

Rank-Sum Test

The rank-sum test is a distribution-free (or nonparametric) alternative to the two sample t-test in that the populations involved do not have to be normal. It performs nearly as well as the two-sample t-test when the two populations are normal and much better when there are extreme outliers.

1. Hypotheses:

H_0 : two populations have *same* distribution (*null hypothesis*)

H_a : population 1 tends to have larger values than population 2 (*Alternative 1*)

H_a : population 1 tends to have smaller values than population 2 (*Alternative 2*)

H_a : the two populations are different (*Alternative 3*)

2. Test Statistic T (rank-sum statistic): T is obtained by the following steps

1. list all observations from both samples in increasing order
2. label these ordered observations from 1 to $n_1 + n_2$; call this label the “order”
3. identify which sample each observation came from (Sample 1 or Sample 2)
4. create a variable called “ranks”. The ranks are usually just the “order” except when there are ties (duplicate values appearing the data). The ranks for tied observations are taken to be the average of the corresponding orders.
5. sum up the ranks for all observations from Sample 1, call this T

3. P-value

SAS computes the exact p-value and the corrected normal p-value based on the test statistic T use `npairway` procedure.

R also computes these p-values, use `wilcox.test` function.

Finding a p-value by a standard normal approximation:

A standard normal distribution, or Z -distribution, is a t -distribution with degrees of freedom ∞ . (Basically, we use similar ideas from t -tests to find p-values.)

If H_0 is true, the test statistic

$$Z = \frac{T - \text{Mean}(T)}{\text{SD}(T)}, \quad \text{where} \quad \text{Mean}(T) = n_1 \bar{R} = n_1(n_1 + n_2 + 1)/2, \quad \text{SD}(T) = s_R \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

approximately follows a standard normal distribution. (Above \bar{R} is the sample average and s_R is the sample standard deviation for all the $(n_1 + n_2)$ ranks combined.)

H_a *Alternative 1* p-value = % of Z -distribution values larger than our Z
i.e., area under t -curve with $df = \infty$ to right of Z

H_a *Alternative 2* p-value = % of Z -distribution values smaller than our Z
i.e., area under t -curve with $df = \infty$ to left of Z

H_a *Alternative 3* p-value = $2 \times$ % of Z -distribution values larger than our $|Z|$
i.e., $2 \times$ area under t -curve with $df = \infty$ to right of $|Z|$

4. Interpretation of p-value and Conclusion: as usual