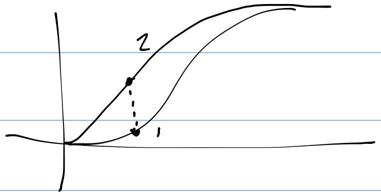
Midterm 1 no longer 25th

If experimental units (ex. New employees), are not randomly selected from a larger population of units, the inferences we can make from our randomization procedure will apply to the effects of the treatments (ex. Training methods), but only as the treatments affect the units in the study.
However, if the units are initially selected at random from larger populations of units, then inferences may be drawn about the effects of the treatments as applied to the larger populations.
********** Midterm

Let F_1(x) and F_2(x) be the cdf's of the two populations.
The null hypothesis is:
 H_0 : F_1(x) = F_2(x)
In other words, the two distributions are identical under the null hypothesis.

A one-sided alternative hypothesis is:
 H_a : F_1(x) <= F_2(x),
where strict inequality occurs for at least one x.
What does this alternative hypothesis mean about observations for treatment 1 vs treatment 2? Do treatment 1 observations tend to be larger or smaller than treatment 2 observations?

Probability that our variable will take a value less than the specified value



There is a higher probability that 2 will have a value less than specified
There is a higher probability that 1 will have a value grater than specified

The average values of 1 will be greater than the average values of 2

The two-sided alternative hypothesis is not F_1(x) != F_2(x)
This would allow, for instance, the possibility that the means are the same but the variances are different.
The permutation test using the difference between the means as the statistic is not designed for this possibility.
The two-sided alternative is in fact:
 H_a : F_1(x) <= F_2(x) or F_1(x) >= F_2(x), for all x
With strict inequality occurring for at least one x.

The mean is the most commonly used measure of the center of a set of numbers. It seems natural then to use the difference between two means as a statistics for comparing two treatments.
What are some problems through using means? Spread/asymmetry of data. One tail could be extreme compared to other.

That's right! The median and trimmed means!

Prior to computers, and even with certain basic computer software now, permutation tests were limited by prohibitive computations.

For example, if two treatments each have 8 observations.
(16 C 8) = 12,870 possible two-sample data sets
Or strikingly if just two more observations are added to each treatment:

(20 C 10) = 184, 756 possible two-sample data sets

As the number of observations in each treatment increases, it is easy to see how quickly the possible number of two-sample data sets increases.
Fortunately, there is a "simple" way to obtain an approximate *p*-value for the permutation test in these cases.
....

1. Compute the difference of means, D_obs, between the two treatments for the observed data.
2. Create
3. E
4. Compute the difference of means for the shuffled data set.
5. Repeat the procedure a predetermined number of times (ex. 1000).

For an upper-tail test, the fraction of the sampled mean differences that are greater than or equal to the observed difference D_obs is an approximate *p*-value for the permutation ttest.
(Similarly, we may obtain approximate lower-tail or two-tail *p*-values.

Accuracy of the Procedure
If among all possible permutations of the data the fraction of mean differences greater than or equal to D_obs is p (that is, if the true p-value is p), then the theory of hte binomial distribution tells us that an approximate p-value basd on R randomly selected permutations will have about a 95% chance of being within.

$$\pm \sqrt{\frac{p(1-p)}{R}}$$

Of the true p-value.

Let X_1, X_2, ..., X_N denote a set of N observations.
The rank of X_i among the N observations, denoted R(X_i), is given by
R(X_i) = number of X_j's <= X_i

Assume there are m observations for treatment 1 and n observations for treatment 2. Assume no two observations have the same value.

1. Combine the m+n observations into one group, and rank the observations from smallest to largest. Let 1 be the rank of the smallest observation, 2 the rank of the next smallest, and so on. Find the observed rank sum W of treatment 1 (or treatment 2).
2. X
3. For each permutation of the ranks, find the sum of the ranks for treatment 1 (or treatment 2, respectively).
4. Determine the upper-tail, lower-tail, or two-sided p-value as appropriate. For instance an upper tail test is....

Statistical test based on the sum of ranks of one of the treatments will have the same p-value, and reach the same conclusion, as a test based on the difference of mean ranks.
Similar to the permutation test, we can choose to use either the sum of ranks of one of the treatments or the mean difference of ranks to conduct the Wilcoxon Rank-Sum test and would achieve the same result.

Let W_1 denote the sum of the ranks for treatment 1 for example. If we have N = m+n total observations in the data, then the sum of all ranks is?
T = 1 + 2 + ... + N = ?
How do you find the sum of the first 100 integers?
(n)*(n+1) / 2 bookends add $\rightarrow$ $\underset{\downarrow}{1} + 101 = 101 \Big\}\ 50$ or $\frac{100}{2} \rightarrow \frac{n(n+1)}{2}$

Sorlol =101

We can show:

difference of mean ranks = $\frac{W_1}{m} - \frac{W_2}{n}$

$$= \frac{W_1}{m} - \frac{\frac{N(N+1)}{2} - W_1}{n}$$

$$= W_1\left(\frac{1}{m} + \frac{1}{n}\right) - \frac{N(N+1)}{2n}$$

One of these proofs will be on the midterm***

Table A3 of the Appendix contains upper-tail and lower-tail critical values for the Wilcoxon rank sum test

Example 2.4.2 A particular type of herbicide was tested for controlling weeds among strawberry plants. To see any potential damage that the herbicide might do to the strawberry plants, a researcher compared the dry weights of plants treated with the herbicide to the dry weights of untreated plants.
Data were obtained on seven untreated plants and nine treated plants. It is expected that the untreated plants will have larger dry weights than the treated plants.

The sum for the n=7 untreated plants is 84. The 5% critical value for when m=9 and n=7 is 76. Therefore, for a one-sided upper tail test, what can we conclude about the weights of the untreated plants vs. treated plants?
We reject the null hypothesis in favor of the theory that untreated plants have larger dry weights than the treated plants.

Upper-tail vs. Lower-tail (m vs. n)
N = sample that will be used for the test

Our previous method for the Wilcoxon Rank-Sum test is quite clear when all of the data values, and hence the ranks, are unique. What do we do however when observations share the same value? How do we rank these? adjusted ranks

What about the critical value? IF the number of ties is small, then approximate critical values may be obtained from the distribution of the rank-sum statistic without ties (Table A3). We will cover larger samples in Section 2.10.