Statistical Inference

Non-Parametric Tests

Behnam Bahrak

Non-Parametric Statistics

- Parametric statistical methods are always designed for a specific family of distributions such as Normal, χ^2, t-student, … for the parameter of interest.
- Nonparametric statistics does not assume any particular distribution.
- On one hand, nonparametric methods are less powerful because the less you assume about the data the less you can find out from it.
- On the other hand, having fewer requirements, they are applicable to wider applications.

Sign Test

- It refers to a population median M:

$$H_0: M = m$$
$$H_A: M < m \quad \text{or} \quad M > m \quad \text{or} \quad M \neq m$$

- To conduct the sign test, simply count how many observations are above m:

$$S_{obs} = S(X_1, \ldots, X_n) = \#\text{ of } i\ : X_i > m$$

- If H_0 is true and m is the median, then each X_i is equally likely to be above m or below m, and S has Binomial distribution with parameters n and $p = 1/2$.

Sign Test

- For large enough n ($n \geq 20$), we have:

$$Binomial(n\,, 1/2) \sim N(n/2\,, \sqrt{n}/2)$$

- So the test statistic will be:

$$Z = \frac{S_{obs} - n/2}{\sqrt{n}/2}$$

- Note that you may also need to consider continuity correction.

- Computing the p-value and deciding on the rejection of H_0 is the same as parametric tests with a normal distribution of the parameter of interest.

Example

- When Eric learned to drive, he suspected that the median speed of cars on his way to school exceeds the speed limit, which is 30 mph. As an experiment, he drives to school with his friend Eve, precisely at the speed of 30 mph, and Eve counts the cars. At the end, Eve reports that 56 cars were driving faster than Eric, and 44 cars were driving slower. Does this confirm Eric's suspicion?
  - This is a one-sided right-tail test of:

$$H_0: M = 30$$
$$H_A: M > 30$$

Example

- The sign test statistic is $S = 56$, and the sample size is $n = 56 + 44 = 100$.
- Calculate the P-value (don't forget the continuity correction):

$$p\text{-}value = P(S \geq 56) = P\left(Z > \frac{55.5 - 50}{5}\right) = 0.1357$$

- We fail to reject H_0, i.e. we can conclude that Eric and Eve did not find a significant evidence that the median speed of cars is above the speed limit.

Rank

- Rank of any unit of a sample is its position when the sample is arranged in the increasing order.
- In a sample of size n, the smallest observation has rank 1, the second smallest has rank 2, and so on, and the largest observation has rank n.
- If two or more observations are equal, their ranks are typically replaced by their average rank.

$$R_i = \text{rank of the } i\text{-th observation}$$

- $R_i = r$ means that X_i is the r-th smallest observation in the sample.

Behnam Bahrak
bahrak@ut.ac.ir

Example

- Consider a sample

$$3, 7, 5, 6, 5, 4$$

- The smallest observation, $X_1 = 3$, has rank 1.
- The second smallest is $X_6 = 4$; it has rank 2.
- The 3rd and 4th smallest are $X_3 = X_5 = 5$; their ranks 3 and 4 are averaged, and each gets rank 3.5.
- Then, $X_4 = 6$ and $X_2 = 7$ get ranks 5 and 6.
- So, we have the following ranks:

$$R_1 = 1, R_2 = 6, R_3 = 3.5, R_4 = 5, R_5 = 3.5, R_6 = 2$$

Behnam Bahrak
bahrak@ut.ac.ir

Wilcoxon Signed Rank Test

- It refers to a population median M:

$$H_0: M = m$$
$$H_A: M < m \quad \text{or} \quad M > m \quad \text{or} \quad M \neq m$$

1. Consider the distances between observations and the tested value, $d_i = |X_i - m|$.
2. Order these distances and compute their ranks R_i. Notice these are ranks of d_i, not X_i.
3. Take only the ranks corresponding to observations X_i greater than m. Their sum is the test statistic W: $W = \sum_{i:X_i>m} R_i$
4. Large values of W suggest rejection of H_0 in favor of $H_A: M > m$; small values support $H_A: M < m$; and both support a two-sided alternative $H_A: M \neq m$.

Wilcoxon Signed Rank Test

- For convenience, Wilcoxon proposed to use signed ranks, giving a "+" sign to a rank R_i if $X_i > m$ and a "−" sign if $X_i < m$.
- His statistic W then equals the sum of positive signed ranks.
- Wilcoxon showed that if the distribution of X is symmetric about its median M, then for $n \geq 15$, W will have a Normal distribution:

$$W \sim \text{Normal}\left(\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}}\right)$$

- When computing the p-value, you should also consider the continuity correction because W is a discrete random variable.

Wilcoxon Signed Rank Test

1. Test of the median, H_0: $M = m$.

2. Test statistic $W = \sum_{i:X_i>m} R_i$,
   where R_i is the rank of $d_i = |X_i - m|$.

3. For $n \geq 15$:

$$W \sim \mathrm{Normal}\left(\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}}\right)$$

Assumptions: the distribution of X_i is continuous and symmetric

Example

- If an unauthorized person accesses a computer account with the correct username and password (stolen or cracked), can this intrusion be detected?
- Recently, a number of methods have been proposed to detect such unauthorized use:
  - The median of the time between keystrokes, the time a key is pressed, the frequency of various keywords are measured and compared with those of the account owner.
- If there are significant differences, an intruder is detected.
- Assume that a longtime authorized user of the account makes 0.2 seconds between keystrokes.

Example

- The following times between keystrokes were recorded when a user typed the username and password:

  .24, .22, .26, .34, .35, .32, .33, .29, .19, .36,
  .30, .15, .17, .28, .38, .40, .37, .27 seconds

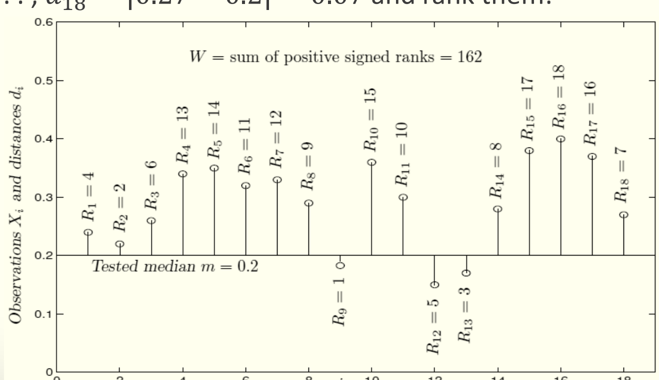
- Is this an evidence of an unauthorized attempt?

- We apply Wilcoxon signed rank test to test

$$H_0: \quad M = 0.2$$
$$H_A: \quad M \neq 0.2$$

Example

- We compute the distances $d_1 = |X_1 - m| = |0.24 - 0.2| = 0.04, \ldots, d_{18} = |0.27 - 0.2| = 0.07$ and rank them:



Example

$$W \sim N\left(\frac{18 \times 19}{4}, \sqrt{\frac{18 \times 19 \times 37}{24}}\right) \Rightarrow W \sim N(85.5\,, 23)$$

- Making a continuity correction (because W is discrete), we find the p-value for this two-sided test:

$$\text{p-value} = 2P(W \geq 162) = 2P\left(Z > \frac{161.5-85.5}{23}\right) = 0.001$$

- Thus Wilcoxon test shows strong evidence that the account was used by an unauthorized person.

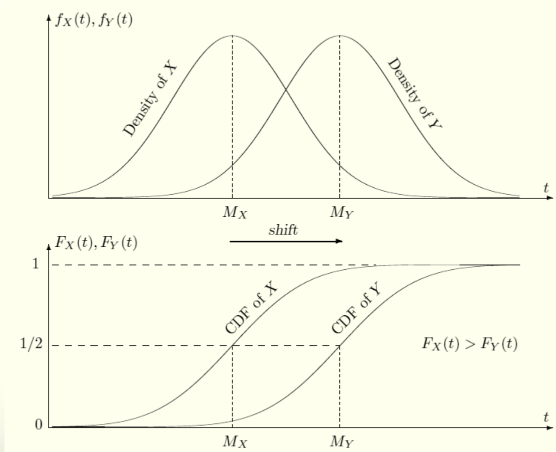
Mann-Whitney-Wilcoxon Rank Sum Test

- Suppose we have samples from two populations, and we need to compare their medians or just some location parameters to see how different they are.
- Wilcoxon signed rank test can be extended to a two-sample problem as follows.
- We are comparing two populations, the population of X and the population of Y. In terms of their cumulative distribution functions, we test:

$$H_0\!: F_X(t) = F_Y(t) \text{ for all } t.$$

- Assume that under the alternative H_A, either Y is stochastically larger than X, and $F_X(t) > F_Y(t)$, or it is stochastically smaller than X, and $F_X(t) < F_Y(t)$.

Mann-Whitney-Wilcoxon Rank Sum Test



Mann-Whitney-Wilcoxon Rank Sum Test

- Observed are two independent samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$.
- Here we compare one sample against the other. Here is how we conduct this test.
  1. Combine all X_i and Y_j into one sample.
  2. Rank observations in this combined sample. Ranks R_i are from 1 to $(n + m)$. Some of these ranks correspond to X-variables, others to Y-variables.
  3. The test statistic U is the sum of all X-ranks.

Mann-Whitney-Wilcoxon Rank Sum Test

- Test of two populations, H_0: $F_X = F_Y$.
- Test statistic $U = \sum_i R_i$, where R_i are ranks of X_i in the combined sample of X_i and Y_i.
- For $n, m \geq 10$, $U \sim N(\frac{n(n+m+1)}{2}, \sqrt{\frac{nm(n+m+1)}{12}})$

Assumptions: the distributions of X_i and Y_i are continuous; $F_X(t) < F_Y(t)$ for all t or $F_X(t) > F_Y(t)$ for all t under H_A.

Example

- Round-trip transit times (pings) are given at two locations. Arranged in the increasing order, they are:

  Location I: 0.0156, 0.0210, 0.0215, 0.0280, 0.0308, 0.0327, 0.0335, 0.0350, 0.0355, 0.0396, 0.0419, 0.0437, 0.0480, 0.0483, 0.0543 seconds

  Location II: 0.0039, 0.0045, 0.0109, 0.0167, 0.0198, 0.0298, 0.0387, 0.0467, 0.0661, 0.0674, 0.0712, 0.0787 seconds
- Is there evidence that the median ping depends on the location?
- Let us apply the Mann-Whitney-Wilcoxon test:

$$H_0: \ F_X = F_Y$$
$$H_A: \ F_X \neq F_Y$$

where X and Y are pings at the two locations.

Example

- We observed samples of sizes $n = 15$ and $m = 12$. Among them, X-pings have ranks 4, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 19, 21, 22, and 23, and their sum is:

$$U_{obs} = \sum_{i=1}^{15} R_i = 213$$

$$E(U|H_0) = \frac{n(n+m+1)}{2} = 210$$

$$Var(U|H_0) = \frac{nm(n+m+1)}{12} = 420$$

Example

- Then compute the P-value for this two-sided test:

$$\text{p-value} = 2P\{U \geq 213\}$$

$$= 2P\{Z \geq \frac{212.5 - 210}{\sqrt{420}}\}$$

$$= 0.9044$$

- We fail to reject H_0: there is no evidence that pings at the two locations have different distributions.