ST 516: Foundations of Data Analytics

Wilcoxon Rank-Sum Test

Wilcoxon Rank-Sum Test: Introduction

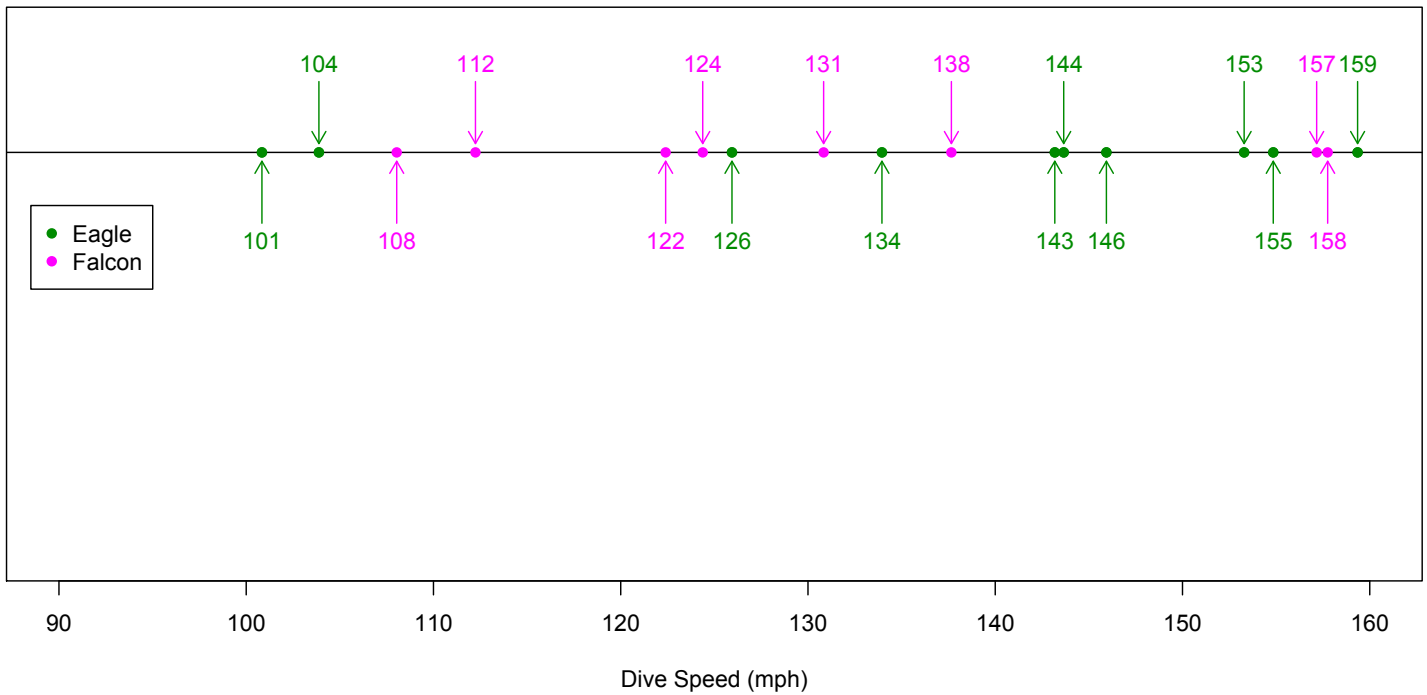
The Wilcoxon rank-sum test is used to assess whether there is a 'difference' between two population distributions. We will clarify what kind of 'differences' the test detects at the end of this lecture.

This test is known by several other names: Rank-sum test; Wilcoxon test; Mann-Whitney U test; Wilcoxon-Mann-Whitney test.

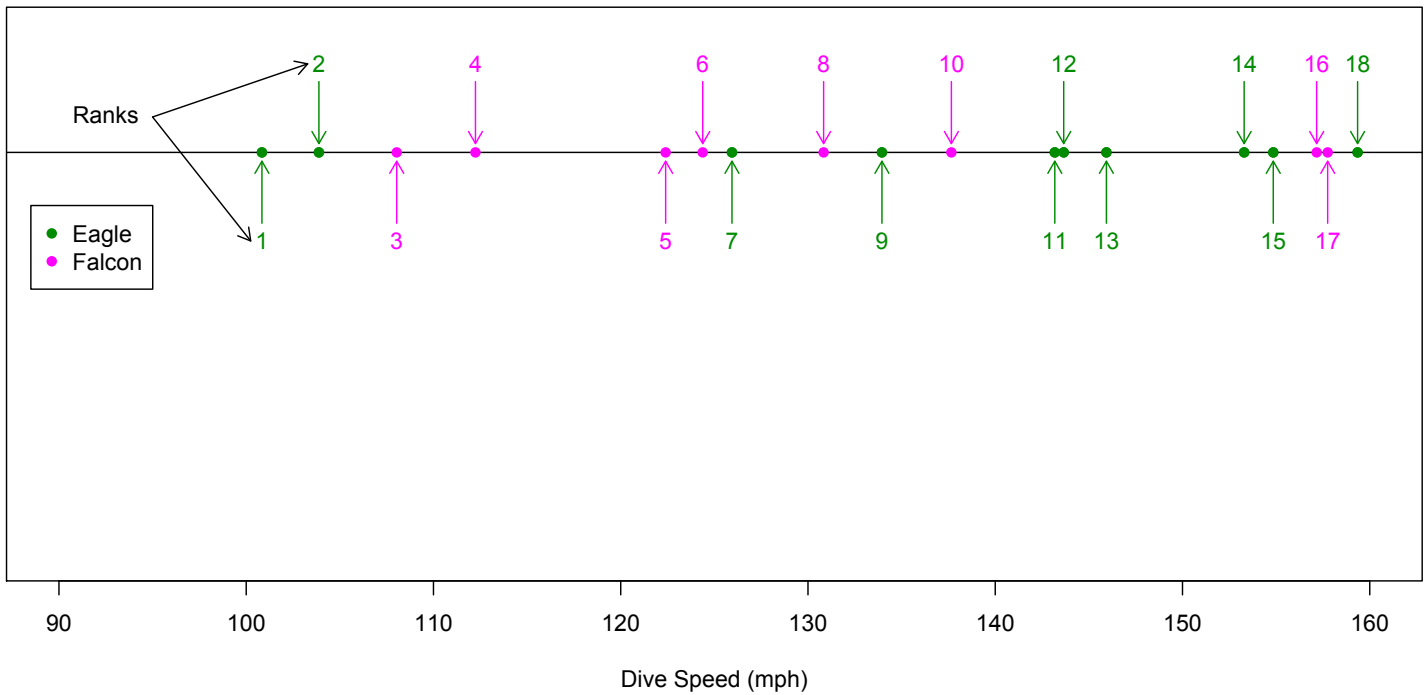
Wilcoxon rank-sum test summary:

- Setting: Two independent samples of independent observations
  - Sample of size n_1 from population 1: $X_{11}, X_{12}, \ldots, X_{1n_1}$
  - Sample of size n_2 from population 2: $X_{21}, X_{22}, \ldots, X_{2n_2}$
- Procedure:
  1. Combine the samples (keeping track of which observation comes from which sample)
  2. Rank the observations from smallest to largest
  3. Add up the ranks that correspond to observations in the smaller group to obtain the test statistic T.

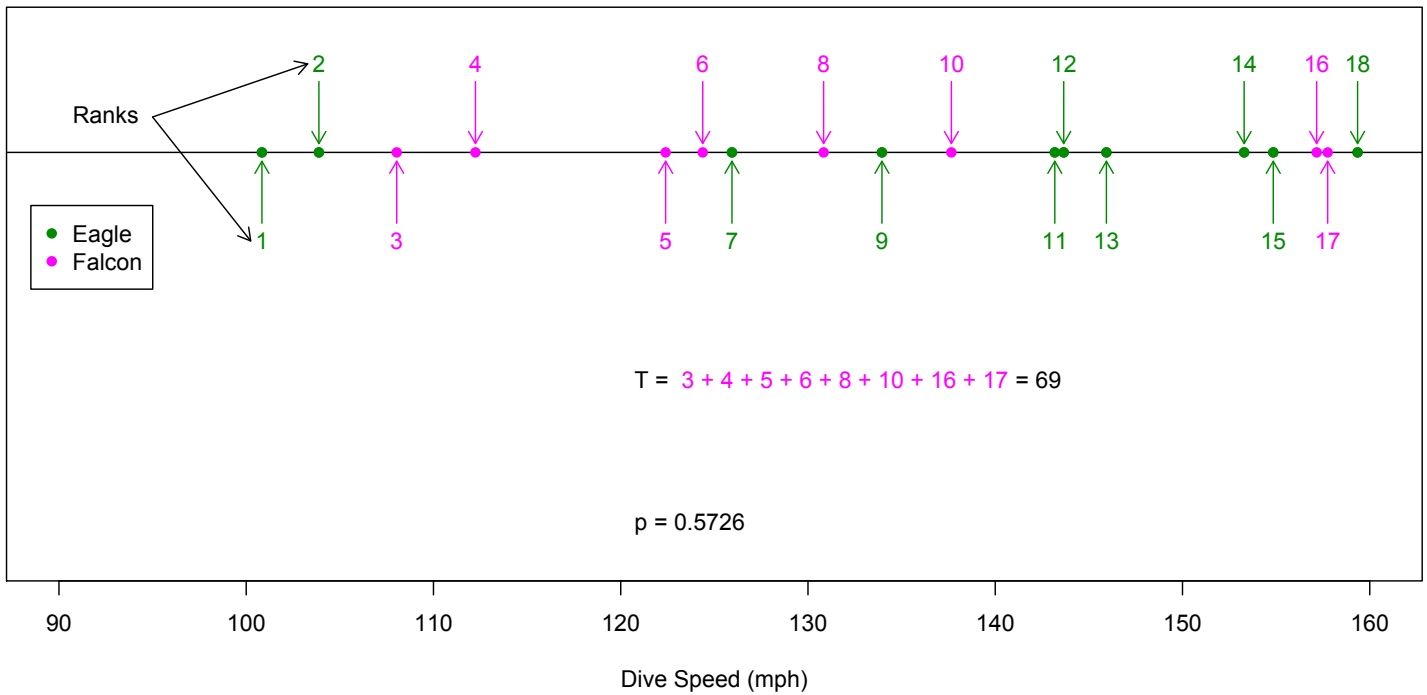
Wilcoxon Rank-Sum Test: Example



Wilcoxon Rank-Sum Test: Example



Wilcoxon Rank-Sum Test: Example



Reference Null Distribution

How do we decide whether a resulting rank-sum test statistic is 'significant'? How do we compute p-values?

We use a *permutation* approach to decide how unusual an observed test statistic is if really all of the observations came from the same distribution.

- It is unlikely that all of the small observations (or all of the large observations) would belong to Group 1 if the two populations have the same distribution.
- *Permute* the group assignments and recompute the test statistic for each permutation to obtain the null distribution.

Reference Null Distribution: Example

Suppose we observe the following data:

Group 1: $4, 17, 12$

Group 2: $13, 18, 15, 20$

1. What is the Wilcoxon rank-sum test statistic T?
2. What is the reference null distribution for T?
   That is, what is the distribution of the values of T if each rank has the same chance of being assigned to Group 1?

Reference Null Distribution: Example

Group	Observed value	Rank
1	4	1
1	17	5
1	12	2
2	13	3
2	18	6
2	15	4
2	20	7

$T = 1 + 5 + 2 = 8$

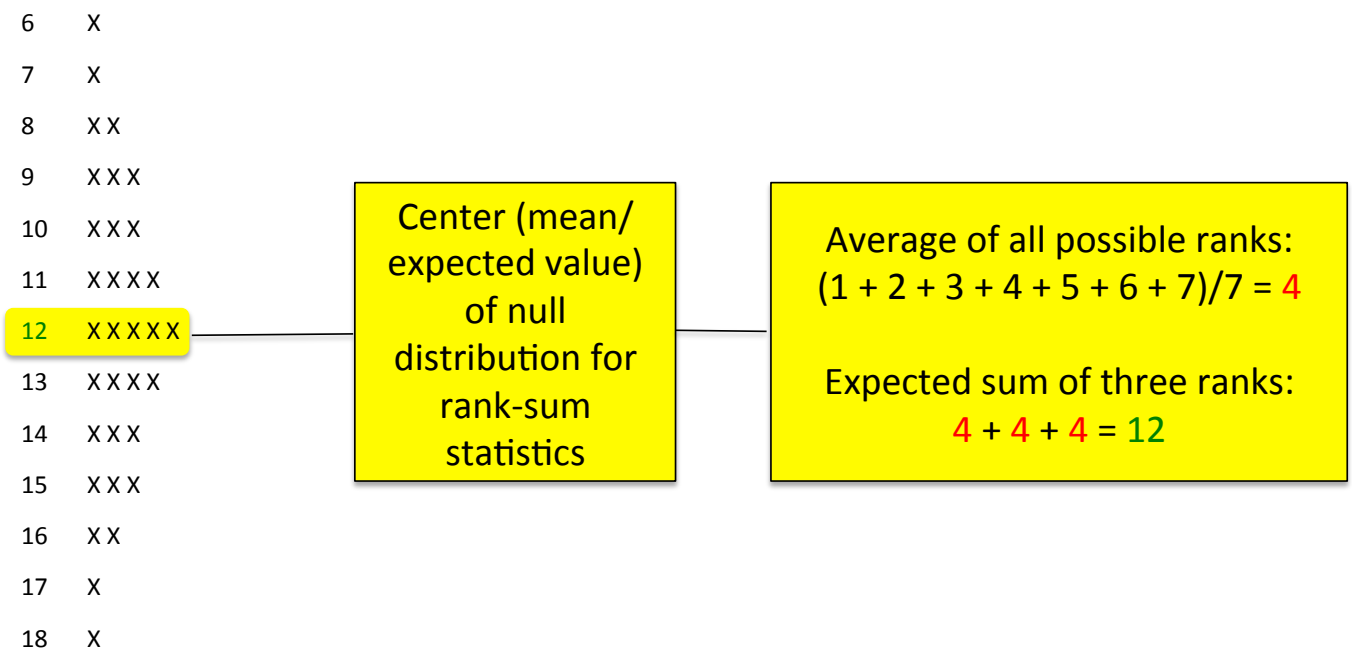
Reference Null Distribution: Example

To obtain the reference null distribution, we list all possible ways that three ranks could be chosen from the ranks $1,\ldots,7$ for Group 1, and we find the corresponding value of T_{perm}:

	T_{perm}		T_{perm}		T_{perm}		T_{perm}		T_{perm}
1, 2, 3	6	1, 3, 6	10	1, 6, 7	14	2, 4, 7	13	3, 5, 6	14
1, 2, 4	7	1, 3, 7	11	2, 3, 4	9	2, 5, 6	13	3, 5, 7	15
1, 2, 5	8	1, 4, 5	10	2, 3, 5	10	2, 5, 7	14	3, 6, 7	16
1, 2, 6	9	1, 4, 6	11	2, 3, 6	11	2, 6, 7	15	4, 5, 6	15
1, 2, 7	10	1, 4, 7	12	2, 3, 7	12	3, 4, 5	12	4, 5, 7	16
1, 3, 4	8	1, 5, 6	12	2, 4, 5	11	3, 4, 6	13	4, 6, 7	17
1, 3, 5	9	1, 5, 7	13	2, 4, 6	12	3, 4, 7	14	5, 6, 7	18

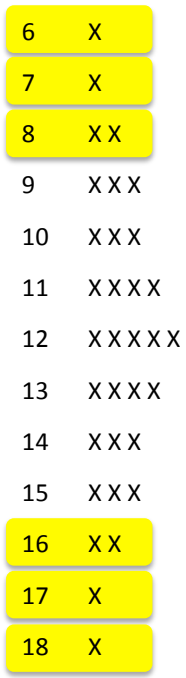
Reference Null Distribution: Example

The resulting reference null distribution is displayed as a sideways barplot here:



Reference Null Distribution: Example

Reference distribution:



How likely is it that *if the two populations were identical* we would see a value at least as *extreme* as the observed test statistic value $T = 8$?

Here, as usual, 'extreme' means far from the center of the reference distribution, which is the expected value of T: $\mathrm{E}(T) = 12$.

- There are 8 permutations that lead to a rank-sum statistic at least as far from 12 as $T = 8$.
- There are 35 total ways to pick 3 ranks from 7 observations.
- Therefore, the p-value corresponding to this observation $T = 8$ is $8/35 = 0.2286$.

Reference Null Distribution for Large Samples

When sample sizes are large, it becomes very time-consuming to enumerate all possible ways to permute the ranks.

We have a few options to find the reference distribution with larger sample sizes:

- Use tabled critical values based on the permutation distribution–available for reasonably small sample sizes, or
- Select a random sample from all possible permutations of ranks, or
- Use a Normal approximation to the reference distribution, as given on the next few slides.

Normal Approximation to Null Distribution

The null distribution of the rank-sum statistic T can be approximated as

$$T \dot{\sim} \text{Normal}\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right)$$

We can use this normal approximation to *standardize* T (subtract its mean, divide by the square-root of its variance) and compare the resulting z-statistic to a standard normal distribution:

$$Z = \frac{T - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}} \dot{\sim} \text{Normal}(0,1)$$

Normal Approximation to Null Distribution: Example

Suppose we observe the following data:

$$\begin{aligned} &\text{Group 1:} \quad 4, 17, 12 \\ &\text{Group 2:} \quad 13, 18, 15, 20 \end{aligned}$$

Then we compute the z-statistic as follows:

$$\begin{aligned} \mathsf{E}(T) &= \frac{n_1(n_1+n_2+1)}{2} = \frac{3(8)}{2} = 12 \\ \mathsf{Var}(T) &= \frac{n_1 n_2(n_1+n_2+1)}{12} = \frac{4(3)(8)}{12} = 8 \\ Z &= \frac{T-\mathsf{E}(T)}{\sqrt{\mathsf{Var}(T)}} = \frac{8-12}{\sqrt{8}} = -1.414 \end{aligned}$$

Normal Approximation to Null Distribution: Example

This z-statistic of -1.414 corresponds to a p-value of

```
2*(1 - pnorm(abs(-1.414)))
```

```
##  0.1573
```

Not that this is slightly different from the p-value computed using the exact permutation distribution (0.2286). For small sample sizes like this, we can expect some difference between the exact and normal approximation p-values; for larger sample sizes the p-values will be closer.

Wilcoxon Rank-Sum Test: Other Issues

Some other issues that arise in performing the Wilcoxon rank-sum test:

- How do we handle ties in observed values: how should we rank tied values?
    - And how do ties affect the reference distribution?
- Normal approximation in small samples: continuity correction
- **What question have we answered?** What is the specific null hypothesis that is rejected or not rejected?

Tied Observations

We handle ties using the following approach:

1. Assign ranks to observations as usual, arbitrarily deciding which tied observation gets the smaller rank
2. Average the ranks assigned to tied values

Example:

Group	1	1	2	2	2	1	2	2	1
Observed Value	3	6	6	8	9	12	18	18	18
Rank	1	2	3	4	5	6	7	8	9

Tied Observations

We handle ties using the following approach:

1. Assign ranks to observations as usual, arbitrarily deciding which tied observation gets the smaller rank
2. Average the ranks assigned to tied values

Example:

Group	1	1	2	2	2	1	2	2	1
Observed Value	3	6	6	8	9	12	18	18	18
Rank	1	2	3	4	5	6	7	8	9
Adjusted Rank	1	2.5	2.5	4	5	6	8	8	8

Tied Observations

We handle ties using the following approach:

1. Assign ranks to observations as usual, arbitrarily deciding which tied observation gets the smaller rank
2. Average the ranks assigned to tied values

Example:

Group	1	1	2	2	2	1	2	2	1
Observed Value	3	6	6	8	9	12	18	18	18
Rank	1	2	3	4	5	6	7	8	9
Adjusted Rank	**1**	**2.5**	2.5	4	5	**6**	8	8	**8**

Test statistic value: $T = 1 + 2.5 + 6 + 8 = 17.5$

Reference Null Distribution with Ties

The reference distribution for T changes if there are tied observations:

- The permutation approach still works
  - Permutations of ranks preserve the tied ranks
  - ...But tabled values of the reference distribution critical values *will not* be correct, because these tables assume no ties
- If samples sizes are *large* and the number of ties is *small*, the normal approximation will still be approximately valid.
- However, if the number of ties is large relative to the sample sizes, the normal approximation will not be very good, and should not be used.

Continuity Correction

Some sources recommend a 'continuity correction' to the normal approximation for the distribution of T:

- Add 0.5 to the observed value of T if you are computing a lower probability (that is, use `pnorm(T - 0.5)`).
- Subtract 0.5 from the observed value of T if you are computing an upper probability (that is, use `1 - pnorm(T + 0.5)`).
- Most of the time, we are computing a *two-sided* p-value, so we would use the corrected p-value `2*(1 - pnorm(abs(T) + 0.5))`.

The continuity correction slightly improves the normal approximation for small sample sizes, and the default in the R function `wilcox.test()` is to use the continuity correction. See the *Statistical Sleuth*, pp. 92 – 93 for more discussion.

What Does the Rank-Sum Test Answer?

The most important question for us to address regarding the Wilcoxon Rank-Sum test is:

What question do we answer when we perform a Wilcoxon Rank-Sum test?

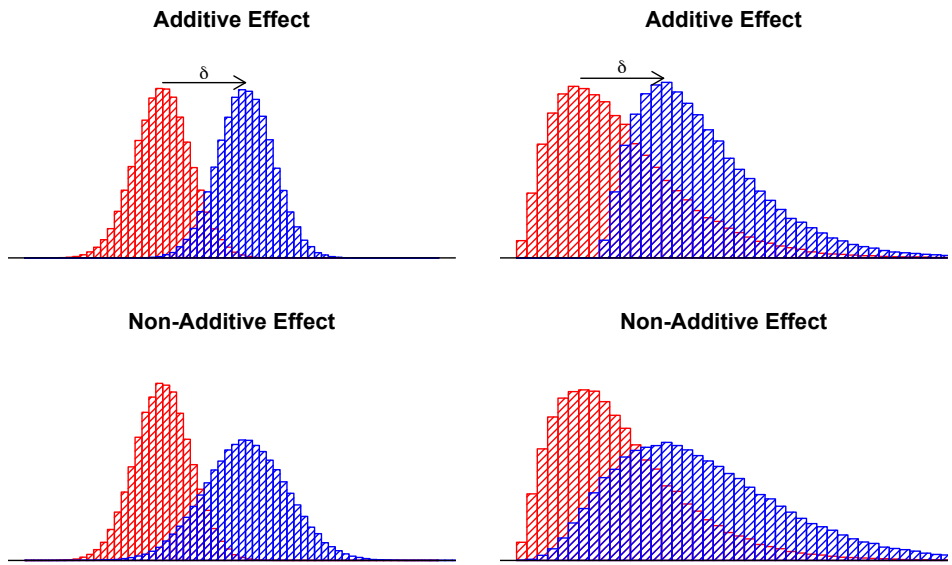
Recall that the t-test answers questions about the value or differences in values of population means, and the sign test and Mood's test answer questions about the value(s) of the population median(s).

It is a bit more complicated to specify what question the Wilcoxon Rank-Sum test answers: we need to consider whether or not a particular assumption is reasonable.

What Does the Rank-Sum Test Answer?

Some sources assume that the only possible difference between the two populations is an **additive effect**.

- An **additive effect** means that the population distribution for Group 1 is just a shift of the population distribution for Group 1.
- This means that the *shapes* and *scales* of the distributions cannot change at all—only the *location* differs between the two populations.
- This is also sometimes referred to as a **location shift** hypothesis (or just *shift*).



What Does the Rank-Sum Test Answer?

You will likely encounter people, references, and literature that claim that the Wilcoxon Rank-Sum test is a test that compares population medians. **THIS IS NOT ENTIRELY TRUE.**

- If you *are assuming the additive effect model*, then the location shift of δ between the two populations is a difference in medians...but it is also a difference in means, 95th percentiles, 10th percentiles, maxima, minima...any measure of location.
- If you *are not assuming the additive effect*, then the Wilcoxon does not tell you anything about a comparison of medians. Instead, it can be considered a test of the hypothesis $H_0 : P(X > Y) = \frac{1}{2}$ where X is a randomly chosen value from population 1, and Y is a randomly chosen value from population 2.