NON-PARAMETRIC STATISTICS: ONE AND TWO SAMPLE TESTS

Non-parametric tests are normally based on **ranks** of the data samples, and test hypotheses relating to **quantiles** of the probability distribution representing the population from which the data are drawn. Specifically, tests concern the **population median**, η, where

$$\Pr[\text{ Observation } \leq \eta\,] = \frac{1}{2}$$

The **sample median**, x_{MED}, is the mid-point of the sorted sample; if the data $x_1, \ldots, x_n$ are sorted into **ascending** order, then

$$x_{\text{MED}} = \begin{cases} x_m & n \text{ odd}, n = 2m+1 \\ \dfrac{x_m + x_{m+1}}{2} & n \text{ even}, n = 2m \end{cases}$$

1. ONE SAMPLE TEST FOR MEDIAN: THE SIGN TEST

For a single sample of size n, to test the hypothesis $\eta = \eta_0$ for some specified value η_0 we use the **Sign Test.**. The test statistic S depends on the alternative hypothesis, H_a.

(a) For **one-sided** tests, to test

$$\begin{aligned} H_0 &: \eta = \eta_0 \\ H_a &: \eta > \eta_0 \end{aligned}$$

we define test statistic S by

$$S = \text{Number of observations } \textbf{greater than } \eta_0$$

whereas to test

$$\begin{aligned} H_0 &: \eta = \eta_0 \\ H_a &: \eta < \eta_0 \end{aligned}$$

we define S by

$$S = \text{Number of observations } \textbf{less than } \eta_0$$

If H_0 is **true**, it follows that

$$S \sim \text{Binomial}\left(n, \frac{1}{2}\right)$$

The p-value is defined by

$$p = \Pr[X \geq S]$$

where $X \sim \text{Binomial}(n, 1/2)$. The rejection region for significance level α is defined implicitly by the rule

$$\text{Reject } H_0 \text{ if } \alpha \geq p.$$

The Binomial distribution is tabulated on pp 885-888 of McClave and Sincich.

(b) For a **two-sided** test,

$$\begin{aligned} H_0 &: \eta = \eta_0 \\ H_a &: \eta \neq \eta_0 \end{aligned}$$

we define the test statistic by

$$S = \max\{S_1, S_2\}$$

where S_1 and S_2 are the counts of the number of observations less than, and greater than, η_0 respectively. The p-value is defined by

$$p = 2\,\Pr[X \geq S]$$

where $X \sim \text{Binomial}(n, 1/2)$.

Notes :

1. The only assumption behind the test is that the data are drawn independently from a continuous distribution.

2. If any data are equal to η_0, we **discard** them before carrying out the test.

3. **Large sample approximation.** If n is large (say $n \geq 30$), and $X \sim \text{Binomial}(n, 1/2)$, then it can be shown that

$$X \stackrel{\sim}{\approx} \text{Normal}(np, np(1-p))$$

   Thus for the sign test, where $p = 1/2$, we can use the test statistic

$$Z = \frac{S - \dfrac{n}{2}}{\sqrt{n \times \dfrac{1}{2} \times \dfrac{1}{2}}} = \frac{S - \dfrac{n}{2}}{\sqrt{n} \times \dfrac{1}{2}}$$

   and note that if H_0 is true,

$$Z \stackrel{\sim}{\approx} \text{Normal}(0, 1).$$

   so that the test at $\alpha = 0.05$ uses the following critical values

$$\begin{array}{llll} H_a : \eta > \eta_0 & \text{then} & C_R = 1.645 \\ H_a : \eta < \eta_0 & \text{then} & C_R = -1.645 \\ H_a : \eta \neq \eta_0 & \text{then} & C_R = \pm 1.960 \end{array}$$

4. For the large sample approximation, it is common to make a **continuity correction**, where we replace S by $S - 1/2$ in the definition of Z

$$Z = \frac{\left(S - \dfrac{1}{2}\right) - \dfrac{n}{2}}{\sqrt{n} \times \dfrac{1}{2}}$$

Tables of the standard Normal distribution are given on p 894 of McClave and Sincich.

2. TWO SAMPLE TESTS FOR INDEPENDENT SAMPLES: THE MANN-WHITNEY-WILCOXON TEST

For a two **independent** samples of size n_1 and n_2, to test the hypothesis of **equal population medians**

$$\eta_1 = \eta_2$$

we use the **Wilcoxon Rank Sum Test**, or an equivalent test, the **Mann-Whitney U Test**; we refer to this as the

Mann-Whitney-Wilcoxon (MWW) Test

By convention it is usual to formulate the test statistic in terms of the **smaller** sample size. Without loss of generality, we label the samples such that

$$n_1 > n_2.$$

The test is based on the **sum of the ranks** for the data from sample 2.

EXAMPLE : $n_1 = 4, n_2 = 3$ yields the following ranked data

SAMPLE 1	0.31	0.48	1.02	3.11
SAMPLE 2	0.16	0.20	1.97	

SAMPLE	2	2	**1**	**1**	**1**	2	**1**
	0.16	0.20	**0.31**	**0.48**	**1.02**	1.97	**3.11**
RANK	1	2	**3**	**4**	**5**	6	**7**

Thus the rank sum for sample 1 is

$$R_1 = 3 + 4 + 5 + 7 = 19$$

and the rank sum for sample 2 is

$$R_2 = 1 + 2 + 6 = 9.$$

Let η_1 and η_2 denote the medians from the two distributions from which the samples are drawn. We wish to test

$$H_0 \ : \ \eta_1 = \eta_2$$

Two related test statistics can be used

- **Wilcoxon Rank Sum Statistic**

$$W = R_2$$

- **Mann-Whitney U Statistic**

$$U = R_2 - \frac{n_2(n_2+1)}{2}$$

We again consider three alternative hypotheses:

$$\begin{aligned} H_a &: \ \eta_1 < \eta_2 \\ H_a &: \ \eta_1 > \eta_2 \\ H_a &: \ \eta_1 = \eta_2 \end{aligned}$$

and define the rejection region separately in each case.

Large Sample Test

If $n_2 \geq 10$, a large sample test based on the Z statistic

$$Z = \frac{U - \dfrac{n_1 n_2}{2}}{\sqrt{\dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

can be used. Under the hypothesis $H_0 \ : \ \eta_1 = \eta_2$,

$$Z \stackrel{\cdot}{\sim} \text{Normal}(0, 1)$$

so that the test at $\alpha = 0.05$ uses the following critical values

$$\begin{array}{lll} H_a \ : \ \eta_1 > \eta_2 & \text{then} & C_R = -1.645 \\ H_a \ : \ \eta_1 < \eta_2 & \text{then} & C_R = 1.645 \\ H_a \ : \ \eta_1 \neq \eta_2 & \text{then} & C_R = \pm 1.960 \end{array}$$

Small Sample Test

If $n_1 < 10$, an **exact** but more complicated test can be used. The test statistic is R_2 (the sum of the ranks for sample 2). The null distribution under the hypothesis $H_0 \ : \ \eta_1 = \eta_2$ can be computed, but it is complicated.

The table on p. 832 of McClave and Sincich gives the critical values (T_L and T_U) that determine the rejection region for different n_1 and n_2 values up to 10.

- **One-sided tests:**

  $$\begin{array}{lll} H_a \ : \ \eta_1 > \eta_2 & \text{Rejection Region is} & R_2 \leq T_L \\ H_a \ : \ \eta_1 < \eta_2 & \text{Rejection Region is} & R_2 \geq T_U \end{array}$$

  These are tests at the $\alpha = 0.025$ significance level.

- **Two-sided tests:**

  $$H_a \ : \ \eta_1 \neq \eta_2 \quad \text{Rejection Region is} \quad R_2 \leq T_L \text{ or } R_2 \geq T_U$$

  This is a test at the $\alpha = 0.05$ significance level.

Notes :

1. The only assumption is are needed for the test to be valid is that the samples are independently drawn from two continuous distributions.

2. The sum of the ranks across **both** samples is

   $$R_1 + R_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

3. If there are **ties** (equal values) in the data, then the rank values are replaced by **average** rank values.

DATA VALUE	0.16	0.20	0.31	0.31	0.48	1.97	3.11
ACTUAL RANK	1	2	3	3	5	6	7
AVERAGE RANK	1	2	3.5	3.5	5	6	7