Statistics for EES and MEME
5. Rank-sum tests

Dirk Metzler

April 7, 2026

Wilcoxon's rank sum test

is also called

Mann-Whitney U test

References

[1] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1**:80–83.

[2] Mann, H. B., Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18**:50–60.

Contents

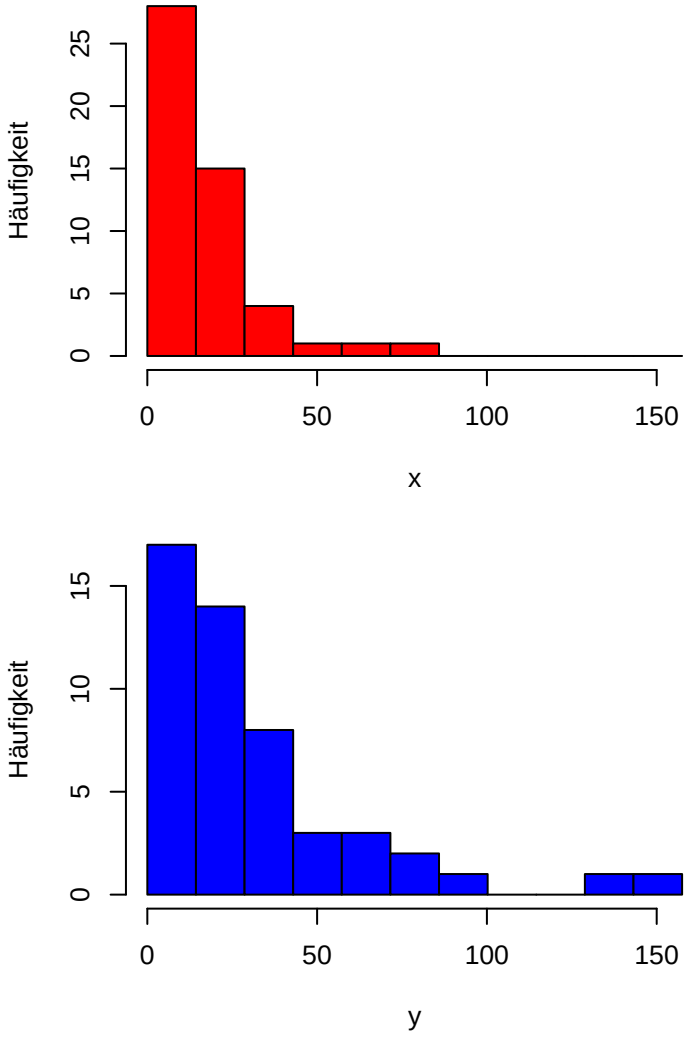
Contents

1 Motivation

For (more or less) bell-shaped symmetrically distributed observations
(or for sufficiently large sample sizes)
we can apply the t-Test for the null hypothesis $\mu_1 = \mu_2$:
The t-statistic is then nearly Student-t-distributed.

This can be different especially for very asymmetric long-tailed distributions

Assume that we have to compare these distributions:



Examples

- Waiting times
- Distances of dispersion
- Frequencies of cell types

Wanted: a test that does not assume a distribution

Such tests are called *non-parametric*

2 Wilcoxon-test for independent samples

Observations: Two samples

$$X : x_1, x_2, \ldots, x_m$$

$$Y : y_1, y_2, \ldots, y_n$$

Test the *null hypothesis*: that X and Y come from the same population

alternative: X "typically larger" than Y or Y "typically larger" than X

Idea

Observations:

$$X : x_1, x_2, \ldots, x_m$$

$$Y : y_1, y_2, \ldots, y_n$$

- Sort all observations by size.

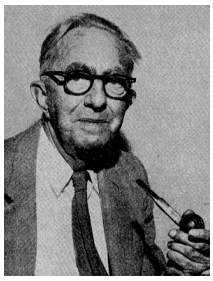
- Determine the *ranks* of the m X-values among all $m+n$ observations.
- If the null hypothesis is true, than the m X-ranks are randomly chosen from $\{1, 2, \ldots, m+n\}$.
- Compute the sum of the X-ranks and check if it is untypically small or large compared to sum of random ranks.

Wilcoxon's rank-sum statistic

Observation:

$X : x_1, x_2, \ldots, x_m$

$Y : y_1, y_2, \ldots, y_n$



Frank Wilcoxon, 1892-–1965

$$W = \text{Sum of the } X\text{-ranks} - (1 + 2 + \cdots + m)$$

is called

Wilcoxon's rank-sum statistic

Wilcoxon's rank-sum statistic

Note:

$$W = \text{Sum of the } X\text{-ranks} - (1 + 2 + \cdots + m)$$

We could also use the sum of the Y-ranks, because

$$\begin{aligned} &\text{Sum of the } X\text{-ranks} + \text{Sum of the } Y\text{-ranks} \\ &= \text{Sum of all ranks} \\ &= 1 + 2 + \cdots + (m+n) = \frac{(m+n)(m+n+1)}{2} \end{aligned}$$

A *small* example

- Observations:

$$\begin{aligned} X &: 1.5,\ 5.6,\ 35.2 \\ Y &: 7.9,\ 38.1,\ 41.0,\ 56.7,\ 112.1,\ 197.4,\ 381.8 \end{aligned}$$

- Pool observations and sort: 1.5, 5.6, 7.9, 35.2, 38.1, 41.0, 56.7, 112.1, 197.4, 381.8
- Determine ranks: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
- rank-sum: $W = 1 + 2 + 4 - (1 + 2 + 3) = 1$

Interpretation of W

X-Population smaller $\implies$ W small:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $\quad W = 0$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $\quad W = 1$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $\quad W = 2$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $\quad W = 2$

X-population larger $\implies$ W large:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $\quad W = 21$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $\quad W = 20$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $\quad W = 19$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $\quad W = 19$

Significance

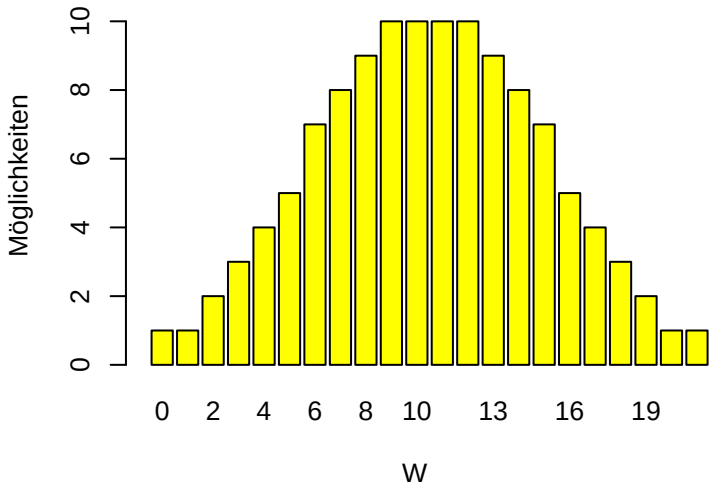
Null hypothesis:
X-sample and Y-sample were taken from the same distribution

The 3 ranks of the X-sample 1 2 3 4 5 6 7 8 9 10
could just as well have been any 3 ranks 1 2 3 4 5 6 7 8 9 10

There are $\frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$ possibilities.

(In general: $\frac{(m+n)(m+n-1)\cdots(n+1)}{m(m-1)\cdots 1}) = \frac{(m+n)!}{n!m!} = \binom{m+n}{m}$ possibilities)

Distribution or the Wilcoxon statistic $(m = 3, n = 7)$[1ex]



Under the null hypothesis all rank configurations are equally likely, thus

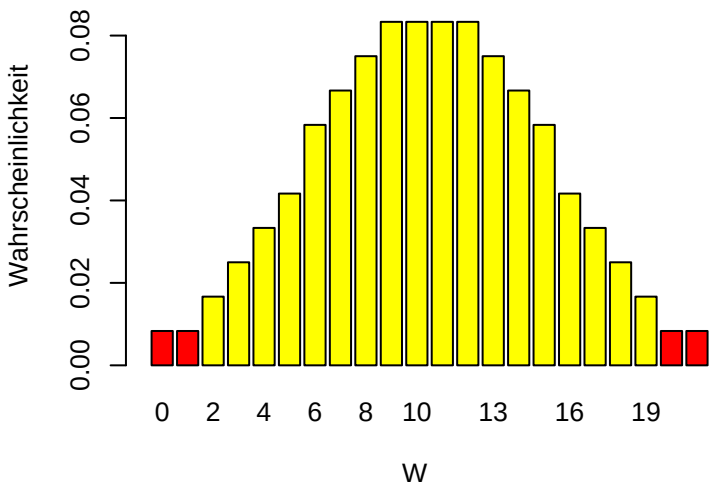
$$\mathbb{P}(W = w) = \frac{\text{number of possibilities with rank-sum statistic } w}{120}$$

We see in our example: 1.5, 5.6, 7.9, 35.2, 38.1, 41.0, 56.7, 112.1, 197.4, 381.8

$$W = 1$$

$$\begin{aligned}\mathbb{P}(W \leq 1) + \mathbb{P}(W \geq 20) &= \mathbb{P}(W = 0) + \mathbb{P}(W = 1) + \mathbb{P}(W = 20) + \mathbb{P}(W = 21) \\ &= \tfrac{1+1+1+1}{120} \doteq 0.033\end{aligned}$$

Distribution of the Wilcoxon statistic $(m = 3, n = 7)$[1ex]



For our example ($W = 1$):

$$p\text{-value} = \mathbb{P}(\text{such an extreme } W) = 4/120 = 0.033$$

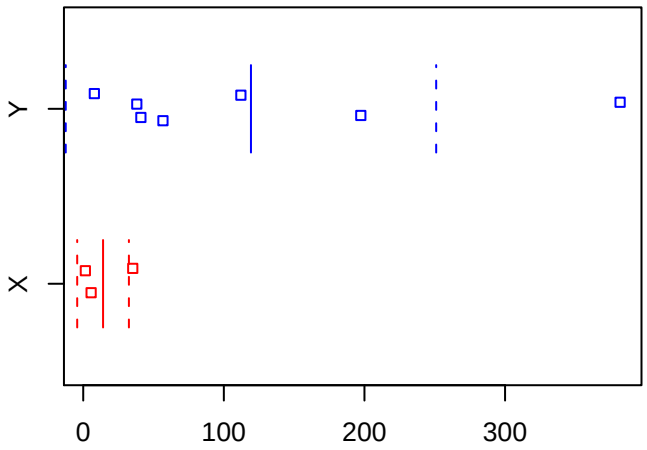
We *reject* the *null hypothesis*, that the distributions of X and Y were equal, on the 5%-level.

Wilcoxon test in R with `wilcox.test`:

```
> x
[1]  1.5  5.6 35.2
> y
[1]   7.9  38.1  41.0  56.7 112.1 197.4 381.8
> wilcox.test(x,y)

      Wilcoxon rank sum test

data:  x and y
W = 1, p-value = 0.03333
alternative hypothesis: true location shift is
not equal to 0
```

Let's compare to the t-Test:

```
> x
[1]  1.5  5.6 35.2
> y
[1]   7.9  38.1  41.0  56.7 112.1 197.4 381.8
> t.test(x,y)

     Welch Two Sample t-test

data:  x and y
t = -2.0662, df = 6.518, p-value = 0.08061
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -227.39182   17.02039
sample estimates:
mean of x mean of y
  14.1000  119.2857
```



3 Application to Hipparion tooth measurements

Remember Welch's t-test

We examine one trait in two populations:

$$\begin{array}{cc} \text{Population} & 1 \quad 2 \\ \text{mean} & \mu_1 \quad \mu_2 \end{array}$$

Null hypothesis: $\mu_1 = \mu_2$

We draw samples from the two populations with means $\overline{x}_1 \quad \overline{x}_2$

To check the null hypothesis H_0, we compute the *t-Statistic*[1ex]

$$t = \frac{\overline{x}_1 - \overline{x}_2}{f} \quad \text{with} \quad f^2 = \left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2$$

p-value under H_0: $p \approx \mathbb{P}(|T_g| \geq |t|)$ (g=(estimated) degrees of freedom depends on n_1, n_2, s_1, s_2)

If the distributions are not close to normal, *Wilcoxon's rank-sum Statistic* can be used.

```
> t.test(md[Art=="africanum"],md[Art=="libycum"])

     Welch Two Sample t-test

data:  md[Art == "africanum"] and md[Art == "libycum"]
t = -3.2043, df = 54.975, p-value = 0.002255
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.1025338 -0.9453745
sample estimates:
mean of x mean of y
 25.91026  28.43421

> wilcox.test(md[Art=="africanum"],md[Art=="libycum"])

     Wilcoxon rank sum test with continuity correction

data:  md[Art == "africanum"] and md[Art == "libycum"]
W = 492, p-value = 0.01104
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(md[Art == "africanum"], md[Art == "libycum"]) :
  kann bei Bindungen keinen exakten p-Wert Berechnen
```

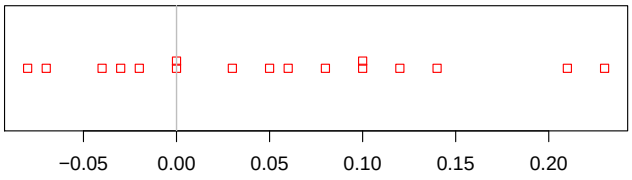
IMPORTANT!

Is the Wilcoxon test really appropriate here? Not clear because its null hypothesis is that the data come from the same distribution, not just that the means are equal. If we want to test whether the means are different but allow the standard deviations to be different (like in the assumptions of Welch's t-test), the Wilcoxon test cannot be applied!

4 Wilcoxon signed-rank test

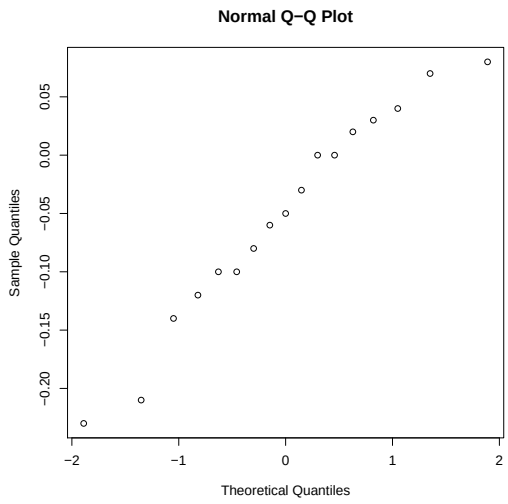
Remember the pied flycatchers?

$$x := \text{“green length”} - \text{“blue length”}$$



t-Test with R

```
> x <- length$green-length$blue
> t.test(x)

        One Sample t-test

data:  x
t = 2.3405, df = 16, p-value = 0.03254
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.004879627 0.098649784
sample estimates:
 mean of x
0.05176471
```



Wilcoxon test for paired samples

Wilcoxon signed rank test:

given independent samples $X_1, X_2, \ldots, X_n$ from an unknown distribution.

null hypothesis: distribution of X_i is symmetric around 0.0.

Application: paired samples $(Y_1, Z_1), \ldots, (Y_n, Z_n)$. null hypothesis: Y_i and Z_i have same distribution.

Apply Wilcoxon signed rank test to

$$X_1 = Y_1 - Z_1, X_2 = Y_2 - Z_2, \ldots, X_n = Y_n - Z_n.$$

`wilcox.test(flycatchers$blue,flycatchers$green, paired=TRUE)` and `wilcox.test(flycat` both give:

```
     Wilcoxon signed rank test with continuity correction

data:  flycatchers$blue and flycatchers$green
V = 22.5, p-value = 0.03553
alternative hypothesis: true location shift is not equal to 0

Warning messages:
1: In wilcox.test.default(flycatchers$blue, flycatchers$green, paired = TRUE) :
  kann bei Bindungen keinen exakten p-Wert Berechnen
2: In wilcox.test.default(flycatchers$blue, flycatchers$green, paired = TRUE) :
  kann den exakten p-Wert bei Nullen nicht berechnen
```

statistic: sum of signed ranks.

signed rank: take rank of absolute value, equipped with sign of original value

example:

values:	-2.6	-2.5	-2.3	-1.3	-0.6	1.6	2.2	6.1
absolute values:	2.6	2.5	2.3	1.3	0.6	1.6	2.2	6.1
ranks:	7	6	5	2	1	3	4	8
signed ranks:	-7	-6	-5	-2	-1	3	4	8

Test statistik: $3 + 4 + 8 = -6$

Some of what you should be able to explain

- When should you apply non-parametric tests?
- Wilcoxon test: requirements and null hypothesis.
- The statistics of the Wilcoxon test and the signed-rank test.
- How to calculate the p value for the Wilcoxon test.