Inference Lab 1: the Bootstrap

Data Science Team

Today’s Activity

Today’s activity on the bootstrap will start with a demo of how to code a bootstrap for one or two variables. Start a new RNotebook and code your own bootstrap functions as we go. Then, with a partner, complete as many of the following questions as time allows. (Don’t forget to upload to Canvas.)

```r
library(dplyr)
library(ggplot2)
movies <- read.csv("http://s3.amazonaws.com/dcwoods2717/movies.csv")
glimpse(movies)
```

Observations: 2,961
Variables: 11
$ title <fct> Over the Hill to the Poorhouse, The Broadw...
$ genre <fct> Crime, Musical, Comedy, Comedy, Comedy, An...
$ director <fct> Harry F. Millarde, Harry Beaumont, Lloyd B...
$ year <int> 1920, 1929, 1933, 1935, 1936, 1937, 1939, ...
$ duration <int> 110, 100, 89, 81, 87, 83, 102, 226, 88, 14...
$ gross <int> 3000000, 2808000, 2300000, 3000000, 163245...
$ budget <int> 100000, 379000, 439000, 609000, 1500000, 2...
$ cast_facebook_likes <int> 4, 109, 995, 824, 352, 229, 2509, 1862, 11...
$ votes <int> 5, 4546, 7921, 13269, 143086, 133348, 2918...
$ reviews <int> 2, 107, 162, 164, 331, 349, 746, 863, 252,...
$ rating <dbl> 4.8, 6.3, 7.7, 7.8, 8.6, 7.7, 8.1, 8.2, 7....

Bootstrapping a Univariate Statistic

Task Choose a genre and create a data frame that only contains that data for that genre (but not one with a trivial number of observations). Use this genre throughout.

Task Choose a variable and bootstrap the mean. Now calculate a 95% confidence interval using the bootstrap principle (the mean, sample standard deviation, and sample size) and using the quantile function on the bootstrap replicates. How similar are they? If they are very similar, can you think of a scenario where the two methods would produce very different results? If they are very different, can you think of a scenario where they would be very different?

Question Now bootstrap the median. Which do you expect to be more robust (stable across bootstrap simulations), the mean or the median, and why?

Bootstrapping a Bivariate Statistic

Task Choose two continuous variables and bootstrap the correlation and bootstrap Pearson’s correlation. Plot the histograms along with the 95% confidence interval (from the quantiles). Also add a vertical line for the correlation in the full sample. Do you conclude the correlation is positive, negative, or that there is no clear evidence of correlation?
Permutation Test—a different test statistic

Today, we’ll use a simulation technique to test the hypothesis that two continuous variables are independent. This is our null hypothesis.

Recall that we need to construct a reference distribution for our statistic when the null is true. How can we simulate under the null? One option would be to permute the order of one of the variables and recompute the test statistic. To estimate the reference distribution, we need to do this many times. This is called a permutation test.

We can use the permutation test to estimate p-values, the probability of obtaining a test as extreme as the one observed under the null. For example, if we observe a sample correlation of 0.43, we can ask what are the chances of observing a correlation of 0.43 or more if the two variables are independent. If that value is small (typically, \( p < 0.05 \)), we reject the null since it is implausible that the two are uncorrelated. (Uncorrelated means no linear relationship; independent means no relationship at all.)

The key difference between the permutation test and the bootstrap is that the bootstrap samples with replacement while the permutation test samples without replacement.

**Task** Create a copy of your code for the bivariate bootstrap and call the new function something like permutation. Modify the code so that (a) sampling is done without replacement and (b) only y is affected (permutation tests scramble the order of y but do not change how many times each observation appears). Also, (c) make sure any references to boot are replaced in the code with something more appropriate like perm.

**Question** What is your Null Hypothesis?

**Question** Do you reject the null hypothesis as implausible? Plot a histogram of the correlations with a vertical line for the sample correlation. Also, make sure the labels are clear and include the \( p - value \) from the permutation test in the subtitle.

Choose a different genre

In any remaining time, repeat the above analyses for a different genre, making brief notes as to which things change and which don’t.