CompSci 116: A/B Testing

Jeff Forbes

March 5, 2019

Plan for The Week

- Assignments
  - Project 2
  - Final Project
- A/B Testing: Comparing two random samples
- Percentiles & Confidence Intervals
- Lab on Thursday: The Bootstrap

Final Project

- Demonstrate proficiency in techniques from class applied to dataset of your choosing
- Data!
  - Domain: ?
  - Scale: Large enough
    - Categorical and numerical variables
    - More than 100 observations
  - Accessible & Manageable

Project: Making a Notebook

1. Go to `https://jupyterhub.cs.duke.edu`
2. Make a folder (if desired)
3. Create a new Python 3 Notebook
4. Copy preamble code from any previous notebook
5. Refer to the Markdown Cheatsheet for how to add links, lists,headers, etc.
6. Upload data as CSV in the same directory as notebook

Deflategate!

EXECUTIVE SUMMARY

On January 18, 2015, the New England Patriots and Indianapolis Colts played in the AFC Championship Game at Gillette Stadium in Foxborough, Massachusetts to determine which team would advance to Super Bowl XLIX. During the first half of the game, a question was raised by the Colts concerning the inflation level of the footballs being used by the Patriots. As a result, at halftime, members of the officiating crew assigned to the game, overseen by a senior officiating supervisor from the National Football League (the "NFL" or the "League"), tested the air pressure of footballs being used by each of the Patriots and the Colts. All eleven of the Patriots game balls tested measured below the minimum pressure level of 12.5 pounds per square inch ("psi") allowed by Rule 2 of the Official Playing Rules of the National Football League (the "Playing Rules") on both of two air pressure gauges used to test the balls. The four Colts balls tested each measured within the 12.5 to 13.5 psi range permitted under the Playing Rules on at least one of the gauges used for the tests.

A/B Testing

- Two random samples:
  - Sample A
  - Sample B

- Question: Are they drawn from the same underlying distribution?

- Answer by A/B testing

The Hypotheses

- Null:
  - The two samples are drawn from the same underlying population distribution; they look like two random draws from the same set.
- Alternative:
  - The samples are drawn from different distributions; they don’t look like random draws from the same set.

Permutation Test

- **Null:** The two samples are drawn randomly from the same underlying distribution.
- If the null is true, all rearrangements of the variable values among the two samples are equally likely. So:
  - compute the observed test statistic
  - then shuffle the attribute values and recompute the statistic; **repeat**; compare with the observed statistic

Smoking and Birth Weights

- Measured the birth weights of many babies and whether the mom smoked.
- **Null:** The distribution of birth weights is the same for both smoker and non-smoker moms.
- How to simulate the distribution of the test statistic, if the null hypothesis is true?

Hypothesis Test

- **Null:** The distribution of birth weights is the same for both smoker and non-smoker moms.
- If the null is true, we can model each birth as:
  - randomly choose a birth weight
  - randomly choose whether the mom is a smoker or non-smoker
- How do we know the distribution of birth weights?

The Bootstrap

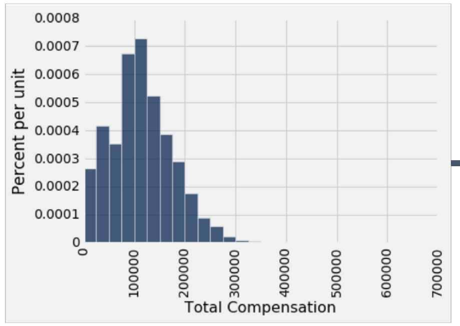
- A technique for simulating repeated random sampling

- All that we have is the original sample
  - … which is large and random
  - Therefore, it probably resembles the population

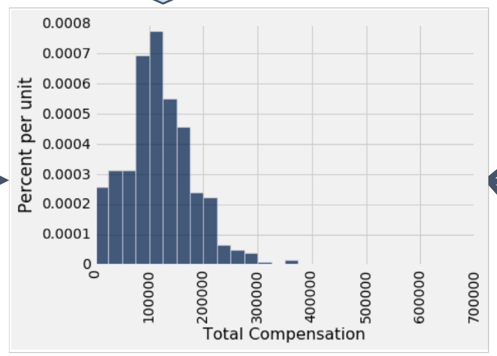
- So we sample at random from the original sample!

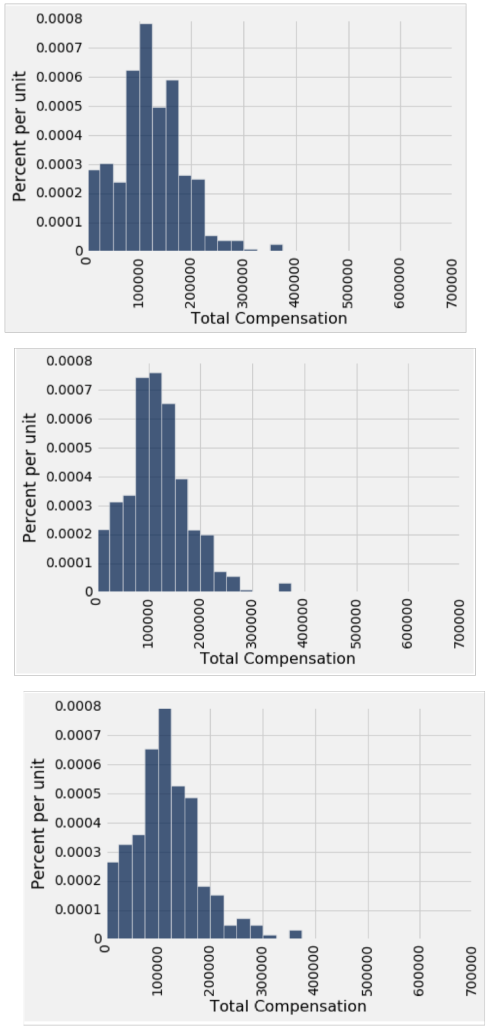
Why the Bootstrap Works

population



sample





resamples

All of these look pretty similar, most likely.

Key to Resampling

- From the original sample,
  - draw at random
  - with replacement
  - as many values as the original sample contained

- The size of the new sample has to be the same as the original one, so that the two estimates are comparable

(Demo)

Deflategate

- **Null**: Each group is like a sample drawn at random without replacement from all 15 footballs.
- **Alternative**: The Patriots' values are too *low* for them to look like a random sample from the 15 balls.

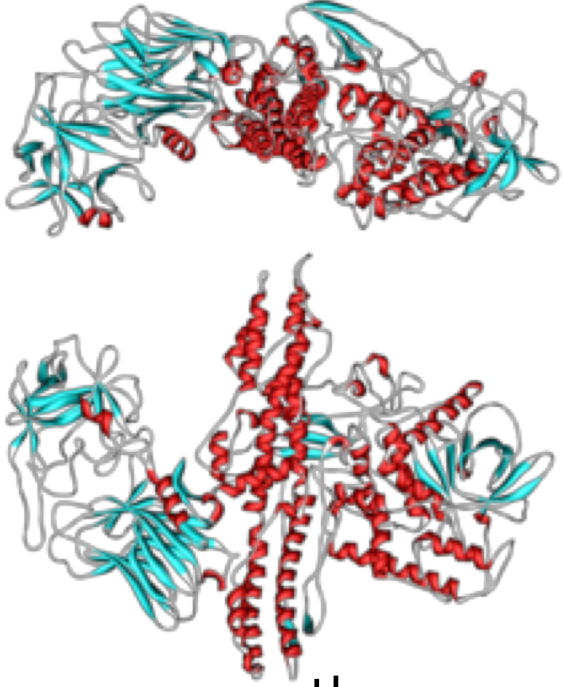
- **Test Statistic**: Colts' average - Patriots' average
  - *P*-value direction: ?

Causality

- Use Randomized Control Experiments to establish causality
  - Sample A: control group
  - Sample B: treatment group
- If the treatment and control groups are selected at random, then you can make causal conclusions.
- Any difference in outcomes between the two groups could be due to
  - chance
  - the treatment

(Demo)

Botox for Back Pain!



- 31 patients
  - 16 injected with Botulinim toxin A
  - 15 injected with Saline (placebo)
- **Null**: Distribution of 31 potential *control* scores is the same as the distribution of all 31 potential *treatment* scores
- **Alternative**: Distribution of 31 potential *control* scores is different from the distribution of all 31 potential *treatment* scores
- **Test Statistic**: *Distance* between the two proportions of relief in each group
  - | *control* proportion – *treatment* proportion |

Computing Percentiles

The 80th percentile is the value in a set that is at least as large as 80% of the elements in the set

For `s = [1, 7, 3, 9, 5]`, `percentile(80, s)` is 7

The 80th percentile is ordered element 4: `(80/100) * 5`

Percentile

Size of set

For a percentile that does not exactly correspond to an element, take the next greater element instead

The `percentile` Function

- The *p*th percentile is the value in a set that is at least as large as *p*% of the elements in the set
- Function in the `datascience` module:

```
percentile(p, values)
```

- **`p`** is between 0 and 100
- Returns the *p*th percentile of the array

http://bit.ly/FoDS-s19-0305