Distribution-free confidence intervals and the bootstrap

Patrick Breheny

November 12

Introduction

- In the previous lecture, we discussed nonparametric tests, but avoided any discussion of nonparametric confidence intervals; intervals are the subject of today's lecture
- We will discuss two general approaches to constructing distribution-free confidence intervals:
  - Inverting nonparametric hypothesis tests
  - A more modern, computationally-intensive approach known as the "bootstrap"

Inverting the Wilcoxon rank sum test

- We have inverted hypothesis tests to construct confidence intervals several times in this course
- This begs the natural question: if we flip the MWW test around, do we get a confidence interval for something? If so, what?
- Before answering that question, we first need to generalize our description of the MWW test to include testing for differences other than zero

Testing nonzero location shifts

- Consider introducing a “shift” parameter Δ in which we modify all the observations in group 1 by adding Δ to them prior to carrying out the Wilcoxon rank sum test
- In other words, the Wilcoxon rank sum test proceeds exactly as usual, but the data in group has been modified so that x_i becomes $x_i + \Delta$ (the data in group 2 is left alone)
- Then, as we have seen several times, we could carry out such a test for all values of Δ and collect all the non-rejected values into an interval for the shift in location between the two groups
- Note: Such an interval is typically referred to as “semiparametric” rather than “nonparametric” in the sense that we had to introduce the parameter Δ in order to carry out the test

The location shift confidence interval

- For the tailgating data, this procedure produces the confidence interval $[0.57, 7.51]$ for Δ
- In words, illegal drug users seem to follow the car in front of them about 1-7 meters closer than drivers who do not use illegal drugs
- It is worth noting that we could also obtain a point estimator $\hat{\Delta}$ by solving for the value of Δ such that $p = 1$
- For the tailgating data, $\hat{\Delta} = 4.3$; note that this is not necessarily equal to the difference in medians, which for the tailgating data was 5.0

The bootstrap

- A different approach to making nonparametric confidence intervals is the *bootstrap*
- Although the theory underlying the bootstrap (why it works, and when it doesn't) is a deep and complex subject, the idea behind it is simple
- We'll first illustrate the idea using the tailgating data to obtain a nonparametric confidence intervals for the difference in median following times, then say a few words about why it works

Bootstrap procedure: Difference in medians

- To "bootstrap" a sample, we simply place all 55 observed following distance values for the illegal drug user group in an urn and randomly draw 55 observations back out again (with replacement)
- Calculate the median for this "bootstrapped" sample
- Do the same for the non-illegal drug user group, and calculate the difference in medians
- Repeat the above a large number of times (say, $B = 10,000$), obtaining a long list of differences in medians
- The (percentile) bootstrap confidence interval is the interval that contains the middle 95% of this list of values

Bootstrap results: Tailgating study

- For the tailgating study, this interval is (1.1, 7.6); similar to the Wilcoxon interval from earlier, although not identical, since the assumptions that go into the two approaches are different
- The great virtue of the bootstrap, like that of the permutation test, is its versatility – this same technique can be used to obtain nonparametric confidence intervals for almost any other quantity one cares to define
- For this reason, Casella & Berger (2002) call it "perhaps the single most important development in statistical methodology in recent times"

Derivation of bootstrap

- Suppose we are interested in deriving the distribution of estimate $\hat{\theta} = \theta(\mathbf{x})$
- It's actual distribution $P(\hat{\theta} \in A)$ is given by

$$\int \cdots \int 1\{\theta(\mathbf{x}) \in A\} dF(x_1) \cdots dF(x_n)$$

- There are two problems with evaluating this expression directly
- The first is that we do not know F; a natural solution to this problem is to plug in the empirical CDF, $\hat{F}$:

$$\int \cdots \int 1\{\theta(\mathbf{x}) \in A\} d\hat{F}(x_1) \cdots d\hat{F}(x_n)$$

Monte Carlo approach

- The second problem is that this integral is difficult to evaluate
- However, we can approximate this answer instead using *Monte Carlo integration*
- Instead of actually evaluating the integral, we approximate it numerically by drawing random samples of size n from $\hat{F}$ and finding the sample average of the integrand
- This approach gives us the bootstrap
- By the law of large numbers, this approximation will converge to the actual value of the integral as the number of random samples that we draw goes to infinity

Resampling

- What does a random sample drawn from $\hat{F}$ look like?
- Because $\hat{F}$ places equal mass at every observed value x_i, drawing a random sample from $\hat{F}$ is equivalent to drawing n values, with replacement, from $\{x_i\}$
- This somewhat curious phenomenon in which we draw new samples by sampling our original sample is called *resampling*

Bootstrap accuracy

- Thus, the bootstrap works by using $\hat{F}$ to approximate F, and using Monte Carlo integration to approximate the true distribution of $\hat{\theta}$ given by the full integral over $\mathbb{R}^n$
- It's worth pointing out that the accuracy of the bootstrap calculations depends on both B, the number of bootstrap samples, and n, the number of observations
- If B is small, then the Monte Carlo approximation might not be accurate; this is usually easy to fix, because you can always increase B – the only cost is computing time
- If n is small, then $\hat{F}$ might not be a good estimate of F; to fix this, you would actually need to go out and gather more data

Summary

There are two primary ways of constructing confidence intervals without assuming we know what family the distribution of the data belongs to:

- Inverting a nonparametric test; this involves introducing a parameter (such as the location shift Δ) and thus, such intervals are usually referred to as *semiparametric* confidence intervals
- The bootstrap; this involves using the empirical CDF $\hat{F}$ to estimate the true CDF F and Monte Carlo integration to approximate the true n-dimensional integral we are interested in

The above description makes the bootstrap sound complicated, but the idea is actually quite straightforward and extremely versatile