Tutorial 2

Yue Ju, Yanmeng Wang

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

September 23, 2019

Outline

One-sample location problem

Statistical inference on one location parameter from **one sample** is referred to as the *one-sample location problem*.
(i) one sample data (location);
(ii) paired data (X_i, Y_i) (treatment effect).

Sign test and Wilcoxon signed rank test

There are two kinds of tests to test the hypothesis $H_0 : \theta = 0$:
(i) Sign test;
(ii) Wilcoxon signed rank test.
The sign test only utilizes the signs of the data, but not their values, hence it is considered as less efficient for underuse of information from the data.
The Wilcoxon signed rank test is more efficient when it also uses the magnitude of the data, but it has stricter assumptions on the data.

Sign test

Assumption

(i) $X_1, ..., X_n$ are mutually independent;
(ii) $X_1, ..., X_n$ are continuous with a **common median** θ (not necessarily have the same distribution)

Null hypothesis $H_0 : \theta = 0$

Test statistics

B (number of positive X_i's)$= \sum_{i=1}^{n} I_{\{X_i > 0\}} \sim$ Bin(n, 0.5) under H_0

Rejection rule

Reject H_0 against $H_1 : p > p_0$ at α-level (achievable) if $B \geq b_\alpha$, where Pr$(B \geq b_\alpha) = \alpha$ under H_0, $b_\alpha \in \{0, 1, 2, ...\}$.
Similar for left-sided and two sided test.

Normal approximation

$$B^* = \frac{B - 0.5n}{0.5\sqrt{n}} \sim Z \sim N(0, 1) \text{ approximately for large n}$$

Sign test

Remark

(i) If we observe zeros from $X_i's$, a sensible option is to discard those zero values and replace the sample size n by the number of nonzero observations.
(ii)If we wish to test $H_0 : \theta = \theta_0 \neq 0$, then we re-define the test statistic by

B (number of $X_i's$ larger than θ_0)= $\sum_{i=1}^{n} I_{\{X_i>\theta_0\}}$

(iii)Independence is assumed between the differences $Z_1, ..., Z_n$, but not needed within each pair (X_i, Y_i).

Sign test

Estimation of θ

Let $X_{(1)}, ..., X_{(n)}$ be the order statistics of $X_1, ..., X_n$. A nonparametric estimator for the median (treatment effect) θ is given by

$$\tilde{\theta} = \text{median}\{X_i, 1 \le i \le n\} = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd;} \\ \dfrac{X_{(n/2)} + X_{(n/2+1)}}{2} & \text{if is even.} \end{cases}$$

Confidence interval of θ

A 100(1-α)% confidence interval for θ is given by

$$(\theta_L, \theta_U) = (X_{(C_\alpha)}, X_{(n+1-C_\alpha)}) = (X_{(n+1-b_{\alpha/2})}, X_{(b_{\alpha/2})})$$

For large n, C_α can be approximated by

$$C_\alpha \approx \mathrm{E}_0[B] - z_{\alpha/2}\sqrt{\mathrm{Var}_0(B)} = 0.5n - z_{\alpha/2} 0.5\sqrt{n}$$

Wilcoxon signed rank test

Assumption

(i) $X_1, ..., X_n$ are mutually independent;
(ii) $X_1, ..., X_n$ are continuous and **symmetric** about a **common median** θ (not necessarily identical)

Null hypothesis $H_0 : \theta = 0$
Rank Assume no ties among $X_1, ..., X_n$. Let $|X|_{(1)}, ..., |X|_{(n)}$be ordered values of $|X_1|, ..., |X_n|$ .Define the rank R_i of X_i by $R_i = k$ if $|X_i| = |X|_{(k)}$.That is, the X_i with the k^{th} smallest absolute value has rank $R_i = k$.
Test statistics There are several equivalent forms of the Wilcoxon signed rank test statistic. We will consider the following form:

$$T^+ = \sum_{i=1}^{n} R_i \psi_i, \quad \text{where } \psi_i = I_{\{X_i > 0\}}, \ i = 1, \ldots, n.$$

Exact distribution of T^+

$$T^+ = \sum_{i=1}^{n} R_i \psi_i \sim \sum_{i=1}^{n} i\psi_i = \sum_{i=1}^{B} r_i$$

The range of T^+ is $\{$0,1,..., M $\}$ with M $= n(n+1)/2$
By equally likely outcomes in sample space Ω, the distribution of T^+ under H_0 is given by

$$\Pr(T^+ = t) = \frac{\text{Number of } \omega = (r_1, \ldots, r_B) : r_1 + \cdots + r_B = t}{2^n}$$

for t $\in \{0, 1, ..., M\}$, with $T^+ = 0$ if and only if B $=$ 0.

Exact distribution of T^+

Rejection rule: Let $M = n(n+1)/2$ and $\Pr(T^+ \geq t_\alpha) = \alpha$ with $t_\alpha \in \{0,1,\ldots,M\}$ under H_0. Then the Wilcoxon signed rank test rejects H_0 at the α level if

- $T^+ \geq t_\alpha$ against $H_1: \theta > 0$;
- $T^+ \leq M - t_\alpha$ against $H_1: \theta < 0$;
- either $T^+ \geq t_{\alpha/2}$ or $T^+ \leq M - t_{\alpha/2}$ against $H_1: \theta \neq 0$.

The level α is achievable such that $\Pr(T^+ \geq t) = \alpha$ for some $t \in \{0,1,\ldots,M\}$.

Wilcoxon signed rank test

Approximate distribution of T^+

$$T^* = \frac{T^+ - \mathrm{E}_0[T^+]}{\sqrt{\mathrm{Var}_0(T^+)}} = \frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0,1)$$

Approximate rejection rule: Reject H_0 at the α level if

- $T^* \geq z_\alpha$, or $T^+ \geq \mathrm{E}_0[T^+] + z_\alpha \sqrt{\mathrm{Var}_0(T^+)}$, against $H_1: \theta > 0$;
- $T^* \leq -z_\alpha$, or $T^+ \leq \mathrm{E}_0[T^+] - z_\alpha \sqrt{\mathrm{Var}_0(T^+)}$, against $H_1: \theta < 0$;
- $|T^*| \geq z_{\alpha/2}$ against $H_1: \theta \neq 0$, where T^* is defined in (2.11).

Wilcoxon signed rank test

Ties: assign the average rank to tied values.
This does not affect mean of T^+, but the variance reduces to

$$\mathrm{Var}_0(T^+) = \frac{n(n+1)(2n+1)}{24} - \frac{1}{48}\sum_{j=1}^{g} t_j(t_j-1)(t_j+1),$$

where g is the number of groups with tied ranks, and t j is the number of tied ranks in group j, j = 1,...,g.

Wilcoxon signed rank test

Symmetry of T^+

$$T^+ = \sum_{i=1}^{n} \psi_i R_i \sim \sum_{i=1}^{n} (1-\psi_i) R_i = \sum_{i=1}^{n} R_i - \sum_{i=1}^{n} \psi_i R_i = \frac{n(n+1)}{2} - T^+ = M - T^+$$

Equivalent versions

$$W = \sum_{i=1}^{n} \operatorname{sgn}(X_i) R_i = T^+ - T^- = 2T^+ - M, \quad \text{where } \operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{if } x < 0. \end{cases}$$

Wilcoxon signed rank test

Estimation of the median

Let $W_{(1)} \le W_{(2)} \le \cdots \le W_{(M)}$ be ordered values of the $M = n(n+1)/2$ averages of (X_i, X_j) (known as the *Walsh averages*):

$$W_{ij} = \frac{X_i + X_j}{2}, \quad i \le j = 1, \ldots, n.$$

The median (treatment effect) θ can be estimated by

$$\hat{\theta} = \text{median}\left\{ \frac{X_i + X_j}{2}, i \le j = 1, \ldots, n \right\}$$

$$= \begin{cases} W_{((M+1)/2)} & \text{if } M \text{ is odd;} \\ \dfrac{W_{(M/2)} + W_{(M/2+1)}}{2} & \text{if } M \text{ is even.} \end{cases}$$

Confidence interval

a $100(1-\alpha)\%$ confidence interval for θ is given by

$$(\theta_L, \theta_U) = (W_{(C_\alpha)}, W_{(M+1-C_\alpha)}) = (W_{(M+1-t_{\alpha/2})}, W_{(t_{\alpha/2})})$$

C_α can be approximated by

$$C_\alpha \approx \mathrm{E}_0[T^+] - z_{\alpha/2}\sqrt{\mathrm{Var}_0(T^+)} = \frac{n(n+1)}{4} - z_{\alpha/2}\sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Question 1

Let $X_1, \ldots, X_n$ be independent continuous random variables with $\Pr(X_i < \theta) = 0.5$ for a real number θ, $i = 1, \ldots, n$. Based on the data observed from $X_1, \ldots, X_n$:

(a) the sign test is appropriate to test the null hypothesis $H_0 : \theta = 0$ without any other conditions.

(b) if $X_1, \ldots, X_n$ are identically distributed, then the Wilcoxon signed-rank test is more efficient than the sign test for $H_0 : \theta = 0$.

(c) a nonparametric confidence interval of θ can be obtained from the order statistics of the sample $X_1, \ldots, X_n$.

Question 2

Given the following paired data (X_i, Y_i), $i = 1, \ldots, 5$:

$$(3.2, 5.6), (4.8, 3.5), (5.2, 6.5), (2.6, 3.9), (2.5, 4.9)$$

(a) Obtain the exact distribution of the Wilcoxon signed rank statistic T^+ based on

$$Z_i = Y_i - X_i, \quad i = 1, \ldots, 5,$$

conditional on any ties in $|Z_1|, \ldots, |Z_5|$.

(b) Assume that $(Z_1, \ldots, Z_5)$ have symmetric distributions with a common median θ. Test $H_0 : \theta = 0$ against $H_1 : \theta > 0$ at the 10% level of significance using the exact p-value for the test.