Mathematics HL Internal Assessment

Making Formula 1 Predictions

Mathematic method of making predictions for Formula 1 Races

By: Karim Zakir

Introduction

During the race weekend, Formula 1 drivers and teams prepare to perform well during the Sunday race. They prepare during the three practice sessions lasting four hours in total. The drivers get familiarized with the racetrack and the engineers optimize the car. The first and second practice sessions (FP1 and FP2 respectively) both happen on a Friday; the third and final practice session (FP3) happens on a Saturday. Afterwards, qualifying takes place, determining the starting position of the drivers at the Sunday race.

Although the practice sessions do not influence the championship standings, the drivers are still ranked based on their quickest lap time. Based on the results, the experts and the fans analyze the pace of the car and make predictions for the race. The predictions based on the practice sessions can be straightforward; if a driver achieved P6 in the practice session, they are predicted to achieve P6 in the race (these predictions will be referred to as $s(x)$ in the rest of the exploration). This may be inaccurate though. In this exploration, I aim to determine the most accurate predictions using statistics and probability. Since my family always predicts the races during the race weekend, I want to make better predictions than them using mathematical analysis. My predictions will be compared to the standard straight forward predictions mentioned above to determine the most accurate one.

If a driver finishes x^{th}, the position will be stated as Px

To make the predictions, I will only consider FP3, since the teams use the FP1 and the FP2 mostly to test new components and/or gather data, making FP3 the most accurate representation of the race. FP3 results will be the x variable. My predictions for the 2019 season will be based off the 2014-2018 seasons (5 years), known as the hybrid era, due to the use of hybrid engines. The races from the same era will provide the most accurate predictions.

The entire F1 history will be considered later in the investigation to potentially predict the 2019 races.

The y variable will be the mean finishing position of the driver. The predictions will take several variables: $f(x)$, $g(x)$, and $s(x)$. The data for the seasons will be taken from the official Formula 1 website. Due to the amount of data, the calculations will be done using a programming language Python and its libraries. These calculations will be verified by doing the same calculations on a smaller sample.

If a driver crashes or is disqualified during a race or an FP3 without setting a time, they will not be considered for that race weekend. Finally, the drivers who finished after P20 will not be considered, since there are only 20 drivers in the 2019 season.

Analysis

Before making the predictions, it is first important to establish the correlation between the FP3 position and the position in the race.

Firstly, I will plot each FP3 position against the corresponding mean finish position of the drivers for the respective FP3 position.

There will be several major variables used throughout this exploration:

- x = FP3 Position
- y = Mean race result
- r = Pearson's correlation
- $\bar{X}$ = Probability of mean finish race result
- $f(x)$ = First set of predictions
- $g(x)$ = Second set of predictions
- $s(x)$ = straightforward predictions
- c = measure of accuracy

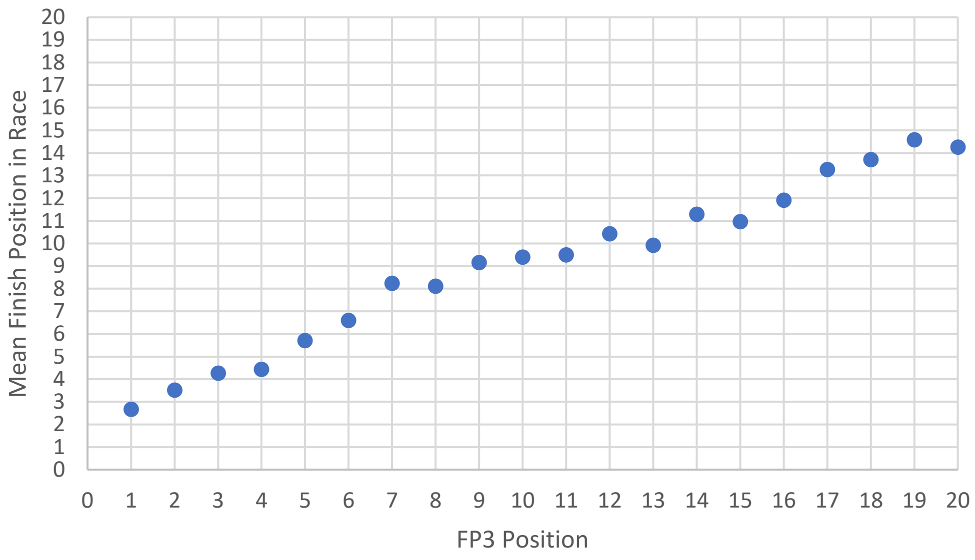
The mean finish position for each FP3 position will be calculated as follows:

$$y = \frac{\sum_{i=1}^{n} p_i}{n}$$

Where p_i is the final position of the driver in each race and n is the total number of races where the driver finishes. The results (**appendix I**) were plotted in a chart.

Instead of mean finish position, other measurements of central tendency, such as mode finish position, could be used to make the predictions; however, using mode, it is impossible to construct probability models, such as the normal distribution model explored later in the investigation.

Fig. 1: FP3 position and mean finishing position in the races from (2014-2018)



At first glance, it may seem that since there is a strong correlation then the straightforward predictions (P6 in FP3 = P6 in the race) become the most accurate. This is not true, since the mean finish position for P18 for example is approximately P13. The possible inaccuracy of the straightforward predictions will be shown later again in the exploration.

Visually, the correlation on figure 1 is extremely strong. This gives us reason to establish this correlation numerically using the Pearson correlation coefficient. In

When finding a relationship between variables x (FP3 position) and y (mean race result), it is also possible to use covariance. In this case, Pearson correlation coefficient was used due to its simplicity and utility.

Throughout the exploration, I will mostly round to 3 significant figures, since this makes the most sense for the figures, such as position and probabilities, presented in this exploration. This establishes consistency. It can be assumed that all values to 3 significant figures have been approximated.

order to obtain the coefficient, we must use the Pearson coefficient formula, which I obtained online[1]. The formula is:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

After determining the necessary values (**appendix II)** and substituting them into the equation, we obtain the following calculation:

$$r = \frac{403.39}{\sqrt{665}\sqrt{251.38}}$$

$$r = \frac{403.39}{25.79 \times 15.85}$$

$$r = \frac{403.39}{408.77}$$

To ensure accuracy, I decided to round to 2 decimal places only in the calculation. The result is in 3 significant figures.

$r = 0.987$. Since the value of r is very close to 1, it confirms the **strong positive correlation** seen in the scatter plot. This allows us to move forward with the exploration, since we established that the FP3 results do correlate with the race results.

Predictions

To make predictions, I will use the normal distribution model, since this model will also provide me with exact probabilities for each possible position. We use the following formula to obtain a normal distribution curve:

It is possible to use the above chart to make predictions by taking the average finishing position for FP3, but it will not give an exact probability the same way a normal distribution curve does.

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

When using this formula, it is assumed that the population is distributed normally. This is not the case, since races can be highly unpredictable.

[1] Stangroom, J., 2020. *Social Science Statistics.* [Online]
Available at: https://www.socscistatistics.com/tests/pearson/
[Accessed 28 October 2019].

Instead, it is possible to use the central limit theorem (CLT). The CLT provides the probability distribution of the sample (2019 season) mean; the mean values could be used to make predictions for the overall season. If those values are used, it is most likely that the drivers on average will finish in those positions. To normally distribute the sample mean for 2019, we must use the following formula[2]:

The analysis only considers the position. It does not consider the skill of a driver or the strength of the team's car, meaning the predictions should not solely rely on the probabilities given by the CLT.

$$\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$$

Where μ is the mean finishing position for each FP3 position, σ^2 is the variance, and n is the number of races from the 2019 season where the driver who finished in x position in FP3 also finished the race.

The process will be completed for each position (**appendix III**); as an example, I will use FP3 P5 to demonstrate.

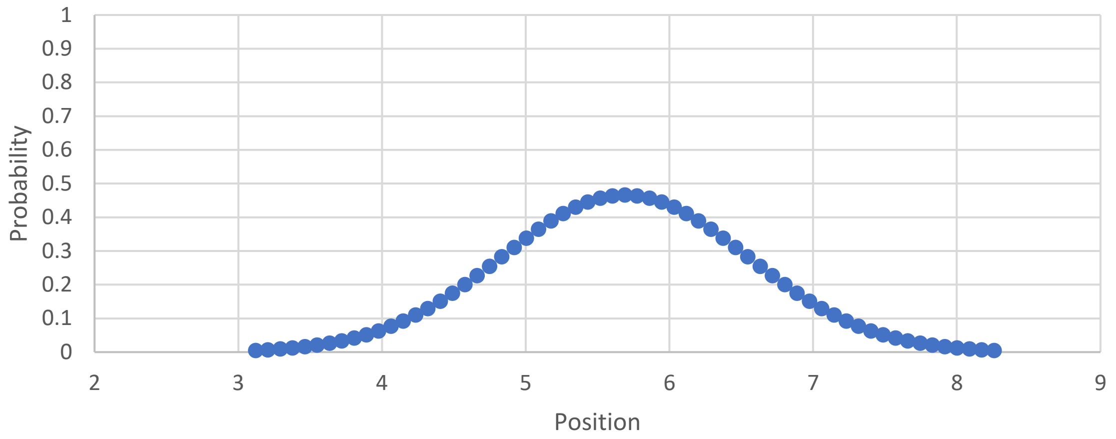
Over the last 5 years, the drivers who finished P5 in FP3 finished on average in the 5.69th position in the race, making $\mu = 5.69$. The variance for their positions is 13.2, meaning $\sigma^2 = 13.2$. This value was calculated by using the variance formula[3]:

$$\sigma^2 = \overline{(x^2)} - (\bar{x})^2$$

During the 2019 season, there were 18 races in which the drivers who finished P5 in the FP3 finished the race, meaning $n = 18$. Using the following values, the following normal distribution curve is obtained.

[2] Paul Fannon, V. K. B. W. S. W., 2012. The normal distribution. In: *Mathematics Higher Level for the IB Diploma.* Cambridge: Cambridge University Press, p. 777.
[3] Paul Fannon, V. K. B. W. S. W., 2012. Measures of Spread. In: *Mathematics Higher Level for the IB Diploma.* Cambridge: Cambridge University Press, p. 691.

Fig. 2: Probability of an average finshing position in 2019 for P5



As seen on figure 2, the most probable whole position is P6; however, even for this position the probability is relatively low – 0.436. According to the model, P5 is not the most probable position, despite it being the most obvious and straightforward choice. When making a prediction for a driver who finishes P5 in FP3, it is most probable to predict P6, since the probability that the drivers will finish in that position on average is the highest.

Another important reason to use normal distribution and the CLT is that if we simply took the mean finishing values and rounded them to use as predictions, several positions would have the exact same predicted finishing position. It is impossible to have two or more cars finishing in the same position. By knowing the probabilities, it is possible to assign different predictions based on the probabilities given by the model. However, for the sake of accuracy and exploration, some of the predictions made in this investigation might have several cars in one position.

Assigning Predictions

There are two methods that I can use to assign predictions using the probabilities found above. These methods will be compared to each other as well as the straightforward predictions to determine the most accurate method.

One method is to assign each finishing position an FP3 position. Each finishing position will be assigned a prediction by calculating which FP3 position has the highest probability of finishing in that position. To elaborate, we will take a predicted position of P1 for example and determine which FP3 position has the highest probability of finishing in that position. In the case of P1, the position is P1 in the FP3, since the probability of a driver who finished P1 in FP3 finishing P1 in the race is 0.0174. In comparison, P2's probability of finishing in P1 is only 0.00148. P1 is now eliminated for other predictions, meaning that even if it has a higher probability of finishing in P2 than P2 itself, it will not be considered. The predictions will be referred to as $f(x)$.

Another method is to take the most probable position for each of the FP3 positions and predict that position. This method is different, since it treats the probabilities and applies them differently. This may mean that some predictions may repeat, but theoretically, a higher probability may lead to higher accuracy. The predictions will be labelled $g(x)$.

Using the methods above, we obtain the following predictions:

x	$f(x)$	$g(x)$
1	1	3
2	2	4
3	3	4
4	4	4
5	5	6
6	6	7
7	8	8
8	7	8
9	9	9
10	20	9
11	19	9
12	18	10
13	10	10
14	17	11
15	11	11
16	12	12
17	13	13
18	14	14
19	15	15
20	16	14

The first method ($f(x)$) of predicting is somewhat problematic. Firstly, some predictions are assigned even if their probability is low. For example, the driver who finished P10 in FP3 is predicted to finish P20. The probability of this happening is 7.18×10^{-34}. This is because the probabilities for P10 are highly spread out (high standard deviation), meaning that other positions have higher probabilities of finishing P10 than P10 itself.

Interestingly, for $f(x)$, the first positions (1-6) follow the straightforward predictions; however, other positions are mostly different, confirming that there are different methods of predicting.

There is also a problem with using the second method ($g(x)$). Some positions are predicted the same outcome which is impossible during the race. Additionally, some outcomes are not covered, meaning there are no predictions for positions like P1 or P2, although there are drivers who will finish in those positions.

Accuracy

There are several ways to measure the accuracy of my predictions. Firstly, I will measure how far my predictions are from the actual finishing position on average. I will measure the same value for $s(x)$ and compare these values. To measure the value for a single position, I will use the absolute error and mean absolute error formula[4]. The formula for c, a variable that will represent accuracy, is:

$$c = \frac{\sum(|p_p - p_r|)}{n}$$

Where p_p is the most probable position, p_r is the actual position, and n is the number of races where the driver finished in the respective position. The results of the formula for each position can be seen in **appendix IV** for all three sets of predictions.

It is possible to apply the following formula to measure the average accuracy of the predictions for every position. To do so, we use the following formula:

$$\bar{c} = \frac{\sum_{p=1}^{20} c}{n}$$

Although this does provide a measure of accuracy, the measurement may not be the best. If $c = 1$, it may indicate a high degree of accuracy, even if the driver never finished in their predicted position, meaning the model is inaccurate, even with a low c value. This problem will be addressed using the next method.

[4] Statistics How To, 2016. *Absolute Error & Mean Absolute Error (MAE).* [Online] Available at: https://www.statisticshowto.datasciencecentral.com/absolute-error/ [Accessed 15 November 2019].

Where n is the number of positions, in this case 20. Using the following formula, we obtain $\bar{c} = 3.17$ for $f(x)$ and $s(x)$, while $g(x)$ predictions have a value of $\bar{c} = 3.095$, making $g(x)$ predictions slightly more accurate than the other ones.

Interestingly, $\bar{c}$ is approximately 3 for all of the predictions, which is the range which the normal distribution curve covers for most of the positions. This means that the predictions should not be necessarily done for the highest probability, but for the values within the ± 3 range of the mean. The range is the representation of other external non-mathematical factors, such as driver skill, team car, and team strategy. The accuracy of the model including the range will be considered later.

Another way to measure accuracy would be to calculate the amount of times that the predictions are exactly right. Using this value, it is possible to calculate the probability that the prediction will be correct with the following formula:

$$P(C) = \frac{n(A)}{n(N)}$$

Where $n(A)$ is the number of times the prediction was correct, and $n(N)$ is the total number of races. $n(N)$ can take several values. It can equal the total number of races where a driver with a prediction finishes or the total number of races in general; both will be considered. The results for each position can be seen in **appendix V.**

To calculate the average probability, the following formula is used:

$$\overline{P(C)} = \frac{\sum_{i=1}^{20} P(C)_i}{20}$$

Where i is a position.

Using the formula, the $g(x)$ predictions are the most accurate again. The probability of the predictions being right overall is 0.147 assuming the driver finishes the race. For other predictions, the probability is 0.128, assuming the driver finishes. The same pattern applies without the assumption that the driver finishes. The $g(x)$

predictions have the probability of 0.129 of being right; the $f(x)$ and $s(x)$ predictions have the probability of 0.112. This makes the $g(x)$ predictions more accurate, since they are more likely to be right.

As mentioned above, the range ±3 includes outside factors. It is interesting to see how likely the predictions are to be right if the positions within the range are included using the above formula. This method will be applied to the $f(x)$ and $g(x)$ predictions. They will not be applied to the $s(x)$ predictions, since there is no range that is mathematically shown/supported.

Results for the individual positions are seen in **appendix VI**. When the formula is applied to $f(x)$, the $\overline{P(C)} = 0.672$, assuming the driver finishes; $\overline{P(C)} = 0.590$ without making the assumption that the driver finishes. For $g(x)$, $\overline{P(C)} = 0.697$ assuming the driver finishes; $\overline{P(C)} = 0.614$ without the assumption. This shows that if the predictions are made within the range, it is much more likely that they will be correct. It is important to include outside factors into the model, while staying within the range to maximize results.

Evaluation

Firstly, using the CLT is problematic when making predictions for the races. I used the 2019 races to determine the mean, meaning I could easily calculate the sample (n in the CLT equation). This is not possible to do with complete precision, since it is impossible to predict with certainty whether a driver will finish the race. It is possible to overcome this issue, though, by calculating the probability that a driver will finish the race. This probability can be multiplied by the total number of races in a particular season, which will give the best approximation of the value of n.

When the model was made, only the hybrid era was considered part of the population. This makes the model inaccurate for other eras. This is important especially today, since the FIA (the sports association regulating F1) is planning to introduce new regulations in 2021, which will drastically change the sport. This will make the predictions established in this exploration possibly inaccurate. It is possible to determine that by considering the current population as the sample mean and the whole F1 history as the population. Using the current population as a sample mean, it is possible to calculate the possible population mean using the confidence interval. To do so, we will use a 95% confidence interval, since a higher interval will give a wider range, although it is unlikely that the mean finishing position varied greatly throughout eras. We can determine the possible mean range for the whole history of F1 using the confidence interval formula, obtained in a book[5]:

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}} \times z_{\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} \times z_{\frac{\alpha}{2}}\right]$$

Using this formula on P5, the values will be $\bar{x} = 5.69$, $\sigma = 3.63$, and $n = 85$, since the values were obtained from 85 races. Once these values are substituted into the equation, the range is [4.92, 6.46]. This makes P5 a possible mean and the most probable position predicted, which supports the straightforward predictions. The range also shows that if the predictions made in this exploration are used in other eras, they may not be accurate.

Additionally, there are other assumptions that I made in the model that make it inaccurate. I have assumed that the driver will not crash and finish the race; this is not true in all cases, since the motorsport is highly unpredictable and accidents,

[5] Josip Harcet. Lorraine Heinrichs, P. M. S. M. T.-S., 2014. Confidence Intervals for the mean. In: *Mathematics Higher Level: Statistics and Probability.* Oxford: Oxford University Press, p. 86.

small or big, can happen. There are several ways to overcome this. Firstly, we can calculate the probability that a driver in a specific FP3 position finishes by taking the number of races finished and dividing it by the total number of races that happened in 2014-2018 seasons. The probability that a driver will finish should be calculated by the probability obtained in the CLT. Although this method is consistent with the model, since it only considers the position and not the driver, it is inaccurate. The FP3 position does not corelate with driver's reliability the same way that the FP3 position and finishing position do. Another method must be used.

Instead of relying solely on position, a way to incorporate driver skill into the model is by calculating their reliability, more specifically the probability that the driver will finish the race. This could be done by dividing the number of races the driver finished by the total number of races the driver entered. The probability should then be multiplied by the probability of the prediction to arrive at a final probability. This method, however, may be problematic for new drivers, since their reliability is unknown. Their reliability could be calculated based on their racing history in other championships such as Formula Two, but this also may prove ineffective, since the power of the cars in the championships differ greatly, making some rookie drivers unable to handle the car as well as in the previous championships.

Conclusion

In conclusion, the predictions based on probability, specifically the second method of predictions, were more correct than the straightforward predictions. The accuracy of the predictions can be further improved if we take into consideration the skill of the driver as well as the team itself. In case the driver is skilled, they should be predicted a higher position; the position should be within the aforementioned range, though, to maximize the probability of the prediction being right.

Bibliography

Formula One World Championship Limited, 2019. *Race Results.* [Online]

Available at: https://www.formula1.com/en/results.html

[Accessed 25 October 2019].

Josip Harcet. Lorraine Heinrichs, P. M. S. M. T.-S., 2014. Confidence Intervals for the mean. In: *Mathematics Higher Level: Statistics and Probability.* Oxford: Oxford University Press, p. 86.

Paul Fannon, V. K. B. W. S. W., 2012. Measures of Spread. In: *Mathematics Higher Level for the IB Diploma.* Cambridge: Cambridge University Press, p. 691.

Paul Fannon, V. K. B. W. S. W., 2012. The normal distribution. In: *Mathematics Higher Level for the IB Diploma.* Cambridge: Cambridge University Press, p. 777.

Stangroom, J., 2020. *Social Science Statistics.* [Online]

Available at: https://www.socscistatistics.com/tests/pearson/

[Accessed 28 October 2019].

Statistics How To, 2016. *Absolute Error & Mean Absolute Error (MAE).* [Online]

Available at: https://www.statisticshowto.datasciencecentral.com/absolute-error/

[Accessed 15 November 2019].

Appendix I:

Position	Mean
1	2.655172
2	3.51087
3	4.264368
4	4.421687
5	5.694118
6	6.585366
7	8.227848
8	8.108434
9	9.146341
10	9.391304
11	9.493671
12	10.41558
13	9.914286
14	11.27632
15	10.95385
16	11.89855
17	13.25974
18	13.69737
19	14.57971
20	14.25397

Appendix II:

Averages	
M_x	10.5
M_y	9.087428

x	y	$x-M_x$	$y-M_y$	$(x-M_x)(y-M_y)$	$(x-M_x)^2$	$(y-M_y)^2$
1	2.655172	-9.5	-6.43226	61.10642963	90.25	41.37391
2	3.51087	-8.5	-5.57656	47.40074088	72.25	31.098
3	4.264368	-7.5	-4.82306	36.17294813	56.25	23.26191
4	4.421687	-6.5	-4.66574	30.32731488	42.25	21.76914
5	5.694118	-5.5	-3.39331	18.66320363	30.25	11.51455
6	6.585366	-4.5	-2.50206	11.25927788	20.25	6.260313
7	8.227848	-3.5	-0.85958	3.008529125	12.25	0.738877
8	8.108434	-2.5	-0.97899	2.447484375	6.25	0.958429
9	9.146341	-1.5	0.058913	-0.088369875	2.25	0.003471
10	9.391304	-0.5	0.303876	-0.151938125	0.25	0.092341

11	9.493671	0.5	0.406243	0.203121625	0.25	0.165034
12	10.41558	1.5	1.328152	1.992228375	2.25	1.763988
13	9.914286	2.5	0.826858	2.067145625	6.25	0.683695
14	11.27632	3.5	2.188892	7.661122875	12.25	4.791249
15	10.95385	4.5	1.866422	8.398900125	20.25	3.483532
16	11.89855	5.5	2.811122	15.46117238	30.25	7.902408
17	13.25974	6.5	4.172312	27.12002963	42.25	17.40819
18	13.69737	7.5	4.609942	34.57456688	56.25	21.25157
19	14.57971	8.5	5.492282	46.68439913	72.25	30.16516
20	14.25397	9.5	5.166542	49.08215138	90.25	26.69316

Appendix III:

Position	Mean	STD	Variance	No. Races 2019	CTL STD
1	2.655172	2.537292	6.437851	17	0.615384
2	3.51087	3.010949	9.065815	17	0.730262
3	4.264368	3.485718	12.15023	18	0.821592
4	4.421687	3.052869	9.320012	17	0.74043
5	5.694118	3.628667	13.16723	16	0.907167
6	6.585366	4.157252	17.28275	17	1.008282
7	8.227848	3.832885	14.69101	19	0.879324
8	8.108434	3.625829	13.14664	18	0.854616
9	9.146341	3.213098	10.324	18	0.757334
10	9.391304	3.452269	11.91816	16	0.863067
11	9.493671	3.587276	12.86855	14	0.95874
12	10.41558	4.014286	16.11449	16	1.003571
13	9.914286	3.238101	10.4853	17	0.785355
14	11.27632	3.602217	12.97596	13	0.999075
15	10.95385	3.402346	11.57596	16	0.850587
16	11.89855	3.722527	13.8572	18	0.877408
17	13.25974	2.970923	8.826384	17	0.720555
18	13.69737	3.229116	10.42719	14	0.863018
19	14.57971	2.96254	8.776641	18	0.698277
20	14.25397	3.654287	13.35381	13	1.013517

Appendix IV:

Position	f(x)	g(x)	s(x)
1	2.11	2.00	2.11
2	2.32	2.74	2.32
3	2.00	2.74	2.00

4	2.74	2.74	2.74
5	2.44	3.22	2.44
6	3.22	2.52	3.22
7	3.05	3.05	2.52
8	2.52	3.05	3.05
9	3.50	3.50	3.50
10	4.00	3.50	2.94
11	3.65	3.50	3.19
12	3.25	2.94	3.50
13	2.94	2.94	2.95
14	5.16	3.19	3.40
15	3.19	3.19	3.82
16	3.50	3.50	3.65
17	2.95	2.95	5.16
18	3.40	3.40	3.25
19	3.82	3.82	3.65
20	3.65	3.40	4.00

Appendix V:

Position	f(x) Finish	f(x) No Finish	g(x) Finish	g(x) No Finish	s(x) Finish	s(x) No Finish
1	0.211	0.190	0.250	0.238	0.211	0.190
2	0.105	0.095	0.263	0.238	0.105	0.095
3	0.250	0.238	0.263	0.238	0.250	0.238
4	0.263	0.238	0.263	0.238	0.263	0.238
5	0.222	0.190	0.222	0.190	0.222	0.190
6	0.222	0.190	0.143	0.143	0.222	0.190
7	0.100	0.095	0.100	0.095	0.143	0.143
8	0.143	0.143	0.100	0.095	0.100	0.095
9	0.100	0.095	0.100	0.095	0.100	0.095
10	0.067	0.048	0.100	0.095	0.111	0.095
11	0.000	0.000	0.100	0.095	0.063	0.048
12	0.063	0.048	0.111	0.095	0.056	0.048
13	0.111	0.095	0.111	0.095	0.105	0.095
14	0.105	0.095	0.063	0.048	0.200	0.143
15	0.063	0.048	0.063	0.048	0.118	0.095
16	0.056	0.048	0.056	0.048	0.050	0.048
17	0.105	0.095	0.105	0.095	0.105	0.095
18	0.200	0.143	0.200	0.143	0.063	0.048
19	0.118	0.095	0.118	0.095	0.000	0.000
20	0.050	0.048	0.200	0.143	0.067	0.048

Appendix VI:

Position	f(x) Finish	f(x) No Finish	g(x) Finish	g(x) No Finish
1	0.842	0.762	0.900	0.857
2	0.895	0.810	0.842	0.762
3	0.900	0.857	0.842	0.762
4	0.842	0.762	0.842	0.762
5	0.778	0.667	0.722	0.619
6	0.722	0.619	0.714	0.714
7	0.750	0.714	0.750	0.714
8	0.714	0.714	0.750	0.714
9	0.650	0.619	0.650	0.619
10	0.533	0.381	0.650	0.619
11	0.600	0.571	0.650	0.619
12	0.625	0.476	0.722	0.619
13	0.722	0.619	0.722	0.619
14	0.316	0.286	0.563	0.429
15	0.563	0.429	0.563	0.429
16	0.500	0.429	0.500	0.429
17	0.632	0.571	0.632	0.571
18	0.667	0.476	0.667	0.476
19	0.588	0.476	0.588	0.476
20	0.600	0.571	0.667	0.476