Bootstrap Applications

YOUR NAME

26 July, 2020

Exercises

1. Poker

An aspiring poker player recorded her winnings and losses over 50 evenings of play, the data is in the `openintro` package in the object `poker`. The poker player would like to better understand the volatility in her long term play.

a. Load the data and plot a histogram.

b. Find the summary statistics.

c. *Mean absolute deviation* or *MAD* is a more intuitive measure of spread than variance. It directly measures the average distance from the mean. It is found by the formula:

$$mad = \sum_{i=1}^{n} \frac{|x_i - \bar{x}|}{n}$$

Write a function and find the *MAD* of the data.

d. Find the bootstrap distribution of the *MAD* using 1000 replicates.
e. Plot a histogram of the bootstrap distribution.
f. Report a 95% confidence interval on the MAD.
g. ADVANCED: Do you think sample MAD is an unbiased estimator of population MAD? Why or why not?

2. Bootstrap hypothesis testing

Bootstrap hypothesis testing is relative undeveloped, and is generally not as accurate as permutation testing. Therefore in general avoid it. But for our problem in the notes, it may work. We will sample in a way that is consistent with the null hypothesis, then calculate a P-value as a tail probability like we do in permutation tests. This example does not generalize well to other applications like relative risk, correlation, regression, or categorical data.

a. Using the `HELPrct` data set, store the observed value of the difference of means for male and female.
b. The null hypothesis requires the means to be equal. Pick one group to adjust. Subtract the sample mean of this group and then add the sample mean of the other group to each data point in the group. Store in a new object called `HELP_null`.
c. Run `favstats()` to check that the means are equal.
d. On this new adjusted data set, generate a bootstrap distribution of the difference in sample means.
e. Plot the bootstrap distribution and a line at the observed difference in sample means.

f. Find a p-value.
g. How does the p-value compare with those in the notes.

3. Paired data

Are textbooks actually cheaper online? Here we compare the price of textbooks at the University of California, Los Angeles' (UCLA's) bookstore and prices at Amazon.com. Seventy-three UCLA courses were randomly sampled in Spring 2010, representing less than 10% of all UCLA courses. When a class had multiple books, only the most expensive text was considered.

The data is in the `openintro` package in the object `textbooks`.

Each textbook has two corresponding prices in the data set: one for the UCLA bookstore and one for Amazon. Therefore, each textbook price from the UCLA bookstore has a natural correspondence with a textbook price from Amazon. When two sets of observations have this special correspondence, they are said to be **paired**.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations. In `textbooks`, we look at the difference in prices, which is represented as the `diff` variable. It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices.

a. Is this data tidy? Explain.
b. Make a scatterplot of the UCLA price versus the Amazon price. Add a 45 degree line to the plot.
c. Make a histogram of the differences in price.

The hypotheses are: H_0: $\mu_{diff} = 0$. There is no difference in the average textbook price. H_A: $\mu_{diff} \neq 0$. There is a difference in average prices.

d. To use a t distribution, the variable `diff` has to independent and normally distributed. Since the 73 books represent less than 10% of the population, the random sample is independent. Check normality using `qqnorsim()` from the `openintro` package. It generates 8 qq plots of simulated normal data that you can use to judge the `diff` variable.
e. Run a t test on the `diff` variable. Report the p-value and conclusion.
f. Create a bootstrap distribution and generate a 95% confidence interval on the mean of the differences, the `diff` column.
g. If there is really no differences between book sources, the variable `more` is a binomial and under the null the probably of success is $\pi = 0.5$. Run a hypothesis test using the variable `more`.
h. Could you use a permutation test on this example? Explain.