MANA130083.01 Nonparametric *Spring 2025*

Lecture 4

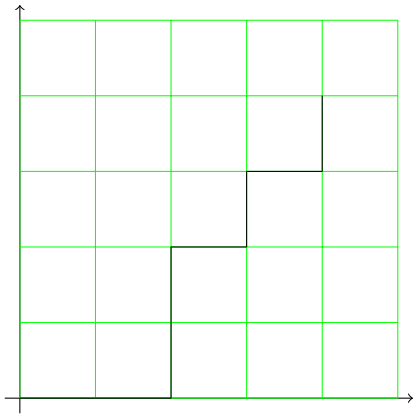
Instructor: Prof. Bowen Gang *Scribes: Yize Wang, Jingyi Zhou*

4.1 Review: Kolmogorov-Smirnov Two Sample Test

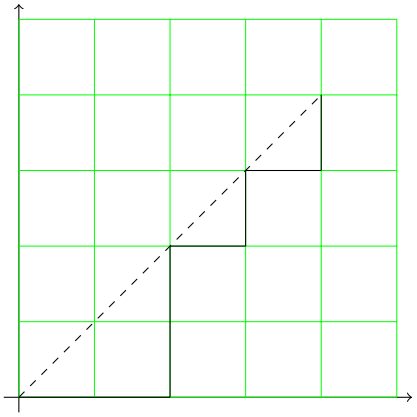
Consider the basic setup: we have observations $X_1, ..., X_m \sim F_X$ and $Y_1, ..., Y_n \sim F_Y$. The null hypothesis is $F_X = F_Y$. The K-S statistic is given by

$$D_{m,n} = \max_x |\hat{F}_X(x) - \hat{F}_Y(x)|$$

where we reject the null when $D_{m,n} > d$ for some d. In order to calculate the p-value $\Pr(D_{m,n} > d|H_0)$, we can either do simulation or use the following formula. To derive the exact formula, we do the following process. We combine the two series of observations together and rank them from smallest to largest. This sequence corresponds to a 'path' from the grid (0,0) to (m,n). Say the sequence is $X_{(1)}, X_{(2)}, Y_{(1)}, Y_{(2)}, X_{(3)}, Y_{(3)}, X_{(4)}, Y_{(4)}$. We may draw the following path, where X moves the point to rightward and Y moves the point upward.



Then all the observed values of $m\hat{F}_X(x)$ and $n\hat{F}_Y(x)$ are respectively the coordinates of all points (u, v) on the path. There are $\binom{m+n}{n}$ possible paths in total. We are actually trying to find $\max|\frac{u}{m} - \frac{v}{n}| = \frac{|nu-mv|}{mn}$, which is equivalent to $\max|\hat{F}_X(x) - \hat{F}_Y(x)|$. We then link the origin point and (m, n) in the graph:



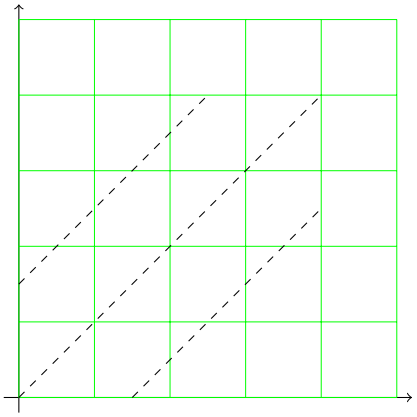
where the dashed line is given by $nx - my = 0$. Further, the vertical distance between (u, v) and $nx - my = 0$ is given by

$$\left| v - \frac{nu}{m} \right|$$

Let $\frac{|nu-mv|}{mn} = d$, it is obvious that

$$\left| v - \frac{nu}{m} \right| = nd$$

Then the p-value is equivalent to calculate the number of paths whose distances from the dashed line are less than or equal to nd. The total number of path is given by $\binom{m+n}{n}$. Again we want to compute the number of paths which have points at a distance from the diagonal not less than nd. Let $A(m, n)$ be the number of paths from $(0, 0)$ to (m, n) which lie entirely within the boundary lines, where the distances between each boundary line and the dashed line is nd.



Then

$$Pr\left(D_{m,n} > d | H_0\right) = 1 - \frac{A(m, n)}{\binom{m+n}{n}}$$

Note that we have the recursive formula

$$A(u, v) = A(u - 1, v) + A(u, v - 1)$$

where the initial condition is given by:

$$A(0, v) = A(u, 0) = 1$$

Then the exact distribution of K-S two sample test is solved.

4.2 The Median Test

The basic setting is still $X_1, ..., X_m \sim F_X$ and $Y_1, ..., Y_n \sim F_Y$. We are interested in the question that $F_X = F_Y$. The naive idea is that under this null, the two samples should have similar distributions. To be more specific, let

$$P_X = Pr\left(X \leqslant \delta\right) \; and \; P_Y = Pr\left(Y \leqslant \delta\right)$$

then under H_0 we have $P_X = P_Y$. Also let u, v be the respective number of X and Y observations less than δ. If u and v are far away from each other, we tend to reject the null. The joint distribution of U, V is given by

$$f_{U,V}(u,v) = \binom{m}{u} P_X^u \left(1 - P_X\right)^{m-u} \cdot \binom{n}{v} P_Y^v \left(1 - P_Y\right)^{n-v}$$

Under H_0, $P_X = P_Y = p$. We can estimate p by calculate $\frac{u+v}{m+n}$. If you take δ to be the median of the combined sample, we tend to reject the null when u differs significantly from $\frac{m}{2}$. To be more specific, let $N = m + n$ and the median t is given by

$$t = \begin{cases} \frac{N}{2}, & if \; N \; even \\ \frac{N+1}{2}, & if \; N \; odd \end{cases}$$

The test statistic is then given by

$$Z = \frac{U - \frac{mt}{N}}{\left(mnt(N-t)/N^3\right)^{\frac{1}{2}}} \xrightarrow{H_0} N(0,1)$$

In order to show the result explicitly, we do the following process:

$$\begin{aligned} z &= \frac{Nu - mt}{\sqrt{mnt(N-t)/N}} \\ &= \frac{nu - m(t-u)}{\sqrt{mnN\left(\frac{t}{N}\right)\left(1 - \frac{t}{N}\right)}} \\ &= \frac{\frac{u}{m} - \frac{v}{n}}{\sqrt{\left(\frac{u+v}{N}\right)\left(1 - \frac{u+v}{N}\right)\frac{N}{mn}}} \\ &\approx \frac{\frac{u}{m} - \frac{v}{n}}{\sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{m} + \frac{1}{n}\right)}} \to N(0,1) \end{aligned}$$

Note that there is an extra incremental randomness here: we do not know the median at first. We estimate the median by using pooled sample median.

4.3 The Control Median Test

The basic setup is still based on $X_1, ..., X_m \sim F_X$ and $Y_1, ..., Y_n \sim F_Y$ and we want to test $H_0 : F_X = F_Y$. Here we let

$$V = \#\; of \; X \; less \; than \; the \; median \; of \; Y$$

where we tend to reject the null if v differs from $\frac{m}{2}$ too much. For simplicity, we assume n is odd: $n = 2r + 1$ then

$$Pr\left[V = j | H_0\right] = \frac{\binom{m+r-j}{m-j}\binom{j+r}{j}}{\binom{m+2r+1}{m}}$$

In practice, we use the following approximation instead:

$$z = \frac{v - \frac{m}{2}}{\sqrt{m(m+n)/4n}} \xrightarrow{D} N(0,1)$$

4.4 The Mann Whitney U Test

Let $U = \#$ of times a Y proceeds an X in the combined ordered arrangement of two independent random samples $X_1, ..., X_m$ and $Y_1, ..., Y_n$. To be more specific, let

$$D_{ij} = \begin{cases} 1, \ if \ Y_j < X_i \\ 0, \ if \ Y_j > X_i \end{cases}$$

then the test statistic is given by

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

For example, the ordered sample is

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

Here are two X's that proceed $Y_{(1)}$ and $Y_{(2)}$ and one X that proceeds $Y_{(3)}$. So $U = 2 + 2 + 1 = 5$. You can imagine that if X and Y has different distributions, say $F_X > F_Y$ or $F_Y < F_X$, U will be either too small or too large:

$$X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}, Y_{(1)}, Y_{(2)}, Y_{(3)} \quad \text{(U=0)}$$
$$Y_{(1)}, Y_{(2)}, Y_{(3)}, X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)} \quad \text{(U=12)}$$

So what is the null distribution of U? Under null hypothesis,

$$f_U(u) = Pr\,(U = u) = \frac{r_{m,n}(u)}{\binom{m+n}{n}}$$

where $r_{m,n}(u) = \#$ of distinguishable arrangements such that in each sequence the $\#$ of times a Y proceeds an X is exactly u. We can calculate $r_{m,n}(u)$ recursively:

$$r_{m,n}(u) = r_{m,n-1}(u) + r_{m-1,n}(u - n)$$

where the initial condition is given by

$$\begin{cases} r_{i,j}(u) = 0, & \forall u < 0 \\ r_{i,0}(0) = 1, & \\ r_{0,i}(0) = 1, & \\ r_{i,0}(u) = 0, & \forall u \neq 0 \\ r_{0,i}(u) = 0, & \forall u \neq 0 \end{cases}$$

Now we have solved the exact distribution and we will derive the asymptotic distribution as follows. Let

$$p = Pr\,(Y < X)$$

then

$$\begin{aligned}
\mathbb{E}[D_{ij}] &= p = \mathbb{E}\left[D_{ij}^2\right] \\
Var[D_{ij}] &= p(1-p) \\
Cov[D_{ij}, D_{hk}] &= 0 \; for \; i \neq h, j \neq k \\
Cov[D_{ij}, D_{ik}] &= \underbrace{\mathbb{E}[D_{ij}D_{ik}]}_{0-1 \; variable} - \underbrace{\mathbb{E}[D_{ij}]\,\mathbb{E}[D_{ik}]}_{p^2}
\end{aligned}$$

where we can also let

$$p_1 = \mathbb{E}[D_{ij}D_{ik}] = Pr(Y_j < X_i \cap Y_k < X_i)$$

Similarly

$$Cov[D_{ij}, D_{hj}] = p_2 - p^2$$

where

$$p_2 = Pr(X_i > Y_j \cap X_h > Y_j) = \mathbb{E}[D_{ij}D_{hj}]$$

In order to calculate p_1 and p_2, we can do the following integration:

$$\begin{aligned}
p_1 &= Pr(Y_j < X_i \; and \; Y_k < X_i) \\
&= Pr(Y_j \text{ and } Y_k < X_i) \\
&= \mathbb{E}_{X_i} \Pr(Y_j \text{ and } Y_k < X_i | X_i) \\
&= \mathbb{E}_{X_i}[F(X_i)]^2 \\
&= \int_{-\infty}^{\infty} [F_Y(x)]^2 \underbrace{dF_X(x)}_{f_X(x)dx}
\end{aligned}$$

Under the null, $F_Y(x) = F_X(x)$ thus

$$\begin{aligned}
\int_{-\infty}^{\infty} [F_Y(x)]^2 \, dF_X(x) &= \int_{-\infty}^{\infty} [F_X(x)]^2 \, dF_X(x) \\
&= \frac{1}{3}
\end{aligned}$$

Similarly

$$p_2 \overset{H_0}{=} \frac{1}{3}$$

Finally we have

$$\begin{aligned}
\mathbb{E}[U] &= \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbb{E}[D_{ij}] = mnp \\
Var[U] &= \sum\sum Var[D_{ij}] + \sum\sum\sum\sum Cov[D_{ij}, D_{hk}] \to mn[p - p^2(N-1) + (n-1)\frac{1}{3} + (m-1)\frac{1}{3}] \\
&= \frac{mn(N+1)}{12} \quad (\text{under } H_0, p = \frac{1}{2})
\end{aligned}$$

where $N = m + n$. And the normalized distribution is asymptotically normal:

$$Z = \frac{U - \frac{mn}{2}}{\sqrt{\frac{mn(N+1)}{12}}} \to N(0,1)$$

If we have ties in our observations, say, $X_i = Y_j$, we can think about using a modified test:

$$D_{ij}^{\star} = \begin{cases} 1, & if\ X_i > Y_j \\ 0, & if\ X_i = Y_j \\ -1, & if\ X_i < Y_j \end{cases}$$

and similarly

$$U = \sum \sum D_{ij}^{\star}$$

4.5 Linear Rank Statistics

4.5.1 Definition

The basic setup is still: $X_1, ..., X_m \sim F_X$ and $Y_1, ..., Y_n \sim F_Y$. The null hypothesis is $F_X = F_Y$. For linear rank statistic, we need to construct vector z, where

$$z = (z_1, z_2, ..., z_N), N = m + n$$

For each element of z,

$$z_i = \begin{cases} 1, & if\ i^{th}\ sample\ is X \\ 0, & if\ i^{th}\ sample\ is Y \end{cases}$$

where we actually mean combined ordered sample we we say 'sample' here. For example, if

$$(X_1, X_2, X_3, X_4) = (2, 9, 3, 4)$$
$$(Y_1, Y_2, Y_3) = (1, 6, 10)$$

then the combined ordered sample is given by

$$1, 2, 3, 4, 6, 9, 10$$
$$Y_1, X_1, X_3, X_4, Y_2, X_2, Y_3$$

and the corresponding z is

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

Linear rank statistic is a statistic of the form

$$T_N(z) = \sum_{i=1}^{N} a_i z_i$$

4.5.2 Properties of Linear Rank Statistic

Theorem. *Under $H_0 : F_X = F_Y$, we have*

$$\mathbb{E}[z_i] = \frac{m}{N},\ Var[z_i] = \frac{mn}{N^2}$$
$$cov[z_i, z_j] = \frac{-mn}{N^2(N-1)}$$

The proof is immediate from $z_i \sim Ber\left(\frac{m}{N}\right)$. For the covariance,

$$\begin{aligned} cov\left[z_i, z_j\right] &= \mathbb{E}\left[z_i z_j\right] - \mathbb{E}\left[z_i\right]\mathbb{E}\left[z_j\right] \\ &= Pr\left(z_i = 1 \& z_j = 1\right) - \left(\frac{m}{N}\right)^2 \\ &= \frac{\binom{m}{2}}{\binom{N}{2}} - \left(\frac{m}{N}\right)^2 \\ &= \frac{m(m-1)}{N(N-1)} - \left(\frac{m}{N}\right)^2 \\ &= \frac{-mn}{N^2(N-1)} \end{aligned}$$

Theorem. *Under H_0,*

$$\begin{aligned} \mathbb{E}\left[T_N\right] &= m\sum \frac{a_i}{N} \\ Var\left[T_N\right] &= \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N} a_i^2 - \left(\sum_{i=1}^{N} a_i\right)^2\right] \end{aligned}$$

The proof is immediate from

$$\begin{aligned} \mathbb{E}\left[T_N\right] &= \mathbb{E}\left[z_i\right]\sum a_i \\ Var\left[T_N\right] &= Var\left(\sum a_i z_i\right) = \sum a_i^2 Var\left[z_i\right] + \sum\sum_{i\neq j} a_i a_j cov\left[z_i, z_j\right] \end{aligned}$$

Theorem. *If $B_N = \sum_{i=1}^{N}$, $T_N = \sum_{i=1}^{N} a_i z_i$, then*

$$cov\left(B_N, T_N\right) = \frac{mn}{N^2(N-1)}\left[N\left(\sum a_i b_i\right) - \left(\sum a_i\right)\left(\sum b_i\right)\right]$$

The proof is also immediate because both B_N and T_N are linear rank statistics, which can be seen as $\vec{a}\cdot\vec{z}$.

Definition. *$T_N(z)$ is symmetric if*

$$Pr\left[T_N(z) - \mu = k\right] = Pr\left[T_N(z) - \mu = -k\right]$$

where $\mu = \mathbb{E}\left[T_N(z)\right]$. This definition is actually: the p.m.f. is symmetric.

The question is, how can we tell a linear rank statistic is symmetric? Suppose we observe that for every z, a conjugate z' exists s.t. whenever $T_N(z) = \mu + k$, we have $T_N(z') = \mu - k$. Then the frequency of $T_N(z) = \mu + k$ is the same as that of $\mu - k$ (i.e., $T_N(z')$). So the distribution is symmetric. The condition for symmetry is then $T_N(z) + T_N(z') = 2\mu$. Similarly we have the following theorem.

Theorem. *The null distribution of $T_N(z)$ is symmetric about μ whenever $a_i + a_{N-i+1} = c$ for $i = 1, 2, ..., N$, where a_i can be seen as a weight of z_i.*

The proof is as follows: given $z = (z_1, z_2, ..., z_N)$, let $z' = (z'_1, z'_2, ..., z'_N)$, where $z'_i = z_{N-i+1}$. Then you can

verify that

$$\begin{aligned} T_N(z) + T_N(z') &= \sum a_i z_i + \sum a_i z_{N-i+1} \\ &= \sum_{i=1}^{N} a_i z_i + \sum_{j=1}^{N} a_{N-j+1} z_j \\ &= \sum_{j=1}^{N} (a_i + a_{N-j+1}) z_j \\ &= c \sum z_j = cm \ (constant) \end{aligned}$$

Theorem. *The null distribution of $T_N(z)$ is symmetric about its mean for any wet of weight if $m = n = \frac{N}{2}$.*

The proof is also to find a conjugate. Define conjugate z' with $z'_i = 1 - z_i$. Then

$$\begin{aligned} T_N(z) + T_N(z') &= \sum a_i z_i + \sum a_i (1 - z_i) \\ &= \sum a_i z_i + \sum a_i - \sum a_i z_i = \sum a_i \ (constant) \end{aligned}$$

Theorem. *The null distribution of $T_N(z)$ is symmetric about its mean if N is even and the weights are $a_i = i$ for $i \leqslant \frac{N}{2}$ and $a_i = N - i + 1$ for $i > \frac{N}{2}$.*

That is to say, $a = (1, 2, ..., \frac{N}{2}, \frac{N}{2}, \frac{N}{2} - 1, \frac{N}{2} - 2, ..., 3, 2, 1)$. The proof is similar as previous ones. In practice, we also use the following asymptotic distribution a lot.

$$\frac{T_N(z) - \mathbb{E}[T_N(z)]}{\sqrt{Var[T_N(z)]}} \to N(0, 1)$$

4.5.3 Linear Rank Tests for the Location Problem

For this part, the null is given by $F_X = F_Y$ and the alternative hypothesis is given by $F_Y(x) = F_X(x - \theta)$ where $\theta \neq 0$. This is a shift of location and that is why it is called as a location problem/test. In parametric context, if F is normal, we use t-test. In non-parametric case, we can use the Wilcoxon rank sum test. The test statistic is given by

$$W_N = \sum_{i=1}^{N} i z_i$$

We will reject the null if W_N is too big or too small. This is a type of linear rank statistic where the weight is $a_i = i$. It is symmetric because here we have $a_i = a_{N-i+1} = c$. We also know its asymptotic distribution.

4.5.3.1 The Exact Distribution of Wilcoxon Rank Sum Statistic

Let $r_{m,n}(k) = \#$ of arrangements of $m * X$ and $n * Y$ s.t. the sum of X rank is k. Then

$$r_{m,n}(k) = \underbrace{r_{m-1,n}(k - N)}_{N^{th} \ position \ is \ X} + \underbrace{r_{m,n-1}(k)}_{N^{th} \ position \ is \ Y}$$

This recursive process is quite similar to that of Mann Whitney U test ($r_{m,n}(u) = r_{m,n-1}(u) + r_{m-1,n}(u-n)$. You can actually prove that they are equivalent. Recall the Mann Whitney U statistics is given by $U = \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij}$, which is equal to

$$\sum_{i=1}^{m} \underbrace{(D_{i1} + D_{i2} + \cdots + D_{in})}_{\#\ of\ values\ of\ j\ for\ which\ Y_j < X_i}$$

which can also be considered as the rank of X_i reduced by n_i where $n_i = \#$ of $X's \leqslant X_i$. Then

$$\begin{aligned}
U &= \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij} \\
&= \sum_{i=1}^{m} (r(X_i) - n_i) \\
&= \sum_{i=1}^{m} r\left(X_i\right) - \underbrace{(n_1 + n_2 + \cdots + n_m)}_{permutation\ of\ 1,2,\ldots,m} \\
&= \sum_{i=1}^{m} r\left(X_i\right) - (1 + 2 + \cdots + n_m) \\
&= \sum_{i=1}^{m} \underbrace{r\left(X_i\right)}_{\sum iz_i} - \underbrace{\frac{m}{2}(m+1)}_{constant}
\end{aligned}$$

So the Mann Whitney U statistic and the Wilcoxon rank sum test only differ by a constant.

4.5.3.2 Terry-Heffding (Normal Scores) Test

Here the test statistic is given by

$$\sum_{i=1}^{N} \mathbb{E}\left[\xi_{(i)}\right] z_i$$

where ξ_i is the i^{th} order statistic of standard normal distribution. If F_X and F_Y happen to follow normal distribution, $H_0 : F_X = F_Y \sim N(\mu, \sigma^2)$ and $H_a : F_X \sim N(\mu_1, \sigma^2), F_Y \sim N(\mu_2, \sigma^2)$, then the Terry-Hoeffding is asymptotically optimal, which means it will have higher power than other rank tests. This is because $\mathbb{E}\left[\xi_{(i)}\right]$ is "more representative" of the raw data than the rank. To be more specific,

$$corr\left(\mathbb{E}\left[\xi_{(i)}\right], X_i\right) \geqslant corr\left(i, X_i\right)$$

The pitfall of Terry-Hoeffding is that it is difficult to compute $\mathbb{E}\left[\xi_{(i)}\right]$. Then we can replace the test statistic by

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

where $\Phi(\cdot)$ is the CDF for $N(0,1)$ and $\frac{i}{N+1}$ can be seen as a quantile. Then $\Phi^{-1}\left(\frac{i}{N+1}\right)$ can be viewed as an approximate of $\mathbb{E}\left[\xi_{(i)}\right]$.