Tutorial 3

Yue Ju, Yanmeng Wang

School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen

September 26, 2020

Outline

Two-sample Location Problem

- **Two-sample data**:
  $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ from two independent subjects (different from paired data)
- **Main problems**:
  (1) Is there a significant difference between the distributions of $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$?
  (2) What is the difference?

Two-sample Location Problem

Basic Assumptions

1. $X_1, \cdots, X_m$ are i.i.d. with common cdf F; $Y_1, \cdots, Y_n$ are i.i.d. with common cdf G.
2. $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ are mutually independent.
3. $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ are continuous random variables.

- **Location-shift Model**:

$$G(t) = F(t - \Delta) \text{ for all } t \in \mathbb{R}$$
$$\Longleftrightarrow Y \sim X + \Delta \text{ (Not } Y = X + \Delta\text{)},$$

  where Δ is known as $\boxed{\text{location shift}}$ or $\boxed{\text{treatment effect}}$.
- $\Delta = 0$ represents no difference in treatment effects between X and Y; $\Delta > (<) 0$ represents a greater(smaller) effect of Y and X in the sense of stochastic order.

Wilcoxon rank sum test

- **Null hypothesis:** $H_0 : \Delta = 0$
- **Y-Ranks:**
  Order $N = n + m$ observations $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ in ascending order. S_j denotes the rank of Y_j, $j = 1, \cdots, n$. $S_1, \cdots, S_n$ are referred as Y-ranks.
- **Test statistic:** $W = \sum_{j=1}^{n} S_j$ (the sum of Y-ranks)
- **Exact distribution of W under H_0:**

$$\Pr(W = w) = \frac{No.\ of (s_1, \cdots, s_n) : s_1 + \cdots + s_n = w}{\begin{pmatrix} N \\ n \end{pmatrix}},$$

where $M_1 \leq w \leq M_2$ with $M_1 = \frac{n(n+1)}{2}$ and $M_2 = mn + \frac{n(n+1)}{2}$.

Wilcoxon rank sum test

- **Mean and variance of W under H_0:**

$$\mathrm{E}_0[W] = \frac{n(m+n+1)}{2}$$
$$\mathrm{Var}_0[W] = \frac{mn(m+n+1)}{12}.$$

- **Symmetry of W:**
  W is symmetric about $\mathrm{E}_0[W]$, which is also the median of W under H_0.

- **Rejection rule:**
  Let $\Pr(W \geq w_\alpha) = \alpha$ under H_0. The Wilcoxon rank sum test rejects $H_0 : \Delta = 0$ at the α level if
  - $W \geq w_\alpha$ against $H_1 : \Delta > 0$
  - $W \leq n(m+n+1) - w_\alpha$ against $H_1 : \Delta < 0$
  - either $W \geq w_{\alpha/2}$ or $W \leq n(m+n+1) - w_{\alpha/2}$ against $H_1 : \Delta \neq 0$

Wilcoxon rank sum test

- **Asymptotic distribution of W under H_0:**

$$W^* = \frac{W - \mathrm{E}_0[W]}{\sqrt{\mathrm{Var}_0[W]}} = \frac{W - n(m+n+1)/2}{\sqrt{mn(m+n+1)/12}} \sim \mathcal{N}(0,1)$$

- **Approximate rejection rule:**
  Reject $H_0 : \Delta = 0$ at the α level if
  - $W^* \geq z_\alpha$ against $H_1 : \Delta > 0$
  - $W^* \leq -z_\alpha$ against $H_1 : \Delta < 0$
  - $|W^*| \geq z_{\alpha/2}$ against $H_1 : \Delta \neq 0$
- **Ties:**
  Assign the average rank to tied values.
  $\mathrm{E}_0[W]$ is unchanged, while the variance is reduced to

$$\mathrm{Var}_0[W] = \frac{mn(m+n+1)}{12} - \frac{mn}{12N(N-1)} \sum_{j=1}^{g} t_j(t_j-1)(t_j+1),$$

where g is the number of groups with tied ranks, t_j is the number of tied points in jth group.

- **Equivalent test statistic: the Mann-Whitney statistic**
  1. No ties:

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} I_{\{X_i<Y_j\}} = W - \frac{n(n+1)}{2}$$

$$\mathrm{E}_0[U] = \frac{mn}{2}$$

$$\mathrm{Var}_0[U] = \frac{mn(m+n+1)}{12}$$

  2. Ties occur among $X_1, \cdots, X_m, Y_1, \cdots, Y_n$:

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(I_{\{X_i<Y_j\}} + \frac{1}{2} I_{\{X_i=Y_j\}} \right) = W - \frac{n(n+1)}{2}$$

Note that ties within $X_1, \cdots, X_m$ or $Y_1, \cdots, Y_n$ do not affect the value of U, neither they affect the value of W (but affect their variances).

Wilcoxon rank sum test

- Estimation of the location shift

$$\hat{\Delta} = \text{median}\left\{ Y_j - X_i, \begin{array}{l} i = 1, \cdots, m \\ j = 1, \cdots, n \end{array} \right\}$$

$$= \begin{cases} U_{((mn+1)/2)} & \text{if } mn \text{ is odd} \\ \frac{U_{(mn/2)}+U_{(mn/2+1)}}{2} & \text{if } mn \text{ is even} \end{cases},$$

where $U_{(1)} \leq U_{(2)} \leq \cdots U_{(mn)}$ are ordered values of $(Y_j - X_i)$'s.

- A $100(1-\alpha)\%$ confidence interval for Δ is

$$(\Delta_L, \Delta_U) = (U_{(C_\alpha)}, U_{(mn+1-C_\alpha)}) = (U_{(C_\alpha)}, U_{(u_{\alpha/2})}),$$

Exact C_α:

$$C_\alpha = mn + 1 + \frac{n(n+1)}{2} - w_{\alpha/2} = mn + 1 - u_{\alpha/2}$$

For large m and n, the approximated C_α:

$$C_\alpha \approx \frac{mn}{2} - z_{\alpha/2}\sqrt{\frac{mn(m+n+1)}{12}}.$$

Wilcoxon rank sum test

- It is worth to note that the test statistic in R is the Mann-Whitney statistic U, not the Wilcoxon rank sum W.

Example 2. $m = 10$, $n = 5$:

```
x<-c(1.46, 0.80, 0.83, 1.64, 1.89, 1.04, 0.73, 1.91, 1.38, 1.45)
y<-c(0.88, 0.74, 1.15, 1.21, 0.90)

> wilcox.test(y, x, alternative = "less")

        Wilcoxon rank sum test

data:  y and x
W = 15, p-value = 0.1272
alternative hypothesis: true location shift is less than 0
```

$U = 15$, $W = 15 + 15 = 30$, $p\text{-value} = \Pr(U \le 15) = \Pr(W \le 30) = 0.1272$ for $H_1: \Delta < 0$

Question 1

The following two samples are extracted from a study:

$(X_1,\ldots,X_m) = (41.7,\ 35.4,\ 34.3,\ 32.4,\ 29.1,\ 27.3,\ 18.9,\ 6.6,\ 5.2)$

$(Y_1,\ldots,Y_n) = (100.0,\ 67.6,\ 65.9,\ 64.7,\ 39.6,\ 31.0)$

Assume the location-shift model for the two samples with location shift Δ.

(a) Calculate the approximate p-value of testing $H_0: \Delta = 0$ against $H_1: \Delta > 0$ by the Wilcoxon rank sum test, and explain its implication.

(b) Determine the exact p-value of the problem in part (a) by counting the number of $(b_1,\ldots,b_n)$ from the ranks of combined $X_1,\ldots,X_m,Y_1,\ldots,Y_n$ such that

$$b_1 + \cdots + b_n \geq w = \text{observed value of the Wilcoxon rank sum statistic } W$$

(or $b_1 + \cdots + b_n \leq 2\mathrm{E}_0[W] - w$ due to the symmetric distribution of W).

Compare the exact p-value with the approximate p-value obtained in part (a).

(c) Estimate the location-shift parameter Δ based on the differences between the two sample: $\{Y_j - X_i, i = 1,\ldots,m, j = 1,\ldots,n\}$.

(d) Find an approximate 95% confidence interval of Δ based on the Wilcoxon rank sum statistic.

Question1

The ordered values of $\{Y_j - X_i\}$ are shown in the following table,

$U_{(1)} \le U_{(2)} \le \cdots \le U_{(54)}$											
1	-10.7	10	7.2	19	25.9	28	33.5	37	45.8	46	62.4
2	-4.4	11	10.5	20	29.3	29	34.4	38	47.0	47	64.6
3	-3.3	12	12.1	21	30.4	30	35.2	39	48.7	48	65.7
4	-2.1	13	12.3	22	30.5	31	35.6	40	58.1	49	67.6
5	-1.4	14	20.7	23	31.6	32	36.8	41	58.3	50	70.9
6	1.9	15	23.0	24	32.2	33	37.4	42	59.3	51	72.7
7	3.7	16	24.2	25	32.3	34	38.5	43	59.5	52	81.1
8	4.2	17	24.4	26	33.0	35	38.6	44	60.7	53	93.4
9	5.3	18	25.8	27	33.3	36	40.3	45	61.0	54	94.8