Biomedical Engineering Degree

4. Nonparametric Methods

Felipe Alonso Atienza

✉felipe.alonso@urjc.es

@FelipeURJC

Escuela Técnica Superior de Ingeniería de Telecomunicación
Universidad Rey Juan Carlos

References

1. R. Bernard. *Fundamentals of Biostatistics*. Ed.: Thompson. Chapter 9

Outline

Statistical inference taxonomy

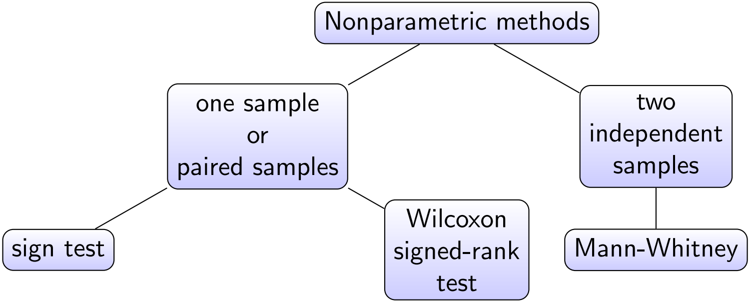
1. **Parametric methods**: require assumptions about probability distribution and associated parameters of the population
    - Normal distribution, the population mean, and standard deviation as its parameters
2. **Nonparametric methods**: do not assume that the population distribution has a particular form

When to use nonparametric methods?

1. The underlying **probability distribution may be unknown** or known to be different from what the parametric method requires
2. The **sample size may be very small** so that it's impossible to test whether parametric assumptions are met.
   - Using a parametric test when assumptions are not met may have severe effects
3. **Ordinal data** (like surveys, scales, etc.) where you cannot calculate a mean and standard deviation in a meaningful way
4. There may be **no parametric technique available** at all for the specific question at hand

Nonparametric methods equivalence

- For most parametric tests, there's an equivalent nonparamatic test.



- Instead of comparing the sample mean(s), we compare the sample **median**(s) (rank-based methods)

Rank-based methods

- Pros:
  - They work for ordinal data too
  - They are insensitive to outliers in the data
  - Robust to assumptions violation. In this case, the reported confidence intervals or significance may not be very accurate, but it won't be far off the real value
- Cons:
  - The power of a parametric test is always higher than an equivalent nonparametric test
  - Therefore, if there is a choice and assumptions are met, a parametric test is preferred

Outline

1. Introduction
2. Sign test
3. Wilcoxon Signed Rank Test
4. Mann-Whitney Test

Sign test

- Case of use: one sample, or two (paired) samples
- Not a rank-based method

Haemoglobin levels (in g/dl) were sampled from ten females vegetarians, to assess the prevalence of anemia

$$\text{sample} = [12.3, 13.1, 11.3, 10.1, 14.0, 13.3, 10.5, 12.3, 10.9, 11.9]$$

We were asked whether the **median**[a] haemoglobin level for female vegetarians is less than 13.0 g/dl

[a] median is a nonparametric measure of the center location of a distribution

- We want to test the hypothesis

$$H_0 : \eta = 13 \quad \text{vs} \quad H_1 : \eta < 13$$

- So, if the null hypothesis is true, how many observation in the sample would you expect to have a level under 13?

Sign test

- Given that we have 10 samples, and if the median value is 13, we would expect:
  - 5 observations below 13 (negative)
  - 5 observations above 13 (positive)
- What do we have in our example?
  - sample $= [12.3, 13.1, 11.3, 10.1, 14.0, 13.3, 10.5, 12.3, 10.9, 11.9]$
  - So, 7 out of 10 observations are below 13 ... seems like an extreme case, but how extreme? Can we reject H_0 in this case?
- We can answer to this question by calculating the probability of having "7 successes of more out of 10", which can be calculated using the binomial distribution

$$p = P(X \geq 7) = \texttt{binom(n = 10, p = 0.5).sf(6)} = 0.172$$

- and this p can be understood as a $p-$value. Thus,

$$\boxed{\text{We fail to reject the null hypothesis } H_0 \text{ at a significance level of } 0.05}$$

Sign test: example

- What if we had 8 out of 10 observations below 13, would we reject H_0?
- What if we had 9 out of 10 observations below 13, would we reject H_0?

- $p = P(X \geq 8) =$ `binom(n = 10, p = 0.5).sf(7)` $= 0.055$
- $p = P(X \geq 9) =$ `binom(n = 10, p = 0.5).sf(8)` $= 0.011$

Another example

A Netflix movie has been rated above 3 stars[a] by 7 out of 10 viewers, can we conclude that this is a good movie?

[a] stars range between 1 and 5

Outline

Wilcoxon Signed Rank Test

- Case of use: one sample, or two (paired) samples
- Rank-based method
- It uses the distance from the median level for comparison

Haemoglobin levels (in g/dl) were sampled from ten females vegetarians, to assess the prevalence of anemia

$$\text{sample} = [12.3, 13.1, 11.3, 10.1, 14.0, 13.3, 10.5, 12.3, 10.9, 11.9]$$

As before, we want to test the hypothesis

$$H_0 : \eta = 13 \quad \text{vs} \quad H_1 : \eta < 13$$

Wilcoxon Signed Rank Test

- We rank our observations

x	$x - \eta$	$\|x - \eta\|$	rank	signed rank
12.3	-0.7	0.7	3.5	-3.5
13.1	$+0.1$	0.1	1	$+1$
11.3	-1.7	1.7	7	-7
10.1	-2.9	2.9	10	-10
14.0	$+1.0$	1.0	5	$+5$
13.3	$+0.3$	0.3	2	$+2$
10.5	-2.5	2.5	9	-9
12.3	-0.7	0.7	3.5	-3.5
10.9	-2.1	2.1	8	-8
11.9	-1.1	1.1	6	-6

- Then, we sum the positive and the negative ranks:
  - $T(+) = 8$
  - $T(-) = 47$
- and we take **the smallest of these two** as our statistic $T = 8$

Wilcoxon Signed Rank Test

We need to compare our T statistic with a critical value T_c:

- If n is small, we can extract T_c from the exact distribution table

Wilcoxon Signed Rank (exact) Test

- if $T \leq T_c$ we reject H_0
- if $T > T_c$ we accept H_0

    - In our example $T_c = 10$, so

$$\boxed{\text{We REJECT the null hypothesis } H_0 \text{ at a significance level of } 0.05}$$

Wilcoxon Signed Rank Test

We need to compare our T statistic with a critical value T_c:

- if n is large, we can approximate T_c using the normal distribution (z score)

$$z = \frac{|T - \frac{1}{4}n(n+1)|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Wilcoxon Signed Rank Test (normal approx.)

- if $z > z_{1-\alpha}$ we reject H_0 (one-sided)
- if $z > z_{1-\alpha/2}$ we reject H_0 (two-sided)
- Otherwise, we accept H_0

  - In our example $z = 1.99$, and $z_{1-0.05} = 1.645$ so

$\boxed{\text{We REJECT the null hypothesis } H_0 \text{ at a significance level of } 0.05}$

Outline

Mann-Whitney Test

- Also known as Wilcoxon rank sum test
- Case of use: two independent samples
- Rank-based method

Haemoglobin levels (in g/dl) were sampled from ten females vegetarians and eight male vegetarians. Is there evidence of a difference in medians Hg levels?

$$\begin{aligned} \text{females} &= [10.1, 10.5, 10.9, 11.3, 11.9, 12.3, 12.3, 13.1, 13.3, 14.0] \\ \text{males} &= [10.8, 11.5, 11.8, 12.1, 12.8, 13.2, 13.5, 14.1] \end{aligned}$$

We want to test the hypothesis

$$H_0 : \eta_F = \eta_M \quad \text{vs} \quad H_1 : \eta_F \neq \eta_M$$

Mann-Whitney Test

$$\begin{aligned} \text{females} &= [10.1, 10.5, 10.9, 11.3, 11.9, 12.3, 12.3, 13.1, 13.3, 14.0] \\ \text{males} &= [10.8, 11.5, 11.8, 12.1, 12.8, 13.2, 13.5, 14.1] \end{aligned}$$

- First, we rank our data from smallest to largest

$$\begin{aligned} \text{females} &= [1, 2, 4, 5, 8, 10.5, 10.5, 13, 14, 17] \\ \text{males} &= [3, 6, 7, 9, 12, 14, 16, 18] \end{aligned}$$

- Then, we calculate the sum of the ranks
  - $T_F = 86$
  - $T_M = 85$
- and the **one with the fewer observations** constitutes our T statistic.

Mann-Whitney Test

- The expecting value of T, given that the null hypothesis is true is given by

$$\mathsf{E}[T] = \frac{1}{2}n_1(n_1 + n_2 + 1) = \frac{1}{2}8(8 + 10 + 1) = 76$$

where n_1 is the number of observations in the smaller sample, and n_2 is the number of observation in the larger sample.

So, is 76 extreme enough, compared to 85, to reject our null hypothesis?

Mann-Whitney Test

- If n_1, n_2 is small, we can compare T to the exact distribution table

Mann-Whitney (exact) Test

- if T lies out the interval of critical values, then we reject H_0
- Otherwise, we accept H_0

  - In our example $85 \in (53, 99)$, so

$$\boxed{\text{We ACCEPT the null hypothesis } H_0 \text{ at a significance level of } 0.05}$$

Mann-Whitney Test

- If n_1, n_2 is large, we can approximate using the normal distribution (z score)

$$z = \frac{|T - \mathsf{E}[T]|}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}}$$

Wilcoxon Signed Rank Test (normal approx.)

- if $z > z_{1-\alpha}$ we reject H_0 (one-sided)
- if $z > z_{1-\alpha/2}$ we reject H_0 (two-sided)
- Otherwise, we accept H_0

  - In our example $z = 0.8$, and $z_{1-0.025} = 1.96$ so

$\boxed{\text{We ACCEPT the null hypothesis } H_0 \text{ at a significance level of } 0.05}$