# Non-Parametric Statistics

# Motivations

Some experiments yield response measurements that defy exact quantification. Examples are

- Rank the teaching performance of four teachers
- Taste Characteristics of five brands of chocolate

Experiments generate response measurement that can be ordered (ranked), but it is impossible to make statements such as “teacher A is twice as good as teacher B”

These kind of data are particularly evident in social science and studies of consumer preference. For such data, non-parametric statistics is applied

Non-parametric statistical procedures are also particularly useful in making inferences in situations where serious doubts exists about the assumptions that underlie standard methodology.

Examples

- Unclear whether the random variables are i.i.d. and hence unclear whether the central limit theorem can be applied
- In t-test of the equality of means, it is assumed that the unknown variance is equal. It is unsure to what extent this is approximately true
- Different normality tests give contradicting results

# Parametric and Non-Parametric Methods

Parametric method: apply to problems where the distributions from which the samples are taken are specified except for the values of a finite number of parameters

Example: all the methods that you have learnt so far

Non-parametric method: no standard definition, but most agree that it does not assume a specific probability distribution

Below are some examples of non-parametric tests

# Sign Test

Let $X_1, \ldots, X_n$ denote a sample from a continuous distribution $F$ and suppose that we are interested in testing the hypothesis that the median of $F$, call it $m$, is equal to $m_0$

Null hypothesis $H_0$: $m = m_0$

Alternative hypothesis $H_1$: $m \neq m_0$

Let

$$I_i = \begin{cases} 1 & if\ X_i < m_0 \\ 0 & if\ X_i \geq m_0 \end{cases}$$

$I_1, \ldots, I_n$ are independent Bernoulli random variable with parameter $p = F(m_0)$

If $m_0$ is the true median, it should be equally likely for $X_i < m_0$ and $X_i \geq m_0$, hence $p = 0.5$. So it is equivalent to test

Null hypothesis $H_0$: $p = 0.5$

Alternative hypothesis $H_1$: $p \neq 0.5$

Suppose $T = \sum_{i=1}^{n} I_i$ is observed, if $T$ is very different from $n/2$, we have reason to reject the null hypothesis

Since $T$ has a binomial distribution $Bin\ (n, 0.5)$

p-value = 2 min $(P\{Bin(n, 0.5) \leq T\}, P\{Bin(n, 0.5) \geq T)$

We accept the null hypothesis if the p-value is larger than level of significance $\alpha$ and vice versa

Example

Suppose we are interested in testing whether a recently instituted industrial safety program has had an effect on the number of man-hours lost to accidents. For each of 10 plants, the data consisted of the pair $(X_i, Y_i)$, which represents respectively the average weekly loss at plant $i$ before and after the program.

Let $Z_i = X_i - Y_i$, $i = 1, \ldots, 10$

If the program has any effect, $Z_i$, $i = 1, \ldots, 10$, would be a sample from a distribution whose median is $0$.

If the distribution is known to be normal, the paired t test can be used. Assume the distribution is unknown, we use the sign test

Let the $Z_i$ be 7.5, -2.3, 2.6, 3.7, 1.5, -0.5, -1, 4.9, 4.8, 1.6

$T = 3$

Since $T < 5$,

p-value = $2\ P\{Bin(10,0.5) \leq 3\}$

$= 2\left(\sum_{i=0}^{3}\binom{10}{i}(0.5)^{10}\right) = 0.344$

If $\alpha = 0.05$, then since p-value = 0.344 > 0.05, the null hypothesis is accepted. This means that the evidence is consistent with the hypothesis that the safety program has no effect on the number of man-hours lost to accidents

Another way to interpret this is that $T = 3$ is close to $\frac{10}{2} = 5$. It is reasonable that a fair dice ($p = 0.5$) will generate 3 heads in 10 trials, though we expect 5 heads

Another way of thinking about this is that the rejection region is $T = 0, 1, 2, 3, 7, 8, 9, 10$, which makes the Type I error (= p-value) very large. We only accept the hypothesis if $T = 4, 5, 6$. This is unlikely to be a good hypothesis test

# For large $n$

When $n$ is large, the binomial distribution can be approximated by a normal distribution with mean $np$ and variance $np(1-p)$.

The standard normal random variable is

$$Z = \frac{T - np}{\sqrt{np(1-p)}}$$

Put in $p = 0.5$

$$Z = \frac{T - n/2}{0.5\sqrt{n}}$$

Reject $H_0$ if $|Z| > z_{\alpha/2}$

# Signed Rank Test

This test not only considers the sign, but also the magnitude of the difference. If the median is $m_0$, $X - m_0$ should be +ve and –ve with the same probability. In other words, if the absolute difference is $|X - m_0|$, it has equal probability of being +ve or -ve

Let $X_1, \ldots, X_n$ denote a sample from a continuous distribution $F$

Re-order $X_1, \ldots, X_n$ according to their absolute values $|X_i - m_0|$. The smallest absolute value has rank 1, etc. This gives a weighting to the data.

Define

$$I_j = \begin{cases} 1 & jth\ ranked\ data\ smaller\ than\ m_0 \\ 0 & otherwise \end{cases}$$

Sign test uses the test statistic $\sum_{j=1}^{n} I_j$

Signed rank test uses the test statistic $T = \sum_{j=1}^{n} jI_j$.

The physical meaning is to give more weighing to larger weights to data values that are further away from $m_0$

Reject the hypothesis if $T$ is very small or very large

Example

$X_1 = 4.2 \quad X_2 = 1.8 \quad X_3 = 5.3 \quad X_4 = 1.7$

Hypothesize the median $m_0 = 2$

| $X_i$ | $X_i - m_0$ | $\lvert X_i - m_0 \rvert$ | Rank |
|---|---|---|---|
| 4.2 | 2.2 | 2.2 | 3 |
| 1.8 | -0.2 | 0.2 | 1 |
| 5.3 | 3.3 | 3.3 | 4 |
| 1.7 | -0.3 | 0.3 | 2 |

$T = 1 + 2 = 3$

The signed rank test can be used to test whether $m_0$ is the median or whether the distribution is symmetric about $m_0$

# Treating Special Cases

1. Absolute values equal to 0 eliminated, and $n$ is reduced
2. If two or more absolute differences are tied for the same rank, then the average of the ranks that would have been assigned to these differences is assigned to each member of the tied group

   (e.g. if two absolute differences are tied for ranks 3 and 4, then each receives rank 3.5, and the next higher absolute difference is assigned rank 5)

# For large $n$

When $n$ is large (guideline $n > 25$), $T$ will be approximately normal distributed (why?) when the null hypothesis is true, and

$$E(T) = \frac{n(n+1)}{4} \qquad Var(T) = \frac{n(n+1)(2n+1)}{24}$$

$$Z = \frac{T - E(T)}{\sqrt{V(T)}}$$

Then we can test the hypothesis using the standard normal distribution

# Mann-Whitney $U$ Test

Sign test and signed rank test can be applied to $n$ observations of the form $(X_i, Y_i)$.

Mann-Whitney test is also called rank sum test, Wilcoxon test or $U$ test. It deals with the following situation:

$X_1, \ldots, X_n$ denote a sample of measurable values of $n$ items produced by method 1. $Y_1, \ldots, Y_m$ is the corresponding value of $m$ items produced by method 2.

Let $F$ and $G$ denote the distribution functions of the two samples respectively

Null hypothesis $H_0$: $F = G$

Alternative hypothesis $H_1$: $F \neq G$

Procedure

Order the $n + m$ data values. Give the smallest data value rank 1, the second smallest rank 2, …, and the $(n + m)$th value rank $n + m$.

For $i = 1, \ldots, n$, let

$$R_i = rank\ of\ the\ data\ value\ X_i$$

Rank sum test utilizes the test statistic $T$ equal to the sum of the ranks from the first sample

$$T = \sum_{i=1}^{n} R_i$$

Example

An experiment designed to compare two treatments yield the following data

Treatment 1: 65.2 67.1 69.4 78.2 74 80.3

Treatment 2: 59.4 72.1 68 66.2 58.5

The ordered values are 58.5, 59.4, 65.2*, 66.2, 67.1*, 68, 69.4*, 72.1, 74*, 78.2*, 80.3*

(* means from sample 1)

$$T = 3 + 5 + 7 + 9 + 10 + 11 = 45$$

Suppose that we desire a significance level $\alpha$ test of $H_0$. If the observed value of $T$ is $T = t$, then $H_0$ should be rejected if either

$$P\{T \leq t\} \leq \frac{\alpha}{2} \qquad or \qquad P\{T \geq t\} \leq \frac{\alpha}{2}$$

$P\{T \geq t\}$ can be expressed in the "$\leq$" form, as follows:

$$P\{T \geq t\} = 1 - P\{T < t\} = 1 - P\{T \leq t - 1\}$$

It follows that $H_0$ should be rejected if

$$P\{T \leq t\} \leq \frac{\alpha}{2} \qquad or \qquad P\{T \leq t - 1\} \geq 1 - \frac{\alpha}{2}$$

Assume $H_0$ is true, let $P(n, m, k)$ be the probability that $T \leq k$

Since if $H_0$ is true, $F = G$, the ranks should be random. Consider

Case 1: the largest of the $(n + m)$ values is found in sample 1

Case 2: the largest of the $(n + m)$ values is found in sample 2

Case 1 occurs with probability $n/(n+m)$. If Case 1 occurs, then

a) the largest value in sample 1 has rank $(n+m)$

b) the sum of ranks of the remaining $(n-1)$ values $\leq k-(n+m)$. This occurs with probability $P(n-1, m, k-n-m)$

Case 2 occurs with probability $m/(n+m)$. If Case 2 occurs, then

a) the largest value in sample 2 has rank $(n+m)$

b) the sum of ranks of the $n$ values in sample 1 $\leq k$. This occurs with probability $P(n, m-1, k)$

Hence

$$P(n,m,k) = \frac{n}{n+m}P(n-1,m,k-n-m) + \frac{m}{n+m}P(n,m-1,k)$$

This formula can be solved recursively, starting with the boundary condition

$$P(1,0,k) = \begin{cases} 0 & k \leq 0 \\ 1 & k > 0 \end{cases} \qquad P(0,1,k) = \begin{cases} 0 & k < 0 \\ 1 & k \geq 0 \end{cases}$$

to obtain $P(n,m,t-1)$ and $P(n,m,t)$, where $t$ is the observed value of $T$

# Example

- In the previous example,

$$P(5,4,k) = \frac{6}{11}P(4,4,k-9) + \frac{5}{11}P(5,3,k)$$

# For moderately large $n$ and $m$

When $n$ and $m$ are moderately large ($n > 7, m > 7$), $T$, under $H_0$, will be approximately normally distributed. It can be shown that

$$E[T] = \frac{n(n+m+1)}{2}$$

$$Var[T] = \frac{nm(n+m+1)}{12}$$

$$Z = \frac{T - E(T)}{\sqrt{V(T)}}$$

Then we can test the hypothesis using the standard normal distribution

# Physical meaning of the $U$ test

If the hypothesis is rejected, it means that the two distributions $F$ and $G$ are different. This has the useful implications that two random processes $X$ and $Y$ are different.

The $U$ test is useful as we do NOT need to specify the actual distribution of $X$ and $Y$ (c.f. assume normal distribution)

# References

Text book Ch. 12