Guesstimating

Alex Sverdlov
alex@theparticle.com

1 Bootstrapping

Bootstrapping is a technique that is used to calculate error bounds on a measurement.

Suppose we are conducting a field study of income levels, and ask N random people their income (where N is generally *small*—talking to physical people is expensive).

We wish to publish the *mean* income level by region, but worried about reproducibility: if another researcher did the same experiment (asking potentially different N random people the same question), what are the chances they would end up with a similar *mean*?

One way to overcome this is to conduct multiple experiments, and take an average of the results that each such experiment generates. This dataset of averages is often normally distributed, so error bounds can be looked up (or calculated) from the normal distribution. The trouble is we only have data from 1 experiment:

$$D = \langle X_1, X_2, X_3, \ldots, X_N \rangle$$

With bootstrapping, we can guestimate the error bounds by "simulating" multiple experiments, and then seeing what fraction of such experiments falls within a certain distance of our calculated mean.

We define a function *sample*, which returns a list of N elements, where each element is randomly chosen from D (with replacement; some elements from D may be selected more than once). We then create another list S where each entry is: $mean(sample(D))$. This in essence gives us the dataset of averages we lacked.

One way to use S is to view it as the histogram of where the *mean* shows up in various sampling "experiments". Using this histogram, we can note which range includes middle 95% of the data. Out of 20 future experiments, 19 will fall into the same range—assuming our original D was not outrageously "special".

This setup can be applied to other measurables—not just *mean*.

2 Permutation tests

Using a similar idea to Bootstraping, we can determine if there is a significant difference between two (or more) datasets. Often, the datasets in question refer to some experiment we are conducting: the control group, and the treatment group.

For example, suppose we wish to test if a vaccine works to prevent covid-19. We take N volunteers, inject random half of them with our vaccine, and the rest with a placebo. Nobody except a database knows who is getting the vaccine and who is getting the placebo. Over the following few weeks/months, we monitor all the test subjects—and note if they test positive for covid-19.

If the vaccine is very effective, we expect the positive counts for vaccinated population to be low, or zero. Similarly, if the vaccine is not effective at all, we would expect the positive counts for both populations to be more or less the same—any differences in results to be caused by chance.

With a permutation test, we rank the results we got to the "pure chance" results we generate by shuffling the "test positive" label and recording the rates per population.

3 German tank problem

Suppose we observe tank serial numbers: 19, 40, 42 and 60. How many tanks are there? Obviously there are at least 60 tanks.

Can there be 120 tanks? If we assume there is nothing special about our observations (that we didn't just observe the first few tanks from the assembly line), then for there to be 120 tanks, we would've had to observe 4 tanks all below the median of all tanks produced—that is not impossible, but unlikely.

Can the total be 60 tanks? What are the chances of us randomly observing the exact last tank produced?

There are several ways of solving this problem, the approach presented is in the theme of this lecture: random sampling.

For each tank number T from 1 to 1000 (or some other domain specific range), randomly pick 4 serial numbers and note the highest H: We treat this as vote from H to T. Repeat this a sufficient number of times to build up a histogram of T values for each H. Since we observed 60 as the highest, we look for the median T value for $H = 60$.

With the above setup, we end up with median around 77, and middle 95% of the histogram in range 67 to 96 (in other words, we have a lot of confidense that our estimate is likely *ok*).

4 Doomsday argument

The above solution to the German tank problem can be applied in *simpler* scenarios: we find a widget with serial number 100, with 90% confidence, what's the maximum number of widgets?

We can perform a similar sampling experiment, but this problem is much simpler—reverse the perspective: we find a widget with serial number "100". If we are not particularly lucky, than this "100" is in the middle 90% of all the widget serial numbers, which leads to:

- if 100 is the 95th percentile, then there are a total of around 106 widgets.

- if 100 is the 5th percentile, then there are a total of around 2000 widgets.

Either way, with 90% confidence, the total number of widgets is between 106 and 2000.

If we just want "better than guessing" outcome, we can assume that 100 is in the interquartile range (IQR, the middle 50% of all widget serial numbers). The range gets smaller: with 50% confidense, we can guestimate number of widgets to be between 133 and 400.

So what's so "doomsday" about this? If we pretend that every human that is born has a serial number, we can apply this argument to humanity—we estimate the number of people that have ever lived, and with any confidense we want, estimate the total number of humans that will ever live. Some folks go farther and project population growth by years, and come up with a year the last human will be born.

The numbers vary depending on what estimates we use, and what confidense level we choose to guestimate at, but the basic idea is that if we assume we are not "special" (that we are not in the first 0.01% of all human's history), then humanity's end is likely nearer than most folks are comfortable with. e.g. will humanity still exist in a million years? How about in 100k years?