14-Nonparametric Statistics

ENSY SILVER[1]

Monday 14th September, 2020

[1]Thanks to my family, my friend and freedom.

1 Introductione

Sometimes, when we do not know the exact distribution, we use nonparametric method to test hypotheses.

2 The sign test

The simplest test is called the sign test.

Theorem 2.1. *Let $y_1, y_2, \cdots, y_n$ be a random sample of size n from any continuous distribution having median $\tilde{\mu}$, where $n \geq 10$. Let k denote the number of y_i' s greater than $\tilde{\mu}_0$, and let $z = \frac{k-n/2}{\sqrt{n/4}}$. Z has approximately a standard normal distribution.*

3 The signed rank test

The signed rank test is based on the magnitudes, and directions, of the deviations of the y_i' s from μ_0. Let $|y_1 - \mu_0|, |y_2 - \mu_0|, \cdots, |y_n - \mu_0|$ be the set of absolute deviations of the y_i's from μ_0. These can be ordered from smallest to largest, and we can define r_i to be the rank of $|y_i - \mu_0|$. Associated with each ri will be a sign indicator, z_i, where

$$z_i = \begin{cases} 1, & \text{if } y_i - \mu_0 > 0 \\ 0, & \text{if } y_i - \mu_0 < 0 \end{cases}$$

The signed rank statistic, w, is defined to be the linear combination

$$e = \sum_{i=1}^{n} r_i z_i$$

4 Wilcoxon test

Theorem 4.1. *Let $y_1, y_2, \cdots, y_n$ be a set of independent observations drawn, respectively, from the continuous and symmetric (but not necessarily identical) pdfs $f_{Y_i}(y)$, $i = 1, 2, \cdots, n$. Suppose that each of the $f_{Y_i}(y)$' s has the same mean μ. If $H_0 : \mu = \mu_0$ is true, the pdf of the data's signed rank statistic, $p_W(w)$, is given by*

$$p_W(w) = P(W = w) = \frac{1}{2^n} \cdot c(w)$$

where $c(w)$ is the coefficient of $e^W t$ in the expansion of

$$\prod_{i=1}^{n} (1 + e^{it})^2$$

The proof is in the book, we omit it. Then, we deduce the Wilcoxon signed rank test.

Theorem 4.2. *When $H_0 : \mu = \mu_0$ is true, the mean and variance of the Wilcoxon signed rank statistic, W, are given by*

$$E(W) = \frac{n(n+1)}{4}$$

and

$$\mathrm{Var}(W) = \frac{n(n+1)(2n+1)}{24}$$

Also, for $n > 12$, the distribution of

$$\frac{W - [n(n+1)]/4}{\sqrt{n(n+1)(2n+1)/24}}$$

can be adequately approximated by the standard normal pdf, $f_Z(z)$.

An extended version of Wilcoxon signed rank test has a more complicated proof, but we give the theorem directly.

Theorem 4.3. *Let $x_1, x_2, \cdots, x_n$ and $x_{n+1}, x_{n+2}, \cdots, x_{n+m}$ be two independent random samples from $f_X(x)$ and $f_Y(y)$, respectively, where the two pdfs are the same except for a possible shift in location. Let r_i denote the rank of the i^{th} observation in the combined sample (where the smallest observation is assigned a rank of 1 and the largest observation, a rank of $n+m$). Let*

$$w' = \sum_{i=1}^{n+m} r_i z_i$$

where z_i is 1 if the ith observation comes from $f_X(x)$ and 0, otherwise. Then

$$E(W') = \frac{n(n+m+1)}{2}$$

$$\mathrm{Var}(W') = \frac{nm(n+m+1)}{12}$$

and

$$\frac{W' - n(n+m+1)/2}{\sqrt{nm(n+m+1)/12}}$$

has approximately a standard normal pdf if $n > 10$ and $m > 10$.

5 The Kruskal-Wallis test

For k-sample problem, one common nonparametric test is the Kruskal-Wallis test. First, we define the R_{ij} to be the rank corresponding to Y_{ij}. Then, we have the theorem.

Theorem 5.1. *Suppose $n_1, n_2, \cdots, n_k$ independent observations are taken from the pdfs $f_{Y_1}(y), f_{Y_2}(y), \cdots, f_{Y_k}(y)$, respectively, where the $f_{Y_i}(y)$'s are all continuous and have the same shape. Let μ_i be the mean of $f_{Y_i}(y)$, $i = 1, 2, \cdots, k$, and let $R_{.1}, R_{.2}, \cdots, R_{.k}$ denote the random sums associated with each of the k samples. If $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is true,*

$$B = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_{.j}^2}{n_j} - 3(n+1)$$

has approximately a χ^2_{k-1} distribution and H_0 should be rejected at the α level of significance if $b > \chi^2_{1-\alpha,k-1}$.

6 The Friedman test

For block data, we hvae Friedman test. Suppose $k(\geq 2)$ treatments are ranked independently within b blocks. Let $r_{\cdot j}$, $j = 1, 2, \cdots, k$, be the rank sum of the j^{th} treatment. The null hypothesis that the population medians of the k treatments are all equal is rejected at the α level of significance(approximately) if

$$g = \frac{12}{bk(k+1)} \sum_{j=1}^{k} r_{.j}^2 - 3b(k+1) \geq \chi^2_{1-\alpha,k-1}$$

7 Randomness test

Theorem 7.1. *Let W denote the number of runs up and down in a sequence of n observations, where $n > 2$. If the sequence is random, then*

1. $E(W) = \frac{2n-1}{3}$.
2. $\mathrm{Var}(W) = \frac{16n-29}{90}$.
3. $\frac{W-E(W)}{\sqrt{\mathrm{Var}(W)}} = Z$, *when* $n \geq 20$.