

Notes: Non-Parametric Statistics

Non-Parametric Statistics

Distribution-free tests

- do not depend on the specific shape of the population distribution (do not require a normal distribution)
- based on either quantitative or categorical data (computations usually done with categorical)

Wilcoxon Signed-Rank Test

Analyzes data from a matched-sample experiment.

Uses quantitative data.

Number of matched pairs is 10 or more.

Assumes the *differences* between matched samples have a symmetric distribution (shape of the two populations is the same).

Focus is on determining if there is a difference between the two *medians*.

Examples of "matched samples" would be: the same person using two different methods to do something; times for the same delivery request using two different services; golf scores in two different rounds, same people.

Two-tail test

H_0 : median of A – median of B = 0

H_a : median of A – median of B \neq 0

Upper tail test

H_0 : median of A – median of B \leq 0

H_a : median of A – median of B $>$ 0

To test for the difference between set A and set B:

- 1) Calculate A – B for each pair, then the absolute value of A – B.
- 2) Eliminate pairs with differences of zero, adjust n (the number of pairs).
- 3) Rank the absolute differences. For repeated values, add raw ranks and divide by number of repeated values; assign same average rank to each repeated value.
- 4) Give each rank the sign of the original difference.
- 5) T^+ = sum of positive signed ranks.

The distribution of T^+ is approximated by a normal distribution for $n \geq 10$, where:

$$\mu_{T^+} = \frac{n(n+1)}{4} \quad \sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} \quad z = \frac{T^+ - \mu_{T^+}}{\sigma_{T^+}}$$

CORRECTION FACTOR: T^+ is discrete, approximated by the normal (continuous) distribution.

If $T^+ = 40$, then actually $39.5 \leq T^+ \leq 40.5$.

Apply the continuity correction factor to err on the safe side.

$$\text{Area in upper tail } (40 > \mu): \quad P(T^+ \geq 40) = P\left(z \geq \frac{39.5 - \mu_{T^+}}{\sigma_{T^+}}\right)$$

$$\text{Area in lower tail } (40 < \mu): \quad P(T^+ \leq 40) = P\left(z \leq \frac{40.5 - \mu_{T^+}}{\sigma_{T^+}}\right)$$

Example

H_0 : median of A – median of B = 0

H_a : median of A – median of B \neq 0 $\alpha = .05$

Gas Prices						
Avis	Budget	A - B	A - B	Rank	- Rank	+Rank
1.58	1.39	0.19	0.19	10		10
1.6	1.55	0.05	0.05	6		6
1.53	1.55	-0.02	0.02	3.5	-3.5	
1.55	1.51	0.04	0.04	5		5
1.57	1.58	-0.01	0.01	1.5	-1.5	
1.8	1.74	0.06	0.06	7		7
1.62	1.6	0.02	0.02	3.5		3.5
1.69	1.6	0.09	0.09	8		8
1.75	1.59	0.16	0.16	9		9
1.55	1.54	0.01	0.01	1.5		1.5
n = 10						T⁺ = 50

$$\mu_{T^+} = \frac{10(10+1)}{4} = 27.5$$

$$\sigma_{T^+} = \sqrt{\frac{10(10+1)(2*10+1)}{24}} = 9.8107$$

$$z = \frac{49.5 - 27.5}{9.8107} = 2.24$$

$$P(z \geq 2.24) = .0125$$

$$p\text{-value} = 2(.0125) = .025 < .05 \rightarrow \text{reject } H_0$$

Test about the median of a symmetric population.

This test can also be used to test a stated population median by comparing it with a sample set of values. For example, if the median SAT score is stated as 500, then 500 would be used as every "B" value to compute all the differences from a list of sample values ("A" values). The test would reveal any significant difference from the stated median.

Kruskal-Wallis Test

Tests three or more populations to see if one or more tend to provide smaller or larger values compared to the others.

Can be used with ordinal or quantitative data.

No assumption of normal distributions.

H_0 : all populations are identical

H_a : not all populations are identical

If all the populations have the same shape, the hypotheses could be stated:

H_0 : all medians are equal

H_a : not all medians are equal

To test:

- 1) Combine data from all samples and assign ranks. For repeated values, add raw ranks and divide by number of repeated values; assign same average rank to each repeated value.
- 2) Split the data with their ranks back into their original samples.
- 3) Add the ranks for each sample: R_1, R_2 , etc.
- 4) Compute the test statistic H . The test statistic H has approximately a χ^2 distribution, as long as sample sizes ≥ 5 . If $H > \chi^2_{\alpha, df}$, reject the null. This is always an upper tail test.

degrees of freedom (df) = $k - 1$

k = number of populations/samples

$$H = \left[\frac{12}{n_T(n_T + 1)} \sum_{i=1}^k \frac{R_i^2}{n_i} \right] - 3(n_T + 1)$$

n_i = number of observations in sample i

n_T = total number of observations in all samples

R_i = sum of the ranks for sample i

Example

$k = 3$ $df = 2$ $\alpha = .01$

$n_T = 21$ $\chi^2_{.01, 2} = 9.210$

$$H = \frac{12}{(21)(22)} \left[\frac{93^2}{8} + \frac{33^2}{7} + \frac{105^2}{6} \right] - 3(22)$$

$$= 13.849 > 9.210$$

→ reject the null

From the χ^2 table: $p\text{-value} < .005$

Exact $p\text{-value}$ is .000983

Group A		Group B		Group C	
Data	Rank	Data	Rank	Data	Rank
23	3	16	1	75	14.5
49	9	19	2	75	14.5
55	10	29	4	78	16
58	11	30	5	96	19
62	12	31	6	98	20
66	13	33	7	102	21
79	17	40	8		
88	18				
$n_1 = 8$	$R_1 = 93$	$n_2 = 7$	$R_2 = 33$	$n_3 = 6$	$R_3 = 105$

Looking at the data, Group B tends to produce the lower values to make a significant difference.