STAT 425: Introduction to Nonparametric Statistics **Winter 2026**

Lecture 3: Permutation, Rank Correlation, and Dependence Test

Instructor: Yen-Chi Chen

3.1 Permutation Test

Permutation test is a power method for conducting a two-sample test. The idea is very simple – given a test statistic, we compute its distribution under $H_0 : P_X = P_Y$ by permuting the two samples. Let $\mathcal{S}_X = \{X_1, \cdots, X_n\}$ and $\mathcal{S}_Y = \{Y_1, \cdots, Y_m\}$ be the two samples. Now we consider a summary statistic of both samples, M_X and M_Y, such summary statistic can be the median of each sample, a particular quantile value of each sample, or some measure of deviation such as interquartile range or standard deviation. The key is: we want to use the same summary statistic for both sample.

Let $T(\mathcal{S}_X, \mathcal{S}_Y) = M_X - M_Y$ be the difference between the two summary statistics. We will use $T(\mathcal{S}_X, \mathcal{S}_Y)$ as our a test statistic. Clearly, under H_0 the test statistic $T(\mathcal{S}_X, \mathcal{S}_Y)$ should be close to 0 since we are using the same summary of both samples.

To see the distribution of $T(\mathcal{S}_X, \mathcal{S}_Y)$ under H_0, we generate new samples $\mathcal{S}_X^*$ and $\mathcal{S}_Y^*$ by the following permuting procedure:

1. Pulling both sample together to form a joint sample $\mathcal{S}_{XY} = \{X_1, \cdots, X_n, Y_1, \cdots, Y_m\}$.
2. Randomly permuting the elements in $\mathcal{S}_{XY}$ and split them into two samples $\mathcal{S}_X^*$ and $\mathcal{S}_Y^*$ and each with n and m observations, respectively. Namely, the new $\mathcal{S}_Y^*$ may contain several observations that were originally in $\mathcal{S}_X$.
3. Treating $\mathcal{S}_X^*$ and $\mathcal{S}_Y^*$ as the original sample, compute the test statistic $T(\mathcal{S}_X^*, \mathcal{S}_Y^*)$.
4. Repeat the above two procedures several times, record the value of test statistic of each time.

After repeating the above permutation procedure B times, we obtain B values of the test statistic $T^{*(1)}, \cdots, T^{*(B)}$. Then we compute the p-value of testing H_0 as

$$\mathsf{pvalue} = \frac{1}{B}\sum_{\ell=1}^{B} I\left(|T^{*(\ell)}| \geq |T(\mathcal{S}_X, \mathcal{S}_Y)|\right)$$

A power of permutation test is that we are free to use any summary statistic, which makes it very flexible.

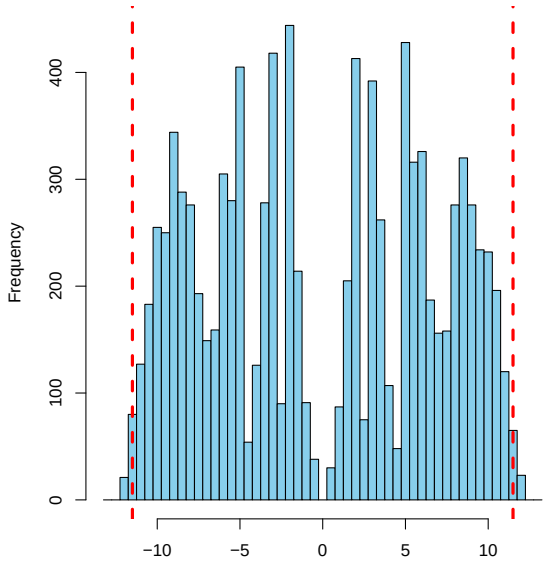
Example 1. Consider the example we have visited in Lecture 1:

$$\begin{aligned} \mathcal{S}_X &= -2, -1, -1, 0, -2, -1, 0, -1, -2, 100 \\ \mathcal{S}_Y &= 7, 13, 11, 5, 14, 9, 8, 10, 12, 11. \end{aligned} \tag{3.1}$$

Now we assume that our test statistic is the median difference, i.e.,

$$T(\mathcal{S}_X, \mathcal{S}_Y) = \mathsf{med}(\mathcal{S}_X) - \mathsf{med}(\mathcal{S}_Y) = -1 - 10.5 = -11.5.$$

The following picture shows the distribution of $T(\mathcal{S}_X, \mathcal{S}_Y)$ a 10000 times permutation:



The two red vertical lines show the observed value of $T(\mathcal{S}_X, \mathcal{S}_Y)$ and $-T(\mathcal{S}_X, \mathcal{S}_Y)$. The histogram displays the distribution of $T(\mathcal{S}_X, \mathcal{S}_Y)$ under H_0 based on permutation. To compute the p-value, we need to count the number of cases where $|T^*|$ is less than or equal to $|T(\mathcal{S}_X, \mathcal{S}_Y)|$. This is the same as counting the number of T^* outside the two vertical red lines. There are totally 189 cases satisfying this condition, so the p-value is 0.0189.

Why permutation test works? It works because under H_0, the two samples are from the same distribution. Thus, randomly exchanging the elements in the two samples should give us a new set of data from the same distribution.

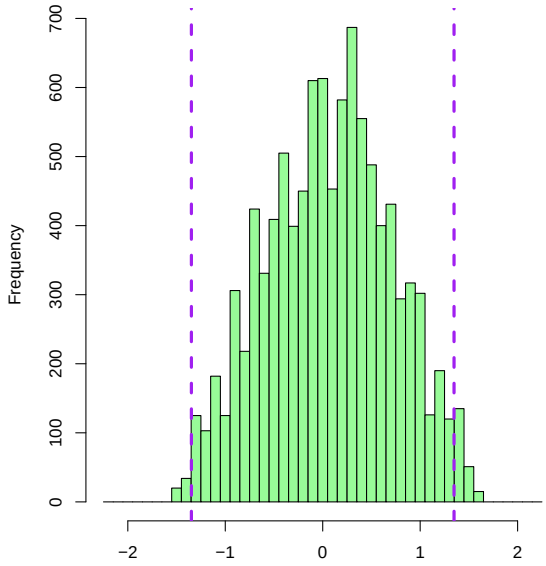
Example 2. Here is another example where the permutation test is applied to two samples with different sizes. Assume we have two samples

$$\begin{aligned} \mathcal{S}_X &= -1.8, 0.4, 3.2, -2.3, -0.2, 0.3, 1.4, -0.5, 4.0, -0.3 \\ \mathcal{S}_Y &= 0.2, 0.5, -0.2, -0.5, 0.9, -1.2, 0.4, 0.0, 0.5, 0.2, 1.0, -0.6. \end{aligned} \tag{3.2}$$

This time we choose our test statistic as the difference between sample standard deviation

$$T(\mathcal{S}_X, \mathcal{S}_Y) = s_X - s_Y = 1.99 - 0.64 = 1.35.$$

After permuting the data 10000 times, we obtain the following distribution:



Again, the two vertical lines indicate the value of $T(\mathcal{S}_X, \mathcal{S}_Y)$ and $-T(\mathcal{S}_X, \mathcal{S}_Y)$. By counting the number of permutations leading to $|T^*| \geq |T(\mathcal{S}_X, \mathcal{S}_Y)|$, we obtain a p-value 0.0259.

3.2 Rank Correlation

In the previous few lectures, we are working on two-sample test problems. Now we will move to a new problem in statistics: analyzing a bivariate random sample. Assume that each of our observation has two variable X and Y and we have n observations. The goal is to understand how the two variables associated with each other. Our data can be described as a bivariate random sample

$$(X_1, Y_1), \cdots, (X_n, Y_n).$$

A common approach to investigate the relationship between X and Y is through their correlation coefficient, also known as the Pearson's correlation:

$$\widehat{r}_{XY} = \frac{s_{XY}^2}{s_X s_Y},$$

where

$$s_{XY}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}_n \bar{Y}_n$$

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2$$

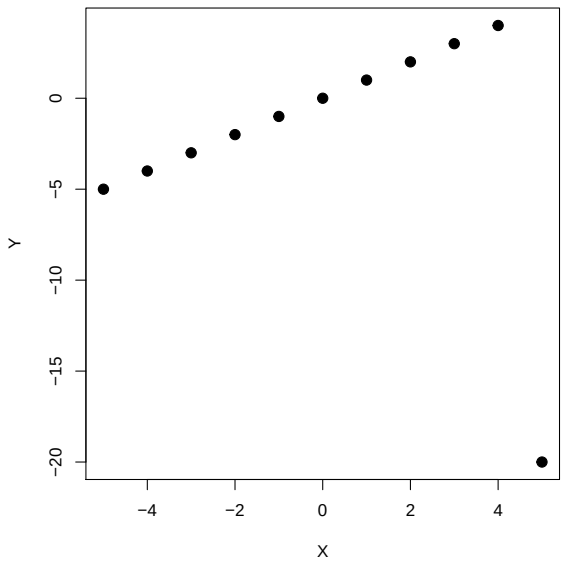
$$s_Y^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - (\bar{Y}_n)^2.$$

The correlation coefficient measures the linear relationship between the two variables. However, similar to the sample average, the correlation coefficient is vulnerable to the outliers.

Example 3. For example, consider a bivariate data

$$(X, Y) = (-5, -5), (-4, -4), \cdots, (3, 3), (4, 4), (5, -20).$$

The scatter plot is as follows:



Apparently, the two variables are highly correlated except the last observation. However, the correlation coefficient is about -0.07 because of the last observation.

3.2.1 Spearman's ρ

Now we introduce a robust approach of calculating the correlation between two variables. This robust approach is called the rank correlation and as you can expect, we will make a rank transformation. For observations' first variable $(X_1, \cdots, X_n)$, we compute their ranks, denoting as $R_1, \cdots, R_n$. Similarly, we calculate the ranks of the second variables, denotes as $S_1, \cdots, S_n$.

The *Spearman's* ρ (also known as the Spearman's coefficient) is

$$\widehat{\rho} = \widehat{r}_{RS}.$$

Namely, we use the correlation coefficients of the ranks as a measurement of correlation. In the previous example, the Spearman's ρ is 0.5, reflecting the result we have seen.

Similar to the rank test or sign test, the Spearman's ρ is robust to outliers. This is a common feature of all rank-based approach – they are robust to outliers. Actually, there are approaches of using median to compare correlation coefficients using the contingency table method (we will learn it in the next lecture).

When there is no ties within X's and within Y's, there is a simple formula for calculating $\widehat{\rho}$:

$$\widehat{\rho} = 1 - \frac{6\sum_{i=1}^{n}(R_i - S_i)^2}{n^3 - n}. \tag{3.3}$$

To see this, we need to use two facts:

$$\begin{aligned}
s_{XY}^2 &= \frac{1}{n}\sum_{i=1}^n X_iY_i - \bar{X}_n\bar{Y}_n \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n) \\
&= \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n (X_i - X_j)(Y_i - Y_j) \\
s_X^2 &= \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \\
&= \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n (X_i - X_j)^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
\widehat{\rho} &= \frac{s_{RS}^2}{s_R s_S} \\
&= \frac{\sum_{i=1}^n\sum_{j=1}^n (R_i - R_j)(S_i - S_j)}{\sqrt{\sum_{i=1}^n\sum_{j=1}^n (R_i - R_j)^2 \sum_{i=1}^n\sum_{j=1}^n (S_i - S_j)^2}} \\
&= \frac{\sum_{i=1}^n\sum_{j=1}^n (R_i - R_j)(S_i - S_j)}{\sum_{i=1}^n\sum_{j=1}^n (R_i - R_j)^2}
\end{aligned}$$

because $\sum_{i=1}^n\sum_{j=1}^n (R_i - R_j)^2 = \sum_{i=1}^n\sum_{j=1}^n (S_i - S_j)^2$. We first expand the numerator:

$$\begin{aligned}
\sum_{i=1}^n\sum_{j=1}^n (R_i - R_j)(S_i - S_j) &= \sum_{i=1}^n\sum_{j=1}^n R_iS_i - \sum_{i=1}^n\sum_{j=1}^n R_jS_i - \sum_{i=1}^n\sum_{j=1}^n R_iS_j + \sum_{i=1}^n\sum_{j=1}^n R_jS_j \\
&= 2n\sum_{i=1}^n R_iS_i - 2\sum_{i=1}^n\sum_{j=1}^n R_iS_j \\
&= 2n\sum_{i=1}^n R_iS_i - 2\underbrace{\sum_{i=1}^n R_i}_{=n(n+1)/2}\underbrace{\sum_{j=1}^n S_j}_{n(n+1)/2} \\
&= 2n\sum_{i=1}^n R_iS_i - \frac{n^2(n+1)^2}{2}.
\end{aligned}$$

Now we consider $\sum_{i=1}^n (R_i - S_i)^2$:

$$\sum_{i=1}^n (R_i - S_i)^2 = 2\sum_{i=1}^n R_i^2 - 2\sum_{i=1}^n R_iS_i.$$

Combing the above two equations together, we obtain

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}(R_i-R_j)(S_i-S_j) &= 2n\sum_{i=1}^{n}R_iS_i - \frac{n^2(n+1)^2}{2}\\ &= 2n \underbrace{\sum_{i=1}^{n}R_i^2}_{n(n+1)(2n+1)/6} - n\sum_{i=1}^{n}(R_i-S_i)^2 - \frac{n^2(n+1)^2}{2}\\ &= \frac{1}{6}n^2(n^2-1) - n\sum_{i=1}^{n}(R_i-S_i)^2.\end{aligned}$$

Because

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}(R_i-R_j)^2 &= 2n\sum_{i=1}^{n}R_i^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n}R_iR_j\\ &= 2n \underbrace{\sum_{i=1}^{n}R_i^2}_{=n(n+1)(2n+1)/6} - 2 \underbrace{\sum_{i=1}^{n}R_i}_{=n(n+1)/2} \sum_{j=1}^{n}R_j\\ &= \frac{1}{6}n^2(n^2-1).\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}\widehat{\rho} &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(R_i-R_j)(S_i-S_j)}{\sum_{i=1}^{n}\sum_{j=1}^{n}(R_i-R_j)^2}\\ &= \frac{\frac{1}{6}n^2(n^2-1) - n\sum_{i=1}^{n}(R_i-S_i)^2}{\frac{1}{6}n^2(n^2-1)}\\ &= 1 - \frac{6\sum_{i=1}^{n}(R_i-S_i)^2}{n(n^2-1)},\end{aligned}$$

which is equation (3.3).

3.2.2 Kendall's τ

In addition to the Spearman's ρ, there is another nonparametric approach of measuring the correlation between the two variables called the Kendall's τ. The idea of Kendall's τ is based on comparing pairs of observations. For a pair of observations, say i-th and j-th observations, we say they are *concordant* if either (1) $X_i < X_j$ and $Y_i < Y_j$ or (2) $X_i > X_j$ and $Y_i > Y_j$. Namely, the (i, j) ordering is the same in both variables. We say this pair os *discordant* if either (1) $X_i < X_j$ and $Y_i > Y_j$ or (2) $X_i > X_j$ and $Y_i < Y_j$. Note that if there is an equality, it is neither concordant or discordant.

We have n observations, so there are $n(n-1)/2$ distinct pairs. Let n_c and n_d denote the number of concordant and discordant pairs, respectively. The Kendall's τ is

$$\widehat{\tau} = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}. \tag{3.4}$$

Using the data in Example 3, the Kendal's τ is $\frac{35}{55} \approx 0.64$ because there are 45 concordant pairs and 10 discordant pairs (please check this).

The intuition of Kendall's τ is: When the two variables are positively correlated, there should be more concordant pairs than the discordant pairs. On the other hand, if the two variables are negatively correlated, more discordant pairs will be observed than the concordant pairs.

Note that if we define

$$A_{ij} = \mathsf{sgn}(R_j - R_i), \quad B_{ij} = \mathsf{sgn}(S_j - S_i),$$

then the Kendall's τ can be written as

$$\widehat{\tau} = \frac{\sum_{i,j=1}^n A_{ij} B_{ij}}{\sqrt{\sum_{i,j=1}^n A_{ij}^2 \sum_{i,j=1}^n B_{ij}^2}}. \tag{3.5}$$

Think about why equation (3.4) and (3.5) are the same.

3.3 Independence Test

The correlation (coefficient) is a common quantity for describing the interaction between two random variables. But being correlated is just a special case of being dependent. In many situation, we may want to know if two variables are dependent or not.

In this situation, we want to test

$$H_0 : X \text{ and } Y \text{ are independent.} \tag{3.6}$$

Of course, this null hypothesis implies

$$H_0 : r_{XY} = 0$$

so sometimes people test the dependence by testing if the correlation coefficient is significantly different from 0.

However, one has to be very careful: the two variables can be highly dependent but has 0 correlation. The following is one example:

$$(X, Y) = (-3, 9), (-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4), (3, 9).$$

In fact, they are dependent because they are from $Y = X^2$. But if you compute their correlation coefficient, you will obtain $r_{XY} = 0$! Thus, sometimes the correlation may provide us insufficient information about the dependence.

There are many approaches for testing the dependence, we will focus on the two methods we have talked about: the Spearman's ρ and Kendall's τ.

Under H_0 of (3.6), the Spearman's ρ has an asymptotic distribution

$$\widehat{\rho} \approx N\left(0, \frac{1}{n-1}\right)$$

when the sample size is large. Thus, we can compute the p-value using

$$\mathsf{pvalue} = 2 \times \Phi\left(-\sqrt{|\widehat{\rho}|^2 \cdot (n-1)}\right),$$

where $\Phi(t)$ is the CDF of the standard normal distribution.

In the case of Kendall's τ, under H_0 of (3.6),

$$\widehat{\tau} \approx N\left(0, \frac{2(2n+5)}{9n(n-1)}\right).$$

Therefore, the p-value can be computed using

$$\mathsf{pvalue} = 2 \times \Phi\left(-\sqrt{\frac{|\widehat{\tau}|^2 \cdot 9n(n-1)}{2(2n+5)}}\right).$$

Although testing the correlation coefficient may not be enough for determining the dependence in general, being correlated and being dependent are equivalent when X and Y follow a bivariate normal distribution. The two random variables X, Y follow a bivariate normal distribution if their PDF is

$$p(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-r^2}} \exp\left(-\frac{1}{2(1-r^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2r\cdot(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right)\right),$$

where $\mu_X = \mathbb{E}(X), \sigma_X^2 = \mathsf{Var}(X)$ and $r = r_{XY}$ is the correlation coefficient. It is easy to see that if $r = 0$, then

$$\begin{aligned} p(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right) \\ &= \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}\right) \times \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right) \\ &= p(x)p(y), \end{aligned}$$

which implies the independence.