Introduction to the Bootstrap

This lab will introduce you to the basic bootstrap. First, you will do a small example by hand. Then, you will use `R` to investigate how the bootstrap works.

Estimating a population mean

Often when we have a sample from a population, we are interested in estimating the mean of the population itself. The mean of the sample gives us an estimate of the mean of the entire population.

The sample mean is what we call a *point estimate* of the population mean. It gives us a single value as an estimate for the true population mean. But, ideally, we would also like to know *how close* this estimate is to the actual value!

1. Take a random sample of size 25 from a normal distribution with mean 10 and standard deviation 3. (We'll call this a normal(10,2) distribution from now on.) What's the true mean of the population? What's the estimate from your sample? Is your estimate exactly correct? Is it close?

   ```
   mysample = rnorm(25, 10, 3)
   mean(mysample)
   ```

In order to consider how close our estimate is to the true value, we can consider the sampling distribution of the sample mean. The *sampling distribution of the sample mean* tells us how the sample mean will vary from sample to sample for repeated samples from the population. Statistical theory (which you saw in SM239) tells us that the sampling distribution of the sample mean has standard deviation given by

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}},$$

where σ is the standard deviation of the population. Most of the time (about 95% of the time), the sample mean will be within 2 of these standard deviations from the true population mean.

2. What is the standard deviation of the sample mean for a sample of size 25 from a normal(10,3) distribution? Is your estimate from #1 within 2 of these standard deviations from the true mean?

3. Investigate the sampling distribution of the sample mean further by taking 1000 samples from this normal distribution. Make a histogram of the means of these 1000 samples. Verify that about 95% of the time the sample mean will be within $\pm\ 2\sigma_{\overline{X}}$ of the true sample mean.

```
samplemeans = rep(NA,1000)
for (i in 1:1000) {

mysample.i = rnorm(25, 10, 3)
samplemeans[i] = mean(mysample.i)

}
hist(samplemeans)
sum(samplemeans >= 10-2*3/sqrt(25) & samplemeans <= 10+2*3/sqrt(25))/1000
```

Generally we don't know the true population mean and standard deviation, which is why we estimate them from the sample. So we generally can't calculate $\pm\ 2\sigma_{\overline{X}}$. Instead, we estimate σ with S, the sample standard deviation. This gives us the *standard error* of $\overline{X}$, defined as

$$SE(\overline{X}) = \frac{S}{\sqrt{n}}.$$

We then can form a 95% confidence interval by taking $\overline{X} \pm 2SE(\overline{X})$. We say that we are 95% confident that the true population mean lies in this interval.

4. Using your sample from #1, calculate a 95% confidence interval for the true population mean. Does this interval include the true population mean of 3?

```
x.bar = mean(mysample); s = sd(mysample)
x.bar - 2*s/sqrt(25); x.bar + 2*s/sqrt(25)
```

As you can see, it is pretty straightforward to estimate and form a confidence interval for the population mean using the sample mean. However, the procedure outlined above is only valid if the population is normally distributed OR the sample size is large enough ($n \geq 30$ is a rule of thumb). In addition, we do not always have such a nice analytic procedure for other quantities besides the mean. The bootstrap is a method that can be used form confidence intervals in a more general setting.

The bootstrap

The *bootstrap* estimates standard error by resampling the data in our original sample. The idea is to treat the sample as a substitute for the population. Instead of repeatedly drawing samples of size 100 from the population, we will repeatedly draw *new* samples of size 100 from our original sample. In order to not end up with exactly the same sample every time, we will resample *with replacement*. This means that after a value is drawn, we replace it before drawing again. This way we can draw the same value out more than once! The following questions demonstrate how this resampling process works.

5. List the values that are in your sample from #1. Also report the mean of your sample.

6. Now take a sample of size 25 *with replacement* from your original sample. This is called a bootstrap sample. Did you get any duplicates in your bootstrap sample? Calculate and report the mean of your bootstrap sample.

```
bootsample = sample(mysample, 25, replace=T); bootsample
bootmean = mean(bootsample); bootmean
```

To do a full bootstrap, we would take a large number of bootstrap samples (at least 1000) from the original sample. For each bootstrap sample, we would calculate the mean of the bootstrap samples. The variation in the means of the bootstrap samples gives us an estimate of the variability in the sample mean. That is, we can estimate the standard error of the sample mean using the standard deviation of the bootstrapped sample means. We can then use this to construct a confidence interval for the true population mean.

7. Take 1000 bootstrap samples from your original sample and calculate the mean of each of these bootstrap samples. Make a histogram of these bootstrapped means. How does this compare to the histogram you made in #3?

```
bootmeans = rep(NA, 1000)
for (i in 1:1000) {

bootsample = sample(mysample, 25, replace=T)
bootmeans[i] = mean(bootsample)

}
hist(bootmeans)
```

8. Now calculate an estimate of the standard error. Use this to construct a 95% confidence interval for the true population mean. How does this compare to the interval you constructed in #4?

```
sd(bootmeans)
### use +/- 2*SE to construct a confidence interval
```

Hopefully you have been convinced that bootstrap sampling is a good substitute for the true sampling distribution, which we can only know exactly if we know the values for the population! Since we don't usually know the values for the population, the bootstrap technique can be really valuable!

A real example – backpack weights

Suppose we were interested in estimating the mean weight carried in backpacks. The file backpacks.txt contains measurements on the backpack weights (in lbs) for a sample. We are going to estimate the mean backpack weight using the bootstrap procedure. Note that this time we don't have the actual population mean to compare it to!

9. What is our sample size? Make a histogram of backpack weight. Does it appear to be normally distributed? Based on the answers to these two questions, would we be able to use the procedure in #4 to find a confidence interval?

10. Use the bootstrap to estimate the standard error of the sample mean. Also calculate the standard error using the formula above #4. How does the two standard errors compare?

```
bootmeans = rep(NA, 1000)
for (i in 1:1000) {

bootsample = sample(backpackWT, 20, replace=T)
bootmeans[i] = mean(bootsample)

}
sd(bootmeans)
```

11. Use your bootstrap estimate of standard error to form an approximate 95% confidence interval for the true population mean. Also construct a 95% confidence interval using the method in #4. How do the two intervals compare?

Lab Summary

Report your 95% bootstrap interval for the mean backpack weight. Write a sentence interpreting what this interval means. Also report the 95% confidence interval for the mean backpack weight you calculated using the method in $4. How does the bootstrap interval compare to the more traditional confidence interval?