Lab 2: Two-sample permutation test

This lab focuses on the two-sample permutation test. In this lab, you will first work through a short permutation test by hand. Then you will do a larger test using `R` and look at some of the variations of the permutation test.

Cloud seeding

Scientists have long been interested in whether humans can cause clouds to produce more rainfall. "In one study, researchers in Florida explored whether injecting silver iodide into cumulus clouds would lead to increased rainfall. On each day that was judged to be suitable for cloud seeding, a target cloud was identified and a plane flew through the target cloud in order to seed it. Randomization was used to determine whether or not to load a seeding mechanism and seed the target cloud with silver iodide on that day. Radar was used to measure the volume of rainfall from the selected cloud during the next 24 hours. The volume of rain was measured in volume units of acre-feet, the "height" of rain across 1 acre." (Rossman and Chance, 2006)

A simple example

Suppose the researchers only performed their experiment for 3 days. On two of those days, the cloud was seeded, and on one day the cloud was unseeded. They measured the following rainfalls:

Day1	unseeded	147.8
Day2	seeded	489.1
Day3	seeded	119.0

1. What is the difference in means between the two groups? Do your calculation as seeded - unseeded.

If there is no difference in rainfall between seeded and unseeded clouds (H_0), then all of the rainfalls can be combined into one 'pot.' You can think about stripping the seeded/unseeded labels off of the data, because the labels are meaningless if there is no difference between them. So if H_0 is true, your data really looks like:

Day1	147.8
Day2	489.1
Day3	119.0

A permutation test works under the assumption that the labels/groups have been randomly assigned to the research subjects. In this case, the clouds were randomly assigned to either be "seeded" or "unseeded," so a permutation test is appropriate here. If the clouds were randomly assigned, then any possible random assignment that results in two seeded clouds and one unseeded cloud would be possible.

2. How many possible cloud assignments would results in two seeded clouds and one unseeded cloud? List out these possible assignments. (Be sure to include the assignment that we actually observed.)

3. Calculate the difference in means between the two groups for each assignment you listed in 2. Again, be sure to do your calculation as seeded - unseeded.

By considering all possible assignments of these clouds under the assumption that there is no difference between seeded and unseeded clouds, you are considering the distribution under the null hypothesis, H_0. Remember that a p-value is defined to be

$$p\text{-value} = Pr(\text{something "as or more extreme" than what we observe}|H_0 \text{ true})$$

Since our list of all possible assignments are under the assumption that H_0 is true, we can calculate a p-value for the test. What did we observe? We observed a difference in means of 156.25 (what you calculated in 1.) Since we are hoping to find that cloud seeding results in more rainfall, we are hoping to see a large difference in means (seeded - unseeded). So "as or more extreme" would be differences in means greater than or equal to our observed difference of 156.25. Since each random assignment is equally likely, the p-value is just the proportion of random assignments that have a mean difference as or more extreme than our observed value.

4. What is the p-value in this situation? Based on this p-value, state a conclusion about the relationship between cloud seeding and rainfall.

You have just completed a permutation test. What you are really doing when you perform a permutation test is you are permuting the labels among the research subjects. In this case, you are permuting the seeded/unseeded labels among the three clouds. This is where the permutation test gets its name.

A larger example

Obviously, this is a pretty simple example, and with three data points it is unlikely we would be able to tell if the seeded clouds have higher rainfall than the unseeded clouds. (Such a small data set has very low power.) Let's look at this same example, but using a larger dataset. And let's let `R` do all of the calculations for us!

Our larger dataset includes 10 days worth of this cloud seeding experiment:

```
rainfall treatment
147.8    unseeded
26.1     unseeded
95.0     unseeded
1.0      unseeded
4.1      seeded
119.0    seeded
489.1    seeded
978.0    seeded
255.0    seeded
2745.6   seeded
```

Copy this data into a notepad file, name the file `cloudseeding.txt`, and then read the data into `R` using `cloudseeding = read.table(file.choose(), header=T)`. Be sure to attach the dataset so that we can work with the `rainfall` and `treatment` variables separately. We are going to work through the same steps we did above, using `R`.

5. Calculate the observed difference in means between the groups (seeded - unseeded).

```
teststat.obs = mean(rainfall[treatment == "seeded"]) - mean(rainfall[treatment == "unseeded"])
```

6. How many different assignments of seeded/unseeded are possible? (The command `choose(n,k)` will calculate the number of ways to choose `k` objects from a group of `n` objects.)

7. Have `R` list all possible assignments of the rainfalls to the two groups, with four unseeded clouds and six seeded clouds. You will want to use all of the code listed below, but I've separated it into pieces to explain what each piece does. If you have questions about what it is doing, ASK!

```
unseeded = combinations(10, 4, v=rainfall, set=FALSE, repeats.allowed=FALSE)
```

   (This `unseeded` object lists the rainfall amounts of all possible ways to choose 4 unseeded clouds from the total of 10 clouds. Each row of `unseeded` is one such combination.)

```
seeded = NULL
for(i in 1:210) {
seeded = rbind(seeded, setdiff(rainfall, unseeded[i,]))
}
```

   (Once 4 clouds (and their rainfall totals) have been assigned to the unseeded group, the remaining 6 clouds must be assigned to the seeded group. This code does this for each combination of unseeded clouds, and stores it as `seeded`.)

8. For each possible assignment of seeded/unseeded clouds, calculate the difference in means between the groups.

```
teststat = rep(NA, 210)
for(i in 1:210) {
teststat[i] = mean(seeded[i,]) - mean(unseeded[i,])
}
```

9. Calculate the p-value by calculating the proportion of possible assignments that have a difference in means that is greater than or equal to the observed difference in means. Then state a conclusion about the relationship between cloud seeding and rainfall.

```
teststat >= teststat.obs
sum(teststat >= teststat.obs)
sum(teststat >= teststat.obs)/210
```

Variations on the permutation test

The wonderful thing about the permutation test is that it is not limited to using the difference in means as a test statistic. Because we "create" the distribution of the test statistic under the null hypothesis by considering all possible assignments of the data, we can find a p-value regardless of what our test statistic is, as long as we compute this test statistic for our observed data and for all possible assignments of the data.

10. Repeat steps 5-9 using the difference in medians between the groups as your test statistic. Since you should have all of the code you used for steps 5-9, you will simply need to change the definition of your `teststat.obs` and `teststat` and then run the entire chunk of code at one time. You do not need to answer each step separately – simply complete all of the steps and then state your p-value and a conclusion.

11. Repeat steps 5-9 using the sum of the observations in the seeded group as your test statistic. Again, you should just have to modify your `teststat.obs` and `teststat`.

12. Repeat steps 5-9 using the maximum observation in the unseeded group as your test statistic. In this case, you will have to modify your `teststat.obs` and `teststat`, and you will also have to make a small modification to how the p-value is calculated.

Lab Summary

Please write clearly and in complete sentences.

1. Write a paragraph summarizing your results about the relationship between cloud seeding and rainfall. You should consider all of the analyses you did on the larger cloud seeding dataset. You may want to comment on any differences in conclusion between the different variations of the permutation test.

2. The actual data set from this cloud seeding experiment contains rainfall amounts for clouds on 52 different days, where there are 26 unseeded clouds and 26 seeded clouds. Explain why performing a permutation test on this entire data set would be impossible by hand and quite daunting even for `R`.