BIEB100. Bootstrapping

If you repeatedly sampled from a population and calculated a statistic to summarize each sample (e.g. a proportion, mean, variance, etc.), this list of statistic values would be your sampling distribution. You could plot it using a histogram or a dotplot. Each value in this list is derived from a different sample and is an independent estimate of the population parameter. For example, if you repeatedly sample and collect the means, each mean is an independent *point* estimate of the population mean.

There will most likely be some variation in these sample means simply because the samples are going to be different. We don't know which sample mean is the closest to the true population mean (because we don't know what that value is), but we can give an interval where the bulk of the sample means are concentrated. This is an *interval* estimate. One method to calculate an interval estimate would be to do this repeated sampling, arrange the sample estimates in order, and then pick the interval *(1-a)/2* along the list and *a+(1-a)/2* along the list where *a* is the confidence level. For a 95% confidence interval this would be the interval between the 2.5 percentile and the 97.5 percentile. In terms of a histogram, it would be the interval on the x axis between which 95% of the area of the histogram lies.

Normally, we don't repeatedly sample from a population. *We only take one sample and have to work with that.* We want to know: how far is the true parameter value likely to be from our sample statistic value? If we had that information, we could decide what distances between the statistic and the parameter would be so big that it would be too unlikely that our sample came from a population with that parameter value. Remember that our goal here is to determine what parameter values are consistent with our statistic. Our sample came from some actual population, and we want to estimate the parameter(s) for that population. This isn't about trying to reject an hypothesized null parameter value. It's about trying to figure out what the parameter value could be for the population that our sample actually came from. **Once we have an interval estimate then we can say what the plausible parameter values are. That also means – almost as a byproduct– that we can say what the plausible parameter values *aren't,* which is what inference/hypothesis testing is all about. Getting an interval estimate effectively lets us do inference/hypothesis testing for all possible null values at once.**

To do this we need to get some idea of how far sample statistics are likely to be from the population parameter, and we have to somehow extract that from our sample. Bootstrapping is a way to simulate what this. The key idea behind it is that our sample is your best estimate for what the population actually looks like. The only difference is that our sample is relatively small and the population is really big, perhaps infinite.

Terminology – to try to avoid confusion, the actual sample from the population is your *data*. The sampling process of bootstrapping gives us*bootstrap samples.*

In bootstrapping, we repeatedly sample your *data*. Say we want to make *m* bootstrap samples each of size *n*. We would pick a value from the sample data, record it, and put it back. Then we would repeat this *n-1* more times. This is sampling *with replacement* (because each item in the data can be picked more than once) and gives us a bootstrap sample of size *n.* This is the same thing as using the sample to construct a population by repeating each observational unit an equal but infinite number of times and then taking a sample of size *n* from it [convince yourself that this is true.] Then we would repeat this *m-1* more times to get *m* bootstrap samples. Ideally, *m* would be >1000. Each time we take a bootstrap sample, we calculate the statistic we are interested in (mean, proportion, median, variance, (sum of data)^3, whatever...) and plot that as a dot in our dotplot.

This dotplot is our *bootstrap distribution*. Then we do the percentile method described above. This gives us a confidence interval.