Variability of statistic + Intro to Bootstrap

10/7/24

Where we are going

We are leaving the world of EDA and beginning to enter the world of inference and modeling!

- Want to answer questions about a population, but must rely on a sample
- Collect data from sample –> calculate statistics
- What can we say about the statistics?
- Data are random! So how sure are we about our conclusions?

Statistics starts here!

Inference

Statistical inference is the process of using sample data to make conclusions about the underlying population the sample came from

- **Estimation**: using the sample to estimate a plausible values for the unknown parameter
- **Testing**: evaluating whether our observed sample provides evidence for or against some claim about the population

Estimation questions

- I want to know the true average number of hours of sleep Middlebury students get a night. Based on a sample of students, what might be a "good estimate" of the true average?
- Is the true average number of of hours of Middlebury students get a night less than 7 hours?
  - Questions here are about population parameter (in this case, μ)
  - All we have access to is the data $x_1, x_2, \ldots, x_n$ from which we can calculate some statistics

Activity

- Target population:
- Sampling method:
- Population parameter:
- Statistics we can calculate:

Point estimate

- $\bar{x}$ is often times a sensible estimate for μ
- $\bar{x}$ is an example of a **point estimate**: a single number used to estimate a true but unknown population parameter
  - i.e. a point estimate is a statistic with a specific purpose
  - Other examples include s for σ, observed proportion $\hat{p}$ for true proportion p,
- What are desirable characteristics of a "good" point estimate?
  - Do we believe that $\bar{x} = \mu$?

Variability of statistic

- Two datasets collected under identical procedures will differ. As a result, value of the point estimate we obtain are also different
  - Activity cont.
- Thus, there exists the notion of a **sampling distribution** *of the statistic*: how the statistic behaves under repeated random samples obtained via the same sampling procedure
  - The variability associated with the sampling distribution of the statistic is called the **standard error**
    - Note: “error” ≠ bad
  - This is in contrast to the standard deviation, which describes variability in the individual data points and not the statistic
- Population distribution vs. sample distribution vs. sampling distribution

Sampling distribution

- Of course, sampling distribution of the statistic depends on underlying distribution of the population
- Sometimes, we ***assume*** that the population/data have a very specific behavior, and this allows us to *exactly* define/quantify the sampling distribution
  - We will see this in a couple of weeks
- If we don't want to make assumptions, what do we do?
  - Could conduct a census! That way we can answer any questions we want about the population. But that's impractical…
  - How to obtain more samples cheaply and quickly?

Bootstrap

Bootstrapping is a flexible method that allows us to move forward in an analysis without knowing exactly how the data were generated.

Procedure

1. Assume we have a sample $x_1, x_2, \ldots, x_n$ from the population. Call this sample $\vec{x}$. Note the sample size is n
2. Choose a large number B. For b in $1, 2, \ldots, B$:
    i. Resample: take a sample of size n with *replacement* from $\vec{x}$. Call this set of resampled data $\vec{x}_b^*$
    ii. Calculate: calculate and record the statistic of interest from $\vec{x}_b^*$

At the end of this procedure, we will have a distribution of **resample or bootstrap statistics**

Demonstration

- Activity cont.
- Live code demonstration

Why resample with replacement?

- We want to understand the sampling error of the sampling distribution!
- What would the bootstrap samples $\vec{x}_b^*$ look like if we sampled *without* replacement?
  - Samplign with replacement -> zero variation in the resampled statistics
- Resampling with replacement will give us “new” datasets that are similar to original sample distribution but not exactly the same!

Remarks

- Relies on having a representative original sample!
  - Resampling from initial sample should be roughly equivalent to sampling directly from the population
- Requires computational tools!
  - We need B to be large enough to accurately capture variability. $B = 5000$ or $B = 10000$ sufficient in this class
  - More complex problems will require larger B
- Bootstrapping can fail!
- Bootstrapping is *not* a solution to small sample sizes!!