SDS 220: Homework 5 (Part 2)

Statistical Inference

Instructions

In this homework, we use in R to conduct statistical inference on American public opinions/public health sentiment during the Ebola virus epidemic of 2013–2016. In particular, we will:

- form a bootstrap sampling distribution for the proportion of New Yorkers who favored a mandatory quarantine for close contacts of Ebola patients
- construct 95% confidence intervals

Motivation

The 2013–2016 epidemic of Ebola virus disease, which was largely centered in West African countries such as Guinea, Liberia, and Sierra Leone, caused major loss of life and socioeconomic disruption. It is to-date the most severe outbreak of Ebola virus disease in history. In October 2014, a doctor who had recently been treating Ebola patients in Guinea presented at a New York City hospital with slight fever; he was subsequently diagnosed with Ebola.

Shortly after this diagnosis became public, a poll was taken by *NBC 4 New York/The Wall Street Journal/Marist* of 1,042 New York adults, asking them whether they favored a "mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient."

The survey results are stored in the dataset `ebola_survey`; the following code displays the first six rows of that data set:

```r
ebola_survey |> glimpse()
```

```
Rows: 1,042
Columns: 1
$ quarantine <fct> favor, favor, favor, favor, favor, against, favor, favor, f~
```

Bootstrap Sampling Distributions

Let's say that we want to understand the proportion p of New York City residents who favor a mandatory 21-day quarantine. We will do this using three different datasets, each representing different possible sizes for our initial sample (i.e., our observed data):

- The full dataset, where $n = 1042$
- A moderate-sized subset of the full data with $n = 100$
- A small subset of the full data with $n = 25$

I've created these second two data sets for you below:

```r
set.seed(32023)

# Moderately-sized dataset
mod_survey <- ebola_survey[sample(1:nrow(ebola_survey), 100), ]

# Small dataset
small_survey <- ebola_survey[sample(1:nrow(ebola_survey), 25), ]
```

Confidence Intervals

A confidence interval is a range of plausible values for p that are consistent with the data we happened to observe.

The range of values that we report is deeply connected to the sampling distribution: the *spread* of this distribution (whether measured by a standard deviation or a quantile-based metric) gives a sense of how far away, on average, we expect $\hat{p}$ to be from the truth. If $\hat{p}$ tends to be very close to the truth, then our confidence interval would not need to be very wide: we're reasonably confident that p is within a small interval of our estimate. If $\hat{p}$ is more variable, then our confidence interval would need to be wider to accommodate that: we're less confident that $\hat{p}$ is close to p, so we need a wider confidence interval in order to be (reasonably) confident we've captured p.

The bootstrap distribution approximates the sampling distribution, so we can use the bootstrap distribution to construct confidence intervals for p. We find a $(1-\alpha) \times 100\%$ confidence interval by taking the $\alpha/2$ and $1-\alpha/2$ percentiles of the bootstrap distribution as our lower and upper bounds.

Questions

Exploring the Original Data

Question 1: When *NBC 4 New York/The Wall Street Journal/Marist* conducted this survey, who do you think they wanted to draw conclusions about? All New York City residents? All New Yorkers? All adults in the United States?

Question 2: Do you have any potential concerns about generalizability or non-random sampling in this context?

Question 3: Use R to (a) create a bar plot summarizing the observed data distribution and (b) construct a frequency table showing the number of respondents who "favored" or were "against" imposing a mandatory quarantine.

Point Estimation

Question 4: Use find a point estimate $\hat{p}$ for the true proportion of New York adults who favored a mandatory 21-day quarantine period for close contacts of Ebola patients in each of the three data sets.

Question 5: On average (across repeated samples of the same size from the full population of all New York adults in October 2014), each of these point estimates should equal the truth. In other words, their sampling distributions should be centered at the true proportion. Given this, which of these three point estimates would you feel most comfortable reporting? Why?

Bootstrap Approximation to the Sampling Distribution

Question 6: Bootstrap the `quarantine` column of each of the three datasets, `ebola_survey`, `mod_survey`, and `small_survey`, and then visualize the three bootstrap distributions. Describe the shape of these three distributions. What do you notice?

Question 7: Calculate the sample mean for each of your three bootstrap distributions. What do you notice about the centers of the bootstrap distributions?

Question 8: Calculate the sample standard deviation for each of your three bootstrap distributions. What do you notice about the spreads of the bootstrap distributions?

Question 9: Use the three bootstrap distributions you found above to create a 95% bootstrap confidence interval for the true proportion of New York adults who support a mandatory quarantine. What do you notice about the width of the confidence intervals?

Question 10: Interpret the 95% bootstrap confidence interval based on the full sample of $n = 1042$ New Yorkers.