Estimation, Part 2: The Bootstrap

Confidence Intervals

We saw in the last set of slides that we could construct confidence intervals by knowing something about how the sampling distribution of the parameter that we care about is distributed.

In certain cases (we saw specifically for the case of a mean), that we can know analytically the sampling distribution. (Eg. $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ follows a t-distribution).

Problem: These relied on certain assumptions (population approximately normal, large enough sample size), or relied on us only wanting to estimate particular parameters. Otherwise, our conclusions about the sampling distributions might not hold.

The Bootstrap

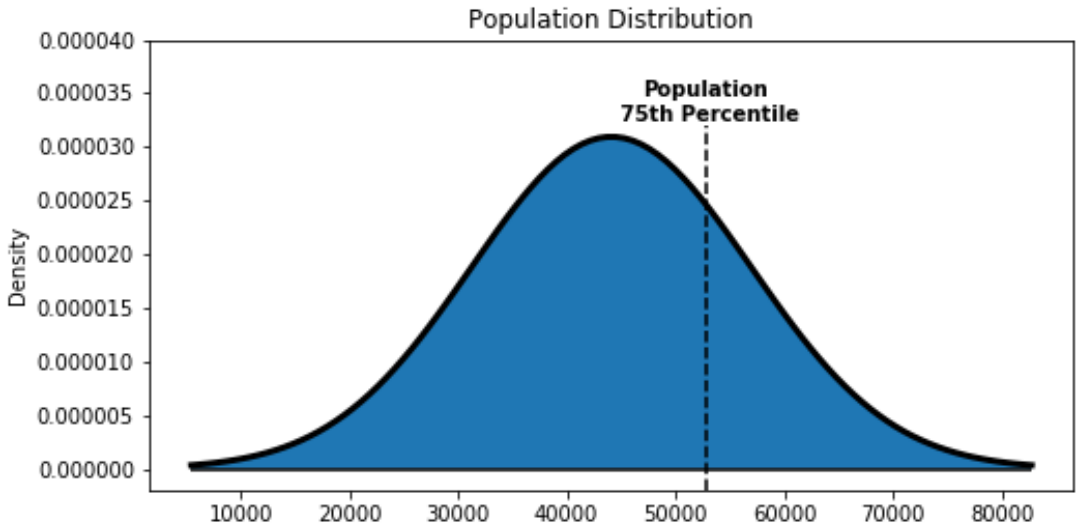
Big Idea: Rather than analytically trying to determine what the sampling distribution is, we can just approximate it using our sample.

We take a sample, and this gives us an approximation of the probability distribution function (pdf) for the population. We can then draw samples from this (by resampling with replacement from our sample) and look at how the relevant sample statistics are distributed.

Once we have a good idea about the sampling distribution, we can use this to construct our confidence interval.

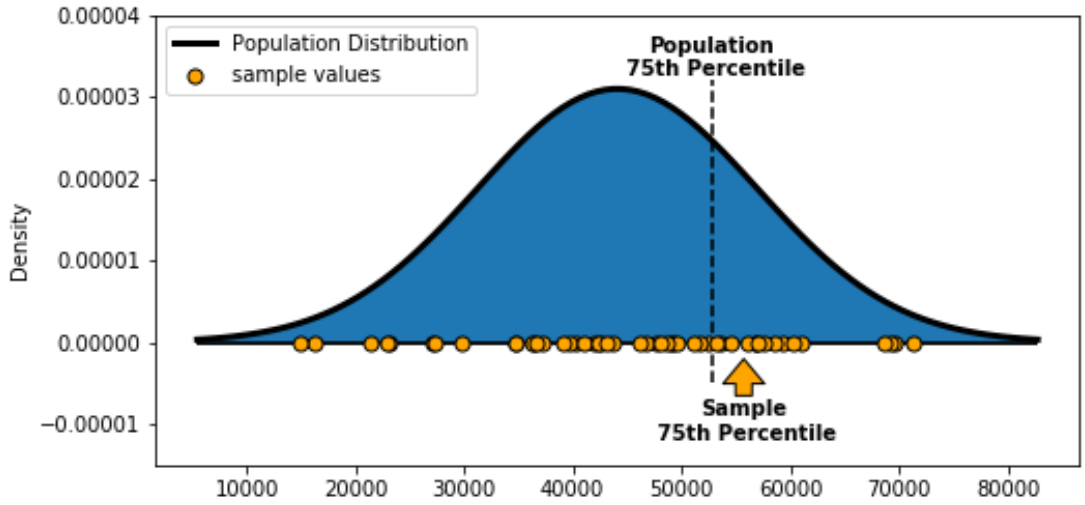
The Bootstrap

Let's look at what this would look like in practice. We'll start by looking at our (unknown) population distribution.
Let's pretend like we want to estimate the 75th percentile of the population values.
Remember that in practice we can't know what this distribution really looks like. We must try to infer it using just a sample.



The Bootstrap

Now, let's take a sample of size 50:



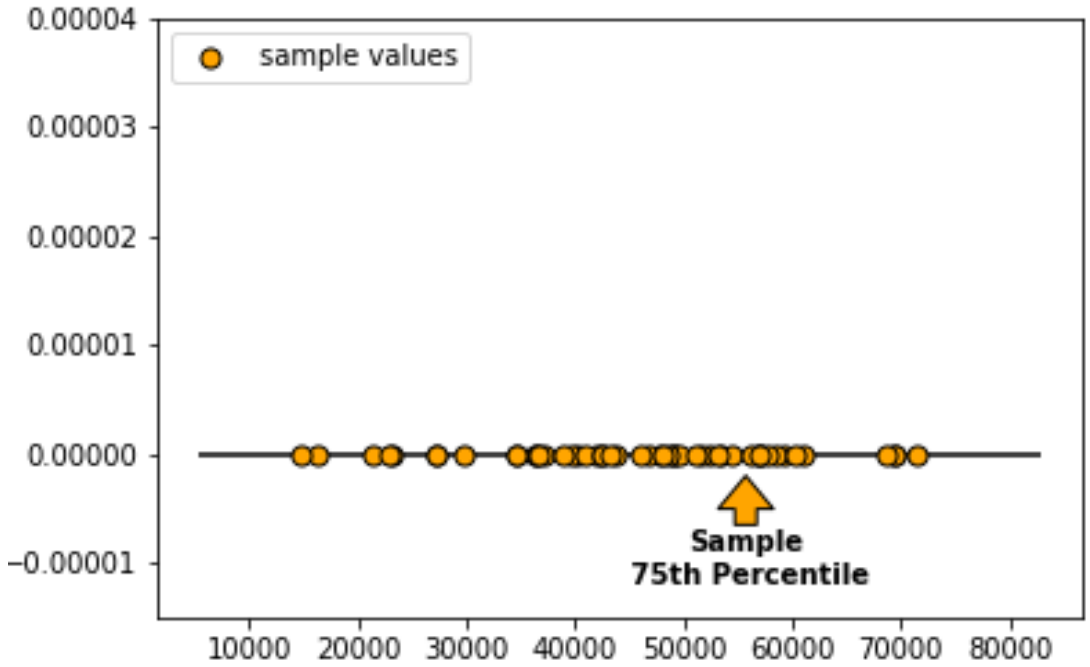
Based on this sample, we can get a *point estimate* of the 75th percentile of the population. It is unlikely to be exactly equal to the true population parameter, and in this case our estimate is slightly too high.

The Bootstrap

How good can we expect this estimate to be?

To know that, we need to understand how much variance there will be in the 75th percentiles across all possible samples.

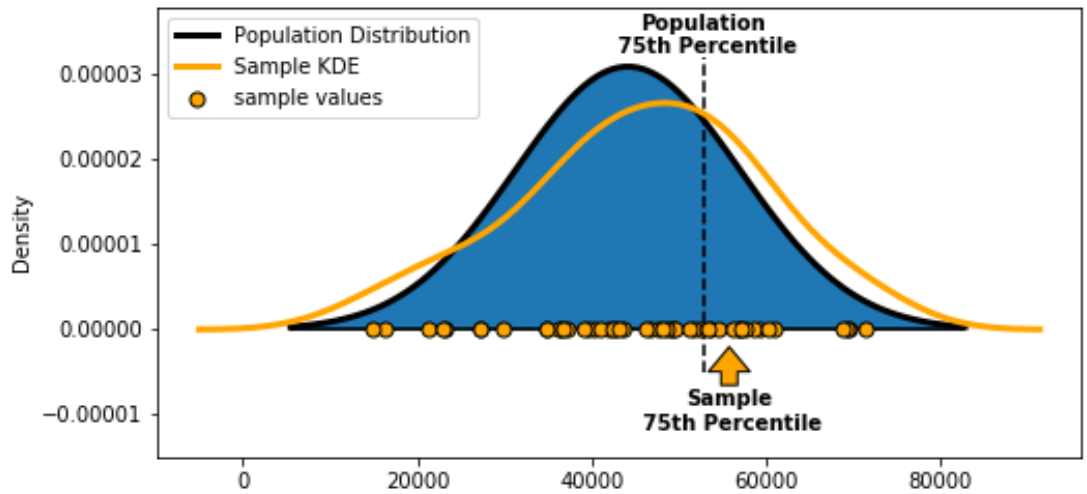
The problem is that we don't know what all possible samples look like. We only have our single sample.



The Bootstrap

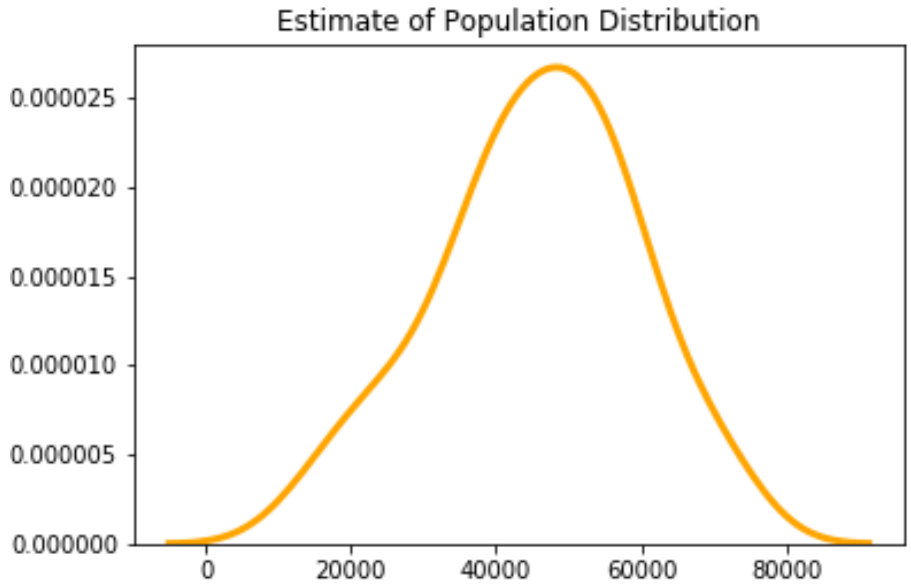
Based on this sample, what is our best estimate of the (unknown) population distribution?

If we want to estimate the population distribution, we could do so by looking at the KDE of the sample:



The Bootstrap

If we want to estimate what possible samples from the population look like, we could estimate it by using the distribution determined by the sample:

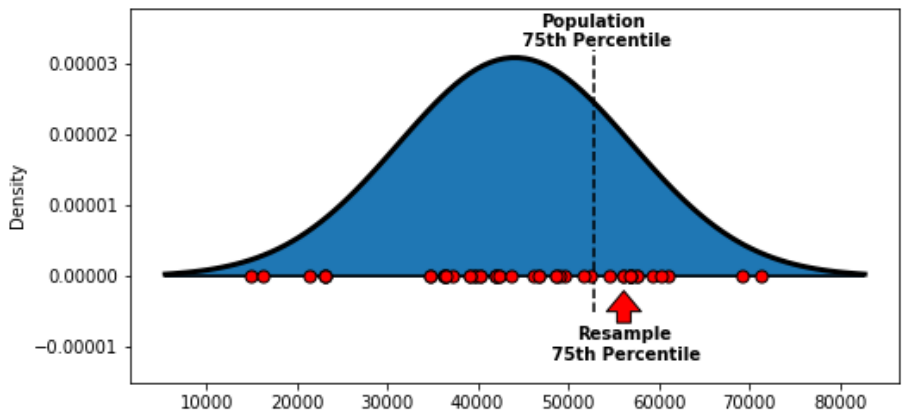


This is usually done by repeatedly *resampling* from our sample. That is, drawing samples of the same size from the samples **with replacement**.

The Bootstrap

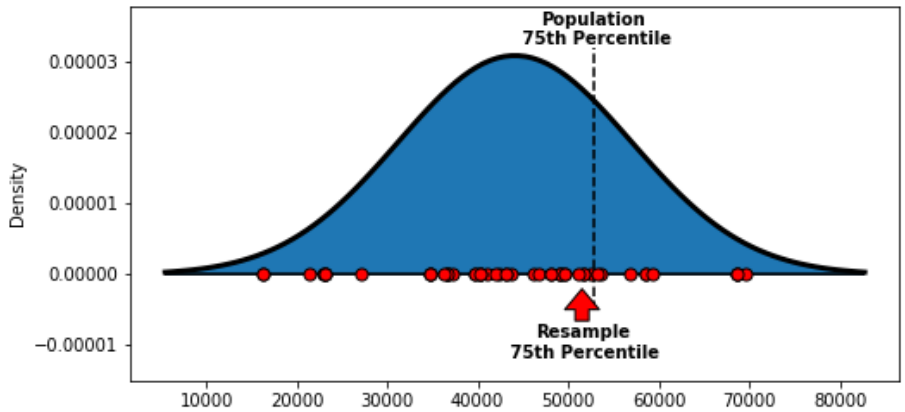
Let's look at how some possible resamples look and where their 75th percentile lands relative to the population 75th percentile.

Here is one possible resample:



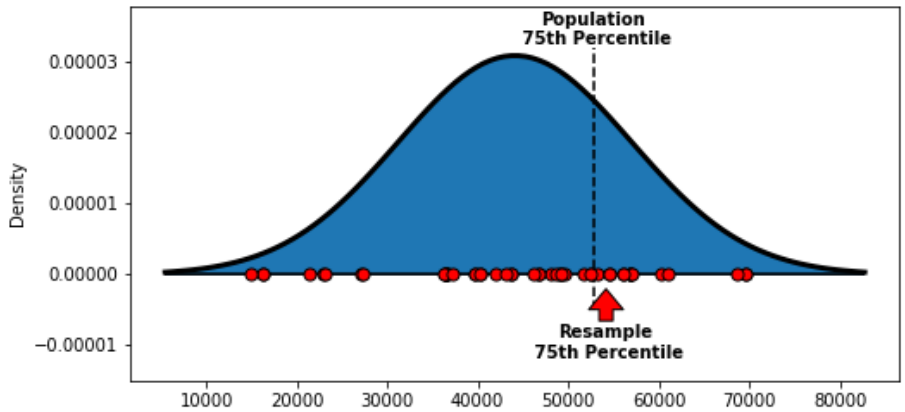
The Bootstrap

And another:



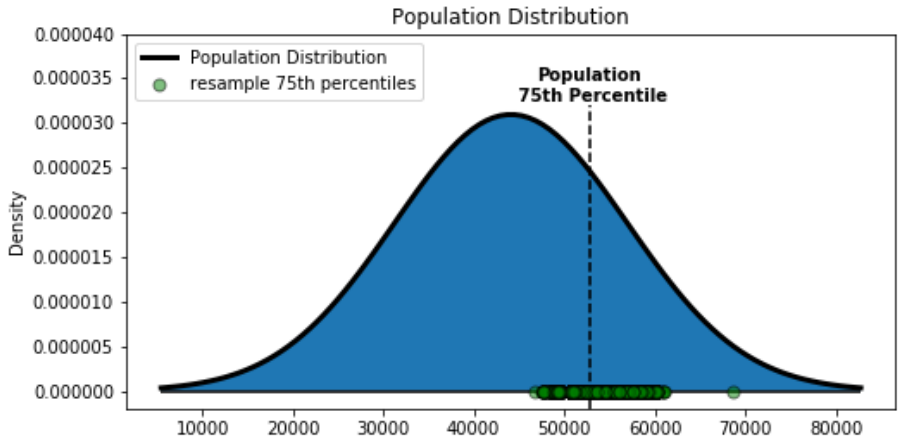
The Bootstrap

And another:



The Bootstrap

If we repeatedly draw resamples a large number of times (10,000 or so), then we can get an estimate of the sampling distribution of the 75th percentile:



What we see is that most of the resample 75th percentiles are clustered close to the population 75th percentile.

NASHVILLE SOFTWARE SCHOOL

The Bootstrap

(After widget demonstration)
General recipe for a 95% confidence interval:

1. Take a sample of size n from the population.
2. Find the statistic of interest for this sample, m (this is the **point estimate**).
3. Draw a large number (say, 10,000) of samples, with replacement, of size n from the original sample.
4. For each sample, calculate the statistic of interest.
5. Find the 2.5th and 97.5th percentile, a and b of the calculated statistics.
6. The 95% bootstrap confidence interval is

$$[m - (b - m), m + (m - a)]$$