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

Lecture 1: Robust Two-Sample Test

Instructor: Yen-Chi Chen

1.1 Introduction

In many problems, we would like to test if two samples are coming from the same population. This is a common scenario in scientific studies. For instance, if a research lab invented a new drug for curing a disease, they will often do a clinical trial to test if this. In a clinical trial, the researchers will recruit a couple of patients and then randomized them into two groups: a control group and a treatment group. For individuals in the control group, they will receive a placebo whereas the individuals in the treatment group will receive the actual drug. After a couple of weeks, the patients will be asked to come back and the researchers will examine disease characteristics of each individuals. Finally, they will compare the responses from the control group versus the treatment group and test if the responses are different.

For simplicity, we assume that we have equal sample size in both control and treatment group. The above problem can be formulated as having two sets of observations

$$\begin{aligned}\mathcal{S}_X &= X_1, \cdots, X_n \sim P_X \\ \mathcal{S}_Y &= Y_1, \cdots, Y_m \sim P_Y,\end{aligned}$$

where X_i's denote the response of individuals in the control group and Y_i's are the response of individuals in the treatment group. The quantities n, m are the sample sizes. Namely, we assume that these responses are from two populations: the population of those taking placebo P_X and the population of those taking the drug P_Y.

If the drug has no influence on the diseases, then there should be no difference between the two populations. Namely, $P_X = P_Y$ if the drug has no effect. Thus, the two-sample test problem is to test

$$H_0 : P_X = P_Y. \tag{1.1}$$

Testing equation (1.1) might be non-trivial because both P_X and P_Y are functions (they are cumulative distribution functions). Thus, many methods will test certain characteristics of the cumulative distribution function. For instance, the t-test[1] (or Z-test) will test

$$H_0 : \mu_X = \mu_Y, \tag{1.2}$$

where μ_X, μ_Y are the mean of P_X and P_Y, respectively. Two-sample tests based on comparing the mean is often called a mean test.

However, testing the mean has a potential problem – when there are outliers, the power will be low. Namely, even the two samples are clearly very different from each other, we often are still unable to reject them. Consider the following very simple example:

$$\begin{aligned}\mathcal{S}_X &= -2, -1, -1, 0, -2, -1, 0, -1, -2, 100 \\ \mathcal{S}_Y &= 7, 13, 11, 5, 14, 9, 8, 10, 12, 11.\end{aligned} \tag{1.3}$$

[1] I assume you all know t-test. If you are not familiar with it, please read some related materials.

The sample average of both sample are 9 but you can clearly see that except the last observation of $\mathcal{S}_X$, all other observations of $\mathcal{S}_X$ are smaller than $\mathcal{S}_Y$. Thus, the t-test or other mean tests fail in this case!

The mean tests are not ideal when the outliers are present. In the next few sections, we are going to introduce a few new methods that are much more robust to outliers.

1.2 Equal sample size: Sign Test

We first introduce a new set of variables $Z_1, \cdots, Z_n$, where $Z_i = Y_i - X_i$ for each $i = 1, \cdots, n$. Under the H_0 in equation (1.1), $P_X = P_Y$ so the difference Z_i is coming from a distribution with 0 mean and 0 median.

Sign test utilizes the fact that the distribution of Z_i has 0 median. This implies that the chance of having positive value and the chance of having negative value of Z_i are the same: 0.5 (we assume that the two samples are from a continuous distribution). Thus, if we count the number of Z_i's being positive, this number should follow a binomial distribution $\mathsf{Bin}(n, 0.5)$ under H_0. Let this number be T.

In the example at the beginning, we have

$$\mathcal{S}_Z = -9, -14, -12, -5, -16, -10, -8, -11, -14, 89.$$

There are only one of them has a positive sign. Thus, $T = 1$ in our case.

If H_0 is correct, the number of positive difference should be around $n/2$ (half of the sample size). The number being very small (≈ 0) or very large ($\approx n$) are both evidences against H_0. A simple way to compute a p-value using T is to calculate the probability of observing a more extreme event again H_0 than what is observed. In our case, we see $T = 1$ so a more extreme event is $T = 0$ and each of them corresponds to a probability

$$P(T = 1) = \frac{\binom{10}{1}}{2^{10}}, \quad P(T = 0) = \frac{\binom{10}{0}}{2^{10}}.$$

However, $T \approx 10$ is also against H_0 so here we need to do a two-sided test. Thus, the p-value will be two times the sum of the above two probabilities, which is

$$\mathsf{pvalue}_1 = 2(P(T = 1) + P(T = 0)) = \frac{22}{1024} \approx 2.15\%.$$

We can use asymptotic normality to carry this test as well. If H_0 is correct, then T will follow $\mathsf{Bin}(n, 0.5)$, which is approximately normal with $N(n/2, n/4)$. Thus, we can use

$$T_* = \frac{T - n/2}{\sqrt{n/4}}$$

as out test statistic. T_* will follow a standard normal distribution so the p-value can be computed using

$$\mathsf{pvalue}_2 = 2P(N(0, 1) \geq |T_*|) \approx 1.14\%.$$

You may notice that p-values are from the two methods are slightly different. The former one is the exact p-value whereas the later one (using normal distribution) is an asymptotic approximation. When the sample size is small, the asymptotic approximation may not be very accurate.

1.3 Equal sample size: Signed-Rank Test

The sign test has a powerful extension that incorporates the information of 'rank'. This extension is called the signed-rank test, also known as Wilcoxon signed-rank test. It assumes a little bit more assumptions about the the distribution of the differences Z_i: symmetric. The signed-rank test requires that the distribution of Z_i is symmetric.

In our case, we have $Z_1, \cdots, Z_n$, n differences. Let $\mathsf{sgn}(Z_i)$ denotes the sign of Z_i ($+1$ or -1). Now we define a new variable $R_i = \mathsf{rank}(|Z_i|)$, where $R_i = 1$ means that Z_i is the smallest value of $|Z_1|, \cdots, |Z_n|$ and $R_i = n$ implies that Z_i is the largest value.

The signed-rank test uses a test statistic

$$W = \sum_{i=1}^{n} \mathsf{sgn}(Z_i) R_i.$$

Namely, W is the signed-rank sum. Under H_0, the median of Z_i is 0, so W should be close to 0. We can then use the limiting distribution of W to test if the median is 0 or not. Note that because we assume that the distribution of Z is symmetric, the mean and median are the same.

Under H_0, the variance of W is

$$\mathsf{Var}(W) = \frac{n(n+1)(2n+1)}{6}$$

and

$$\frac{W}{\sqrt{\mathsf{Var}(W)}} \xrightarrow{D} N(0,1).$$

Intuition of the variance: $W = \sum_{i=1}^{n} \mathsf{sgn}(Z_i) R_i = \sum_{i=1}^{n} Y_i$ with $\mathbb{E}(Y_i) = 0$. Although Y_i's are not independent, they are exchangeable. Then $\mathsf{Var}(W) = \mathbb{E}(W^2) = \sum_{i=1}^{n} \mathbb{E}(Y_i^2) + \sum_{j \neq \ell = 1}^{n} \mathsf{Cov}(Y_j, Y_\ell)$. The covariance is 0 by symmetry and $\mathbb{E}(Y_i^2) = \frac{1}{n} \sum_{i=1}^{n} i^2 = \frac{1}{6}(n+1)(2n+1)$.

Using the asymptotic normality, the p-value can be computed using

$$\mathsf{pvalue} = 2P\left(N(0,1) \geq \left| \frac{(W-0)}{\sqrt{\mathsf{Var}(W)}} \right| \right) = 2P\left(N(0,1) \geq \sqrt{\frac{6W^2}{n(n+1)(2n+1)}} \right).$$

Now, going back to our example:

Z_i	-9	-14	-12	-5	-16	-10	-8	-11	-14	89
$\mathsf{sgn}(Z_i)$	-1	-1	-1	-1	-1	-1	-1	-1	-1	+1
R_i	3	7.5	6	1	9	4	2	5	7.5	10

Thus, $W = -35$, leading to a p-value $2P\left(N(0,1) \geq \sqrt{\frac{6W^2}{n(n+1)(2n+1)}}\right) \approx 7.45\%$.

Although the signed-rank test has a better power of testing H_0 compared to the signed test, the signed test works for general ordinal data whereas signed-rank test will not (think about why).

1.4 Non-equal sample size: Rank-Sum Test

The previous two method assume that the two samples have equal sizes, which may not be true in reality. Here we introduce a novel generalization called the rank-sum test, which does not require the two samples

to be equal in size. It is also called the Mann-Whitney U test, Mann-Whitney-Wilcoxon (MWW), Wilcoxon rank-sum test, and Wilcoxon-Mann-Whitney test. The rank-sum test directly test the null hypothesis $H_0 : P_X = P_Y$.

The idea of the rank-sum test is similar to the signed-rank test but now we do not use the sign. Instead, we pull both samples together and compute the rank of each observation. Then we compute the summation of rank in one sample (the summation of rank in the other sample will also be determined). The summation of ranks will be our test statistics.

Recall that our original two samples are $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$. We pull them together to form a new data $D_1, \cdots, D_{n+m}$ such that $D_i = X_i$ for $i = 1, \cdots, n$ and $D_i = Y_{i-n}$ for $i = n+1, \cdots, n+m$. Essentially, we just concatenating the two samples to form a new one. Then let $S_1, \cdots, S_{n+m}$ be the rank of $D_1, \cdots, D_{n+m}$. Again, a smaller rank means a lower value whereas a higher rank indicates a higher value.

Before we formulate our test statistic, consider the following two statistics:

$$Q_1 = \sum_{i=1}^{n} S_i, \quad Q_2 = \sum_{i=n+1}^{n+m} S_i.$$

Q_1 is the sum of ranks in $\mathcal{S}_X$ while Q_2 is the sum of ranks in $\mathcal{S}_Y$. Note that $Q_1 + Q_2 = \frac{(n+m)(n+m+1)}{2}$.

Now we define two statistics:

$$U_1 = Q_1 - \frac{n(n+1)}{2}, \quad U_2 = Q_2 - \frac{m(m+1)}{2}.$$

We can use any of them as our test statistic because $U_1 + U_2 = n \cdot m$. You can view U_ℓ as the "joint rank sum" (two sample jointly) minus individual group rank-sum (using only one sample). U_ℓ has several nice properties:

$$\mathbb{E}(U_\ell) = \frac{nm}{2}, \quad \mathsf{Var}(U_\ell) = \frac{nm(n+m+1)}{12}[2],$$

and

$$\frac{U_\ell - \mathbb{E}(U_\ell)}{\sqrt{\mathsf{Var}(U_\ell)}} \xrightarrow{D} N(0,1).$$

Thus, we can use this asymptotic distribution to test H_0. Essentially, the p-value will be

$$\mathsf{pvalue} = 2P\left(N(0,1) \geq \left|\frac{U_1 - \mathbb{E}(U_1)}{\sqrt{\mathsf{Var}(U_1)}}\right|\right) = 2P\left(N(0,1) \geq \left|\frac{U_\ell - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}\right|\right).$$

Again, we turn to our example in equation (1.3): Using the above table, $Q_1 = 65$ so $U_1 = 65 - 55 = 10$.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
value	-2	-1	-1	0	-2	-1	0	-1	-2	100	7	13	11	5	14	9	8	10	12	11
S_i	2	5.5	5.5	8.5	2	5.5	8.5	5.5	2	20	11	18	15.5	10	19	13	12	14	17	15.5

Thus, the p-value is

$$\mathsf{pvalue} = 2P\left(N(0,1) \geq \left|\frac{10 - 50}{\sqrt{175}}\right|\right) \approx 0.25\%.$$

Assumptions (Rank-Sum Test).

[2]Note that this result is assuming no ties in the data. When there are ties, there will be some modification for the variance.

- All observations are independent from each other.
- Observations within the same sample are IID.

The rank-sum test requires much weaker assumptions compared to the signed-rank test or signed test. And it also works for ordinal data. However, in general, the rank-sum test is a consistent test (power goes to infinity when the sample size goes to infinity) if $P(X > Y) \neq P(Y > X)$. Namely, the distribution of one dataset is stochastically greater than the distribution of the other.

1.5 Median Test: One Sample Case

We end this section with a simple application of the median. We will consider testing the median in a one-sample problem. Namely, we only have one sample and our goal is to test if the median equals to certain value. Assume that we observe a sample

$$\mathcal{S}_X = -2, -1, -1, 0, -2, -1, 0, -1, -2, 100.$$

Now we want to test

$$H_0 : m_X = 9.$$

How should we test this statement?

If H_0 is true, then the median is 9, so any observation having a value above 9 is about 50%. We can simply compute the ratio of observations whose value is above 9. If H_0 is correct, the ratio should be around 50%. In our case, we see that there is only 1 out of 10 observations whose value is above 9.

Well, this could be cause by purely sampling error. So to see how unlikely our observation is, we compute its p-value. The p-value is the chance of observing a more extreme event than what we have observed, which corresponds to 0 out of 10 and 1 out of 10 observations whose value is above 9.

$$P(0 \text{ out of } 10) = \frac{1}{2^{10}}, \quad P(1 \text{ out of } 10) = \frac{10}{2^{10}}.$$

Notice that we need to do a two-sided test here (think about why), so the p-value will be 2 times the sum of the above two probabilities, which is

$$\mathsf{pvalue} = 2 \times (P(0 \text{ out of } 10) + P(1 \text{ out of } 10)) = \frac{22}{1024} \approx 2.15\%.$$

Essentially, the median test is based on the powerful property that under H_0, any observations being greater than or less than the median value is 0.5. Then we just need to apply combinatorial methods to compute the p-value.

Here is a description about a general form of the median test. Assume we observe $Z_1, \cdots, Z_n \sim P_Z$ and we want to test

$$H_0 : m_Z = m_0.$$

Then we first count the number of observations such that

$$T = \sum_{i=1}^{n} I(Z_i \geq m_0),$$

where $I(\cdot)$ is an indicator function such that if the input is a true statement, it return 1 and 0 otherwise. Namely, T is the number of observations whose value is above or equal to the median. Because both large T and small T are the case where H_0 is unlikely to be true, we use a symmetrized test statistic

$$T_* = \begin{cases} T & \text{if } T \leq n/2 \\ n - T & \text{if } T > n/2. \end{cases}$$

The p-value will then be

$$\mathsf{pvalue} = \frac{2 \times \sum_{\ell=0}^{T_*} \binom{n}{\ell}}{2^n}.$$