1. **Exploration of the Permutation Test**

In last lecture and today’s lecture you learned about the permutation test, used on data of the following form:

\[ X_1, \ldots, X_m \sim_{iid} f(x) \]

and independently

\[ Y_1, \ldots, Y_n \sim_{iid} f(y - \mu) \]

We test against the null that \( \mu = 0 \). Now, if \( f \) had been normal, then instead of a permutation test, we can perform a t-test with the following test statistic:

\[ t = \frac{\bar{Y} - \bar{X}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \]

where \( S_p \), “s-pooled”, is

\[ S_p = \frac{1}{m + n - 2} \left( \sum_{i=1}^{m} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right) \]

This is the form of the test where we’ve assumed equal variances, which comes from the location problem set up. The pdfs for \( X \) and for \( Y \) differ only by a \( \mu \)-sized shift. Then the test statistic \( t \) follows a \( t \) distribution with \( m + n - 2 \) degrees of freedom.

The idea of the permutation test is the following. We can’t assume normality, so we don’t know the distribution of \( t \). However, we can still use \( t \) by creating a “pseudo” null distribution.

The following procedure generates a number \( t^* \) from this pseudo null. From the combined data \( X_1, \ldots, X_m, Y_1, \ldots, Y_n \), shuffle it randomly to create a reordered dataset. Label the first \( m \) entries \( X^* \) and the last \( n \) entries \( Y^* \); then your data is \( X_1^*, \ldots, X_m^*, Y_1^*, \ldots, Y_n^* \). Then define a statistic \( t^* \) as

\[ t^* = \frac{\bar{Y}^* - \bar{X}^*}{S_p^* \sqrt{\frac{1}{m} + \frac{1}{n}}} \]

Repeat this procedure many times (\( N \) times), and obtain values \( t^*_1, \ldots, t^*_N \). Estimate a p-value as follows. For a right tailed test, it is \( \frac{\# \text{ of } t^*_i \geq t}{N} \). For a two tailed test, it is
Math 181C (Abramson) Week 6 Homework (Thursday)

\{\# of \, |t^{*}| \geq |t|\} \cdot N. \quad \text{(This is equivalent to what was stated in class, which says to record how many data lie to the left of } -|t| \text{ and how many lie to the right of } |t|. \)

This defines “a” permutation test. But it doesn’t define “the” permutation test.

Since we no longer have to assume normality, we no longer are really beholden to define \( t \) as we did, which has its form because you must divide by \( S_p \) for it to be \( t \)-distributed.

So what if you did a test based on a different test statistic—namely

\[
t' = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{1}{m} + \frac{1}{n}}}
\]

(or just take \( t' = \bar{Y} - \bar{X} \), they only differ by a constant.) Of course the \( t^* \) is defined analogously. Is there an improvement?

This is a programming question. Your task is to generate \( X \) and \( Y \) data from a continuous distribution so that the null is not true. Then, using a computer, perform a permutation test based on both the \( t \) and \( t' \) statistics, and tell us which one is better—say, with respect to power for a fixed \( \alpha \).

2. **Spearman’s rho**

Suppose your data \((X_1, Y_1) \ldots (X_n, Y_n)\) was converted into a ranked version—that is, you replace the order statistics \(X_{(1)}, \ldots, X_{(n)}\) with \(1, \ldots, n\) and likewise for the \( Y \)’s order statistics. Then, order the data in increasing \( X \) coordinate, so it looks like \((1, R_1), (2, R_2)\ldots, (n, R_n)\) for some integers \( R_i \).

In class, you defined Spearman’s \( \rho \) as simply Pearson’s \( r \) computed with these ranked data.

\[
\rho = \frac{\sum_{i=1}^{n} (i - \bar{i})(R_i - \bar{R})}{\sqrt{\sum (i - \bar{i})^2 \sum (R_i - \bar{R})^2}}
\]

where \( \bar{i} = \bar{R} \) is shorthand for \((n + 1)/2\).

Your task is to complete in full detail what was done only partially in class—show that \( \rho \) can be written

\[
\rho = \frac{12\sum_{i=1}^{n} iR_i}{n(n^2 - 1)} - \frac{3(n + 1)}{n - 1}
\]

3. **Concordance and discordance—intuition**

A set of paired data \((X_1, Y_1) \ldots (X_n, Y_n)\) is said to be perfectly concordant if \( Y = g(X) \) for some strictly monotone increasing function \( g \). It is said to be perfectly discordant if \( Y = h(X) \) for some strictly monotone decreasing function \( h \).

In class, you saw that as \( n \to \infty \), \( \rho \) approaches the value

\[
\rho \to 1 - 6\mathbb{E}[F_X(X_1) - F_Y(Y_1)]^2
\]
in probability.

Using this, you saw justification for why, when testing monotone dependence, that if the null is true (\( \text{Corr}(X,Y) = \rho = 0 \)) that Spearman’s rho asymptotically approaches 0 in probability. You also saw a justification for why, if the data is concordant, that rho approaches 1. (It was a little handwavy, to be honest—it relied on the fact that if \( q(X) \sim \text{Unif}[0,1] \) for a monotone increasing function \( q \), then \( q \) must equal \( F_X \).)

Handwavy, then, is okay. Your task: see if you can come up with a similar argument for why discordant data results in \( \rho \to -1 \).