NKB entry report analysis Apr 2025

y_bib

Outlines of NKB swimmers' performance

Introduction

There is a very well-known fact that male swimmers swim faster than female swimmers. And there is a well known fact that Canadian female swimmers win sufficiently more Olympic medals compared to male swimmers.

We will check what are the trends in the Ottawa biggest chieldrem swimming club named NKB. And also see how age and style preferences effect results. To (possibly:) predict our future winners we will cluster swimmers based on all their results (properly normalized) and see who appeared in the first cluster (but no names here :)).

Data source: NKB SwimMeets Apr 2025 entry report (1470 records, 233 swimmers)

Tools: R for data manipulations, Quarto for report creation

Stats you will see

- Gender and age structure of NKB swimmers
- Plots and box plots describing performance depending on gender, style, distance and their combinations
- Clastering of swimmers based on their best time results
- Hypothesis testing using Shapiro test, Mann-Whitney U test, Permutation test.

Outlines:

- **Gender and Age Structure:** The club has a diverse age range, with swimmers from "10 & Under" to "15 & Over". Male and females swimmers present in all age groups. See **Part 1**.
- **Distance and Style Preferences:** Distances of 400m and more are predominantly assigned to older swimmers (12+), while 50m distances are for younger swimmers (12 and under), with the exception of 50m Free style. 100m Free style is a very popular combination.
- **Free style** consistently appears to be the fastest swimming style across all distances. Part 1.
- **Clustering** of scaled data allows to find the better performing swimmers among all ages.
- **Male swimmers generally exhibit better average and best times across most distances and styles.** That concluded based on **clustering**: The top performance cluster contains significantly more male swimmers (19 in "15 & Over") compared to female swimmers (5 in "15 & Over") - see **Part 2** for details. Series of **hypotheses** testing support the same conclusion - see **Part 3** for details.

Methods

Raw data organization; descriptive statistics; scaling and commutative description for clustering, Shapiro test, permutation test, Mann-Whitney U test.

Part 1

Gender and age structure of NKB swimmers.

Total Swimmers by Age Group and Gender

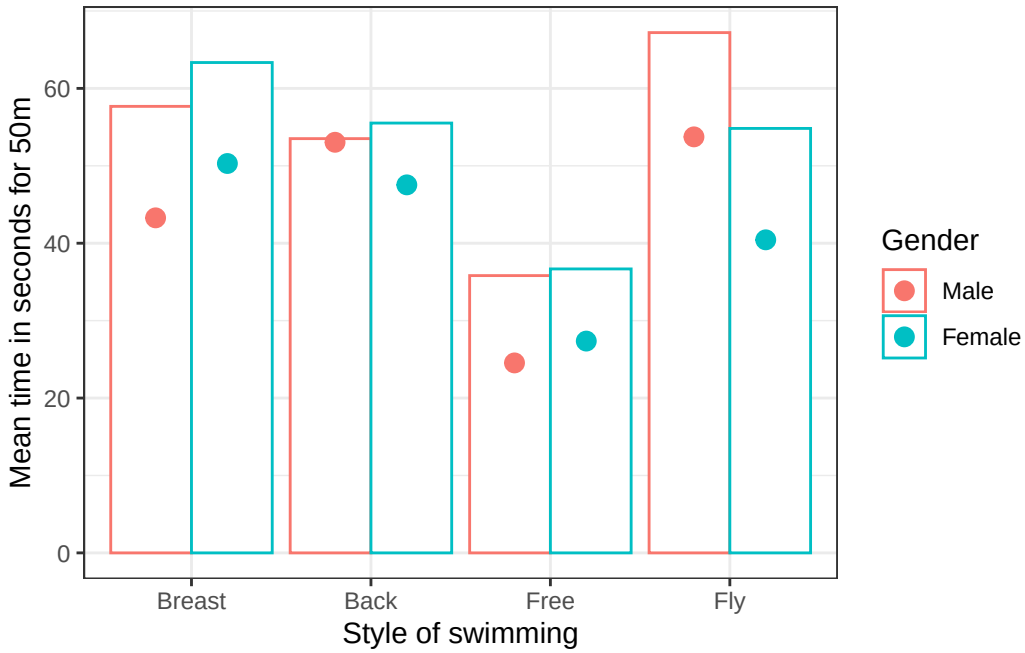
AgeGroup	Male	Female
10&Under	169	153
11-11	74	100
12-12	84	186
13-13	123	93
13-14	22	12
14-14	60	60
15&Over	198	136

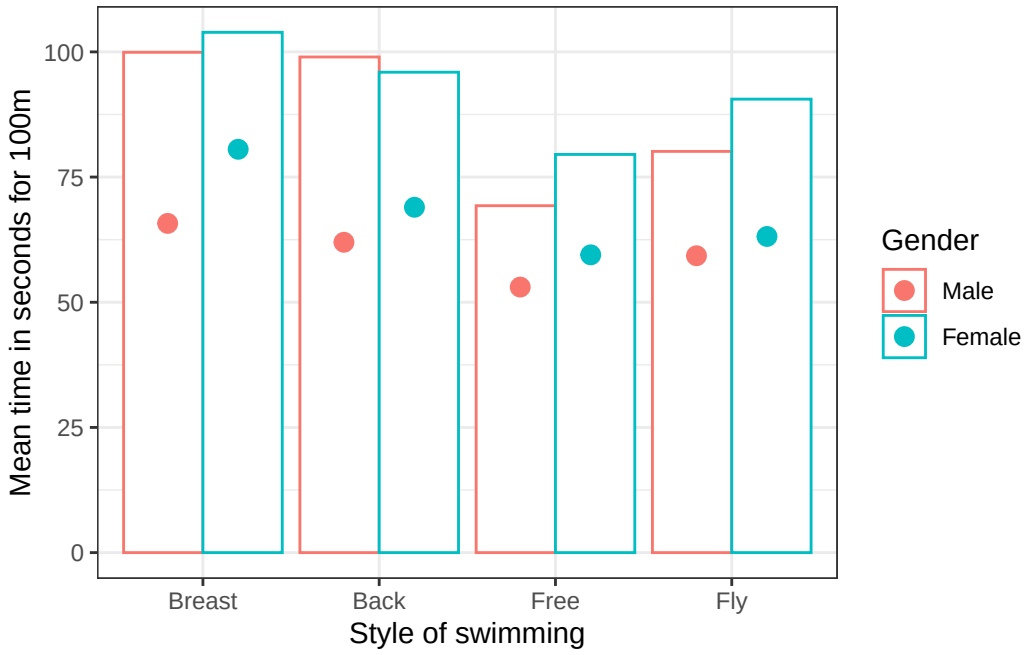
The table below shows what combinations of distance and style are assigned to the swimmers belonging to different age groups. Distances 400m and more are mostly assigned to the swimmers 12 y.o. and older while 50m distances are assigned younger swimmers (12 y.o. or younger) with the exception for 50 m Free style. One more popular combination is 100 m Free style.

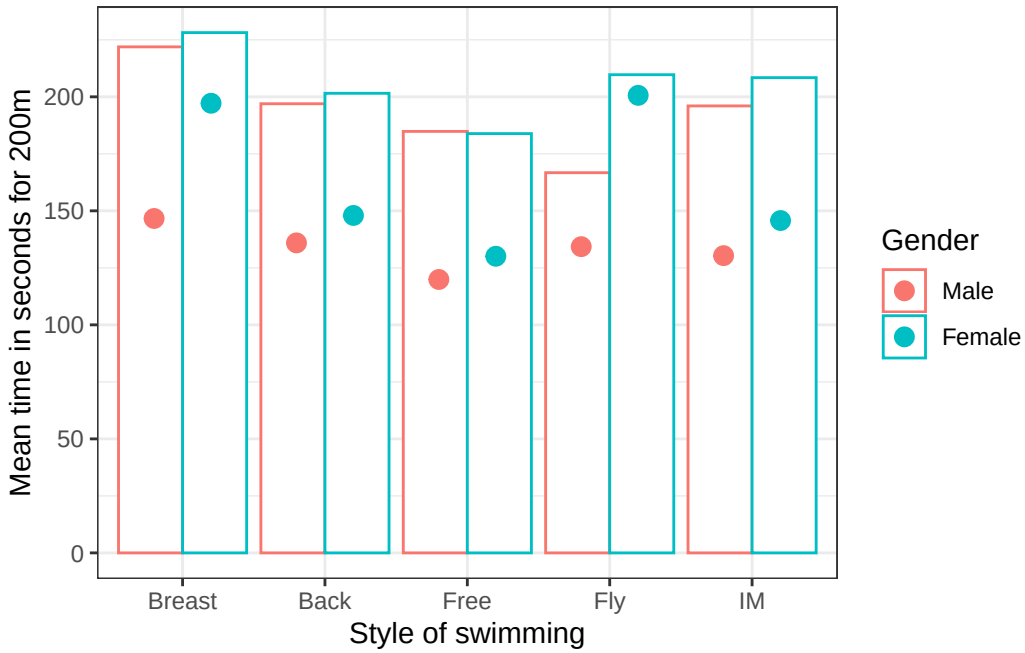
Total Swimmers by Age Group, Distance, and Style

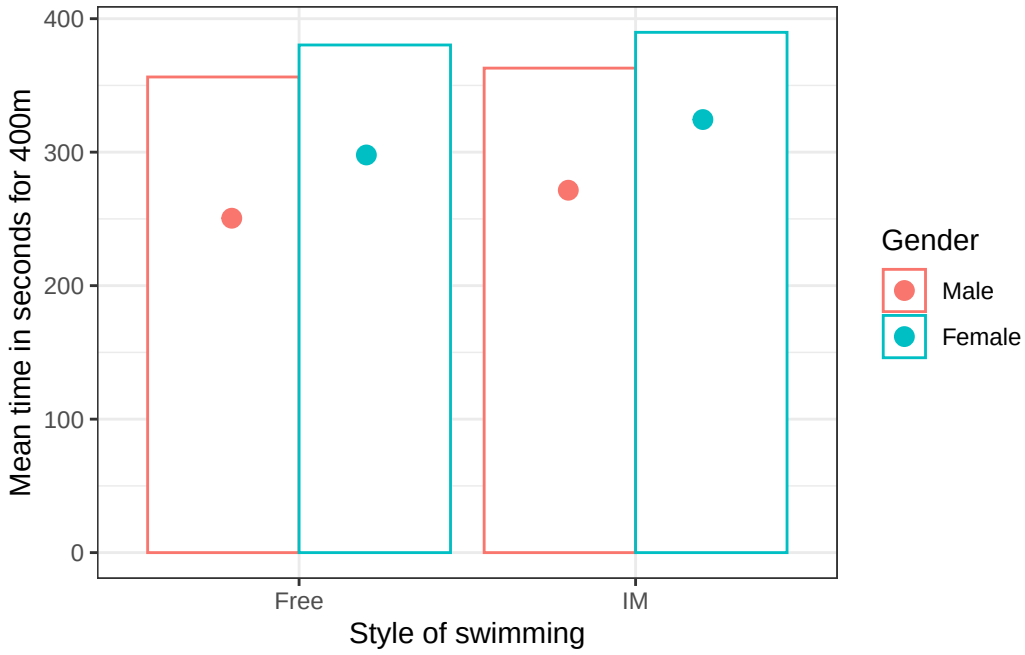
Distance	Style	10&Under	11-11	12-12	13-13	13-14	14-14	15&Over
50	Breast	13	5	6	0	0	0	0
50	Back	27	6	4	0	0	0	0
50	Free	33	20	23	18	0	7	27
50	Fly	25	5	9	0	0	0	0
100	Breast	30	21	31	30	0	12	26
100	Back	32	18	27	11	0	4	25
100	Free	49	23	36	26	34	12	61
100	Fly	3	2	15	19	0	13	20
200	Breast	24	9	16	7	0	7	12
200	Back	21	10	15	11	0	6	25
200	Free	33	21	32	17	0	7	29
200	IM	20	15	18	28	0	16	33
400	Free	5	9	21	27	0	14	18
400	IM	6	3	6	17	0	11	28
800	Free	1	5	9	4	0	5	13
200	Fly	0	2	2	1	0	4	10
1500	Free	0	0	0	0	0	2	7

The diagrams below show male and female NKB swimmers performing in different combinations of distance and style. On each diagram, the columns show mean time (average time among each gender,) and points show the best time depending on style and gender for the chosen distance.


Mean time in seconds for 50m
60
40
20
0
Breast
Back
Free
Fly
Style of swimming
Gender
Male
Female


Mean time in seconds for 100m
100
75
50
25
0
Breast
Back
Free
Fly
Style of swimming
Gender
Male
Female

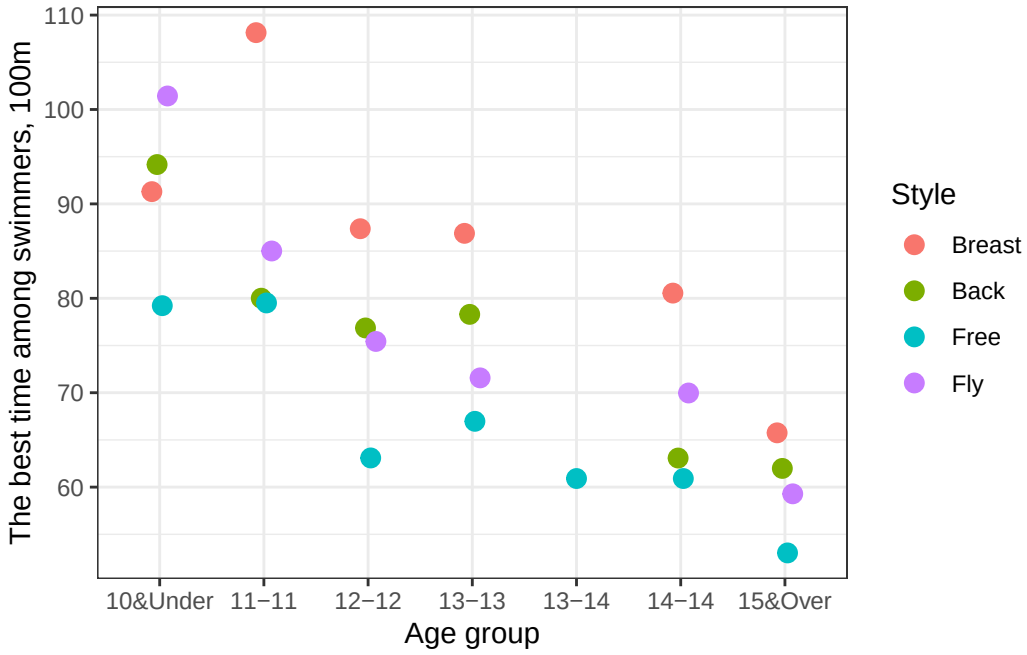




Conclusion: We can conclude that male swimmers show better results, though Free style allows girls to minimize the gap on 50m and on 200m.

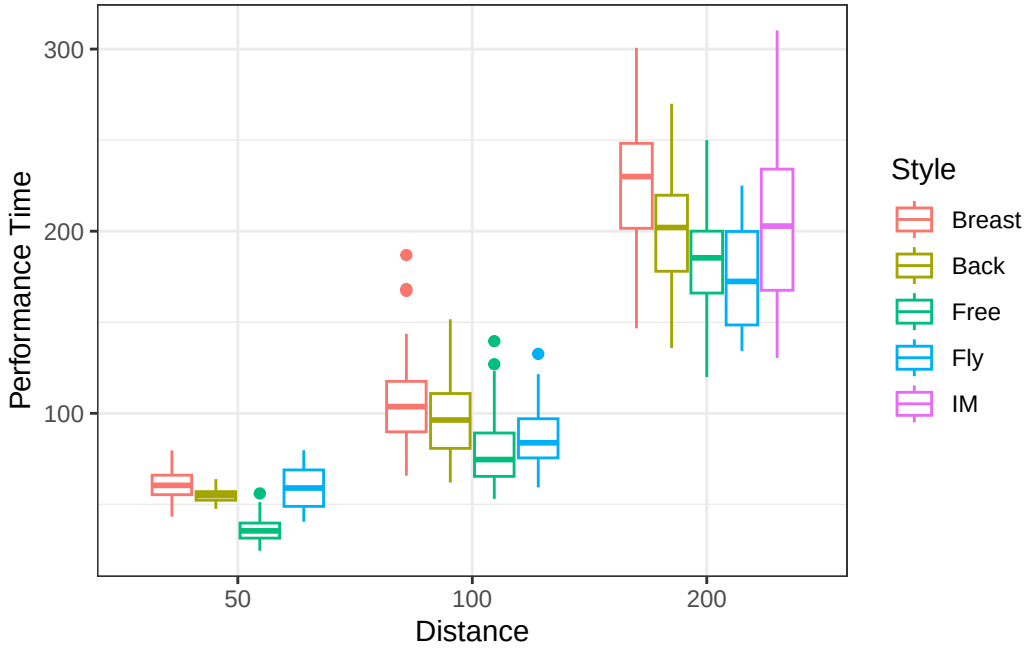
Free style appears to be the fastest on all distances, with the partial exception of 200 m - male swimmers show better average time there.

Minimal times per style on 100m (the genders are not separated)



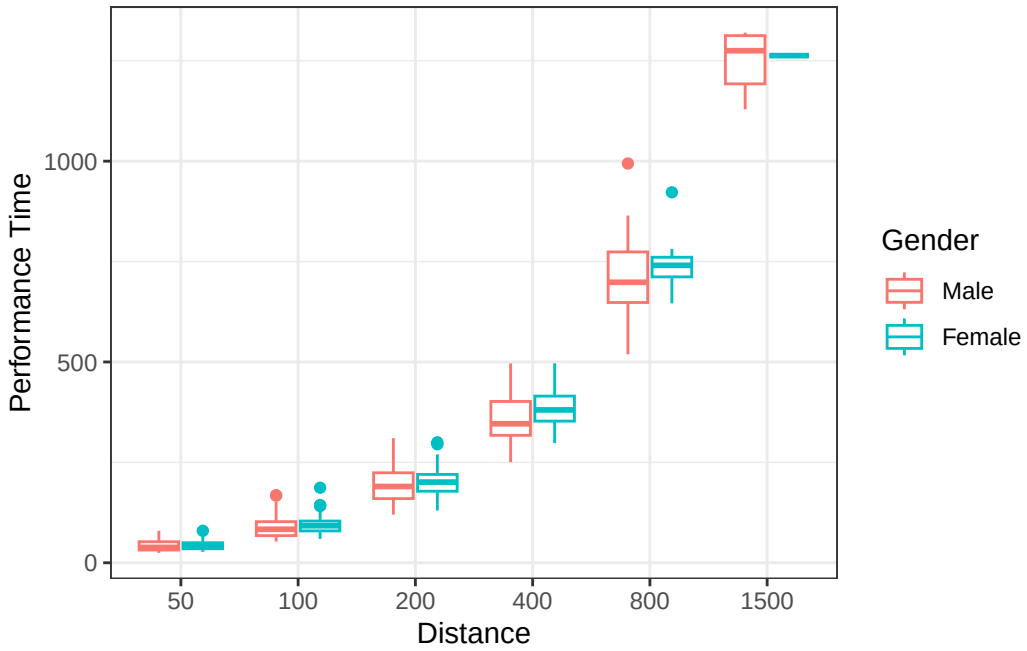
Conclusion: Free style allows the best performance. In 11-11 and 14-14 age groups,, Back style has very close minimal time to Free style.

Distribution of swimmers' results for 50m,100m,200m



Conclusion: We plotted box plots for only three, the most popular distances among swimmers of all age groups. Fly style on 200m was assigned only to 19 swimmers, which may explain the distribution of time results.

Time depending on gender and distance (among all styles)



Conclusion: Male swimmers show better average and minimal results. The difference on 50m is not substantial.

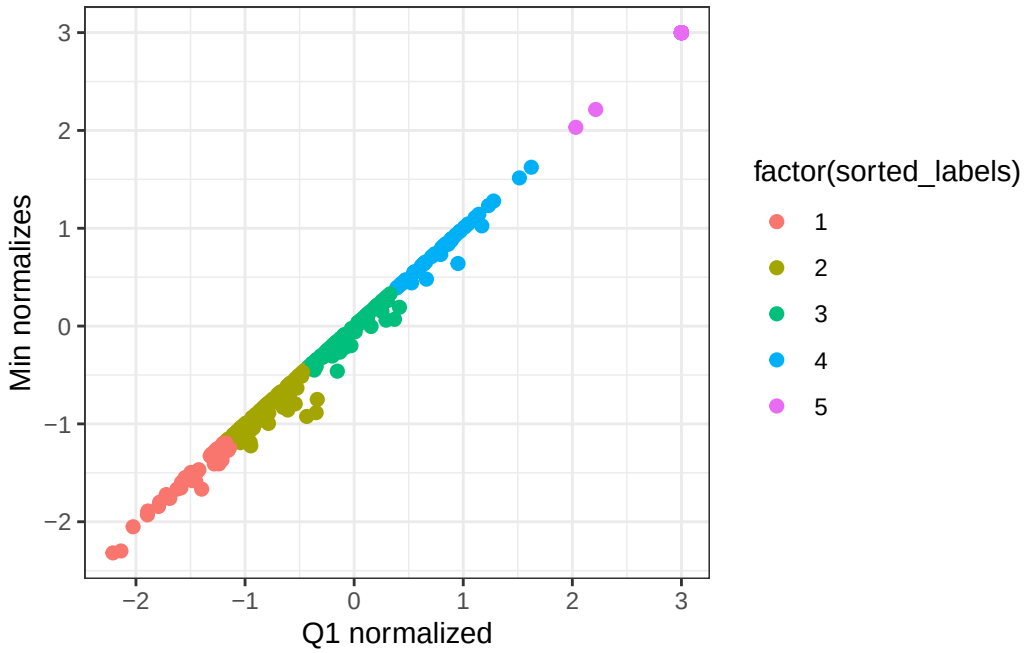
Part 2: Clustering

Clustering based on the best results. All results are scaled first within the distance and style.

The following test shows who appears in the most productive group among all ages and both genders.

We scale results within each combination of style and distance such that the mean is zero and the variance is 1. Such scaling allows to compare performance of swimmers even if the data set does not contain records for the same style and distance combination. (Other scaling may be considered as well - we picked that one as it is widely used for scaling test and exam results.)

We perform K-means clustering method with K=5. Each swimmer is described by two numbers: minimal scaled time and first quartile Q1 of all available scaled time results.



Conclusion: Cluster 1 (red points) represents the swimmers with the best performance (based on minimal times and first quartiles.)

Age and gender structure of Cluster 1:

Swimmer Count in Cluster 1 by Gender and Age Group

Gender	10&Under	11-11	12-12	13-13	13-14	14-14	15&Over
Male	1	0	1	2	1	4	19
Female	1	0	1	1	0	0	5

Conclusion: There are significantly more male swimmers, though some female swimmers are present in Cluster 1. The names of swimmers who appeared in Cluster 1 are available based on the data set, but those names are not listed here.

Part 3: Hypothesis Testing

100m Free style

We check if the gender affects children's time results. To do that, we may consider testing hypotheses on equality of means and first quartiles.

- Group 1: Time results of female swimmers in100m Free style.
- Group 2: Time results of male swimmers in 100m Free style.

Student's t-test or Welch's t-test could work here, but we need to check if the data samples pass the normal distribution test - we apply Shapiro test.

Shapiro-Wilk Normality Test — 100Free by Gender

Gender	W_Statistic	p_value	Interpretation
Female	0.947	0.001	Not normal ($p < 0.05$)
Male	0.936	0.000	Not normal ($p < 0.05$)

Conclusion: It appears based on Shapiro test that both samples do not satisfy the normal distribution test.

Choice of test.

In our case:

- there are exactly two groups;
- the groups are independent (not related);
- the distribution is not normal;
- we need to test the equality of means and equality of the first quartiles;

Based on the list of conditions, we apply **Mann–Whitney U** test and **permutation** test.

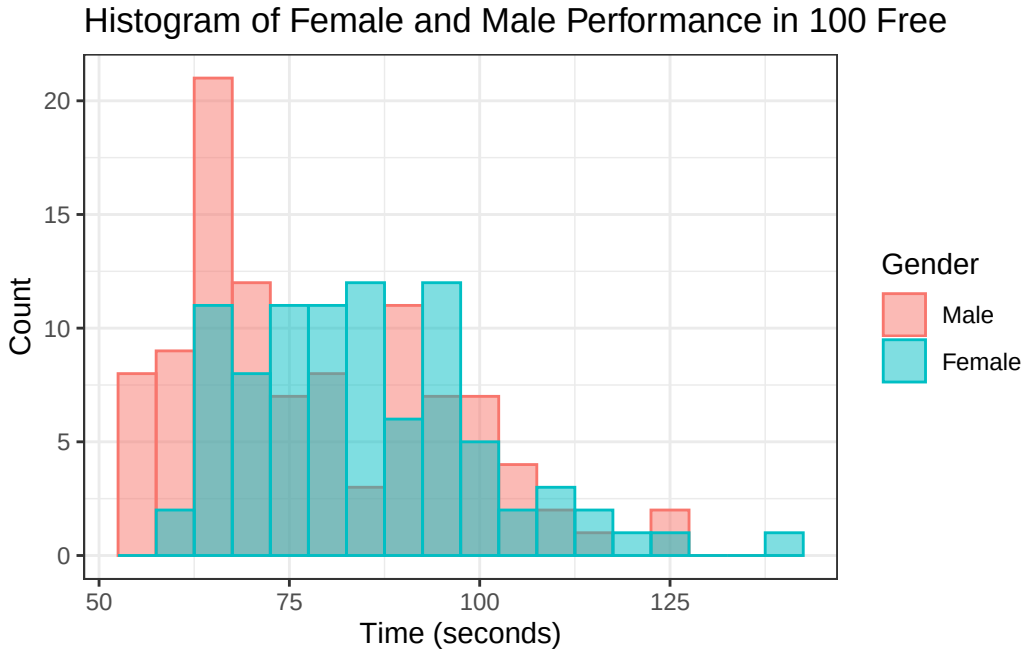
For **Mann–Whitney U** test, we make assumptions: both groups are samples of values of two continuous random variables, the random variables have the same type (shape) of distribution, but maybe different parameter/s (mean in our case).

Permutation test allows us to test also equality of means and of Q1s. It requires the assumption that data points are interchangeable under the null hypothesis.

Before testing hypothesis let us see statistics and distribution of both groups.

Summary Statistics for 100Free by Gender

Statistic	Female	Male
Mean	84.531	78.012
Variance	251.216	285.690
Q1	73.145	63.633



Permutation test, 100m Free style

We test two pair of hypotheses: on equality of means and on equality of first quartiles.

- The null hypothesis is $H_0 = \{\mu_f = \mu_m\}$ and the alternative hypothesis is $H_1 = \{\mu_f \neq \mu_m\}$, where μ_f is mean of time results in 100m Free of female swimmers, μ_m is mean of time results in 100m Free of male swimmers.
- The null hypothesis is $H_0 = \{Q1_f = Q1_m\}$and the alternative hypothesis is $H_1 = \{Q1_f \neq Q1_m\}$, where $Q1_f$ is first quartile of time results in 100m Free of female swimmers, $Q1_m$ is first quartile.of time results in 100m Free of male swimmers.

  P-value received on permutation test for equality of Q1s's: **0.0001**

P-value received on permutation test for equality of means's: **0.0028**

Conclusion: In both cases p-values are small and in both cases the null hypothesis H0 are rejected.

Mann–Whitney U test, 100m Free style

We test hypotheses:

- The null hypothesis is $H_0 = \{\mu_f = \mu_m\}$ and the alternative hypothesis is $H_1 = \{\mu_f \neq \mu_m\}$, where μ_f equls mean of time results in 100m Free of female swimmers, μ_m equals mean of time results in 100m Free of male swimmers.

Wilcoxon Rank-Sum Test — 100Free by Gender

Test	W.Statistic	p.value	Interpretation
Wilcoxon Rank-Sum Test	5,598.000	0.003	Significant difference ($p < 0.05$)

Conclusion: The null hypothesis is rejected by Mann–Whitney U test.

hypotheses testing based on scaled time results

All available time results were first scaled depending on distance and style (see description of scaling in **Part 2**). The scheme of testing is analogous to the presented below but now we consider all and scaled time results.

- Group 1: Scaled time results of female swimmers.
- Group 2: Scaled time results of male swimmers.

We check if the gender affects children swimmers' performance. To do that, we may consider testing hypotheses on the equality of means and of the first quartiles.

Student's t-test or Welch's t-test could work here. First we need to check if the data samples pass the normal distribution test (we apply Shapiro test.)

Statistics of two groups, scaled times

Summary Statistics for Females(Group1) and Males(Group2)

Statistic	Group1	Group2
Mean	0.221	-0.073
Variance	1.136	1.467
Q1	-0.469	-0.999

Shapiro test, scaled times

Shapiro-Wilk Normality Test by Gender

Gender	W_Statistic	p_value	Interpretation
Female	0.919	0.000	Not normal ($p < 0.05$)
Male	0.948	0.000	Not normal ($p < 0.05$)

Conclusion: Scaled time results do not pass Shapiro test on normal distribution.

Permutation test, scaled times

We test two pair of hypotheses: on equality of means and on equality of first quartiles.

- The null hypothesis is $H_0 = \{\mu_f = \mu_m\}$ and the alternative hypothesis is $H_1 = \{\mu_f \neq \mu_m\}$, where μ_f=mean of all scaled time results of female swimmers, μ_m =mean of all scaled time results of male swimmers.
- The null hypothesis is $H_0 = \{Q1_f = Q1_m\}$ and the alternative hypothesis is $H_1 = \{Q1_f \neq Q1_m\}$, where $Q1_f$=first quartile of all scaled time results of female swimmers, $Q1_m$=first quartile of all scaled time results of male swimmers.

Results:

P-value received on permutation test for equality of Q1s's: **0.0010**

P-value received on permutation test for equality of means's: **0.0010**

Conclusion: Both null hypothesis are rejected by permutation test.

Mann–Whitney U test, scaled times

We test hypotheses:

- The null hypothesis is $H_0 = \{\mu_f = \mu_m\}$ and the alternative hypothesis is $H_1 = \{\mu_f \neq \mu_m\}$, where μ_f = mean of all scaled time results of female swimmers, μ_m=mean of all scaled time results of male swimmers.

Wilcoxon Rank-Sum Test — scaled times by Gender

Test	W.Statistic	p.value	Interpretation
Wilcoxon Rank-Sum Test	561,861.500	0.000	Significant difference ($p < 0.05$)

Conclusion: The null hypothesis is rejected by Mann–Whitney U test.