On-Field Insights from Tokyo 2020 Olympic Swimming Events

Ranice TAN Hui Qi

Contents

1. Overview

Sports analytics involves using data and statistical analytics to drive decision making. On-field analytics involves studying the performance, strategy or health and fitness of an athlete or team. The examination of these data can potentially enable athletes and coaches to make better decisions, allowing them to stand a competitive edge.

In this study, we plan to review data from the Tokyo 2020 Olympics data for Swimming events, to discover interesting on-field insights and use statistical analysis to proof these theories.

2. Objective

The objective of this study is to understand the factors affecting the reaction time of Olympic swimmers, as well as determine if reaction time affects the performance of the swimmers. The reaction time across all swimming events will be examined, to identify any relationship between the events, physical attributes or performance and reaction time.

JMP Pro 16 statistical analysis was used for data preparation, analysis and insights discovery using Interactive Data Exploration and Analysis.

3. Data

3.1. Data Used

The data used is the official result report, accessible from the official Tokyo 2020 Olympics webpage. The dataset fetched and combined the reports from all the Swimming events. The list of datasets can be found in Appendix B: Datasets.

3.2. Data Preparation

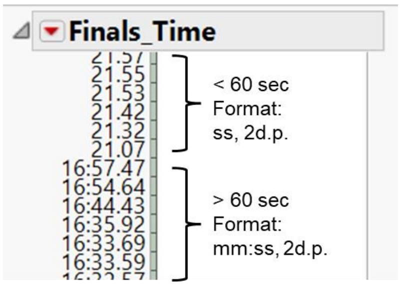
The dataset was imported into JMP Pro to ensure all fields are correctly filled, and all columns are appropriately formatted. A preliminary inspection of the data summary statistics and distribution was also conducted to remove entries with missing data, extreme outliers, and redundant data. The data preparation log is accessible in Finally, a separate study can be done for relay team performance.

Appendix .

3.3. Data Quality

3.3.1. Inconsistent Formatting for ‘Finals_Time’ field

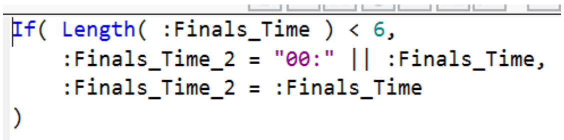
Finals_Time field contains inconsistent formatting for data <60 secs, and ≥ 60 secs. Hence, unable to recognise data and model type as numeric and continuous, and unable to sort the data by ascending order.



```
Finals_Time
21.55
21.53
21.42
21.32
21.07
16:57.47
16:54.64
16:44.43
16:35.92
16:33.69
16:33.59
```

Figure 1: Inconsistency in Formatting in Finals_Time field

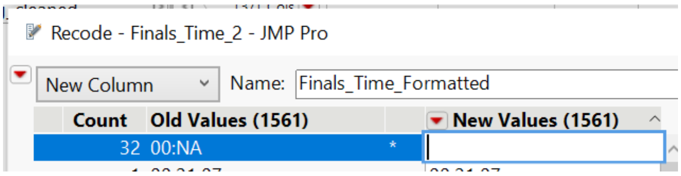
First, formula was written into new column Finals_Time_2 to standardise the formatting of all data to mm:ss, 2 decimal places.



```
If( Length( :Finals_Time ) < 6,
    :Finals_Time_2 = "00:" || :Finals_Time,
    :Finals_Time_2 = :Finals_Time
)
```

Figure 2 Code to standardise formatting in Finals_Time_2 field

Next, Finals_Time_2 was recoded to remove missing (‘NA’) data into new unlocked column Finals_Time_Formatted.



Recode - Finals_Time_2 - JMP Pro

New Column Name: Finals_Time_Formatted

Count	Old Values (1561)	New Values (1561)
32	00:NA	

Figure 3 Recoding of 'NA' data from Finals_Time_2 field

3.3.2. Inaccurate Data and Modelling Types

The variables from the dataset which has incorrect data or modelling types are identified and corrected in the data table below. It is important to input the correct model type so that JMP can infer the appropriate statistic for analysis, without which, may result in poor and errors in analysis.

Table 1 Changes in Data and Model Types for Variables

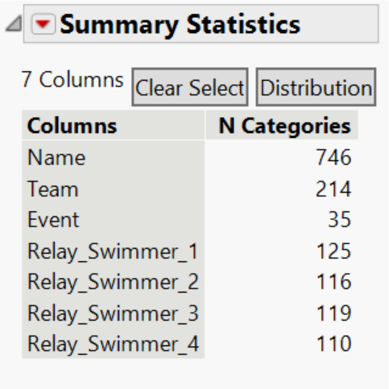
Variable Names	Default		Change To	
	Data	Model	Data	Model
Place	Numeric	Continuous	Numeric	Ordinal
Lane	Numeric	Continuous	Numeric	Nominal
Finals_Time	Character	Nominal	Numeric	Continuous
DQ	Numeric	Continuous	Numeric	Nominal
Exhibition	Numeric	Continuous	Numeric	Nominal
Split_100, 150, 200, 300, 350, 400, 500, 550, 600, 700, 750, 800	Character	Nominal	Numeric	Continuous

3.3.3. Large numbers of categories for Categorical Data

Too many categorical variables may hinder analysis, hence, for variables important for analysis, some changes were made as follows:

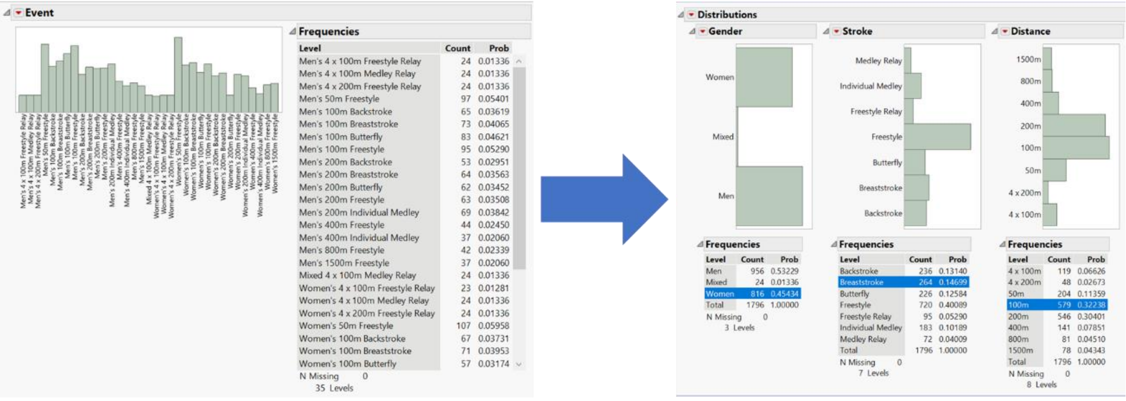
Table 2 Changes made to columns with large categorical variables

Columns	N Categories	Change
Name	746	Information not relevant
Team	214	Removed duplicate entries for teams e.g. BRA and BRA-Brazil.
Event	35	Recoded into Gender, Stroke and Distance columns, of 3, 7 and 8 levels respectively.
Relay_Swimmer_1	125	Information not relevant
Relay_Swimmer_2	116	Information not relevant
Relay_Swimmer_3	119	Information not relevant
Relay_Swimmer_4	110	Information not relevant



Columns	N Categories
Name	746
Team	214
Event	35
Relay_Swimmer_1	125
Relay_Swimmer_2	116
Relay_Swimmer_3	119
Relay_Swimmer_4	110

Figure 4 Summary Statistic for Categorical Columns with large levels



Event — Frequencies

Level	Count	Prob
Men's 4 x 100m Freestyle Relay	24	0.01336
Men's 4 x 100m Medley Relay	24	0.01336
Men's 4 x 200m Freestyle Relay	24	0.01336
Men's 50m Freestyle	97	0.05401
Men's 100m Backstroke	65	0.03619
Men's 100m Breaststroke	73	0.04065
Men's 100m Butterfly	83	0.04621
Men's 100m Freestyle	95	0.05290
Men's 200m Backstroke	53	0.02951
Men's 200m Breaststroke	64	0.03563
Men's 200m Butterfly	62	0.03452
Men's 200m Freestyle	63	0.03508
Men's 200m Individual Medley	69	0.03842
Men's 400m Freestyle	44	0.02450
Men's 400m Individual Medley	37	0.02060
Men's 800m Freestyle	42	0.02339
Men's 1500m Freestyle	37	0.02060
Mixed 4 x 100m Medley Relay	24	0.01336
Women's 4 x 100m Freestyle Relay	23	0.01281
Women's 4 x 100m Medley Relay	24	0.01336
Women's 4 x 200m Freestyle Relay	24	0.01336
Women's 50m Freestyle	107	0.05958
Women's 100m Backstroke	67	0.03731
Women's 100m Breaststroke	71	0.03953
Women's 100m Butterfly	57	0.03174

N Missing 0
35 Levels

Gender — Frequencies

Level	Count	Prob
Men	956	0.53229
Mixed	24	0.01336
Women	816	0.45434
Total	1796	1.00000

N Missing 0
3 Levels

Stroke — Frequencies

Level	Count	Prob
Backstroke	236	0.13140
Breaststroke	264	0.14699
Butterfly	226	0.12584
Freestyle	720	0.40089
Freestyle Relay	95	0.05290
Individual Medley	183	0.10189
Medley Relay	72	0.04009
Total	1796	1.00000

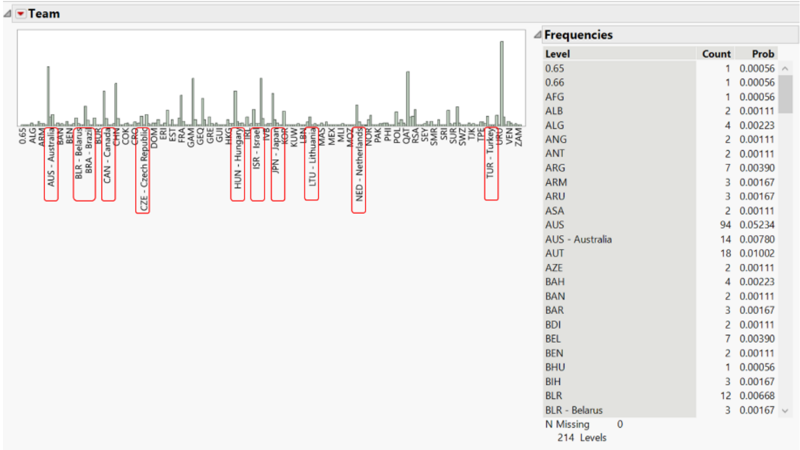
N Missing 0
7 Levels

Distance — Frequencies

Level	Count	Prob
4 x 100m	119	0.06626
4 x 200m	48	0.02673
50m	204	0.11359
100m	579	0.32238
200m	546	0.30401
400m	141	0.07851
800m	81	0.04510
1500m	78	0.04343
Total	1796	1.00000

N Missing 0
8 Levels

Figure 5 Recoding from Event field to Gender, Stroke and Distance fields



Frequencies

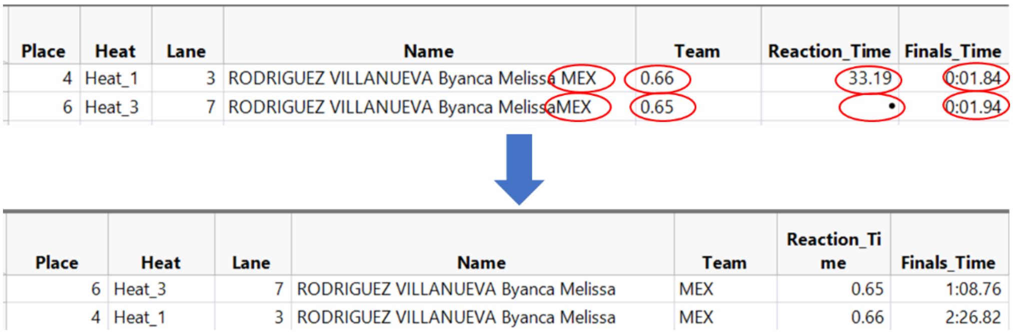
Level	Count	Prob
0.65	1	0.00056
0.66	1	0.00056
AFG	1	0.00056
ALB	2	0.00111
ALG	4	0.00223
ANG	2	0.00111
ANT	2	0.00111
ARG	7	0.00390
ARM	3	0.00167
ARU	3	0.00167
ASA	2	0.00111
AUS	94	0.05234
AUS - Australia	14	0.00780
AUT	18	0.01002
AZE	2	0.00111
BAH	4	0.00223
BAN	2	0.00111
BAR	3	0.00167
BDI	2	0.00111
BEL	7	0.00390
BEN	2	0.00111
BHU	1	0.00056
BIH	3	0.00167
BLR	12	0.00668
BLR - Belarus	3	0.00167

N Missing 0
214 Levels

Figure 6 Team categorical data contains duplicate entries

3.3.4. Incorrect Information Pre-Filled

2 rows by swimmer Rodriguez Villanueva were pre-filled with incorrect information and manually corrected as follows:



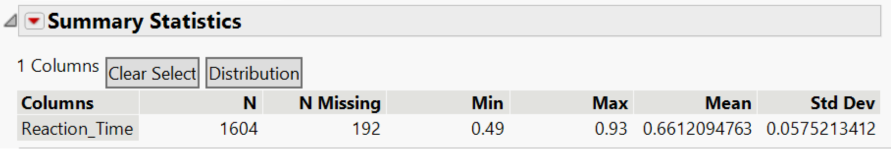
Place	Heat	Lane	Name	Team	Reaction_Time	Finals_Time
4	Heat_1	3	RODRIGUEZ VILLANUEVA Byanca Melissa MEX	0.66	33.19	0:01.84
6	Heat_3	7	RODRIGUEZ VILLANUEVA Byanca MelissaMEX	0.65	•	0:01.94

Place	Heat	Lane	Name	Team	Reaction_Time	Finals_Time
6	Heat_3	7	RODRIGUEZ VILLANUEVA Byanca Melissa	MEX	0.65	1:08.76
4	Heat_1	3	RODRIGUEZ VILLANUEVA Byanca Melissa	MEX	0.66	2:26.82

Figure 7 Manual correction of incorrect data entry

3.3.5. Missing Data

192 out of 1796 of reaction time data is missing from disqualified swimmers and relay events. Hence, limiting study to individual performances only.



Columns	N	N Missing	Min	Max	Mean	Std Dev
Reaction_Time	1604	192	0.49	0.93	0.6612094763	0.0575213412

Figure 8 Summary Statistics for Reaction_Time Column

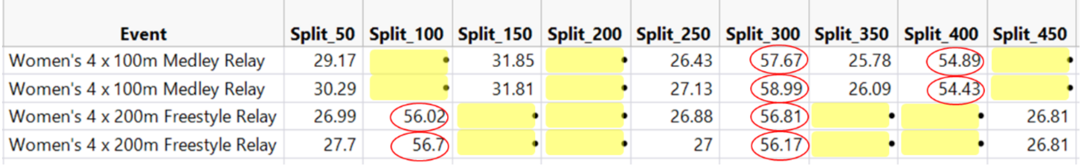
3.3.6. Inconsistent definition for split timing data between events

Definition for some split time is inconsistent throughout the events, hence may lead to confusion if the data is used for calculations or analysis.

Table 3 Inconsistencies in Split time data

Event	Inconsistencies
Individual Events	Time taken for each 50m interval i.e. Split_50: 0-50m, Split _100: 50m-100m
4 X 100m Relay Events	Time taken for each swimmer’s 1st 50m interval, and 1st 100m interval, with some entries missing data. i.e. Split_50: 0-50m, Split _100: 0m-100m
4 X 200m Relay Events	Similar to “4 X 100m Relay Events”, but missing data for 100-200m intervals.

	i.e. Split_50: 0-50m, Split _100: 50m-100m, Split_150: Missing, Split _200: Missing

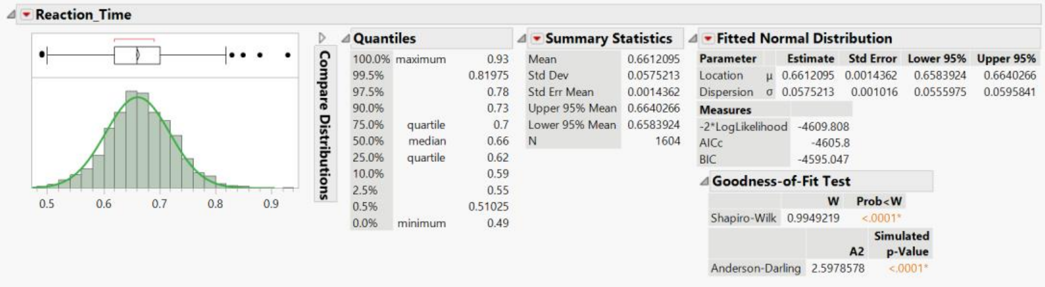


Event	Split_50	Split_100	Split_150	Split_200	Split_250	Split_300	Split_350	Split_400	Split_450
Women's 4 x 100m Medley Relay	29.17	•	31.85	•	26.43	57.67	25.78	54.89	•
Women's 4 x 100m Medley Relay	30.29	•	31.81	•	27.13	58.99	26.09	54.43	•
Women's 4 x 200m Freestyle Relay	26.99	56.02	•	•	26.88	56.81	•	•	26.81
Women's 4 x 200m Freestyle Relay	27.7	56.7	•	•	27	56.17	•	•	26.81

Figure 9 Examples of Relay Events with Definition Discrepancies in Split Timing

3.3.7. Outliers in Reaction Time data

Reaction Time data contains 7 outlier points, which accounts for <0.5% of data, hence may not drastically impact the analysis when removed.



Quantiles

100.0%	maximum	0.93
99.5%		0.81975
97.5%		0.78
90.0%		0.73
75.0%	quartile	0.7
50.0%	median	0.66
25.0%	quartile	0.62
10.0%		0.59
2.5%		0.55
0.5%		0.51025
0.0%	minimum	0.49

Summary Statistics

Mean	0.6612095
Std Dev	0.0575213
Std Err Mean	0.0014362
Upper 95% Mean	0.6640266
Lower 95% Mean	0.6583924
N	1604

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6612095	0.0014362	0.6583924	0.6640266
Dispersion	σ	0.0575213	0.001016	0.0555975	0.0595841

Measures	
-2*LogLikelihood	-4609.808
AICc	-4605.8
BIC	-4595.047

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9949219	<.0001*

	A2	Simulated p-Value
Anderson-Darling	2.5978578	<.0001*

Figure 10 Statistics of Reaction Time Data showing few outliers

3.3.8. Missing parameters

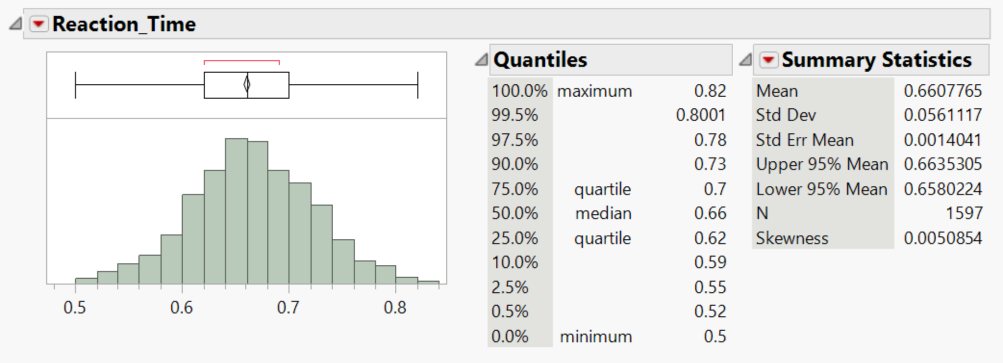
Some parameters that are useful to the study and are missing include time of event, and age of swimmer. Time – morning, afternoon and night may be used to gauge the alertness of the swimmer. Age data may be used to analyse any correlation between reaction time or performance to identify peak performance for an athlete.

4. Data Analysis

4.1. General Insights

4.1.1. Analysis 1: Olympic swimmers have similar mean and median reaction time.

The mean and median reaction times for Olympic swimmers is 0.66s and 0.66 respectively.



Quantiles

100.0%	maximum	0.82
99.5%		0.8001
97.5%		0.78
90.0%		0.73
75.0%	quartile	0.7
50.0%	median	0.66
25.0%	quartile	0.62
10.0%		0.59
2.5%		0.55
0.5%		0.52
0.0%	minimum	0.5

Summary Statistics

Mean	0.6607765
Std Dev	0.0561117
Std Err Mean	0.0014041
Upper 95% Mean	0.6635305
Lower 95% Mean	0.6580224
N	1597
Skewness	0.0050854

Figure 11 Statistics of Reaction Time

4.1.2. Analysis 2: Lanes 3-4 swimmers have a lower reaction time than swimmers in the other lanes

The reaction time of swimmers in Lanes 3, 4 have a median of 0.65s and their means are < 0.66s, whereas swimmers in the other lanes have medians of 0.66s and means > 0.66s. Hence, swimmers in Lanes 3 and 4 seem to be able to react faster than the other lanes.

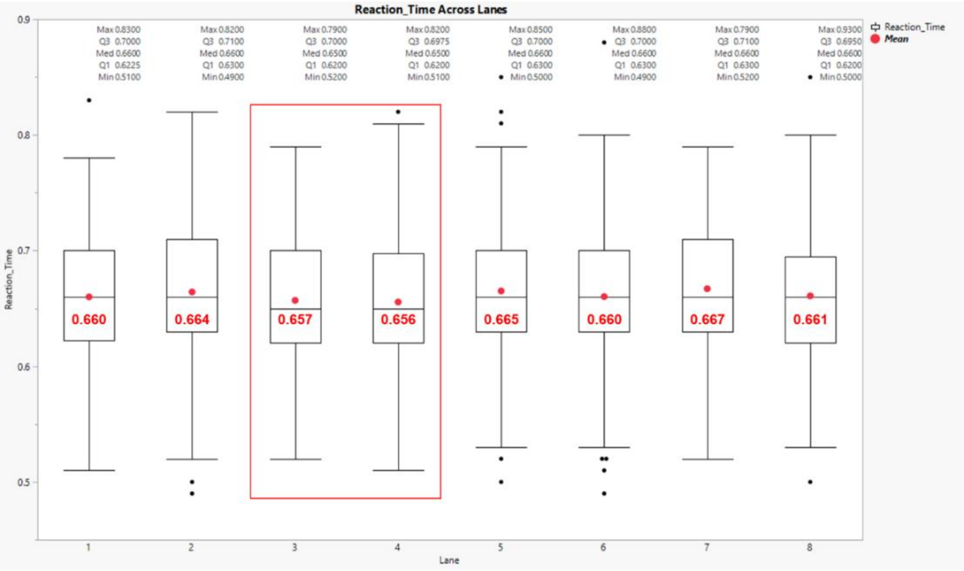
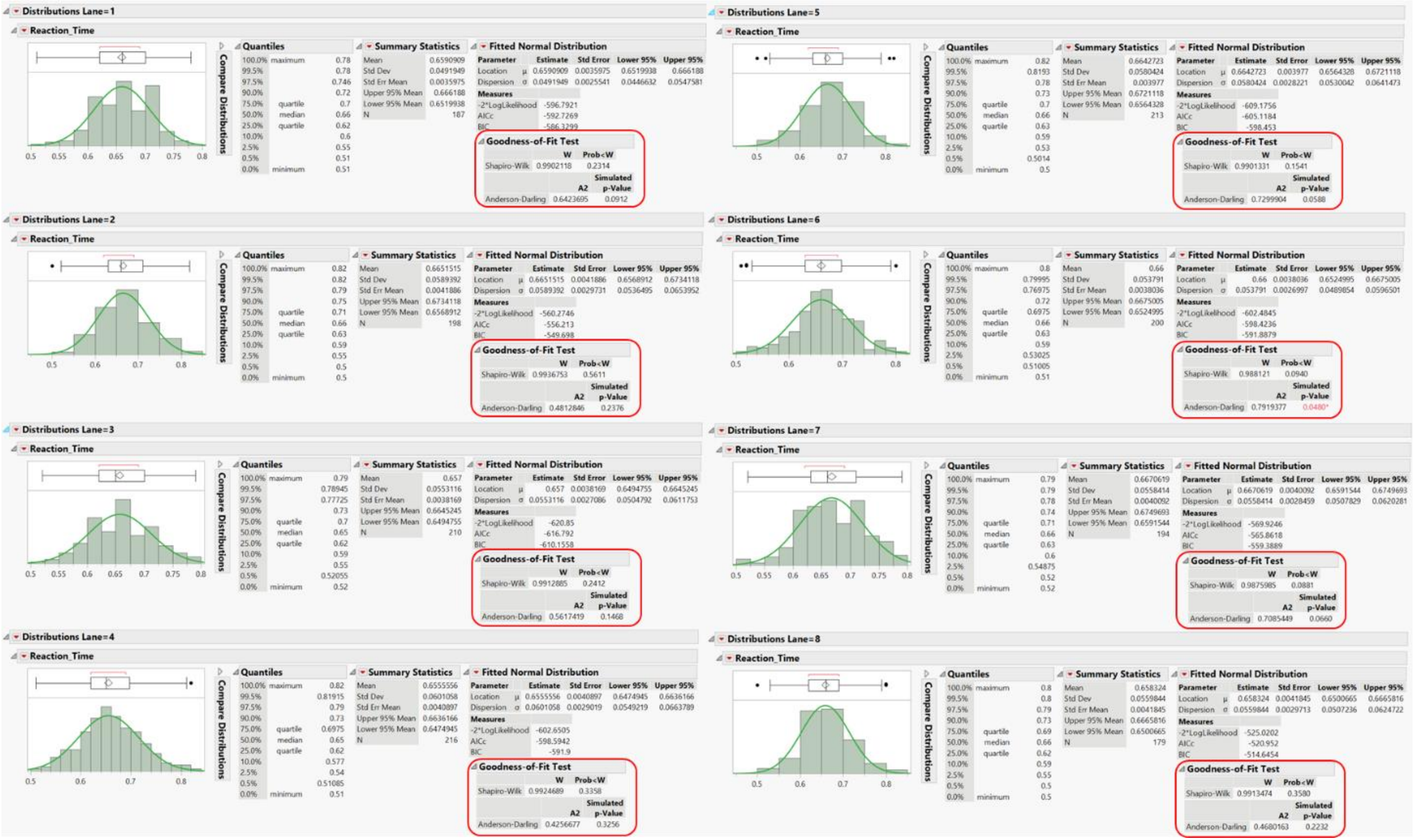


Figure 12 Boxplot of Reaction Time across Lanes

4.1.2.1. Hypothesis 1: Swimmers in Lanes 3 and 4 react faster than the swimmers in the other lanes

We are interested to test if the swimmers in lanes 3 and 4 are truly faster in reacting then the swimmers in the other lanes. A normality test was conducted to determine the appropriate hypothesis testing method. In this hypothesis, we will use Shapiro-Wilk and Anderson-Darling Goodness-of-Fit Test with confidence level of 99% to test:

- H_0: Reaction Time is normally distributed
- H_1: Reaction Time is not normally distributed



```
Distributions Lane=1
Reaction_Time

Quantiles
100.0%  maximum   0.78
99.5%             0.78
97.5%             0.746
90.0%             0.72
75.0%   quartile  0.7
50.0%   median    0.66
25.0%   quartile  0.62
10.0%             0.6
2.5%              0.55
0.5%              0.51
0.0%    minimum   0.51

Summary Statistics
Mean             0.6590909
Std Dev          0.0491949
Std Err Mean     0.0035975
Upper 95% Mean   0.666188
Lower 95% Mean   0.6519938
N                187

Fitted Normal Distribution
Parameter         Estimate   Std Error  Lower 95%  Upper 95%
Location     μ    0.6590909  0.0035975  0.6519938  0.666188
Dispersion   σ    0.0491949  0.0025541  0.0446632  0.0547581
Measures
-2*LogLikelihood  -596.7921
AICc              -592.7269
BIC               -586.3299

Goodness-of-Fit Test
                  W          Prob<W
Shapiro-Wilk      0.9902118  0.2314
                             Simulated
                  A2         p-Value
Anderson-Darling  0.6423695  0.0912
```

```
Distributions Lane=2
Reaction_Time

Quantiles
100.0%  maximum   0.82
99.5%             0.82
97.5%             0.79
90.0%             0.75
75.0%   quartile  0.71
50.0%   median    0.66
25.0%   quartile  0.63
10.0%             0.59
2.5%              0.55
0.5%              0.5
0.0%    minimum   0.5

Summary Statistics
Mean             0.6651515
Std Dev          0.0589392
Std Err Mean     0.0041886
Upper 95% Mean   0.6734118
Lower 95% Mean   0.6568912
N                198

Fitted Normal Distribution
Parameter         Estimate   Std Error  Lower 95%  Upper 95%
Location     μ    0.6651515  0.0041886  0.6568912  0.6734118
Dispersion   σ    0.0589392  0.0029731  0.0536495  0.0653952
Measures
-2*LogLikelihood  -560.2746
AICc              -556.213
BIC               -549.698

Goodness-of-Fit Test
                  W          Prob<W
Shapiro-Wilk      0.9936753  0.5611
                             Simulated
                  A2         p-Value
Anderson-Darling  0.4812846  0.2376
```

```
Distributions Lane=3
Reaction_Time

Quantiles
100.0%  maximum   0.79
99.5%             0.78945
97.5%             0.77725
90.0%             0.73
75.0%   quartile  0.7
50.0%   median    0.65
25.0%   quartile  0.62
10.0%             0.59
2.5%              0.55
0.5%              0.52055
0.0%    minimum   0.52

Summary Statistics
Mean             0.657
Std Dev          0.0553116
Std Err Mean     0.0038169
Upper 95% Mean   0.6645245
Lower 95% Mean   0.6494755
N                210

Fitted Normal Distribution
Parameter         Estimate   Std Error  Lower 95%  Upper 95%
Location     μ    0.657      0.0038169  0.6494755  0.6645245
Dispersion   σ    0.0553116  0.0027086  0.0504792  0.0611753
Measures
-2*LogLikelihood  -620.85
AICc              -616.792
BIC               -610.1558

Goodness-of-Fit Test
                  W          Prob<W
Shapiro-Wilk      0.9912885  0.2412
                             Simulated
                  A2         p-Value
Anderson-Darling  0.5617419  0.1468
```

```
Distributions Lane=4
Reaction_Time

Quantiles
100.0%  maximum   0.82
99.5%             0.81915
97.5%             0.79
90.0%             0.73
75.0%   quartile  0.6975
50.0%   median    0.65
25.0%   quartile  0.62
10.0%             0.577
2.5%              0.54
0.5%              0.51085
0.0%    minimum   0.51

Summary Statistics
Mean             0.6555556
Std Dev          0.0601058
Std Err Mean     0.0040897
Upper 95% Mean   0.6636166
Lower 95% Mean   0.6474945
N                216

Fitted Normal Distribution
Parameter         Estimate   Std Error  Lower 95%  Upper 95%
Location     μ    0.6555556  0.0040897  0.6474945  0.6636166
Dispersion   σ    0.0601058  0.0029019  0.0549219  0.0663789
Measures
-2*LogLikelihood  -602.6505
AICc              -598.5942
BIC               -591.9

Goodness-of-Fit Test
                  W          Prob<W
Shapiro-Wilk      0.9924689  0.3358
                             Simulated
                  A2         p-Value
Anderson-Darling  0.4256677  0.3256
```

```
Distributions Lane=5
Reaction_Time

Quantiles
100.0%  maximum   0.82
99.5%             0.8193
97.5%             0.78
90.0%             0.73
75.0%   quartile  0.7
50.0%   median    0.66
25.0%   quartile  0.63
10.0%             0.59
2.5%              0.53
0.5%              0.5014
0.0%    minimum   0.5

Summary Statistics
Mean             0.6642723
Std Dev          0.0580424
Std Err Mean     0.003977
Upper 95% Mean   0.6721118
Lower 95% Mean   0.6564328
N                213

Fitted Normal Distribution
Parameter         Estimate   Std Error  Lower 95%  Upper 95%
Location     μ    0.6642723  0.003977   0.6564328  0.6721118
Dispersion   σ    0.0580424  0.0028221  0.0530042  0.0641473
Measures
-2*LogLikelihood  -609.1756
AICc              -605.1184
BIC               -598.453

Goodness-of-Fit Test
                  W          Prob<W
Shapiro-Wilk      0.9901331  0.1541
                             Simulated
                  A2         p-Value
Anderson-Darling  0.7299904  0.0588
```

```
Distributions Lane=6
Reaction_Time

Quantiles
100.0%  maximum   0.8
99.5%             0.79995
97.5%             0.76975
90.0%             0.72
75.0%   quartile  0.6975
50.0%   median    0.66
25.0%   quartile  0.63
10.0%             0.59
2.5%              0.53025
0.5%              0.51005
0.0%    minimum   0.51

Summary Statistics
Mean             0.66
Std Dev          0.053791
Std Err Mean     0.0038036
Upper 95% Mean   0.6675005
Lower 95% Mean   0.6524995
N                200

Fitted Normal Distribution
Parameter         Estimate   Std Error  Lower 95%  Upper 95%
Location     μ    0.66       0.0038036  0.6524995  0.6675005
Dispersion   σ    0.053791   0.0026997  0.0489854  0.0596501
Measures
-2*LogLikelihood  -602.4845
AICc              -598.4236
BIC               -591.8879

Goodness-of-Fit Test
                  W          Prob<W
Shapiro-Wilk      0.988121   0.0940
                             Simulated
                  A2         p-Value
Anderson-Darling  0.7919377  0.0480*
```

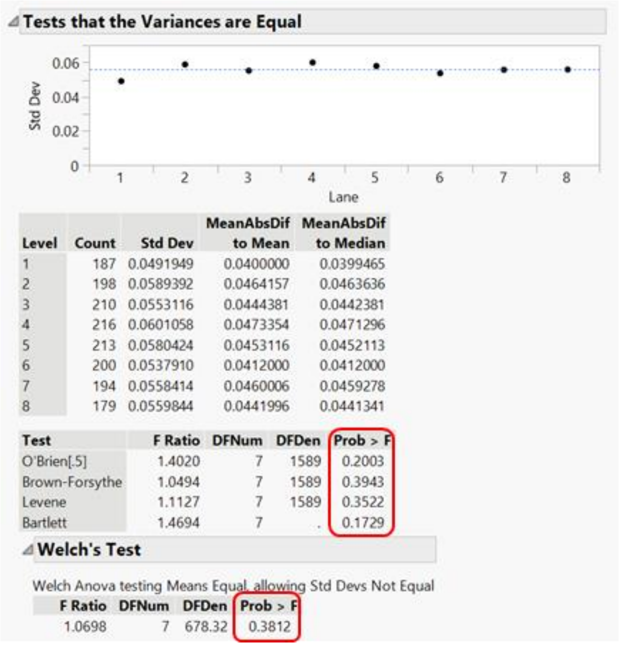
```
Distributions Lane=7
Reaction_Time

Quantiles
100.0%  maximum   0.79
99.5%             0.79
97.5%             0.78
90.0%             0.74
75.0%   quartile  0.71
50.0%   median    0.66
25.0%   quartile  0.63
10.0%             0.6
2.5%              0.54875
0.5%              0.52
0.0%    minimum   0.52

Summary Statistics
Mean             0.6670619
Std Dev          0.0558414
Std Err Mean     0.0040092
Upper 95% Mean   0.6749693
Lower 95% Mean   0.6591544
N                194

Fitted Normal Distribution
Parameter         Estimate   Std Error  Lower 95%  Upper 95%
Location     μ    0.6670619  0.0040092  0.6591544  0.6749693
Dispersion   σ    0.0558414  0.0028459  0.0507829  0.0620281
Measures
-2*LogLikelihood  -569.9246
AICc              -565.8618
BIC               -559.3889

Goodness-of-Fit Test
                  W          Prob<W
Shapiro-Wilk      0.9875985  0.0881
                             Simulated
                  A2         p-Value
Anderson-Darling  0.7085449  0.0660
```

```
Distributions Lane=8
Reaction_Time

Quantiles
100.0%  maximum   0.8
99.5%             0.8
97.5%             0.79
90.0%             0.73
75.0%   quartile  0.69
50.0%   median    0.66
25.0%   quartile  0.62
10.0%             0.59
2.5%              0.55
0.5%              0.5
0.0%    minimum   0.5

Summary Statistics
Mean             0.658324
Std Dev          0.0559844
Std Err Mean     0.0041845
Upper 95% Mean   0.6665816
Lower 95% Mean   0.6500665
N                179

Fitted Normal Distribution
Parameter         Estimate   Std Error  Lower 95%  Upper 95%
Location     μ    0.658324   0.0041845  0.6500665  0.6665816
Dispersion   σ    0.0559844  0.0029713  0.0507236  0.0624722
Measures
-2*LogLikelihood  -525.0202
AICc              -520.952
BIC               -514.6454

Goodness-of-Fit Test
                  W          Prob<W
Shapiro-Wilk      0.9913474  0.3580
                             Simulated
                  A2         p-Value
Anderson-Darling  0.4680163  0.2232
```

Figure 13 Normal Distribution Test for Reaction Times across Swim lanes

Based on the results above, we accept that reaction time for all lanes is normally distributed as smallest p-value 0.0480 (Lane 6) > critical value 0.01. Hence, the hypothesis testing will be conducted using parametric ANOVA and Welch, with 99% confidence level. Our hypothesis is defined as:

- H_0: Reaction Time is same for all lanes
- H_1: Reaction Time is not the same for all lanes



Tests that the Variances are Equal

Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
1	187	0.0491949	0.0400000	0.0399465
2	198	0.0589392	0.0464157	0.0463636
3	210	0.0553116	0.0444381	0.0442381
4	216	0.0601058	0.0473354	0.0471296
5	213	0.0580424	0.0453116	0.0452113
6	200	0.0537910	0.0412000	0.0412000
7	194	0.0558414	0.0460006	0.0459278
8	179	0.0559844	0.0441996	0.0441341

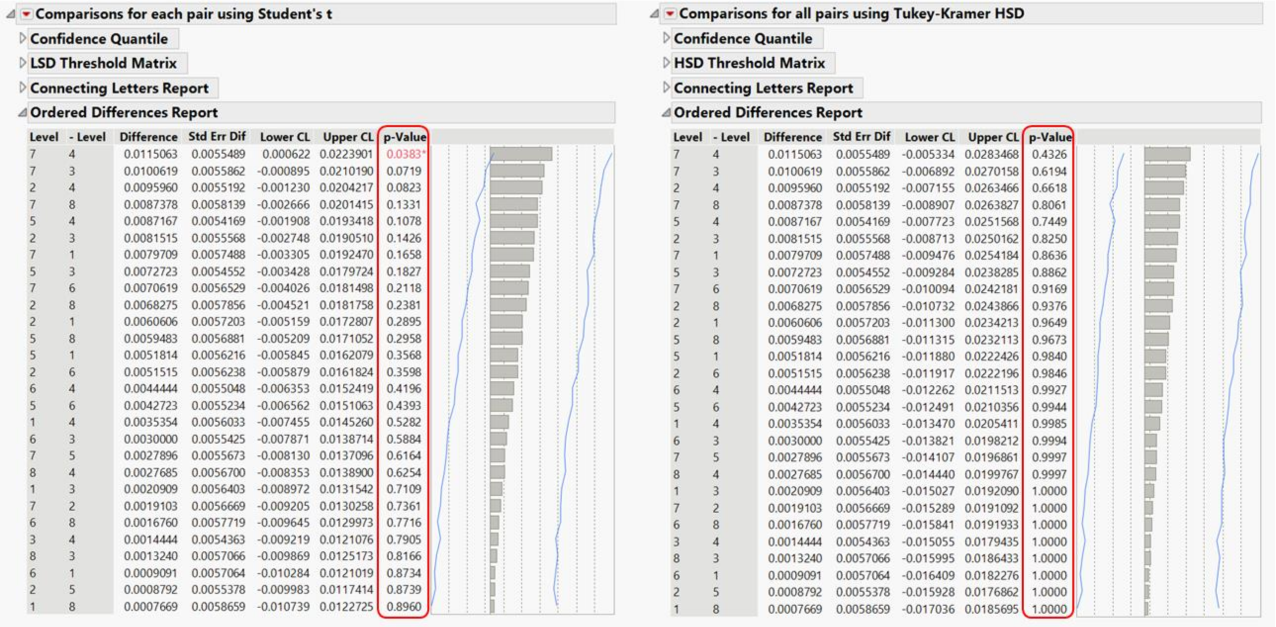
Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	1.4020	7	1589	0.2003
Brown-Forsythe	1.0494	7	1589	0.3943
Levene	1.1127	7	1589	0.3522
Bartlett	1.4694	7	.	0.1729

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
1.0698	7	678.32	0.3812

Figure 14 Welch's Test for Reaction Time across Lanes



Comparisons for each pair using Student's t

Confidence Quantile

LSD Threshold Matrix

Connecting Letters Report

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
7	4	0.0115063	0.0055489	0.000622	0.0223901	0.0383*
7	3	0.0100619	0.0055862	-0.000895	0.0210190	0.0719
2	4	0.0095960	0.0055192	-0.001230	0.0204217	0.0823
7	8	0.0087378	0.0058139	-0.002666	0.0201415	0.1331
5	4	0.0087167	0.0054169	-0.001908	0.0193418	0.1078
2	3	0.0081515	0.0055568	-0.002748	0.0190510	0.1426
7	1	0.0079709	0.0057488	-0.003305	0.0192470	0.1658
5	3	0.0072723	0.0054552	-0.003428	0.0179724	0.1827
7	6	0.0070619	0.0056529	-0.004026	0.0181498	0.2118
2	8	0.0068275	0.0057856	-0.004521	0.0181758	0.2381
2	1	0.0060606	0.0057203	-0.005159	0.0172807	0.2895
5	8	0.0059483	0.0056881	-0.005209	0.0171052	0.2958
5	1	0.0051814	0.0056216	-0.005845	0.0162079	0.3568
2	6	0.0051515	0.0056238	-0.005879	0.0161824	0.3598
6	4	0.0044444	0.0055048	-0.006353	0.0152419	0.4196
5	6	0.0042723	0.0055234	-0.006562	0.0151063	0.4393
1	4	0.0035354	0.0056033	-0.007455	0.0145260	0.5282
6	3	0.0030000	0.0055425	-0.007871	0.0138714	0.5884
7	5	0.0027896	0.0055673	-0.008130	0.0137096	0.6164
8	4	0.0027685	0.0056700	-0.008353	0.0138900	0.6254
1	3	0.0020909	0.0056403	-0.008972	0.0131542	0.7109
7	2	0.0019103	0.0056669	-0.009205	0.0130258	0.7361
6	8	0.0016760	0.0057719	-0.009645	0.0129973	0.7716
3	4	0.0014444	0.0054363	-0.009219	0.0121076	0.7905
8	3	0.0013240	0.0057066	-0.009869	0.0125173	0.8166
6	1	0.0009091	0.0057064	-0.010284	0.0121019	0.8734
2	5	0.0008792	0.0055378	-0.009983	0.0117414	0.8739
1	8	0.0007669	0.0058659	-0.010739	0.0122725	0.8960

Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

HSD Threshold Matrix

Connecting Letters Report

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
7	4	0.0115063	0.0055489	-0.005334	0.0283468	0.4326
7	3	0.0100619	0.0055862	-0.006892	0.0270158	0.6194
2	4	0.0095960	0.0055192	-0.007155	0.0263466	0.6618
7	8	0.0087378	0.0058139	-0.008907	0.0263827	0.8061
5	4	0.0087167	0.0054169	-0.007723	0.0251568	0.7449
2	3	0.0081515	0.0055568	-0.008713	0.0250162	0.8250
7	1	0.0079709	0.0057488	-0.009476	0.0254184	0.8636
5	3	0.0072723	0.0054552	-0.009284	0.0238285	0.8862
7	6	0.0070619	0.0056529	-0.010094	0.0242181	0.9169
2	8	0.0068275	0.0057856	-0.010732	0.0243866	0.9376
2	1	0.0060606	0.0057203	-0.011300	0.0234213	0.9649
5	8	0.0059483	0.0056881	-0.011315	0.0232113	0.9673
5	1	0.0051814	0.0056216	-0.011880	0.0222426	0.9840
2	6	0.0051515	0.0056238	-0.011917	0.0222196	0.9846
6	4	0.0044444	0.0055048	-0.012262	0.0211513	0.9927
5	6	0.0042723	0.0055234	-0.012491	0.0210356	0.9944
1	4	0.0035354	0.0056033	-0.013470	0.0205411	0.9985
6	3	0.0030000	0.0055425	-0.013821	0.0198212	0.9994
7	5	0.0027896	0.0055673	-0.014107	0.0196861	0.9997
8	4	0.0027685	0.0056700	-0.014440	0.0199767	0.9997
1	3	0.0020909	0.0056403	-0.015027	0.0192090	1.0000
7	2	0.0019103	0.0056669	-0.015289	0.0191092	1.0000
6	8	0.0016760	0.0057719	-0.015841	0.0191933	1.0000
3	4	0.0014444	0.0054363	-0.015055	0.0179435	1.0000
8	3	0.0013240	0.0057066	-0.015995	0.0186433	1.0000
6	1	0.0009091	0.0057064	-0.016409	0.0182276	1.0000
2	5	0.0008792	0.0055378	-0.015928	0.0176862	1.0000
1	8	0.0007669	0.0058659	-0.017036	0.0185695	1.0000

Figure 15 Pair comparison tests using student's t and Tukey-Kramer method

The results have a p-values (smallest 0.0383) greater than critical value of 0.01. This statistical evidence can be used to accept the null hypothesis and conclude that the reaction time is consistent across all lanes. Hence, the hypothesis Lanes 3 and 4 swimmers react faster is not true.

4.2. Insights from Gender

4.2.1. Analysis 3: Male swimmers react faster than female swimmers

Across all events, men had lower mean and median reaction times than women. Women swimmers also demonstrated higher variability in their reaction time, as shown by the higher standard deviation.

Table 4 Reaction Times Statistics across Gender

Gender	Reaction Times				
	N	Mean	Median	Std Dev	IQR
Men	872	0.645	0.64	0.053	0.07
Women	732	0.680	0.68	0.057	0.07

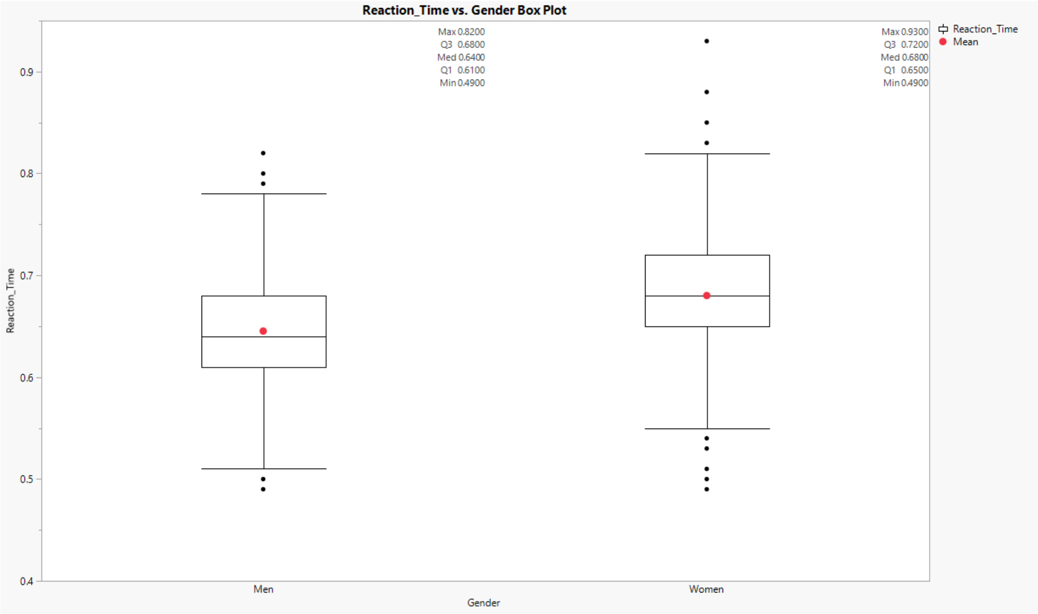
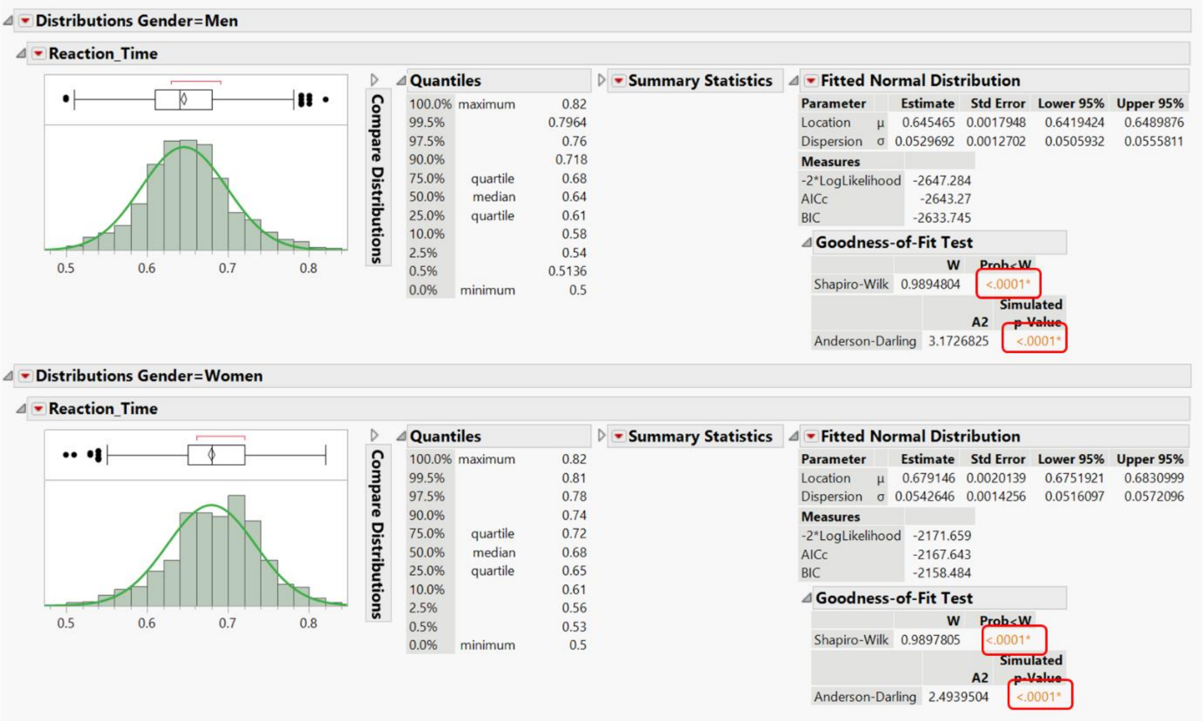


Figure 16 Boxplot of Reaction Time across Genders

4.2.1.1. Hypothesis 2: Males have faster reaction time than Female swimmers

We are interested to test if there is evidence to suggest that the reaction times are different among male and female swimmers. A normality test was conducted to determine the appropriate hypothesis testing method. In this hypothesis, we will use Shapiro-Wilk and Anderson-Darling Goodness-of-Fit Test with confidence level of 95% to test:

- H_0: Reaction Time is normally distributed
- H_1: Reaction Time is not normally distributed



Distributions Gender=Men

Reaction_Time

Quantiles

100.0%	maximum	0.82
99.5%		0.7964
97.5%		0.76
90.0%		0.718
75.0%	quartile	0.68
50.0%	median	0.64
25.0%	quartile	0.61
10.0%		0.58
2.5%		0.54
0.5%		0.5136
0.0%	minimum	0.5

Summary Statistics

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.645465	0.0017948	0.6419424	0.6489876
Dispersion	σ	0.0529692	0.0012702	0.0505932	0.0555811

Measures	
-2*LogLikelihood	-2647.284
AICc	-2643.27
BIC	-2633.745

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9894804	<.0001*

	A2	Simulated p-Value
Anderson-Darling	3.1726825	<.0001*

Distributions Gender=Women

Reaction_Time

Quantiles

100.0%	maximum	0.82
99.5%		0.81
97.5%		0.78
90.0%		0.74
75.0%	quartile	0.72
50.0%	median	0.68
25.0%	quartile	0.65
10.0%		0.61
2.5%		0.56
0.5%		0.53
0.0%	minimum	0.5

Summary Statistics

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.679146	0.0020139	0.6751921	0.6830999
Dispersion	σ	0.0542646	0.0014256	0.0516097	0.0572096

Measures	
-2*LogLikelihood	-2171.659
AICc	-2167.643
BIC	-2158.484

Goodness-of-Fit Test

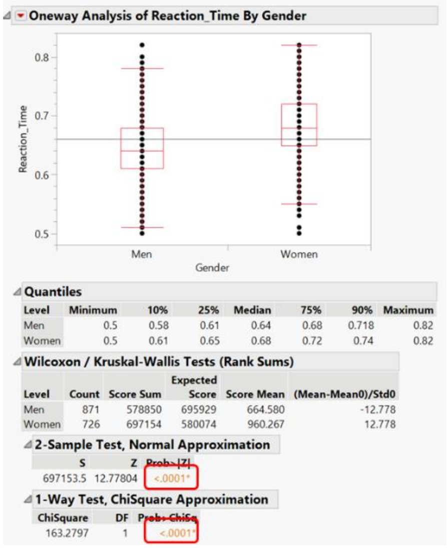
	W	Prob<W
Shapiro-Wilk	0.9897805	<.0001*

	A2	Simulated p-Value
Anderson-Darling	2.4939504	<.0001*

Figure 17 Normal distribution test for Reaction Time across Genders

Based on the results above, we reject that reaction time for all lanes is normally distributed as p-values < 0.0001 which are less than critical value of 0.05. Hence, the hypothesis testing will be conducted using non-parametric Kruskal-Wallis test, with 95% confidence level. Our hypothesis is defined as:

- H_0: Reaction Time is same for both genders
- H_1: Reaction Time is not the same for both genders



Oneway Analysis of Reaction_Time By Gender

Quantiles

Level	Minimum	10%	25%	Median	75%	90%	Maximum
Men	0.5	0.58	0.61	0.64	0.68	0.718	0.82
Women	0.5	0.61	0.65	0.68	0.72	0.74	0.82

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Men	871	578850	695929	664.580	-12.778
Women	726	697154	580074	960.267	12.778

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
697153.5	12.77804	<.0001*

1-Way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
163.2797	1	<.0001*

Figure 18 Kruskal-Wallis test for Reaction Time across Genders

The results have a p-values less than critical value of 0.01. We have evidence to reject the the null hypothesis and conclude that the reaction time is consistent not the same for both genders. Hence, the hypothesis that male swimmers have faster reaction time than female swimmers it True.

4.2.2. Analysis 4: Most of the male swimmers are faster than female swimmers for Freestyle 200m event.

From the boxplot below, male swimmers are about 11s faster than their female counterparts for the Freestyle 200m race. Other than 1 outlier, all other male swimmers are faster than the fastest female swimmer. Women freestyle swimmers have higher variability in their final timing. The distribution of final time for both genders are skewed right, with the means higher than the median.

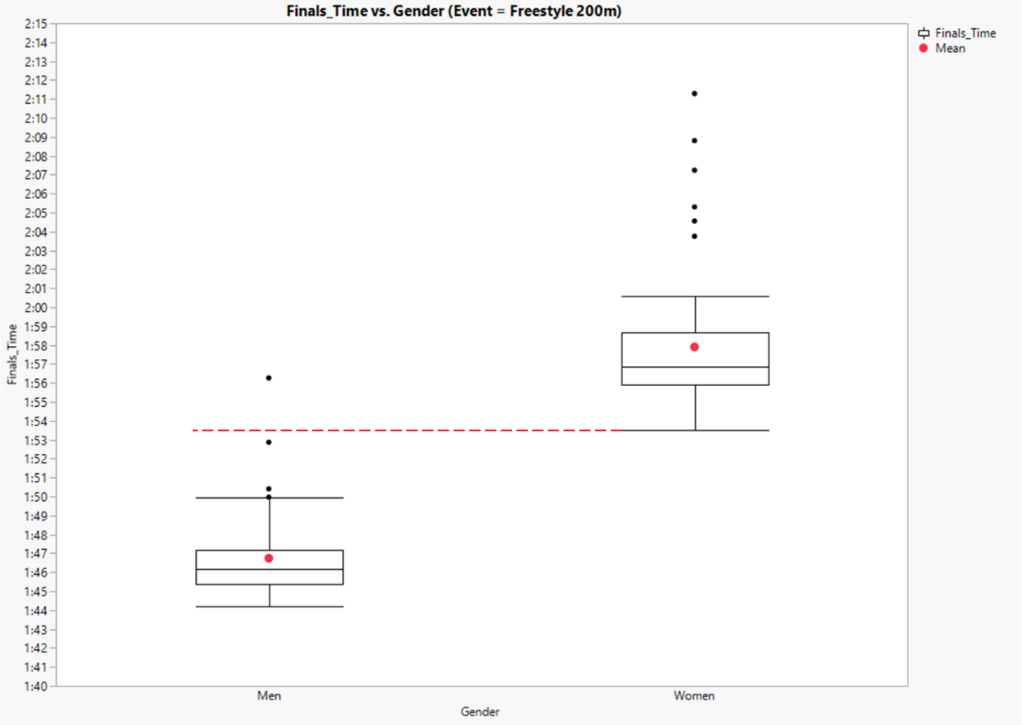


Figure 19 Boxplot of Final Time across Gender for 200m Freestyle Event

4.3. Insights from Events

4.3.1. Analysis 5: Longer distance events tend to have longer reaction times

Events with shorter distances (<400m) tend to have lower mean and median reaction times than the longer distance events (≥ 400m).

Table 5 Statistics of reaction time across Event's distance

Distance	Reaction Times				
	N	Mean	Median	Std Dev	IQR
50m	202	0.667	0.66	0.055	0.07
100m	562	0.642	0.64	0.052	0.07
200m	543	0.653	0.66	0.052	0.07
400m	140	0.699	0.70	0.050	0.07
800m	79	0.716	0.72	0.047	0.07
1500m	77	0.717	0.72	0.059	0.07

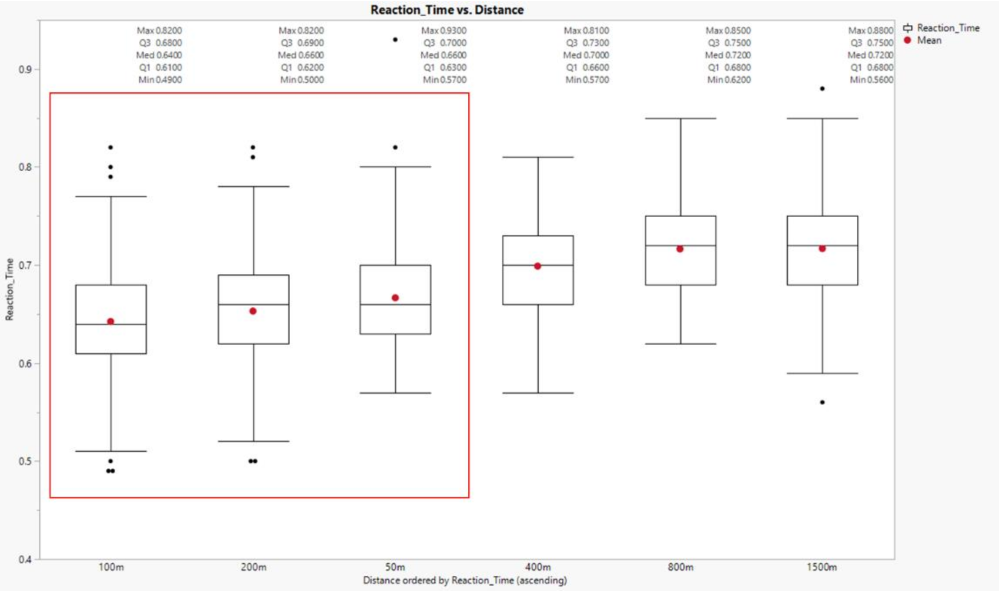


Figure 20 Boxplot of Reaction Time across Distance

4.3.2. Analysis 6: Backstroke swimmers react faster than the other swimmers

Backstroke swimmers have the lowest mean and median reaction times, which indicates they have the fastest reaction time. Breaststroke reaction time distribution is the narrowest inter-quartile range and variability, which may represent more consistency in their reaction.

Table 6 Statistics of Reaction Time Across Event's Stokes

Stroke	Reaction Times				
	N	Mean	Median	Std Dev	IQR
Backstroke	232	0.598	0.60	0.053	0.07
Breaststroke	255	0.667	0.67	0.038	0.05
Butterfly	222	0.653	0.65	0.047	0.07
Freestyle	713	0.680	0.68	0.055	0.08
Individual Medley	182	0.669	0.66	0.050	0.06

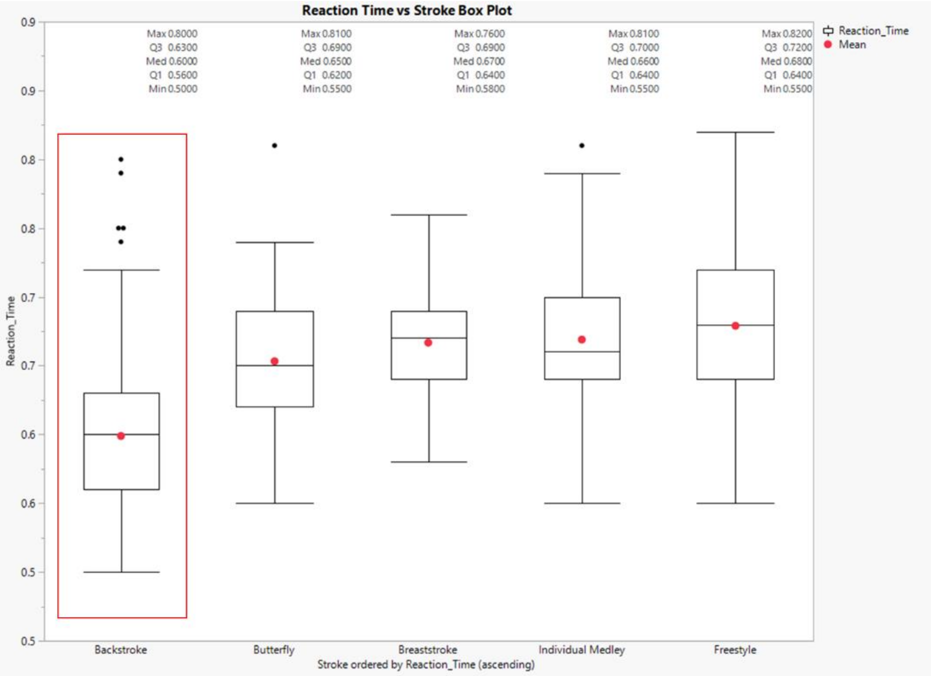


Figure 21 Boxplot of Reaction Time across Strokes

4.3.3. Analysis 7: Freestyle swimmers are the fastest swimmers

Freestyle swimmers have the lowest mean and median final timings, indicating they are faster than the other strokes. Despite backstroke swimmers having the fastest reaction time, they end up with mid-range final times.

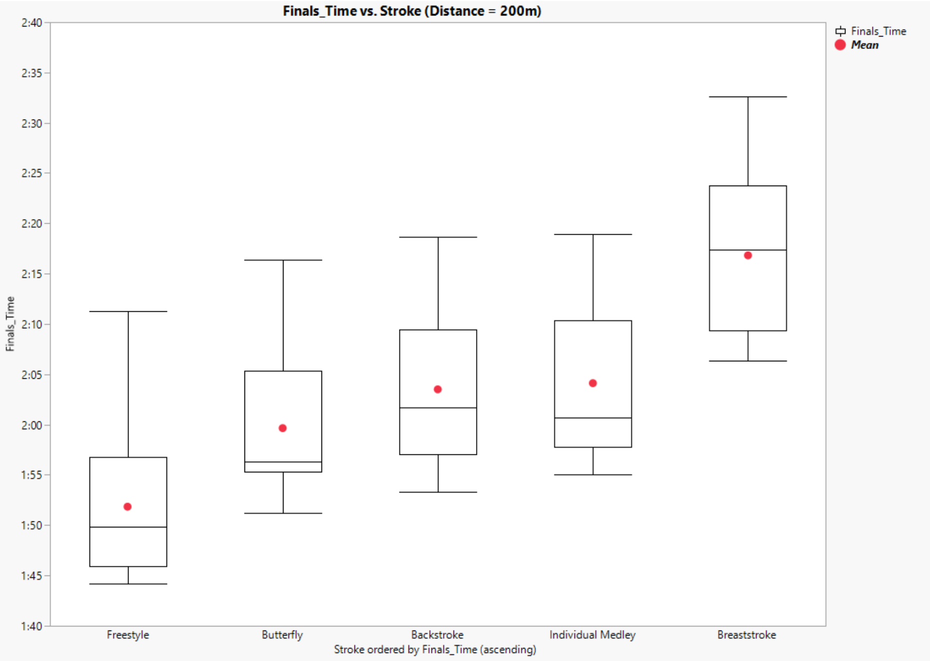


Figure 22 Boxplot of Final Times across Strokes (Limit Distance = 200m)

4.3.4. Analysis 8: Swimmers in Semis event tend to have lower reaction times

Semis event swimmers have the lowest mean and median reaction time of 0.65s, showing that swimmers tend to react slightly faster for the semi events, compared to heats and finals. An unexpected observation is that the reaction time for the finals event has the widest IQR and does not have outliers.

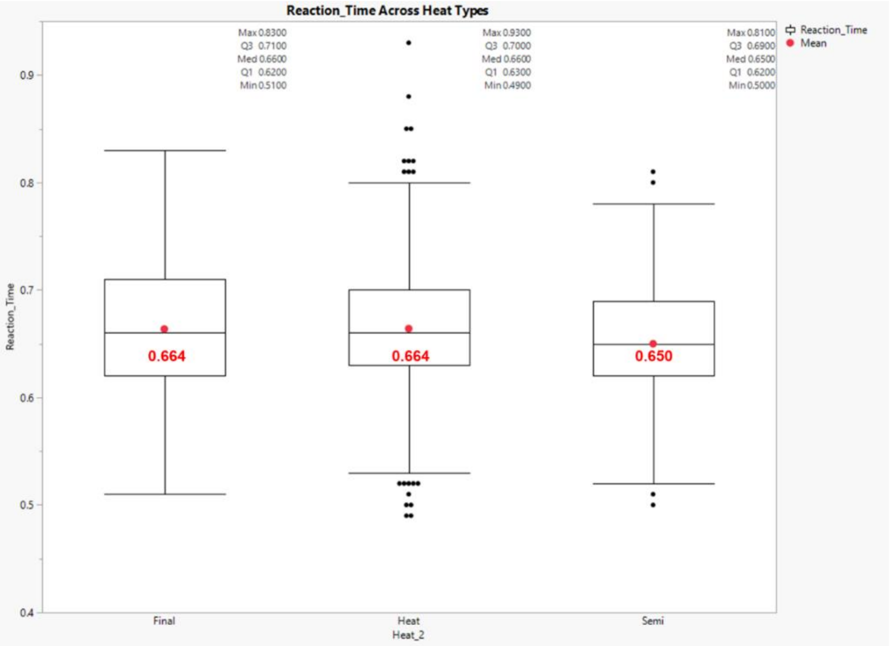


Figure 23 Boxplot of Reaction Time across Heat types

4.4. Insights from Performance

4.4.1. Analysis 9: Reaction time does not have much impact on a swimmer’s placings

The boxplot below shows reaction time fluctuates across swimmers for all placings, with 1st-3rd place swimmers having similar reaction times to 7th-8th place. This may mean that the reaction time may not have a large impact on a swimmer’ performance. Though, in the finals event (2nd boxplot), Gold medalists have the lowest mean (0.649s) and median (0.65s) reaction times, they also have a moderate IQR.

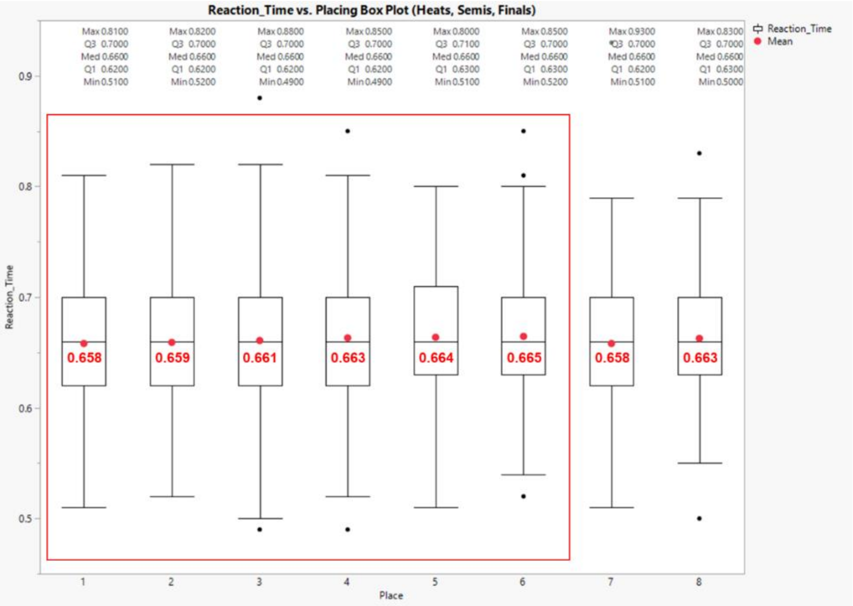


Figure 24 Boxplot of Reaction Time across Placing for Heats, Semis and Finals

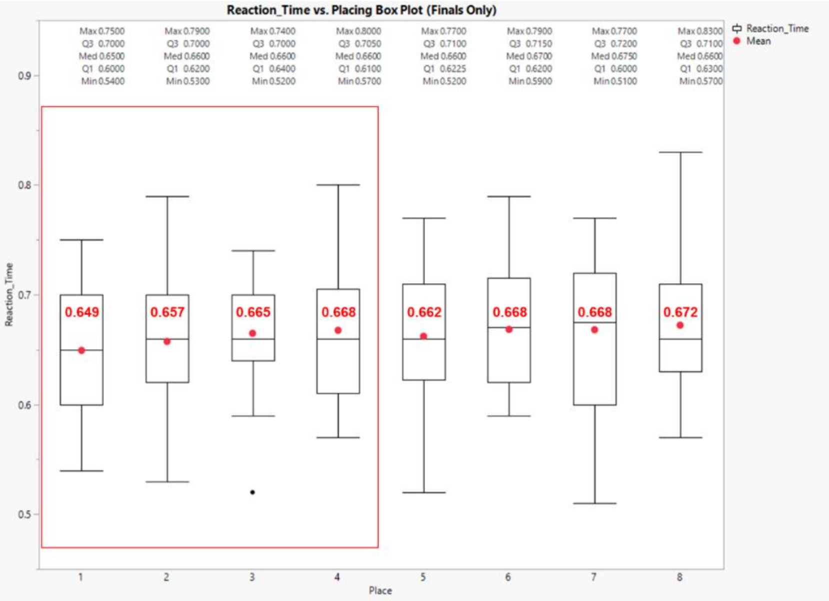
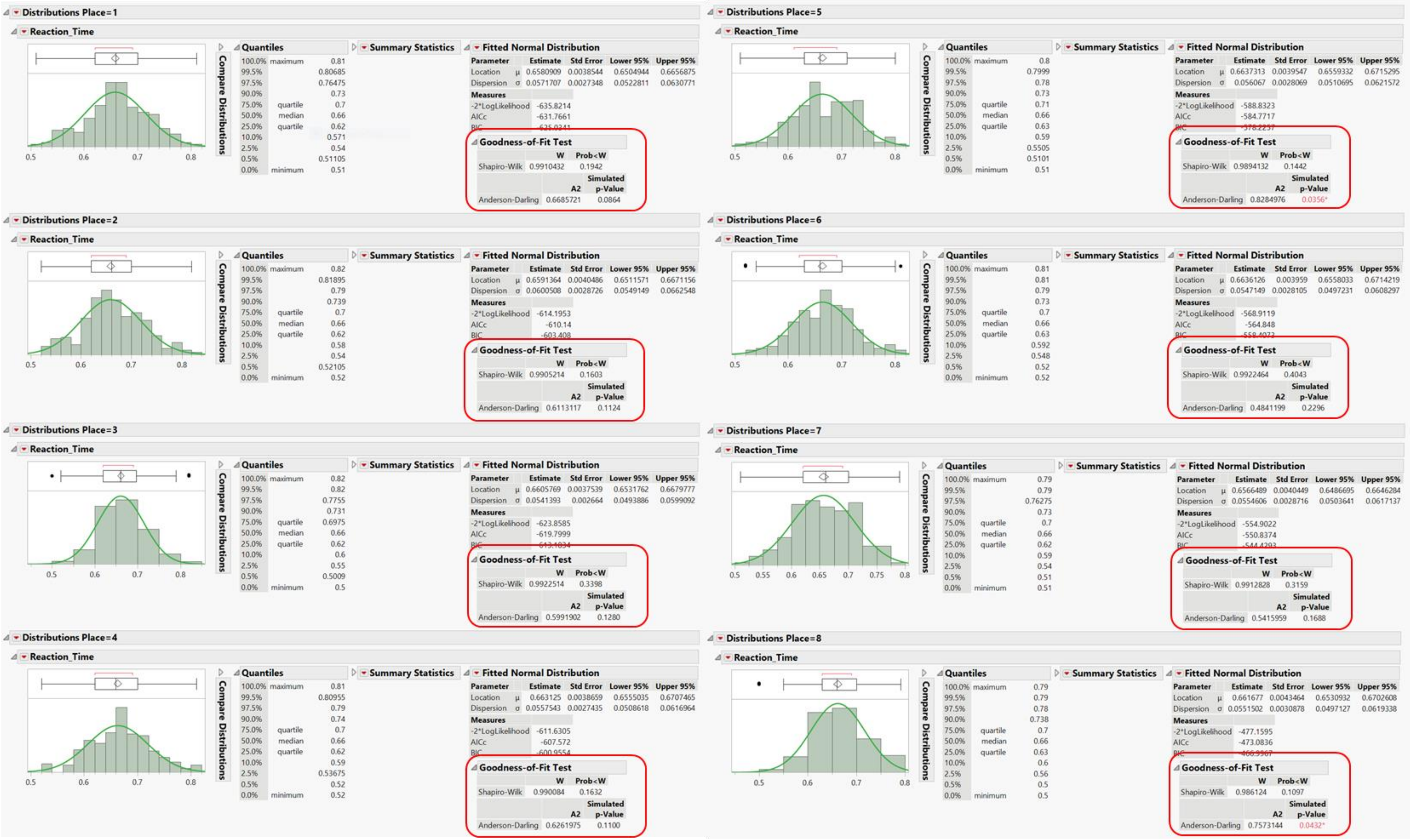


Figure 25 Boxplot of Reaction Time across Placings for Finals events only

4.4.1.1. Hypothesis 3: Swimmers with lower reaction times have better performance

We are interested to test if reaction time impacts the performance of the swimmers. A normality test was conducted to determine the appropriate hypothesis testing method. We will use Goodness-of-Fit Test with confidence level of 99% to test:

- H_0: Reaction Time is normally distributed
- H_1: Reaction Time is not normally distributed



Distributions Place=1

Reaction_Time

Quantiles

100.0%	maximum	0.81
99.5%		0.80685
97.5%		0.76475
90.0%		0.73
75.0%	quartile	0.7
50.0%	median	0.66
25.0%	quartile	0.62
10.0%		0.571
2.5%		0.54
0.5%		0.51105
0.0%	minimum	0.51

Summary Statistics

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6580909	0.0038544	0.6504944	0.6656875
Dispersion	σ	0.0571707	0.0027348	0.0522811	0.0630771

Measures	
-2*LogLikelihood	-635.8214
AICc	-631.7661
BIC	-625.0341

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9910432	0.1942

	A2	Simulated p-Value
Anderson-Darling	0.6685721	0.0864

Distributions Place=2

Reaction_Time

Quantiles

100.0%	maximum	0.82
99.5%		0.81895
97.5%		0.79
90.0%		0.739
75.0%	quartile	0.7
50.0%	median	0.66
25.0%	quartile	0.62
10.0%		0.58
2.5%		0.54
0.5%		0.52105
0.0%	minimum	0.52

Summary Statistics

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6591364	0.0040486	0.6511571	0.6671156
Dispersion	σ	0.0600508	0.0028726	0.0549149	0.0662548

Measures	
-2*LogLikelihood	-614.1953
AICc	-610.14
BIC	-603.408

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9905214	0.1603

	A2	Simulated p-Value
Anderson-Darling	0.6113117	0.1124

Distributions Place=3

Reaction_Time

Quantiles

100.0%	maximum	0.82
99.5%		0.82
97.5%		0.7755
90.0%		0.731
75.0%	quartile	0.6975
50.0%	median	0.66
25.0%	quartile	0.62
10.0%		0.6
2.5%		0.55
0.5%		0.5009
0.0%	minimum	0.5

Summary Statistics

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6605769	0.0037539	0.6531762	0.6679777
Dispersion	σ	0.0541393	0.002664	0.0493886	0.0599092

Measures	
-2*LogLikelihood	-623.8585
AICc	-619.7999
BIC	-613.1834

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9922514	0.3398

	A2	Simulated p-Value
Anderson-Darling	0.5991902	0.1280

Distributions Place=4

Reaction_Time

Quantiles

100.0%	maximum	0.81
99.5%		0.80955
97.5%		0.79
90.0%		0.74
75.0%	quartile	0.7
50.0%	median	0.66
25.0%	quartile	0.62
10.0%		0.59
2.5%		0.53675
0.5%		0.52
0.0%	minimum	0.52

Summary Statistics

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.663125	0.0038659	0.6555035	0.6707465
Dispersion	σ	0.0557543	0.0027435	0.0508618	0.0616964

Measures	
-2*LogLikelihood	-611.6305
AICc	-607.572
BIC	-600.9554

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.990084	0.1632

	A2	Simulated p-Value
Anderson-Darling	0.6261975	0.1100

Distributions Place=5

Reaction_Time

Quantiles

100.0%	maximum	0.8
99.5%		0.7999
97.5%		0.78
90.0%		0.73
75.0%	quartile	0.71
50.0%	median	0.66
25.0%	quartile	0.63
10.0%		0.59
2.5%		0.5505
0.5%		0.5101
0.0%	minimum	0.51

Summary Statistics

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6637313	0.0039547	0.6559332	0.6715295
Dispersion	σ	0.056067	0.0028069	0.0510695	0.0621572

Measures	
-2*LogLikelihood	-588.8323
AICc	-584.7717
BIC	-578.2257

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9894132	0.1442

	A2	Simulated p-Value
Anderson-Darling	0.8284976	0.0356*

Distributions Place=6

Reaction_Time

Quantiles

100.0%	maximum	0.81
99.5%		0.81
97.5%		0.79
90.0%		0.73
75.0%	quartile	0.7
50.0%	median	0.66
25.0%	quartile	0.63
10.0%		0.592
2.5%		0.548
0.5%		0.52
0.0%	minimum	0.52

Summary Statistics

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6636126	0.003959	0.6558033	0.6714219
Dispersion	σ	0.0547149	0.0028105	0.0497231	0.0608297

Measures	
-2*LogLikelihood	-568.9119
AICc	-564.848
BIC	-558.4073

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9922464	0.4043

	A2	Simulated p-Value
Anderson-Darling	0.4841199	0.2296

Distributions Place=7

Reaction_Time

Quantiles

100.0%	maximum	0.79
99.5%		0.79
97.5%		0.76275
90.0%		0.73
75.0%	quartile	0.7
50.0%	median	0.66
25.0%	quartile	0.62
10.0%		0.59
2.5%		0.54
0.5%		0.51
0.0%	minimum	0.51

Summary Statistics

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6566489	0.0040449	0.6486695	0.6646284
Dispersion	σ	0.0554606	0.0028716	0.0503641	0.0617137

Measures	
-2*LogLikelihood	-554.9022
AICc	-550.8374
BIC	-544.4293

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9912828	0.3159

	A2	Simulated p-Value
Anderson-Darling	0.5415959	0.1688

Distributions Place=8

Reaction_Time

Quantiles

100.0%	maximum	0.79
99.5%		0.79
97.5%		0.78
90.0%		0.738
75.0%	quartile	0.7
50.0%	median	0.66
25.0%	quartile	0.63
10.0%		0.6
2.5%		0.56
0.5%		0.5
0.0%	minimum	0.5

Summary Statistics

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.661677	0.0043464	0.6530932	0.6702608
Dispersion	σ	0.0551502	0.0030878	0.0497127	0.0619338

Measures	
-2*LogLikelihood	-477.1595
AICc	-473.0836
BIC	-466.9967

Goodness-of-Fit Test

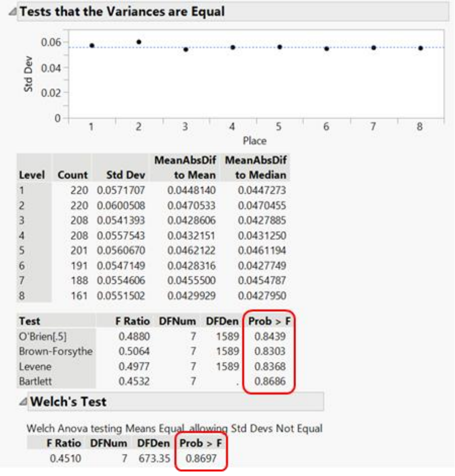
	W	Prob<W
Shapiro-Wilk	0.986124	0.1097

	A2	Simulated p-Value
Anderson-Darling	0.7573144	0.0432*

Figure 26 Normality Test for Reaction Time across Placings

Based on the results above, we accept that reaction time for all placings is normally distributed as smallest p-value 0.0356 (5th place) > critical value 0.01. Hence, the hypothesis testing will be conducted using parametric Welch test, then compare using student's t and Tukey-Kramer with 99% confidence level. Our hypothesis is defined as:

- H_0: Reaction Time is same for all performance
- H_1: Reaction Time is not the same for all performance



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
1	220	0.0571707	0.0448140	0.0447273
2	220	0.0600508	0.0470533	0.0470455
3	208	0.0541393	0.0428606	0.0427885
4	208	0.0557543	0.0432151	0.0431250
5	201	0.0560670	0.0462122	0.0461194
6	191	0.0547149	0.0428316	0.0427749
7	188	0.0554606	0.0455500	0.0454787
8	161	0.0551502	0.0429929	0.0427950

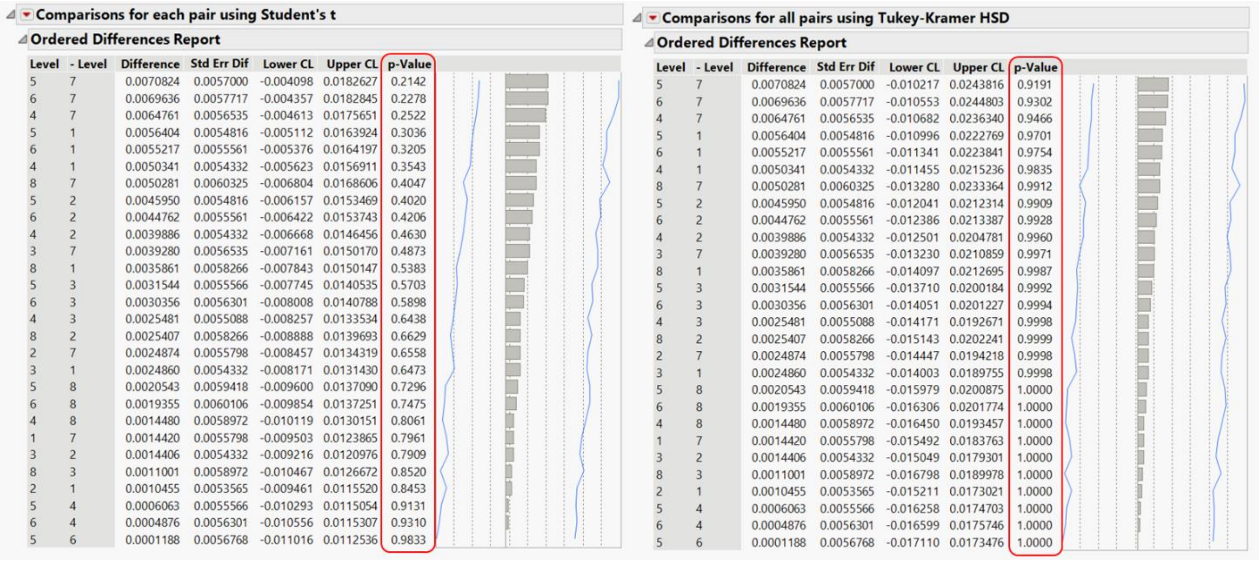
Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	0.4880	7	1589	0.8439
Brown-Forsythe	0.5064	7	1589	0.8303
Levene	0.4977	7	1589	0.8368
Bartlett	0.4532	7	.	0.8686

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
0.4510	7	673.35	0.8697

Figure 27 Welch's test for Reaction Time across performances



Comparisons for each pair using Student's t

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
5	7	0.0070824	0.0057000	-0.004098	0.0182627	0.2142
6	7	0.0069636	0.0057717	-0.004357	0.0182845	0.2278
4	7	0.0064761	0.0056535	-0.004613	0.0175651	0.2522
5	1	0.0056404	0.0054816	-0.005112	0.0163924	0.3036
6	1	0.0055217	0.0055561	-0.005376	0.0164197	0.3205
4	1	0.0050341	0.0054332	-0.005623	0.0156911	0.3543
8	7	0.0050281	0.0060325	-0.006804	0.0168606	0.4047
5	2	0.0045950	0.0054816	-0.006157	0.0153469	0.4020
6	2	0.0044762	0.0055561	-0.006422	0.0153743	0.4206
4	2	0.0039886	0.0054332	-0.006668	0.0146456	0.4630
3	7	0.0039280	0.0056535	-0.007161	0.0150170	0.4873
8	1	0.0035861	0.0058266	-0.007843	0.0150147	0.5383
5	3	0.0031544	0.0055566	-0.007745	0.0140535	0.5703
6	3	0.0030356	0.0056301	-0.008008	0.0140788	0.5898
4	3	0.0025481	0.0055088	-0.008257	0.0133534	0.6438
8	2	0.0025407	0.0058266	-0.008888	0.0139693	0.6629
2	7	0.0024874	0.0055798	-0.008457	0.0134319	0.6558
3	1	0.0024860	0.0054332	-0.008171	0.0131430	0.6473
5	8	0.0020543	0.0059418	-0.009600	0.0137090	0.7296
6	8	0.0019355	0.0060106	-0.009854	0.0137251	0.7475
4	8	0.0014480	0.0058972	-0.010119	0.0130151	0.8061
1	7	0.0014420	0.0055798	-0.009503	0.0123865	0.7961
3	2	0.0014406	0.0054332	-0.009216	0.0120976	0.7909
8	3	0.0011001	0.0058972	-0.010467	0.0126672	0.8520
2	1	0.0010455	0.0053565	-0.009461	0.0115520	0.8453
5	4	0.0006063	0.0055566	-0.010293	0.0115054	0.9131
6	4	0.0004876	0.0056301	-0.010556	0.0115307	0.9310
5	6	0.0001188	0.0056768	-0.011016	0.0112536	0.9833

Comparisons for all pairs using Tukey-Kramer HSD

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
5	7	0.0070824	0.0057000	-0.010217	0.0243816	0.9191
6	7	0.0069636	0.0057717	-0.010553	0.0244803	0.9302
4	7	0.0064761	0.0056535	-0.010682	0.0236340	0.9466
5	1	0.0056404	0.0054816	-0.010996	0.0222769	0.9701
6	1	0.0055217	0.0055561	-0.011341	0.0223841	0.9754
4	1	0.0050341	0.0054332	-0.011455	0.0215236	0.9835
8	7	0.0050281	0.0060325	-0.013280	0.0233364	0.9912
5	2	0.0045950	0.0054816	-0.012041	0.0212314	0.9909
6	2	0.0044762	0.0055561	-0.012386	0.0213387	0.9928
4	2	0.0039886	0.0054332	-0.012501	0.0204781	0.9960
3	7	0.0039280	0.0056535	-0.013230	0.0210859	0.9971
8	1	0.0035861	0.0058266	-0.014097	0.0212695	0.9987
5	3	0.0031544	0.0055566	-0.013710	0.0200184	0.9992
6	3	0.0030356	0.0056301	-0.014051	0.0201227	0.9994
4	3	0.0025481	0.0055088	-0.014171	0.0192671	0.9998
8	2	0.0025407	0.0058266	-0.015143	0.0202241	0.9999
2	7	0.0024874	0.0055798	-0.014447	0.0194218	0.9998
3	1	0.0024860	0.0054332	-0.014003	0.0189755	0.9998
5	8	0.0020543	0.0059418	-0.015979	0.0200875	1.0000
6	8	0.0019355	0.0060106	-0.016306	0.0201774	1.0000
4	8	0.0014480	0.0058972	-0.016450	0.0193457	1.0000
1	7	0.0014420	0.0055798	-0.015492	0.0183763	1.0000
3	2	0.0014406	0.0054332	-0.015049	0.0179301	1.0000
8	3	0.0011001	0.0058972	-0.016798	0.0189978	1.0000
2	1	0.0010455	0.0053565	-0.015211	0.0173021	1.0000
5	4	0.0006063	0.0055566	-0.016258	0.0174703	1.0000
6	4	0.0004876	0.0056301	-0.016599	0.0175746	1.0000
5	6	0.0001188	0.0056768	-0.017110	0.0173476	1.0000

Figure 28 Student's t and Tukey-Kramer pair test for Reaction Time across placings

The results have a p-values (smallest 0.2142) greater than critical value of 0.01. This statistical evidence can be used to accept the null hypothesis and conclude that the reaction time is consistent across placings. Hence, the hypothesis that swimmers with better performance have lower reaction time is false. This evidence suggests that reaction time does not affect the performance of the swimmers to a large extent.

4.4.2. Analysis 10: Swimmers in the centre lanes tend to have higher place finishes

The dot plot below shows that swimmers with better performance seem to be placed in the centre Lanes 3-4, with lane 4 having the highest concentration of Gold medalists. Most podium finishers (1st, 2nd, 3rd) were also competing in Lanes 3-5.

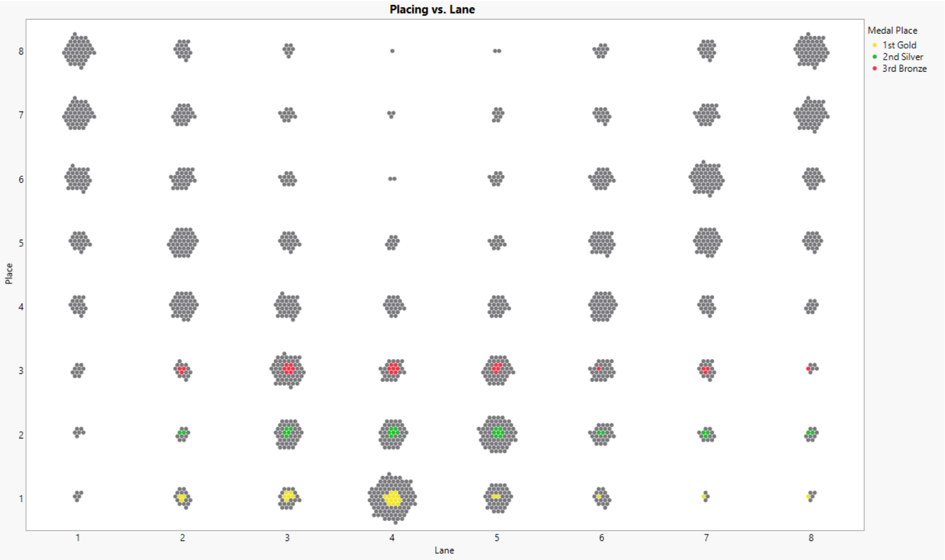


Figure 29 Dotplot of Swimmer's Placing vs Lane

4.4.3. Analysis 11: 96% of Olympic finalists in Lane 4 end up with podium finishes

The heatmap below shows that of the swimmers who compete in Lane 4 finals, 96% end up in the Top 3, and receive medals. There is also a high proportion of Olympic medalists in Lanes 3-5, where >50% of the finalists end up with podium finishes.

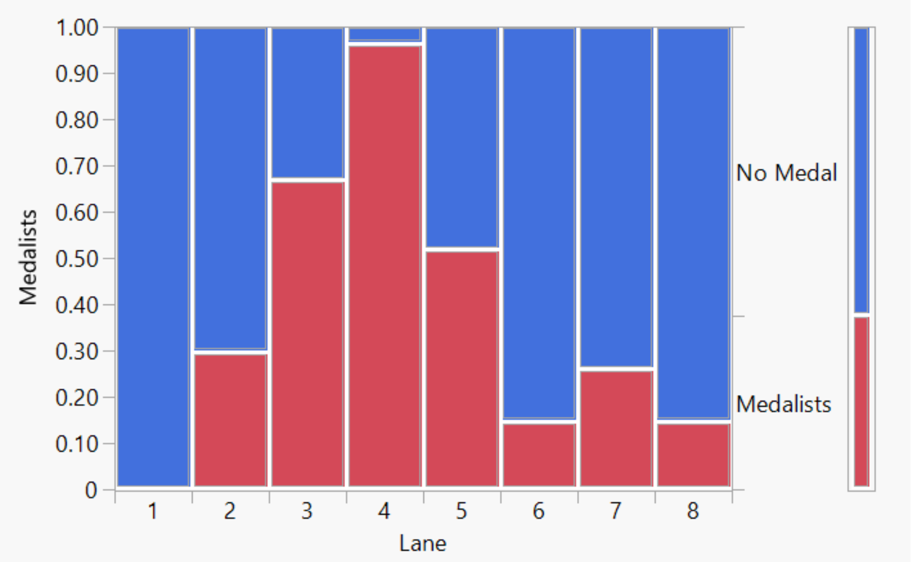


Figure 30 Heatmap of Medalists across Lanes for Olympic Swimming Finalists

4.4.4. Analysis 12: There is little correlation between the Final Time and Reaction Time

The regression between the Final Time and Reaction Time for different events, with distance of 100m, has poor line of fit. The highest R^2 is 0.283, from the Butterfly 100m event.

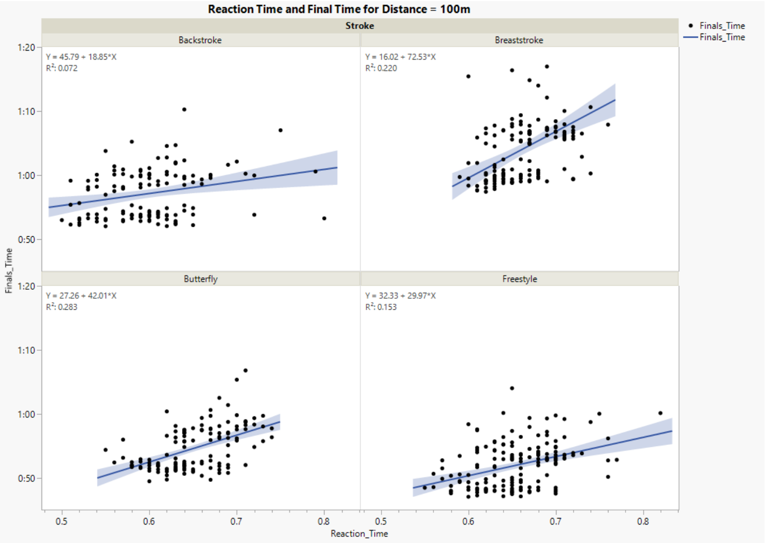
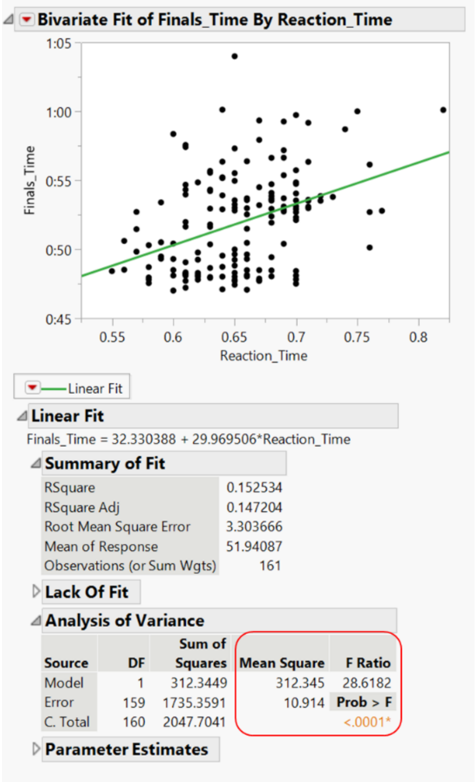


Figure 31 Regression of Final Time and Reaction Time for Event Distance = 100m

4.4.4.1. Hypothesis 4: Final Time increases as Reaction Time increases

In this hypothesis testing, we shortlist the Freestyle 100m event data, to test if there is any relationship between the Final Time (Performance) and Reaction Time of the swimmers. Bivariate fit and analysis, with a 95% confidence level is used to test the hypothesis:

H_0: True slope coefficient is 0.
H_1: True slope coefficient is not 0.



Linear Fit

Finals_Time = 32.330388 + 29.969506*Reaction_Time

Summary of Fit

RSquare	0.152534
RSquare Adj	0.147204
Root Mean Square Error	3.303666
Mean of Response	51.94087
Observations (or Sum Wgts)	161

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	312.3449	312.345	28.6182
Error	159	1735.3591	10.914	**Prob > F**
C. Total	160	2047.7041		<.0001*

Parameter Estimates

Figure 32 Bivariate Fit Test for Final Time and Reaction Time for Freestyle 100m Event

Based on the test result, Final Time and Reaction Time has a p-value (<0.001) less than critical value. Hence, the null hypothesis is rejected. Since the variables are positively correlated (R^2 = 0.15), we can conclude that the Final Time increases as Reaction Time increases, and the hypothesis is true.

5. Interpretation of Analysis Results

5.1. Male swimmers have faster reaction time than female swimmers

Male swimmers have a mean and median reaction time of 0.645s and 0.64s respectively, whereas female swimmers have a mean and median reaction time of 0.680s and 0.68s respectively (Analysis 3: Male swimmers react faster than female swimmers. This is further reiterated in hypothesis testing, whereby the null hypothesis of male and female swimmers having the same reaction time was rejected. (Hypothesis 2: Males have faster reaction time than Female swimmers). This may be due to physical attributes of the human body.

5.2. Reaction time increases across events with longer distances

Swimmers who participate in events with shorter distances (<400m) tend to have lower mean and median reaction 0.64s – 0.67s, as compared to the longer distance events (≥ 400m) of 0.70s – 0.72s Analysis 5: Longer distance events tend to have longer reaction times. This may be because there are other factors that may affect performance for longer distance events.

5.3. Backstroke swimmers have the fastest reaction time but does not final time

Backstroke swimmers have the lowest mean and median reaction times of 0.598s and 0.60s respectively Analysis 6: Backstroke swimmers react faster than the other swimmers . However, amongst all strokes in 200m events, they only have mid-range mean and median final times Analysis 7: Freestyle swimmers are the fastest swimmers. This shows that there are other stroke-related factors impact a swimmers' performance.

5.4. Although the centre lanes tend to churn out more medallists and higher placed finishers, the reaction time across all lanes are consistent

A higher promotion of well performing athletes tend to stem from the center Lanes 3-5 (Analysis 10: Swimmers in the centre lanes tend to have higher place finishes), even with > 50% of finalists in Lanes 3-5 ending up with podium finishes (Analysis 11: 96% of Olympic finalists in Lane 4 end up with podium finishes). However, the hypothesis testing shows that the athletes in the centre do not have faster reaction time, it is consistent across all lanes 1-8 Hypothesis 1: Swimmers in Lanes 3 and 4 react faster than the swimmers in the other lanes. Hence, the reason centre lane swimmers perform better is unlikely due to better reaction time.

5.5. Reaction time does not significantly affect the performance of the swimmers

From the 100m Freestyle event analysis, the Final Time and Reaction Time of a swimmer is positively correlated but has a poor correlation value and line of fit (Analysis 12: There is little correlation between the Final Time and Reaction Time, Hypothesis 4: Final Time increases as Reaction Time increases).

Reaction time also fluctuates across swimmers for all placings, with 1st-3rd place swimmers having similar reaction times to 7th-8th place Analysis 9: Reaction time does not have much impact on a swimmer's placings. The hypothesis testing also provides statistical evidence that swimmers with better placing have lower reaction time is false Hypothesis 3: Swimmers with lower reaction times have better performance. This evidence suggests that reaction time does not impact the performance of the swimmers to a large extent.

6. Recommendation

The results from this study show that generally, the impact of reaction time on the performance of an Olympic swimmer is minor. However, that does not mean that the impact of the reaction time can be entirely neglected, it is dependent on individual performance and other factors. The recommendation for an athlete/coaches would be to compare their individual performance across the data in this study to determine if there are room of improvement.

To improve the study, a suggestion is to widen the scope to collect and analyse other factors like age, start time, stroke distance, stroke speed, turning time and its impact on performance. Finally, a separate study can be done for relay team performance.

7. Appendix

7.1. Appendix A: Data Preparation Change Log

Dataset: Swimming

Item	Variable Name	Issue	Action
1	Place	Should not be continuous modelling type.	Change to nominal modelling type.
2	Lane	Should not be continuous modelling type.	Change to nominal modelling type.
3	DQ	Should not be continuous modelling type.	Change to nominal modelling type.
4	Split_100, 150, 200, 300, 350, 400, 500, 550, 600, 700, 750, 800	Should not be in character data and nominal modelling type.	Change to numeric data type and continuous modelling type.
5	Finals_Time	Inconsistent format for data < 1 min (ss) and data > 1min (min:ss).	Add in new column with formula, Finals_Time_2 to standardise data format (min:ss, 2 decimal place).
6	Finals_Time_2	Column is locked and contains 'NA' information for disqualified swimmers.	Recode into new unlocked column, Finals_Time_recoded and remove 'NA' data.
7	Finals_Time_recoded	Should not be in character data, and nominal modelling type.	Change to numeric data, and continuous modelling type.
8	Finals_Time	Intermediate variable and cannot be used for analysis.	Change naming to Finals_Time_1
9	Finals_Time_recoded	Final variable to be used for analysis.	Change naming to Finals_Time
10	Finals_Time Finals_Time_1 Finals_Time_2	Duplicate columns	Hide and exclude columns for Finals_Time_1 and Finals_Time_2.
11	Event	Contains too many categories, will be difficult for analysis.	Recode and categorised into 'Gender', 'Distance' and 'Stroke' columns.
12	Heat	Contains too many categories, will be difficult for analysis.	Recode into 'Heat_Type', categorised into 'Heat', 'Semi' and 'Final'
13	Team	Field data contains duplicate naming for some teams i.e. 'AUS' and 'AUS-Australia'	Recode and standardise naming.

14	Name, Team, Reaction_Time, Finals_Time, Split	2 rows by swimmer ‘VILLANUEVA Byanca Melissa’ contains data incorrectly	Refer to raw pdf to update the fields.
15	Reaction_Time	Contains missing data for disqualified swimmers and relay events.	Exclude rows with missing Reaction_Time data.
16	Exhibition	Does not contain much information, all rows showing same value. Data is redundant.	Hide and exclude exhibition field.
17	Split	Too many columns for split timing.	Hide columns
18	Relay_Swimmer Relay_Swimmer_Gender	Since relay events are missing reaction time data and not included in this analysis, fields are redundant.	Hide and exclude Relay_Swimmer and Relay_Swimmer_Gender fields

7.2. Appendix B: Datasets

Item	Dataset Name	Brief Description	Usage
1	swimming_cleaned	Main clean dataset	Used for majority of reaction time analysis, that does not need
2	swimming_cleaned_reaction and final time statistics	Statistical mean and median data extracted from main swimming_cleaned dataset.	Used for brief comparison and extraction of numbers.
3	swimming_cleaned_100m Freestyle	Extracted only 100m Freestyle data from main swimming_cleaned data set	Used for bivariate analysis between Final Time and Reaction Time