MANA130083.01 Nonparametric *Spring 2025*

Lecture 5

Instructor: Prof. Bowen Gang *Scribes: Yize Wang, Jingyi Zhou*

5.1 Review: Linear Rank Statistic

Last class we discussed linear rank statistic which is used for test of location problem. Say, we have two series of samples $X_1, ...X_m \sim F_X$ and $Y_1, ..., Y_n \sim F_Y$. We can pool them together and order them from smallest to largest:

$$X_{(1)}, Y_{(1)}, X_{(2)}, Y_{(2)}, Y_{(3)}, Y_{(4)}, X_{(3)}, X_{(4)}, ...$$

which corresponds to a vector z, which denotes samples from F_X as 1 and samples from F_Y as 0:

$$z = (1, 0, 1, 0, 0, 0, 1, 1, ...)$$

5.2 Linear Rank Statistic for the Scale Problem

Here still we have $X_1, ...X_m \sim F_X$ and $Y_1, ..., Y_n \sim F_Y$. Let's also suppose $Var(X) = \sigma_X^2$ and $Var(Y) = \sigma_Y^2$. Here the null hypothesis we would like to test is $H_0 : \sigma_X = \sigma_Y$. Under parametric setting, say, $F_X \sim N\left(\mu_X, \sigma_X^2\right)$ and $F_Y \sim N\left(\mu_Y, \sigma_Y^2\right)$. We then should use the F-statistic:

$$F_{m-1,n-1} = \frac{\frac{1}{m-1}\sum\left(X_i - \bar{X}\right)^2}{\frac{1}{n-1}\sum\left(Y_i - \bar{X}\right)^2}$$

where both the numerator and denominator are sample variances. If the F-statistic is either too large or too small, we should reject the null.

5.2.1 Mood Test

Under non-parametric setting, let's write these things in a more general way.

$$H_0 : F_Y(x) = F_X(x)$$
$$H_a : F_Y(x) = F_X(\theta x)$$

where $\theta \neq 1$. This is a generalization of our previous 'normal' setting. So how can we compare the variance of two non-parametric distributions? A natural idea is to do similar things as before. We can pool the samples together. For example, let's look at the following two cases.

$$X, Y, X, Y, Y, X, X, Y, Y, ... \tag{1}$$
$$X, X, X, X, Y, Y, Y, Y, X, ... \tag{2}$$

In case (2), we can tell that X has larger variation and wider spread. So the variance of X should be greater than that of Y. One way to think of this idea is to define the test statistic as follows:

$$T = \sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)z_i$$

where $N = m + n$ and z_i is the i-th element of the 'z' vector we defined before. Here we are actually comparing how far the ranks of X are from the middle position(rank) $\frac{N+1}{2}$. If this statistic is large, we tend to believe that X has variance greater than Y. However, we have 'offset' problem here:

$$XYXYXY \tag{1}$$
$$XXYYYX \tag{2}$$

You can check that both (1) and (2) have $T = -1.5$. However in (2) obviously X and Y have different variance, which is not reflected in the statistic T. In order to solve this kind of 'offset' problem, we do the square sum here:

$$M_N = \sum_{i=1}^{N} \left(i - \frac{N+1}{2} \right)^2 z_i$$

which is called the Mood test. We reject the null hypothesis $F_Y(x) = Y_X(x)$ when M_N is too large, where the alternative hypothesis is given by $F_{Y-\mu}(x) = F_{X-\mu}(\theta x)$. To determine 'how large is large', we use simulation to estimate the p-value. Also we know that $T = \sum a_i z_i$ is a linear rank statistic, which is asymptotically normal and we have known the mean and variance last class. We can also estimate the p-value according to this property. For the Mood statistic, similarly we have

$$\mathbb{E}\left[M_N\right] = \frac{m(N^2-1)}{12}$$
$$Var\left[M_N\right] = \frac{mn(N+1)(N^2-4)}{180}$$

where m and n are sample size of X and Y respectively and $N = m + n$. Naturally we also have

$$\frac{M_N - \mathbb{E}\left[M_N\right]}{\sqrt{Var\left[M_N\right]}} \xrightarrow{D} N(0,1)$$

5.2.2 Ansari-Bradley-Freund-Barton-David Test

We have different ways to ensure that our test statistic is positive. One way as we discussed before is to do a square sum. Similarly we can also replace the parentheses by absolute value.

$$A_N = \sum_{i=1}^{N} \left| i - \frac{N+1}{2} \right| z_i$$

which is called the Ansari-Bradley-Freund-Barton-David test. This is also a linear rank statistic so we know the mean and variance here and it is asymptotically normal. We can use either asymptotic distribution or simulation to estimate the p-value here.

5.2.3 Siegel-Tukey Test

It is still a linear rank statistic of the form

$$S_N = \sum_{i=1}^{N} a_i z_i$$

where a_i is somehow complicated

$$a_i = \begin{cases} 2i, & i \text{ even and } 1 < i \leqslant \frac{N}{2} \\ 2i-1, & i \text{ odd and } 1 \leqslant i \leqslant \frac{N}{2} \\ 2(N-i)+2, & i \text{ even and } \frac{N}{2} < i \leqslant N \\ 2(N-i)+1, & i \text{ odd and } \frac{N}{2} < i \leqslant N \end{cases}$$

You can refer to the following to understand the idea more clearly.

$$i = 1, 2, 3, 4, 5, ..., \frac{N}{2}, ..., N-4, N-3, N-2, N-1, N$$
$$a_i = 1, 4, 5, 8, 9, ..., N, ..., 10, 7, 6, 3, 2$$

For example, if X has larger variance, most of X's would be in the tail area, which corresponds to small a_i and small S_N. Then we should reject the null. Another advantage of this ranking method is that the ranks are all distinct, i.e., it is a permutation of $1, 2, ..., N$. Therefore, the Siegel-Tukey test is analogous to the Wilcox test. To be more specific, under null the test statistic S_N will follow the same null distribution as the test statistic in Wilcox test ($W_N = \sum_{i=1}^{N} i z_i$). This brings calculation convenience to us because we can use the same p-value table as Wilcox test. So this is the historical reason for Siegel-Tukey test.

Similar to Wilcox test, we also know the mean and variance of S_N under null hypothesis.

$$\mathbb{E}[S_N] = \frac{m(N+1)}{2}$$
$$Var[S_N] = \frac{mn(N+1)}{2}$$

And the standardized statistic also asymptotically follow normal distribution.

5.2.4 Other Useful Tests

Last time we talked about the Terry-Hoeffiding test. The statistic is $\sum_{i=1}^{N} \mathbb{E}\left[\xi_{(i)}\right] z_i$, where $\xi_{(i)}$ is the i-th order statistic of standard normal.

The approximation of Terry is the Van Der Waerden test. The statistic is then $\sum_{i=1}^{N} \Phi^{-1}\left(\frac{i}{N+1}\right) z_i$, where $\Phi^{-1}\left(\frac{i}{N+1}\right)$ is the first order approximation of $\mathbb{E}\left[\xi_{(i)}\right]$. Here we can modify this in order to construct a test for variance.

Note that here we still need to solve the offset problem. For example we can construct the test statistic as $\sum_{i=1}^{N} \left[\mathbb{E}(\xi_{(i)}^2)\right] z_i$, which is the Klotz normal scores test. The corresponding approximation is then $\sum_{i=1}^{N} \left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^2 z_i$.

To sum up, the procedure for non-parametric test is not difficult. It is all about thinking of a certain kind of pattern which is the evidence to reject the null hypothesis. Then we only need to construct a statistic to capture the pattern. The test statistics can be intuitive, flexible or even naive.

We also have the Sukhatame test which is similar to the Mann-Whitney test as discussed last class ($\sum\sum D_{ij}$ where $D_{ij} = 1$ if $Y_j < X_i$). Note that here we need an important assumption for the Sukhatame test: X and Y have the same median. Without loss of generality, we assume they all have $median = 0$. Here $D_{ij} = 1$ when Y is less close to the median than X, i.e., $Y < X < 0$ or $Y > X > 0$.

$$D_{ij} = \begin{cases} 1, & if Y_j < X_i < 0 \text{ or} 0 < X_i < Y_j \\ 0, & otherwise \end{cases}$$

Similarly the test statistic is

$$T = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij}$$

Under null, T depends on a parameter p:

$$p = Pr\left(Y < X < 0,\ or 0 < X < Y\right)$$

Under null, $p = \frac{1}{4}$. The asymptotic distribution of T is given by

$$\mathbb{E}\left[T|H_0\right] = \frac{mn}{4}$$
$$Var\left[T|H_0\right] = \frac{mn(N+4)}{48}$$

T is asymptotically normal:

$$\frac{4\sqrt{3}\left(T - \frac{mn}{4}\right)}{\sqrt{mn(N+4)}} \to N\left(0,1\right)$$

5.3 Test of the Equality of k Independent Samples

The setting becomes k samples in this section:

$$X_{11}, X_{12}, ..., X_{1n_1} \sim F_1$$
$$X_{21}, X_{22}, ..., X_{2n_2} \sim F_2$$
$$\vdots$$
$$X_{k1}, X_{k2}, ..., X_{kn_k} \sim F_k$$

The null hypothesis is

$$H_0 : F_1(x) = F_2(x) = ... = F_k(x)$$

For the alternative hypothesis, firstly we have location alternative: let

$$F_1 = F(x - \theta_1), F_2 = F(x - \theta_2), ..., F_k = F(x - \theta_k)$$

which corresponds to the hypothesis:

$$H_0 : \theta_1 = \theta_2 = ... = \theta_k$$
$$H_a : at\ least\ one\ of\ the\ equality\ is\ false$$

Under parametric setting, $F \sim Normal\ with\ the\ same\ variance$ and we can use F-test then:

$$F = \frac{\sum_{i=1}^{k} n_i \left(\bar{X}_i - \bar{X}\right)^2 / k - 1}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}\right)^2 / N - k}$$

where $N = \sum_{i=1}^{k} n_i$. The numerator is mean square between samples and the denominator is mean square within samples. Under null, $\overset{H_0}{\sim} F_{k-1,N-k}$.

5.3.1 Chi-Square Test

For non-parametric setting, we can construct a test as an extension of the median test. Under the null hypothesis, $\sum_{i=1}^{k} n_i = N$ samples are from the common population. Let the brand median be δ. An observation from any of the k samples as as likely to be above δ as below it. Let $u_i = \#$ of observations in sample number i which are less than δ and $t =$ total # of observations which are less than δ. It is obvious that $t = \sum u_i$. Under null,

$$t = \sum u_i = \begin{cases} \frac{N}{2} & , \ N \ even \\ \frac{N-1}{2} & , \ N \ odd \end{cases}$$

which is directly from the definition that δ is the median.

	sample 1	sample 2	$\cdots$	sample k	Total
$< \delta$	u_1	u_2	$\cdots$	u_k	t
$\geqslant \delta$	$n_1 - u_1$	$n_2 - u_2$	$\cdots$	$n_k - u_k$	$N - t$
Total	n_1	n_2	$\cdots$	n_k	N

Under the null hypothesis, $\mathbb{E}[u_i] \approx \frac{n_i}{2}$. If u_i deviates too much from its expectation, then it is an evidence to reject the null hypothesis. The joint distribution of $u_1, u_2, ..., u_k$ is given by

$$f(u_1, ..., u_k | t) = \frac{\binom{n_1}{u_1}\binom{n_2}{u_2}\cdots\binom{n_k}{u_k}}{\binom{N}{t}}$$

The joint distribution is difficult to evaluate even by computer because there are a lot of factorials. In practice, we can use the Chi-square test. Let $f_{i1} = u_i$, $f_{i2} = n_i - u_i$. Then the corresponding expectations are $e_{i1} = \frac{n_i t}{N} \approx \frac{1}{2} n_i$ and $e_{i2} = \frac{n_i (N-t)}{N} \approx \frac{1}{2} n_i$. Then we can look at the difference between the expectation and the observation: the test statistic is given by

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \overset{H_0}{\sim} \chi^2_{k-1}$$

Similarly we can have an extension of the control median test. Suppose we have independent random sample of size $n_1, n_2, ..., n_k$. Take the first sample as the control, and compare other samples to this sample. Also we choose $q \geqslant 1$ which is number of quantiles: $0 < p_1 < p_2 < \cdots < p_q < 1$. We find the quantiles in the first sample.

$$X_1^{(1)} < X_1^{(2)} < \cdots < X_1^{(q)}$$

Correspondingly we define the intervals:

$$I_j = \left(X_1^{(j)}, X_1^{(j+1)} \right)$$

for $j = 0, ..., q$, where we also let $X_1^{(0)=-\infty}$ and $X_1^{(q+1)} = \infty$ for convenience. Under null, the total number of samples that are in each interval should be similar. So we construct the test statistic as follows:

$$Q = \sum_{j=0}^{q} \pi_j^{-1} \sum_{i=1}^{k} n_i \left(\frac{v_{ij}}{n_i} - \frac{v_j}{N} \right)^2$$

where $v_{ij} =$ count for the i-th sample, the # of observations that belong to interval I_j, $v_j = \sum_{i=1}^{k} v_{ij}$. So $\frac{v_{ij}}{n_i}$ is the individual proportion and $\frac{v_j}{N}$ is the mean proportion. Under null, they should be close to each other. Also $\pi_j = \frac{v_{ij}}{n_i + 1}$. Under null, we have

$$Q \overset{H_0}{\sim} \chi^2_{(k-1)q}$$

provided $\frac{n_i}{N} \to c$, $0 < c < 1$. This test can also be used for scale alternative.

5.3.2 The Kruskal-Wallis One Way ANOVA

The idea is that the sum of all ranks is $1+2+\cdots+N=\frac{N(N+1)}{2}$. Under the null hypothesis, the rank sum should be evenly distributed among each group. For the i-th sample containing n_i observations, the expected sum of ranks is $\frac{N(N+1)}{2}\cdot\frac{n_i}{N}=\frac{n_i(N+1)}{2}$. Let R_i be the actual sum of ranks of group i $R_i=\sum_{j=1}^{n_i} r(X_{ij})$. Then the test statistic is constructed as

$$S=\sum_{i=1}^{k}\left(R_i-\frac{n_i(N+1)}{2}\right)^2$$

Consider

$$H=\frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{1}{n_i}\left[R_i-\frac{n_i(N+1)}{2}\right]^2\to\chi^2_{k-1}$$

which is modified from S. H is constructed in order to have a more beautiful null distribution. Let $\bar{R}_i=\frac{R_i}{n_i}$, then $\mathbb{E}\left[\bar{R}_i\right]=\frac{N+1}{2}$ and $Var\left[\bar{R}_i\right]=\frac{\sigma^2(N-n_i)}{n_i(N-1)}$ where $\sigma^2=\frac{\sum\left[i-\frac{N+1}{2}\right]^2}{N}=\frac{N^2-1}{12}$. We can also calculate $Var\left[\bar{R}_i\right]=\frac{(N+1)(N-n_i)}{12n_i}$ and $Cov\left(\bar{R}_i,\bar{R}_j\right)=\frac{-N+1}{12}$. If n_i is large,

$$z_i=\frac{\bar{R}_i-\frac{N+1}{2}}{\sqrt{\frac{(N+1)(N-n_i)}{12n_i}}}\to N(0,1)$$

Here z_i's are not independent. Slightly modify them,

$$\sum_{i=1}^{k}\frac{N-n_i}{N}z_i^2\to\chi^2_{k-1}$$

This is exactly the H we discussed before. If the null hypothesis is rejected, we are rejecting $H_0: F_1(x)=F_2(x)=\cdots=F_k(x)$. Further, we can compare groups to see which two groups are different:

$$z_{ij}=\frac{\bar{R}_i-\bar{R}_j}{\sqrt{\frac{N(N+1)}{12}\left[\frac{1}{n_i}+\frac{1}{n_j}\right]}}\overset{H_0}{\sim}N(0,1)$$

Because now we have $\frac{k(k-1)}{2}$ hypothesis to check, we should compare the above statistic with $z_{\frac{\alpha}{k(k-1)}}$ instead of $z_{\frac{\alpha}{2}}$.

5.4 Tests Against Ordered Alternatives

Here we also have

$$F_1(x)=F(x-\theta_1), F_2(x)=F(x-\theta_2), ..., F_k(x)=F(x-\theta_k)$$

The null and alternative are given by

$$H_0:\theta_1=\theta_2=\cdots=\theta_k,$$
$$H_a:\theta_1\leqslant\theta_2\leqslant\theta_3\leqslant\cdots\leqslant\theta_k$$
$$\textit{where at least one of } \leqslant \textit{ is strict}$$

You can expand the alternative into $\frac{k(k-1)}{2}$ inequalities:

$$\begin{aligned}
&\theta_1 \leqslant \theta_2, \theta_1 \leqslant \theta_3, \theta_1 \leqslant \theta_4, ..., \theta_1 \leqslant \theta_k \\
&\theta_2 \leqslant \theta_3, \theta_2 \leqslant \theta_4, \theta_2 \leqslant \theta_5, ..., \theta_2 \leqslant \theta_k \\
&\vdots \\
&\theta_{k-1} \leqslant \theta_k
\end{aligned}$$

where at least 1 inequality is strict. We can view it as a collection of $\frac{k(k-1)}{2}$ problems, each of which is a two-sample problem. Recall we have the Mann-Whitney U test in one two-sample problem. Here we can define U_{ij} as the U statistic for sample i and sample j. And we can combine the $U's$ together.

$$B = \underbrace{U_{12} + U_{13} + \cdots + U_{1k}} + \underbrace{U_{23} + U_{24} + \cdots + U_{2k}} + \cdots + \underbrace{U_{k-1,k}}$$

Either B is too large or too small, we tend to reject the null hypothesis.

5.5 Comparison With a Control

Here the null and alternative hypothesis are given by

$$\begin{aligned}
&H_0 : \theta_1 = \theta_2 = \cdots = \theta_k, \\
&H_a : \theta_2 \geqslant \theta_1, \theta_3 \geqslant \theta_1, \cdots, \theta_k \geqslant \theta_1 \\
&\textit{where at least one of the inequalities is strict}
\end{aligned}$$

Suppose θ is the median. We can look at

$$V_i = \sum_{j=1}^{n_i} I\left(X_{ij} < \theta_1\right) \quad (\approx \frac{n_i}{2}, \forall i, \text{under null})$$

where the test statistic is given by

$$V = \min\left(\frac{V_2}{n_2}, \frac{V_3}{n_3}, \cdots, \frac{V_k}{n_k}\right)$$

We tend to reject the null hypothesis when V^+ is too small.

5.6 Measure of Association for Bivariate Samples

Recall the Pearson correlation coefficient:

$$\rho(x, y) = \frac{cov\left(X, Y\right)}{\sqrt{Var\left(X\right) Var\left(Y\right)}}$$

which is invariant under changes of location and scale. If X, Y are independent then $\rho(x, y) = 0$. However, we cannot say X, Y are independent if $\rho(x, y) = 0$. We can say they are independent when $\rho(x, y) = 0$ and they are bivariate normal. Do we have other measures of association? Yes, we do. A good measure should have what properties?

1 For any two independent pairs (X_i, Y_i) and (X_j, Y_j) of random variables which follow this bivariate distribution, the measure will equal $+1$ if the relationship is direct and perfect in the sense that

$$X_i < X_j \textit{ whenever } Y_i < Y_j \textit{ or } X_i > X_j \textit{ whenever } Y_i > Y_j$$

This relation will be referred to as perfect concordance (agreement).

2 For any two independent pairs, the measure will equal -1 if the relationship is indirect and perfect in the sense that

$$X_i < X_j \textit{ whenever } Y_i > Y_j \textit{ or } X_i > X_j \textit{ whenever } Y_i < Y_j$$

This relation will be refereed to as perfect discordance (disagreement).

3 If neither criterion 1 nor criterion 2 is true for all pairs, the measure will lie between the two extremes -1 and $+1$. It is also desirable that, in some sense, increasing degrees of concordance are reflected by increasing positive values, and increasing degrees of discordance are reflected by increasing negative values.

4 The measure will equal zero if X and Y are independent.

5 The measure for X and Y will be the same as for Y and X, or $-X$ and $-Y$, or $-Y$ and $-X$.

6 The measure for $-X$ and Y or X and $-Y$ will be the negative of the measure for X and Y.

7 The measure will be invariant under all transformations of X and Y for which order of magnitude is preserved. i.e., $\rho(X, Y) = \rho(X, lnY) = \rho(X, Y^2)$ as long as $X > Y$, $X > lnY$ and $X > Y^2$.

You can check that Pearson correlation meets the first 6 requirements and does not meet the last one. We here introduce a measure that meets the last requirement. To begin with, we define the following two quantities.

5.6.1 Probability of Concordance

The probability of concordance is defined as

$$\begin{aligned} P_c &= Pr\left\{[(X_i < X_j) \cap (Y_i < Y_j)] \cup [(X_i > X_j) \cap (Y_i > Y_j)]\right\} \\ &= Pr\left[(X_j - X_i)(Y_j - Y_i) > 0\right] \end{aligned}$$

where $\cup$ is a disjoint union.

5.6.2 Probability of Discordance

Similarly we define

$$P_d = Pr\left[(X_i - X_j)(Y_i - Y_j) < 0\right]$$

5.6.3 Kendall's Tau

Then the Kendall's τ is defined as

$$\tau = P_c - P_d$$

Does this new measure meet the 7 requirements? Yes. For the fourth criterion, when X and Y are independent, then

$$\begin{aligned} P_c &= Pr\left(X_i < X_j\right) Pr\left(Y_i < Y_j\right) + Pr\left(X_i > X_j\right) Pr\left(Y_i > Y_j\right) \\ &= Pr\left(X_i > X_j\right) Pr\left(Y_i < Y_j\right) + Pr\left(X_i < X_j\right) Pr\left(Y_i > Y_j\right) \\ &= P_d \end{aligned}$$

The other requirements are satisfied obviously.