Bootstrapping

Week 09: CM1606 Computational Mathematics

Bootstrapping is a statistical method of estimating the accuracy of a sample statistic by repeatedly sampling with replacement from the original sample. The term "bootstrap" comes from the phrase "pulling yourself up by your own bootstraps," which refers to achieving a difficult task through your own efforts.

In statistical analysis, bootstrapping involves taking a random sample of a dataset and using it to create many resamples, which are essentially simulations of possible samples that could have been drawn from the original dataset. These resamples are created by randomly sampling observations from the original dataset with replacement, meaning that the same observation can be selected multiple times.

The resamples are then used to calculate the sample statistic of interest, such as the mean or standard deviation. This process is repeated many times to create a distribution of the sample statistic, from which estimates of the statistic's variability and confidence intervals can be derived.

Bootstrapping can be particularly useful when the underlying distribution of the data is unknown or difficult to model, or when the sample size is small. It is often used in regression analysis, hypothesis testing, and machine learning.

- Imagine we had a new drug to treat an illness, and we gave that drug to 8 different people that had the illness.
- Assume for 5 of these people the drug appeared to help them feel better, but for 3 people the drug appeared to make them feel worse.

- if we calculate the mean of the response to the drug, we get 0.5.
- 0.5 is not a huge improvement but since most of the people (five of eight) improved maybe this drug is better than using no drug at all or maybe these five people all felt better because they were healthier to begin with and maybe these three people all felt worse because they had unhealthy lifestyles.
- It is possible that the reason we got a mean value equal to 0.5 instead of 0 is because of random things that we can't control

Is there anything we can do to decide if the drug works or not?

Yes!

one expensive and time-consuming option would be to replicate the experiment a bunch of times. if we repeat the experiment a bunch of times then we can keep track of each mean value and we will end up with a histogram of mean values.

- Just by looking at this distribution, we can see that mean values close to zero which suggests that the drug does not do anything are relatively likely to occur.
- mean values far from zero indicating that the drug does something are relatively rare.

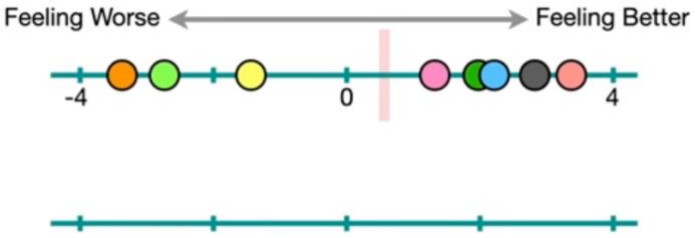
however as said earlier repeating the experiment a bunch of times is both expensive and time-consuming.

Is there something else we can do that is less expensive and time consuming?

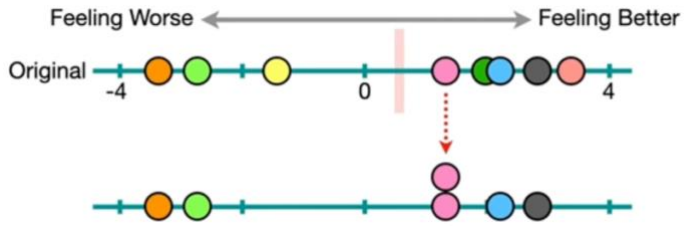
Yes!

we can use bootstrapping instead of replicating the experiment a bunch of times. let's use bootstrapping to get a better sense of which results are likely and which are rare.

- first let's create a new number line.



- Now from the eight original measurements choose one at random and add that value to the new number line. Again, go back to the original eight measurements and choose another value at random and add it to the new number line. Then we repeat that process randomly selecting one of the eight original values for the new number line a total of eight times.
- Note that we can randomly select the same value more than once (randomly selecting data and allowing for duplicates is called sampling with replacement)



Note that, the reason we selected eight measurements for the new number line is because the original data set that we are sampling from contains eight measurements. If we had started with 10 measurements, then we would need to add 10 measurements to the new number line. This new data set that was created using sampling with replacements so that it had the same number of values as the original data set is called a **bootstrapped dataset.**

Now that we have a new bootstrapped dataset, we can calculate the mean now.

Now the bootstrap dataset is different from the original dataset. We get a different mean now. let's add the mean of the bootstrap data set histogram of means.

Do the same thing with a fresh number line and randomly select from the eight original values for the new number line. Add all this means to our histogram.

Keeping track of those calculations is called **bootstrapping.** in other words bootstrapping consists of four steps.

1. Make a bootstrapped data set.
2. Calculate something (in this case we calculated the mean).
3. Keep track of that calculation.
4. Repeat steps 1 to step 3 a bunch of times.

Here are a few examples of how bootstrapping can be used in:

Confidence Intervals: Suppose we want to estimate the average height of all students in a school, but we only have a small sample of 30 students. We can use bootstrapping to generate many resamples from the original sample and calculate the mean height in each resample. From these resamples, we can construct a confidence interval that will give us a range of values that the true population mean is likely to fall within.

Hypothesis Testing: Suppose we want to test if there is a significant difference between two groups, but we don't know if the data follows a normal distribution. We can use bootstrapping to generate many resamples of the data for each group, calculate the difference in means between the resamples, and use this distribution to test for a significant difference.

Regression Analysis: Suppose we want to fit a regression model to a small sample of data, but we are unsure about the distribution of the errors. We can use bootstrapping to generate many resamples of the data, fit the regression model to each resample, and examine the distribution of the model coefficients. This can give us an idea of the variability of the coefficients and help us assess the reliability of our model.

Machine Learning: Bootstrapping can also be used in machine learning algorithms such as random forests and bagging. These algorithms use bootstrapping to generate multiple decision trees or models, which are then combined to make predictions. This can help reduce overfitting and improve the accuracy of the model.