Chapter 9: Applications to Computing

9.7: Bootstrapping (for Hypothesis Testing)

(From "Probability & Statistics with Applications to Computing" by Alex Tsun)

9.7.1 Motivation

We've just learned how to perform a generic hypothesis test, where in our examples we were especially often able to use the Normal distribution and its CDF due to the CLT. But actually, there are tons of specialized other hypothesis tests which won't allow this. For example:

- The t-test for equality of means when variance is unknown.
- The χ^2-test of independence (testing whether to quantities are independent or not).
- The F-test for equality of variances (testing whether or not the variances of two populations are equal or not).

There are many more that I haven't even listed because I probably have never heard of them myself! These three above though involve three distributions we haven't learned yet: the t, χ^2, and F distributions. But because you are a computer scientist, we'll actually learn a way now to completely erase the need to learn each specific procedure, called **bootstrapping**!

9.7.2 The Bootstrap

Bootstrapping is a stellar example of why CS people need to take a course called something like "Probability & Statistics for Computer Scientists". Bootstrapping was invented by Bradley Efron in 1979, who has many accolades largely in part to this particular idea:

- President of the American Statistical Association
- Professor of Statistics at Stanford University
- Founding Editor of the Annals of Applied Statistics
- Won National Science Medal

Disclaimer: I'm not going to teach you everything there is about bootstrapping, just what is necessary for the application of hypothesis testing.

Recall from 8.3 that a p-value is "the probability of, *under the null hypothesis*, of observing a difference at least as extreme." Remember our first application was Probability via Simulation, and since a p-value is just a probability, we will try something very similar! A one sentence summary of bootstrapping:

"The bootstrap provides a way to calculate probabilities of statistics using code."

This application is rather short, so we just need to get through one idea before revealing it!

Example(s)

Main Idea: We have some (not enough) data and want more. How can we "get more"?

Imagine: You have 1000 iid coin flip samples, x_1,...,x_{1000} which are all 1's and 0's. Your boss wants you to somehow get/generate 500 more (independent) samples.

How can you "get more (iid) data" without actually having access to the coin? There are two proposed solutions below: both of which you could theoretically come up with, but only one of which which I expect most of you to guess.

Solution Here are the two ways we might approach this.

1. **Estimate** the parameter p of Ber(p) (e.g., with max-likelihood), then generate more samples.
2. **Resample** the data: sample (uniformly) from the same dataset 500 times, **with** replacement.

In fact, in our scenario, these two are completely equivalent! Why? If for example there were 750/1000 heads and we resample with replacement uniformly, the probability we get a 1 is just 750/1000. If we estimate the parameter to be 750/1000, then each time we also will get a 1 with probability 750/1000. □

However, the resampling method is much more generalizable: if we wanted to get more samples of human heights for example (the exact distribution is completely unknown to us), we would only be able to do the second way! This is the main idea of bootstrapping: "**Sampling with Replacement**",

9.7.3 Bootstrapping for p-values

I think this idea is best illustrated by example, as usual.

Example(s)

A colleague has collected samples of **weights** of labradoodles that live on two different islands: CatIsland and DogIsland. The colleague collects 48 samples from CatIsland, and 43 samples from the DogIsland. The colleague notes ahead of time that she thinks the labradoodles on DogIsland have a higher spread of weights than CatIsland. You are skeptical. *You and your colleague do however agree to assume that their true means are equal.* Here is the data:

CatIsland Labradoodle Weights (48 samples): 13, 12, 7, 16, 9, 11, 7, 10, 9, 8, 9, 7, 16, 7, 9, 8, 13, 10, 11, 9, 13, 13, 10, 10, 9, 7, 7, 6, 7, 8, 12, 13, 9, 6, 9, 11, 10, 8, 12, 10, 9, 10, 8, 14, 13, 13, 10, 11

DogIsland Labradoodle Weights (43 samples): 8, 8, 16, 16, 9, 13, 14, 13, 10, 12, 10, 6, 14, 8, 13, 14, 7, 13, 7, 8, 4, 11, 7, 12, 8, 9, 12, 8, 11, 10, 12, 6, 10, 15, 11, 12, 3, 8, 11, 10, 10, 8, 12

Perform a hypothesis test, computing the p-value using bootstrapping.

Solution Step 5 is the only part where bootstrapping is involved. Everything else is the same as we learned in 8.3!

1. **Make a claim**.

   The spread of labradoodle weights on DogIsland is (significantly) larger than that on CatIsland.

2. **Set up a null hypothesis H_0 and alternative hypothesis H_A.**

   $$H_0 : \sigma_C^2 = \sigma_D^2 \qquad\qquad H_A : \sigma_C^2 < \sigma_D^2$$

   Our null hypothesis is that the spreads are the same, and our alternative is what we want to show. Here, spread is taken to mean "variance".

3. **Choose a significance level α (usually $\alpha = 0.05$ or 0.01).**

   Let's say $\alpha = 0.05$.

4. **Collect data**.

   This is already done for us.

5. **Compute a p-value, $p = \mathbb{P}$(observing data at least as extreme as ours $\mid$ H_0 is true).**

   Here is when we use knowledge of coding to compute our p-value. The idea is probability by simulation: we assume H_0 is true; that is, the variances in both samples $\mathbf{x}$ and $\mathbf{y}$ are the same. That is, we assume there is some global population (a master island if you will), and some seismic event occurred which split the master island into CatIsland and DogIsland (so they have the same variance).

   Because of this, we can combine the two samples into a single one of size $48 + 43 = 91$ (in our case, we've also assumed the means are the same, so this is okay). Then, we repeatedly **bootstrap** this combined sample (let's say 50,000 times): we sample with replacement a sample of size 48, and of size 43, and compute the sample variances of these two samples. Then, we compute the sample proportion of times the difference in variances was at least as extreme, and that's it! See the pseudocode below, and reread these two paragraphs.

Algorithm 1 Bootstrapping for p-value for $H_0 : \sigma_C^2 = \sigma_D^2$ vs $H_A : \sigma_C^2 < \sigma_D^2$

Given: Two samples $\mathbf{x} = [x_1, \ldots, x_n]$ and $\mathbf{y} = [y_1, \ldots, y_m]$.
obs_diff $\leftarrow s_y^2 - s_x^2$ (the difference in sample variances).
combined $\leftarrow$ concat$(x, y) = [x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m]$ (of size $n + m$).
count $\leftarrow 0$.
for $i = 1, 2, \ldots, 50000$ **do** ▷ Any large number is fine.
    $x' \leftarrow$ resample(combined, n) ***with*** replacement. ▷ Sample of size n from combined.
    $y' \leftarrow$ resample(combined, m) ***with*** replacement. ▷ Sample of size m from combined.
    diff $\leftarrow s_{y'}^2 - s_{x'}^2$. ▷ Compute the difference in sample variances.
    if diff $\geq$ obs_diff **then** ▷ This line changes depending on the alternative hypothesis.
        count $\leftarrow$ count $+ 1$.
p-val $\leftarrow$ count$/50000$.

Again, what we're doing is: assuming there was this master island that split into two (same variance), what is the probability we observed a sample of size 48 and a sample of size 43 with variances at least as extreme as we did? That is, if we were to repeat this "separation" process many times, how often would we get a difference so large? We don't have the other labradoodles from the master island, so we bootstrap (reuse our current samples). It turns out this method leads to a good approximation to the true p-value!

It's important to note that the alternative hypothesis is EXTREMELY IMPORTANT. If instead we wanted to assert $H_A : \sigma_C^2 \neq \sigma_D^2$, we would have used absolute values for diff and obs_diff. Also, for example, if we wanted to make a statement about the *means* μ_C and μ_D instead, we would have computed and compared the sample means instead of the sample variances.

It turns out we get a p-value of approximately 0.07. (Try coding this up yourself!)

6. **State your conclusion. Include an interpretation in the context of the problem.**

   Since our p-value of 0.07 was greater than $\alpha = 0.05$, we *fail to reject* the null hypothesis. There is insufficient evidence to show that the labradoodle spreads are different across the two islands.

Actually, this two-sample test for difference in variances is done by an "F-Test of Equality of Variances" (see Wikipedia). But because we know how to code, we don't need to know that!

□

You can imagine bootstrapping for other types of hypothesis tests as well! Actually, bootstrapping is a powerful tool which also has other applications.