Lecture 18

More non-parametric tests

Randomization tests are sometimes called permutation tests (though, they might really be called combination tests). The idea here is that if we are comparing two samples, we want to determine the probability of obtaining a sample difference as extreme as the one we have if we select our two samples randomly from both groups of observations. If the samples are small, we can actually list out all the possible combinations and determine the probability from that sample space, much as is done with small samples from our ranking tests. When the sample sizes become large and it's no longer possible to list every conceivable outcome, we usually switch to simulation techniques to obtain an estimate of the probability.

Let's consider a small example, and then we will return to modeling this in R in a later lecture.

Example. Suppose we have two groups of 4 observations each.

Experiment	3	1	2	7
Control	8	10	11	5

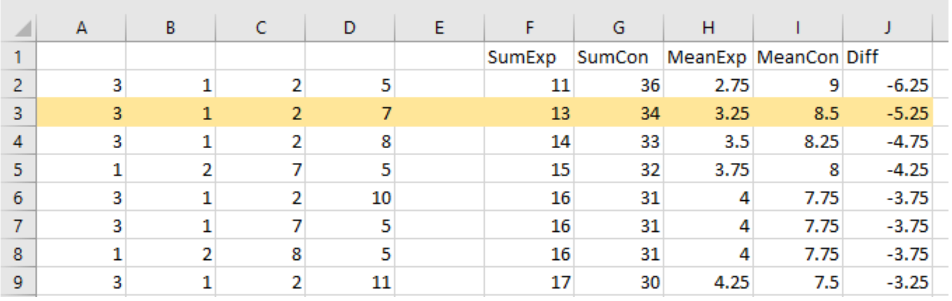
Since we have 8 total observations and 4 to place in the experimental group (the rest will automatically go into the control group). The order doesn't matter since we are going to use combinations to calculate the total number of possible outcomes.

$$\binom{8}{4} = 70$$

I've linked a website below that you can use to list the possible combinations of 4 observations. I've pasted the results below.

{[3,1,2,7], [3,1,2,8], [3,1,2,10], [3,1,2,11], [3,1,2,5], [3,1,7,8], [3,1,7,10], [3,1,7,11], [3,1,7,5], [3,1,8,10], [3,1,8,11], [3,1,8,5], [3,1,10,11], [3,1,10,5], [3,1,11,5], [3,2,7,8], [3,2,7,10], [3,2,7,11], [3,2,7,5], [3,2,8,10], [3,2,8,11], [3,2,8,5], [3,2,10,11], [3,2,10,5], [3,2,11,5], [3,7,8,10], [3,7,8,11], [3,7,8,5], [3,7,10,11], [3,7,10,5], [3,7,11,5], [3,8,10,11], [3,8,10,5], [3,8,11,5], [3,10,11,5], [1,2,7,8], [1,2,7,10], [1,2,7,11], [1,2,7,5], [1,2,8,10], [1,2,8,11], [1,2,8,5], [1,2,10,11], [1,2,10,5], [1,2,11,5], [1,7,8,10], [1,7,8,11], [1,7,8,5], [1,7,10,11], [1,7,10,5], [1,7,11,5], [1,8,10,11], [1,8,10,5], [1,8,11,5], [1,10,11,5], [2,7,8,10], [2,7,8,11], [2,7,8,5], [2,7,10,11], [2,7,10,5], [2,7,11,5], [2,8,10,11], [2,8,10,5], [2,8,11,5], [2,10,11,5], [7,8,10,11], [7,8,10,5], [7,8,11,5], [7,10,11,5], [8,10,11,5]}

I can post the csv file of the calculations on these sets. For each one, the remaining observations go into the control group. We calculate the mean on each set of experiment and control, and then find the differences. In the Excel file I created for this example, I then sorted the data from smallest to largest.



	A	B	C	D	E	F	G	H	I	J
1						SumExp	SumCon	MeanExp	MeanCon	Diff
2	3	1	2	5		11	36	2.75	9	-6.25
3	3	1	2	7		13	34	3.25	8.5	-5.25
4	3	1	2	8		14	33	3.5	8.25	-4.75
5	1	2	7	5		15	32	3.75	8	-4.25
6	3	1	2	10		16	31	4	7.75	-3.75
7	3	1	7	5		16	31	4	7.75	-3.75
8	1	2	8	5		16	31	4	7.75	-3.75
9	3	1	2	11		17	30	4.25	7.5	-3.25

This is a short snapshot of the end of the calculations. To obtain the control group mean, I found that all the observations summed to 47, and so calculated the sum of the missing observations by subtracting from the total.

If the control group really is bigger than the experimental group we'd expect a negative difference of means. If we are doing a one-sided test, then we find the number of observations at or below the value of the difference we found for our original data and divide by the total number of observations to get the P-value. Here, that's $\frac{2}{70} \approx 0.02857$.

If we are doing a two-tailed test, then we'd have to multiply this result by 2 to account for the other tail.

This process becomes less and less feasible as the number of observations grows. Here, we had only 70 possible combinations from 8 observations, but from 12 observations, we have nearly 1000 possible combinations. By the time we have 20 observations, we are over 184,000 combinations.

At that point, we'd turn to simulations.

One advantage of this method is that it does not depend on the distribution of the data, like other non-parametric tests, and we can choose any statistic we wish to estimate, and we don't have to concern ourselves with whether or not there is a theoretical model for the sampling distribution.

It's worth noting here that a permutation test of this type is a little different from bootstrapping. In a permutation test, we take each observation one time and don't allow repetition. In bootstrapping, after sampling, they go back into the pool, so repetition is allowed. That allows us to take different sample sizes than the original data. Bootstrapping also doesn't put all the data into the same pool. We'll spend a whole future lecture on these computational methods coming soon.

In the last class we talked about the Wilcoxon tests. We can construct confidence intervals on these distributions but finding the critical value we need for these situations can be difficult. We need to find a critical value that corresponds to the confidence level. The test statistics (sum of ranks) are normally distributed for large samples and so we can construct confidence intervals on the test statistic, but this is not really the same thing as constructing a confidence interval of the mean or median we are testing.

The critical value for the rank-sum test would be approximately $c \approx \frac{mn}{2} + z_{\alpha/2}\sqrt{\frac{mn(m+n+1)}{12}}$ when both sample sizes are "large".

An easier way to get directly to the confidence interval will be to do simulations or bootstrapping.

In our next lecture (the only one next week), we'll look at distribution-free ANOVA for when we have 3 or more samples to compare.

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://onlinestatbook.com/2/distribution_free_tests/randomization_two.html
3. https://www.mathsisfun.com/combinatorics/combinations-permutations-calculator.html