# STAT 135 Lecture 15

Henry Liev

29 July 2025

**Method 0.1**

$X_1, \ldots, X_n \overset{iid}{\sim} F$
$Y_1, \ldots, Y_m \overset{iid}{\sim} G$
$H_0 : F = G$
$H_1 : F \neq G$
"Treatment does absolutely nothing"
How it works:

(a) Group all the $m + n$ observations together

(b) Rank them by increasing size

(c) Compute sum of ranks for treatment

(d) Reject for extreme values of sum

Under $H_0$: Ranks are random numbers

**Example 0.2**

Two independent samples in a cholesterol-lowering treatment experiment.

| Control | Treatment |
|---|---|
| 103 (2) | 100 (1) |
| 115 (4) | 105 (3) |
| 123 (6) | 121 (5) |
| 131 (7) | |
| sum: 19 | 9 |

**Remark 0.3** (Mann-Whitney)

- For small samples, can compute exactly
- For large samples, there is a normal approximation
  - Approximating a known distribution
  - Safer, more informed use of central limit theorem

**Remark 0.4** (Moments of Mann-Whitney)

$T_Y$ = Sum of ranks of "treatment"
$\mathbb{E}(T_Y)$
$\mathrm{Var}(T_Y)$
For $k = 1, \ldots, m+n$ m are chosen to go into $Y_i$'s (treatment) and n are in $X_i$'s (control)
Let $I_k = \begin{cases} 1 & \text{if } k \text{ is selected in } Y_i\text{'s} \\ 0 & \text{otherwise} \end{cases}$

$$\sum_{k=1}^{m+n} \mathbb{E}(I_k) = m + n\mathbb{E}(I_k) = m$$

$P(I_k \in Y) = \mathbb{E}(I_k) = \frac{m}{m+n}$

$$T_Y = \sum_{k=1}^{m+n} kI_k$$

$$\mathbb{E}(T_Y) = \frac{m}{m+n}\sum_{k=1}^{m+n} k = \frac{m}{m+n}\frac{(m+n)(m+n+1)}{2} = \frac{m(m+n+1)}{2}$$

$$\mathrm{Var}(T_Y) = \mathbb{E}(T_Y^2) - \mathbb{E}^2(T_Y) =$$

$$\mathbb{E}(T_Y^2) = \mathbb{E}\left[\left(\sum_{k=1}^{m+n} k(I_k)\right)^2\right] = \sum_{k=1}^{m+n} k^2\mathbb{E}(I_k^2) + \sum_{i\neq j} ijE(I_jI_i)$$

$$= \frac{m}{m+n}\sum_{k=1}^{m+n} k^2 + \frac{m(m-1)}{(n+m)(n+m-1)}\sum_{i\neq j} ij$$

$$\frac{m}{m+n}\sum_{k=1}^{m+n} k^2 + \frac{m(m-1)}{(n+m)(n+m-1)}\left[\left(\sum_{k=1}^{m+n} k\right)^2 - \sum_{k=1}^{m+n} k^2\right]$$

**Remark 0.5** (Another angle on Mann-Whitney: $U_Y$)

Same setup
$X_1, \ldots, X_n \overset{iid}{\sim} F$
$Y_1, \ldots, Y_m \overset{iid}{\sim} G$
$\pi = P(X < Y)$
$H_0 : \pi = \frac{1}{2}$
$H_1 : \pi \neq \frac{1}{2}$

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

$$U_Y = \hat{\pi} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} 1\{X_i < Y_j\}$$

Define $X_{(i)} = i^{\text{th}}$ smallest of the X's
Define $Y_{(j)} = j^{\text{th}}$ smallest of the Y's
$V_{ij} = P(X_{(i)} < Y_{(j)})$

$$mn\hat{\pi} = \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij}$$

$$\sum_{i=1}^{n} (\# X_j \leq Y_{(i)}) = \sum_{i=1}^{n} \text{Rank}(Y_{(i)}) - i = \sum_{i=1}^{n} \text{Rank}(Y_{(i)}) - \sum_{i=1}^{n} i$$

$U_Y$ is equivalent to $T_Y$

**Method 0.6** (Signed Rank Test)

Nonparametric test for paired data
Have paired observation $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ Pairs are independent of each other
Interested in $D_i = Y_i - X_i, i = 1, \ldots, n$
$H_0 : D$ is symmetric around 0
$H_1 :$ Not symmetric around 0
How to do it:

1. Rank absolute differences $|D_i|$
2. Compute $W_t$ sum of ranks of positive $D$'s
3. Reject for extreme values of $W_t$

**Example 0.7** (Test Prep)

| Before ($X_i$) | After ($Y_i$) | Diff | —Diff— | Rank | Signed Rank |
|---|---|---|---|---|---|
| 25 | 27 | 2 | 2 | 2 | 2 |
| 29 | 25 | -4 | 4 | 3 | -3 |
| 60 | 59 | -1 | 1 | 1 | -1 |
| 27 | 37 | 10 | 10 | 4 | 4 |

Under $H_0$ of symmetric $D$'s, signs of ranks are coin flips
Can compute exact distribution of sum of positive ranks (Approximately Normal)
$W_t = \sum_{k=1}^{n} kI_k \ I_k = \begin{cases} 1 & \text{if } k^{\text{th}} \text{ largest D is positive} \\ 0 & \text{otherwise} \end{cases}$
$I_k$ are coinflips, Bernoulli $(\frac{1}{2})$
$\mathbb{E}(W_t) = \frac{1}{2}\sum_{k=1}^{n} k = \frac{n}{2}$
$\text{Var}(W_t) = \sum_{k=1}^{n} \text{Var}(kI_k) = \sum_{k=1}^{n} k^2 \text{Var}(I_k) = \frac{n(n+1)(2n+1)}{24}$

Takeaways:
For both independent and paired problems have non parametric options with very weak assumptions

- Easy to carry out exactly or with good normal approximation
  - Resistant to outliers
  - Pretty efficient