

Lecture 56: Mann–Whitney–Wilcoxon test: null distribution

Statistical Inference–II ■ Units IV–V: Nonparametric Inference

Kapil Kumar

Department of Statistics
Chaudhary Charan Singh University, Meerut

M.Sc. Statistics, Semester III
Academic Session 2026–2027

This lecture

Previous lecture

Lecture 55: Paired signed-rank test with ties; Hodges–Lehmann estimator.

Today

- Mann–Whitney and Wilcoxon Rank-Sum Tests
 - Rank-Sum Statistic

Two independent samples

Let $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ be independent samples from continuous distributions F and G .

The Mann–Whitney test and Wilcoxon rank-sum test are equivalent forms of the same two-sample rank procedure (Mann and Whitney, 1947; Wilcoxon, 1945).

Pool the $N = m + n$ observations and assign ranks $1, \dots, N$. Let R_X and R_Y be the sums of the ranks belonging to the X and Y samples. Define

$$U_X = R_X - \frac{m(m+1)}{2}, \quad U_Y = R_Y - \frac{n(n+1)}{2}.$$

Equivalently,

$$U_X = \#\{(i, j) : X_i > Y_j\}, \quad U_Y = \#\{(i, j) : Y_j > X_i\},$$

with half weight commonly assigned to ties.

Why U_X and U_Y count pairs

To see this when there are no ties, note that the pooled rank of X_i is

$$\#\{k : X_k \leq X_i\} + \#\{j : Y_j < X_i\}.$$

Summed over i , the first counts give $1 + 2 + \cdots + m = m(m+1)/2$, and the second counts give $\sum_i \sum_j \mathbf{1}(X_i > Y_j)$. Hence

$$R_X = m(m+1)/2 + \sum_{i,j} \mathbf{1}(X_i > Y_j);$$

exchanging the roles of the two samples gives

$$R_Y = n(n+1)/2 + \sum_{i,j} \mathbf{1}(Y_j > X_i).$$

The relation $U_X + U_Y = mn$

Without ties each of the mn pairs (X_i, Y_j) satisfies exactly one of $X_i > Y_j$ and $Y_j > X_i$, so

$$U_X + U_Y = mn, \quad \text{i.e.} \quad U_Y = mn - U_X$$

(Exercise 3.21). Thus

$$\frac{U_X}{mn}$$

estimates a probability of stochastic superiority, $\mathbb{P}(X > Y)$.

Theorem 56.1 (Null distribution of the Mann–Whitney statistic)

Under $H_0 : F = G$ with F continuous, all $\binom{N}{m}$ allocations of m of the ranks $1, \dots, N$ to the X sample are equally likely. Consequently:

- (i) $\mathbb{E}(U_X) = \frac{mn}{2}$ and $\text{Var}(U_X) = \frac{mn(N+1)}{12}$;
- (ii) *the distribution of U_X is symmetric about $mn/2$;*
- (iii) *writing $p_{m,n}(u) = \mathbb{P}(U_X = u)$ for sample sizes m, n ,*

$$p_{m,n}(u) = \frac{m}{N} p_{m-1,n}(u-n) + \frac{n}{N} p_{m,n-1}(u),$$

with $p_{0,n}(0) = p_{m,0}(0) = 1$, which allows exact tables to be built.

Proof.

Under H_0 the N pooled observations are i.i.d. and continuous, so every assignment of the ranks to observations is equally likely. Hence the set of ranks of the X 's is a simple random sample of size m , drawn without replacement, from the finite population $\{1, 2, \dots, N\}$.

(i) This population has mean $(N + 1)/2$ and variance

$$\sigma_P^2 = \frac{1}{N} \sum_{r=1}^N r^2 - \left(\frac{N+1}{2} \right)^2 = \frac{N^2 - 1}{12}.$$

Proof of Theorem: Null distribution of the Mann–Whitney statistic (cont.)

Proof (cont.)

By the standard results for sampling without replacement,

$$\mathbb{E}(R_X) = \frac{m(N+1)}{2},$$

$$\text{Var}(R_X) = m \sigma_P^2 \frac{N-m}{N-1} = m \frac{(N-1)(N+1)}{12} \cdot \frac{n}{N-1} = \frac{mn(N+1)}{12}.$$

Since $U_X = R_X - m(m+1)/2$,

$$\mathbb{E}(U_X) = \frac{1}{2}m(N+1) - \frac{1}{2}m(m+1) = \frac{1}{2}m(N-m) = mn/2,$$

and $\text{Var}(U_X) = \text{Var}(R_X)$.

Proof of Theorem: Null distribution of the Mann–Whitney statistic (cont.)

Proof (cont.)

(ii) Replacing every rank r by $N + 1 - r$ is a one-to-one map of the set of equally likely allocations onto itself. It turns every pair with $X_i > Y_j$ into one with $X_i < Y_j$ and vice versa, so it maps U_X to $mn - U_X$. Hence U_X and $mn - U_X$ have the same distribution.

(iii) Condition on the label of the largest pooled observation. With probability m/N it is an X ; it then exceeds all n Y 's, contributing n to U_X , and the remaining $m - 1$ and n observations again form a null configuration. With probability n/N it is a Y , which is exceeded by no X and contributes nothing, leaving a null configuration of sizes m and $n - 1$. □

Remark 56.2 (Direct derivation of the variance)

Writing $U_X = \sum_{i=1}^m \sum_{j=1}^n I_{ij}$ with $I_{ij} = \mathbf{1}(X_i > Y_j)$, under H_0 we have $\mathbb{E}(I_{ij}) = \frac{1}{2}$ and $\text{Var}(I_{ij}) = \frac{1}{4}$. Two indicators sharing the same X_i satisfy

$$\mathbb{E}(I_{ij}I_{ij'}) = \mathbb{P}(X_i > \max(Y_j, Y_{j'})) = \frac{1}{3},$$

so their covariance is $\frac{1}{3} - \frac{1}{4} = \frac{1}{12}$; similarly for two indicators sharing the same Y_j ; indicators with no observation in common are independent.

Hence

$$\text{Var}(U_X) = \frac{mn}{4} + \frac{mn(n-1)}{12} + \frac{mn(m-1)}{12} = \frac{mn(m+n+1)}{12},$$

in agreement with the Theorem on the null distribution of the Mann–Whitney statistic.

Practice (Lecture Notes, Chapter 3)

- Short-answer questions (Section 3.14): Q10, Q11.
- Exercises (Section 3.15): 3.20, 3.21.

Next lecture

Lecture 57: Mann–Whitney test with ties; normal approximation; Hodges–Lehmann shift estimator.