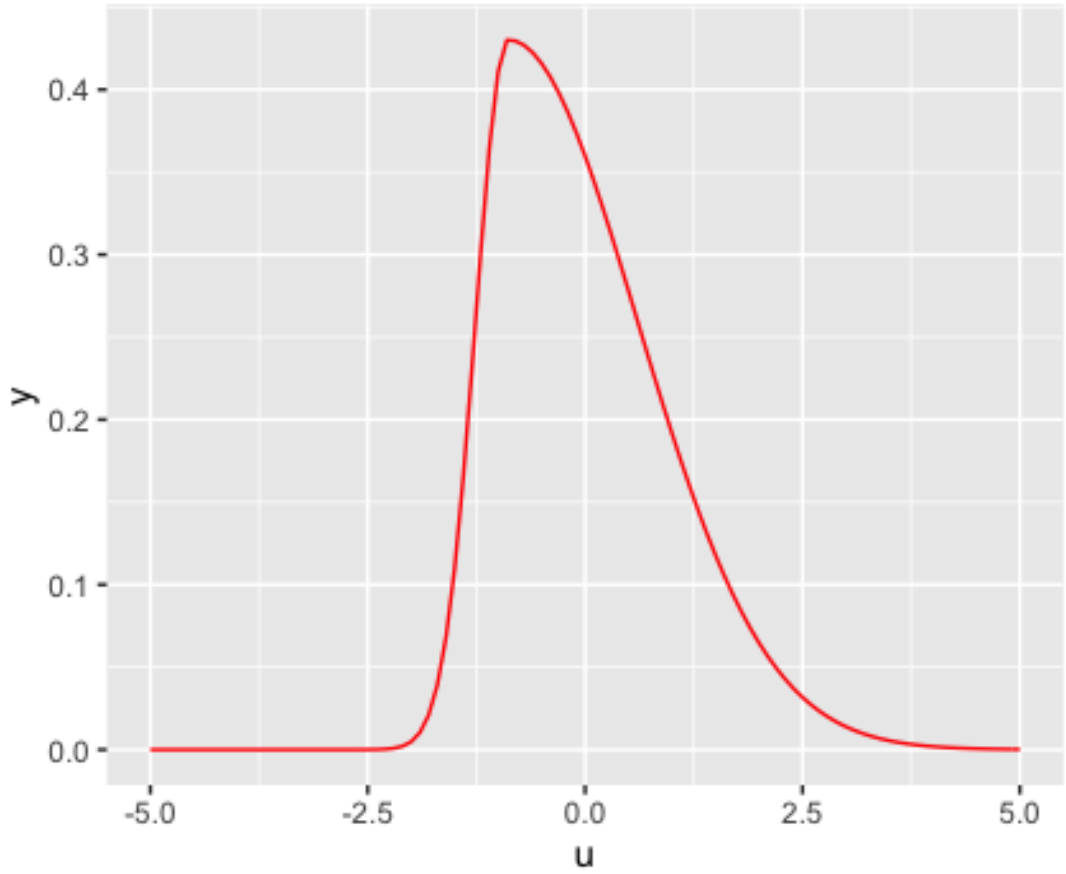
- Use the traditional t confidence interval: statistic ± (t*)SE where t* = the critical value and SE is standard error.
- Use the 2.5th and 97.5th percentiles of the bootstrap distribution.
- Checking if there is agreement between these two methods is a good way to test for normality.



Confidence Intervals Cont.

Calculating confidence intervals when the bootstrap distribution is non-normal and skewed.

- Bootstrap bias-corrected accelerated (BCa)
- Bootstrap tilting.
- More computationally expensive than the methods that assume normality.

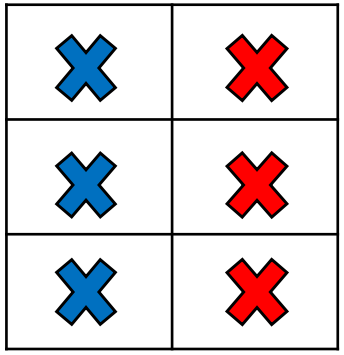


Permutation Testing

- Used to compare between two groups sort of like the two-sample t-test.
  - HO: there is no difference between the two groups
  - HA: there is some difference between the two groups.
- Use resampling to build a permutation distribution for our statistic. This distribution represents the sampling distribution if the null hypothesis is true.
- See where our sample statistic falls on our permutation distribution to calculate a P-value.
- How do we resample from both of our samples in a manner that is consistent with the null hypothesis?

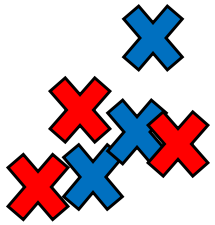
Visual example of Permutation Resampling.

Original 2 samples



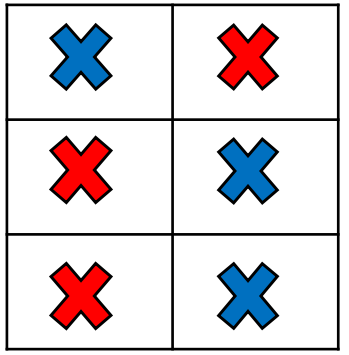
Scramble

Combined sample



Resample

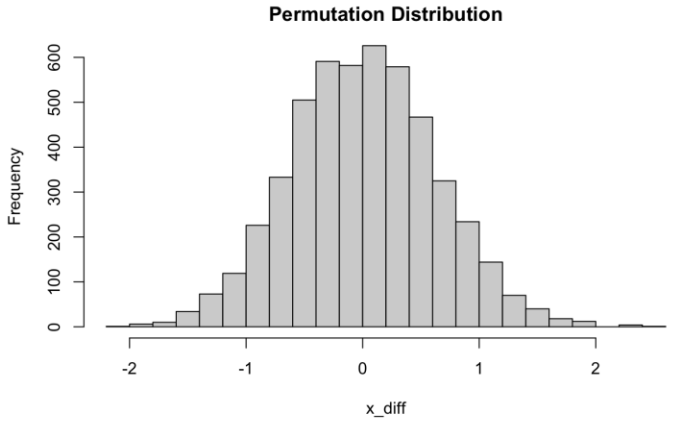
Resamples.



Calculate Statistic

μ_1-μ_2

Repeat 5000X



Permutation Test Recipe

- Take two samples
- Combine both of the samples into one large sample.
- Take two random resamples from this new sample WITHOUT replacement (resample sizes are equal to the original sample sizes).
- Calculate a sample statistic from the two resamples (eg the difference in means).
- Repeat this many many times.
- Build a distribution of your sample statistics.
- This distribution is called the permutation distribution, and it represents what the sampling distribution would be if the null hypothesis were true.
- Locate the difference between the variables of the original sample on the permutation distribution to get your P-values.

Let's Give It A Try.

Assumptions of Permutation Test.

- Random sampling from both populations (of course).
- The populations have identical distributions under the null hypothesis
  - Same mean
  - Same spread (standard deviation)
  - Same shape
- NO assumption of normality.

Additional Resources.

- Companion Chapter 18: Bootstrap Methods And Permutation Tests for the Practice of Business Statistics.