Applied Statistics in R

Elena Parilina

Master's Program
Game Theory and Operations Research

Saint Petersburg State University

2020

Agenda

1. Non-parametric Tests

Non-parametric Tests

Wilcoxon test

Samples:
$X_{[n]} = (X_1, \ldots, X_n)$ and $Y_{[m]} = (Y_1, \ldots, Y_m)$ from continuous random variables with c.d.f. F and G.
$m \leqslant n$
Hypotheses:

- $H_0 : \ F(x) = G(x)$ for any $x \in \mathbb{R}$.
- $H_1 : \ F(x) \geqslant G(x)$ for any $x \in \mathbb{R}$.
- $H_1^{'} : \ F(x) \leqslant G(x)$ for any $x \in \mathbb{R}$.
- $H_1^{''} : \ F(x) \neq G(x)$ for any $x \in \mathbb{R}$.

Wilcoxon test

Two samples are in one:
$Z_{[n+m]} = (X_{[n]}, Y_{[m]})$.
Ordered sample:

$$z_{(1)} < z_{(2)} < \ldots < z_{(m+n)},$$

$z_{(1)} < z_{(2)} < \ldots < z_{(m+n)}$.
Find the ranks of sample $Y_{[m]}$ in $z_{(1)} < z_{(2)} < \ldots < z_{(m+n)}$:

$$rank(Y_1) = s_1,\ \ rank(Y_2) = s_2,\ \ \ldots,\ \ rank(Y_m) = s_m.$$

Statistics is

$$W = \sum_{i=1}^{m} s_i.$$

Mann-Whitney test

Statistics is

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} I\{X_i < Y_j\},$$

where

$$I\{X_i < Y_j\} = \begin{cases} 1, & X_i < Y_j; \\ 0, & X_i > Y_j. \end{cases}$$

Statistics of Wilcoxon test and Mann-Whitney test satisfy formula

$$W = U + \frac{m(m+1)}{2}.$$

wilcox.test

Arguments of function:

- `x`, `y` — samples;
- `alternative` — alternatives `"two.sided"`, `"greater"`, `"less"`. For `"greater"`: $F(x) < G(x)$.
- `exact` is `FALSE` or `TRUE`, in case `TRUE` it will give exact p-value. By default: `exact=FALSE`, exact p-value is calculated if size of each sample is less than 50.
- `correct`: `FALSE` or `TRUE`. In case `TRUE` there is a normal approximation.

wilcox.test

wilcox.test

```
x <- c(0.80, 0.83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64,
0.73, 1.46)
y <- c(1.15, 0.88, 0.90, 0.74, 1.21)
wilcox.test(x, y, alternative = "two.sided")
```

result

```
Wilcoxon rank sum test
data:  x and y
W = 35, p-value = 0.2544
alternative hypothesis:  true location shift is not equal
to 0
```

Sign Test for the population median

Test statistics:

$$ST = \sum_{i=1}^{n} s(x_i - \theta_0),$$

where

$$s(x_i - \theta_0) = \begin{cases} 1, & x_i > \theta_0, \\ 0, & x_i \leqslant \theta_0. \end{cases}$$

SIGN.test

```
library(BSDA)
SIGN.test(x, md = 0,
    alternative = c("two.sided", "less", "greater"),
    conf.level = 0.95)
```

Probability of success in a Bernoulli experiment

Test statistics: $B = \sum_{i=1}^{n} x_i$.

binom.test

```
binom.test(x, n, p = 0.5,
   alternative = c("two.sided", "less", "greater"),
   conf.level = 0.95)
```

Rank-sum Wilcoxon test

Samples
$X_{[n]} = (X_1, \ldots, X_n)$ and $Y_{[n]} = (Y_1, \ldots, Y_n)$ from random variables with continuous c.d.f. $F(x)$ and $G(x)$.
Hypotheses:

- $H_0: \; F(x) = G(x)$ for any $x \in \mathbb{R}$.
- $H_1: \; F(x) \geqslant G(x)$ for any $x \in \mathbb{R}$.
- $H_1': \; F(x) \leqslant G(x)$ for any $x \in \mathbb{R}$.
- $H_1'': \; F(x) \neq G(x)$ for any $x \in \mathbb{R}$.

Rank-sum Wilcoxon test

New variable $z_i = X_i - Y_i$.
Hypotheses:

- $H_0: \ P\{z_i < 0\} = P\{z_i > 0\} = 1/2$.
- $H_1: \ P\{z_i < 0\} > P\{z_i > 0\}$.
- $H_1^{'}: \ P\{z_i < 0\} < P\{z_i > 0\}$.
- $H_1^{''}: \ P\{z_i < 0\} \neq P\{z_i > 0\}$.

Make non-decreasing sample $|z_1|, \ldots, |z_n|$.
Find ranks: $s_1 = rank(|z_1|), \ldots, s_n = rank(|z_n|)$.
Calculate statistics

$$U = \sum_{i=1}^{n} \Psi_i s_i,$$

where

$$\Psi_i = \begin{cases} 1, & z_i > 0; \\ 0, & z_i < 0. \end{cases}$$

Rank-sum Wilcoxon test

wilcox.test

```
wilcox.test(x, y, paired = TRUE, alternative =
"two.sided")
```

- `x`, `y` — samples;
- `alternative` — alternatives `"two.sided"`, `"greater"`, `"less"`. For `"greater"`: $H_1' : \ P\{z_i < 0\} < P\{z_i > 0\}$.
- `paired`: `FALSE` or `TRUE`. A logical value specifying that we want to compute a paired Wilcoxon test

Spearman rank correlation coefficient

Samples: $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$.
Their ranks: $r_1, \ldots, r_n$ and $s_1, \ldots, s_n$.
Let

$$S = \sum_{i=1}^{n} (s_i - r_i)^2.$$

Spearman rank correlation coefficient:

$$\varrho = 1 - \frac{6S}{n^3 - n}.$$

And $|\varrho| \leqslant 1$.

Spearman rank correlation coefficient

Hypotheses:

- H_0: samples are independent.
- H_1: there is a positive correlation.
- H_1': there is a negative correlation.
- H_1'': samples are not independent.

cor.test

```
cor.test(x, y, method = "spearman")
```

Spearman rank correlation coefficient

result

```
 Spearman's rank correlation rho
data:  x and y
S = 10292, p-value = 1.488e-11
alternative hypothesis:  true rho is not equal to 0
sample estimates:
rho
-0.886422
```

rho is the Spearman's correlation coefficient.
The correlation coefficient between x and y are -0.8864 and the p-value is 1.48810^{-11}.

Kendall rank correlation coefficient

Sample (X_1, Y_1), …, (X_n, Y_n).
The procedure is as follow:
Begin by ordering the pairs by the X values.
Then rank the sample by Y: $z_1, \ldots, z_n$.
The new sample is

$$(1, z_1), \ldots, (n, z_n).$$

Let R is a number of inversions in a sample $\{z_1, \ldots, z_n\}$.
In a sample $(4, 3, 1, 2)$ a number of inversions is 5.
Kendall rank correlation coefficient:

$$\tau = 1 - \frac{4R}{n(n-1)}.$$

And $|\tau| \leqslant 1$.

Kendall rank correlation coefficient

cor.test

```
cor.test(x, y, method = "kendall")
```

result

```
 Kendall's rank correlation tau
data:  x and y
z = -5.7981, p-value = 6.706e-09
alternative hypothesis:  true tau is not equal to 0
sample estimates:
tau
-0.7278321
```

tau is the Kendall correlation coefficient.
The correlation coefficient between x and y are -0.7278 and the p-value is 6.70610^{-9}.