The Evolution of Driving Accuracy and Driving Distance on the PGA Tour

Nonparametric Statistics Case Study

Ashley Kapoor, Brad Klassen, Jocelyn Smit, Mark Terpak

Department of Mathematics and Statistics

Faculty of Mathematics and Science Brock University

St. Catharines, Ontario

Dr. Mei-Ling Huang

March 25, 2019

Contents

1 Abstract **1**

2 Statement of a Problem **2**

3 Objective **2**

4 Data **2**

4.1 Summary Statistics 2

4.1.1 Driving Accuracy 3

4.1.2 Driving Distance 3

5 Testing Normality **3**

5.1 Driving Accuracy 4

5.2 Driving Distance 4

6 Correlation Between Variables **4**

7 Nonparametric Methods for Analysis **5**

7.1 Hypothesis Test 5

7.2 Multiple Comparisons 6

7.3 Confidence Intervals 6

8 Parametric Methods for Analysis **7**

8.1 Hypothesis Test 7

8.2 Confidence Interval 8

9 Comparison of Nonparametric and Parametric Methods **8**

9.1 Accuracy Confidence Intervals 9

9.2 Distance Confidence Intervals 10

10 Conclusion **10**

1 Abstract

This case study investigates the evolution of driving accuracy and driving distance over time by analyzing PGA Tour performance during the 2010 to 2018 seasons. The driving accuracy statistic follows a normal distribution, therefore parametric methods were used. Whereas, the driving distance statistic does not follow a normal distribution, so nonparametric methods were used. However, both parametric and nonparametric methods for each variable were conducted for comparison purposes. Although, parametric results for distance and nonparametric results for accuracy were not as reliable since normality conditions of these two variables violate the assumptions of their corresponding aforementioned methods. After comparing the mean values of accuracy and distance from all nine years, it was found that driving distance increased over time, whereas driving accuracy decreased over time. The negative correlation between these variables further shows that most players tend to improve on one variable at a time; the performance of the other is usually sacrificed.

2 Statement of a Problem

No matter the competition being considered, the main objective is to win. At the professional level for a wide variety of sports, there is an added pressure to win with many passionate fans following every game. Without statistics to analyze various factors involved with a given sport, predicting the outcome of sports competitions can become very difficult. After adding in the complexity of new equipment, technology, and increased competition, predicting upcoming performance has grown to be more challenging.

3 Objective

Golf is a sport that generates massive amounts of data with no shortage of opportunity for analysis. This case study will investigate the evolution of driving accuracy and driving distance on the PGA Tour from 2010 to 2018. Studying the trend of the statistics will give athletes a better idea of the areas of their game that are worth investing time to improve. These athletes can use their practice time more efficiently, as they can now focus on improving the factors which will have the largest impact in the future. This study also provides value to non-professional athletes who have an interest in the game of golf. Many people enjoy playing golf, both for leisure and for competitive purposes, and these players can use these findings to improve their skill set. Golf instructors and coaches can use these results to gain a better understanding of how to improve their clients' chance of winning.

4 Data

The data used in this study consists of driving accuracy and driving distance for 439 players over a nine-year period from 2010 to 2018. There are 1678 observations in the data set.

The driving accuracy variable is calculated as the percentage of tee-shots that successfully come to rest in the fairway. The driving distance variable is measured as the average number of yards per measured drive. The players' driving distance is measured to the point at which the ball comes to rest, regardless of whether or not it is in the fairway. Since wind direction could affect the accuracy of these measurements, this had to be taken into account when deciding which holes should be used to document the driving measurements. It was decided that drives would be measured on two holes per round. The two holes selected for measurement were facing in opposite directions of one another, in order to counteract the inaccuracies caused by the wind.

4.1 Summary Statistics

The following table gives the summary statistics for each variable used in the study. The mean accuracy was the highest in 2010, with a value of 63.4%. Whereas, the mean distance reached its

peak in 2018 at 296.70 yards. This suggests there has been an increase in driving distance and a decrease in driving accuracy throughout the years of the study.

4.1.1 Driving Accuracy

Year	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
2010	192	63.37	5.1108	50.15	76.08
2011	186	61.79	5.1294	46.99	75.65
2012	191	61.03	4.7395	47.27	73.00
2013	180	61.29	4.7456	45.58	71.81
2014	177	61.55	4.7976	50.00	75.49
2015	184	61.87	5.1287	50.29	76.88
2016	185	60.23	5.0207	43.02	73.36
2017	190	60.33	5.3067	45.37	72.73
2018	193	61.49	4.9062	47.15	75.19

4.1.2 Driving Distance

Year	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
2010	192	287.51	8.2165	266.40	315.50
2011	186	291.09	8.3430	269.80	318.40
2012	191	290.05	8.3831	268.90	315.50
2013	180	287.92	8.0111	270.50	306.30
2014	177	290.02	8.6961	270.30	314.30
2015	184	290.30	9.1473	270.00	317.70
2016	185	290.90	8.6502	269.70	314.50
2017	190	292.57	9.3664	270.10	316.70
2018	193	296.64	8.1449	278.90	319.70

5 Testing Normality

To assure the results given by nonparametric statistical methods are accurate and can be trusted, the data cannot be normally distributed. The Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests were performed to test the two hypotheses that driving accuracy and driving distance follow a normal distribution. The normality of the two statistics were each studied using the three aforementioned tests.

5.1 Driving Accuracy

The three tests which studied the normality of the driving accuracy data each gave p-values that were above the $\alpha = 0.05$ significance level, therefore we accept the null hypothesis that this data is normally distributed. The normal probability plot for driving accuracy shows a linear pattern, thus it agrees with the assumption of normality for this data.

Variable	Test	Statistic	P-Value
Accuracy	Kolmogornov-Smirnov	0.0128215	0.150
Accuracy	Cramer-von Mises	0.0355478	0.250
Accuracy	Anderson-Darling	0.3335838	0.250

5.2 Driving Distance

The p-values from the three tests for driving distance were all smaller than $\alpha = 0.05$, thus we reject the null hypothesis and conclude this data is not normally distributed. The normal probability plot for driving distance does not show a linear pattern, which further proves this data is not normal.

Variable	Test	Statistic	P-Value
Distance	Kolmogornov-Smirnov	0.0230942	0.029
Distance	Cramer-von Mises	0.1735624	0.012
Distance	Anderson-Darling	1.4128383	0.005

Since the driving accuracy data is normally distributed, and that for driving distance is not, the use of both parametric and nonparametric methods are required for this case study.

6 Correlation Between Variables

To test the correlation between the two variables we have used three methods, Pearson Correlation, Spearman's Rho, and Kendall's Tau. The Pearson and Spearman's Rho correlation coefficients were both approximately -0.53, showing there is a moderately negative correlation between distance and accuracy. When distance increases, accuracy decreases, and vice versa. The Kendall's Tau correlation coefficient of -0.37 showed a slightly weaker negative relationship between driving distance and driving accuracy.

Test	Coefficient	P-Value
Pearson Correlation Coefficient	-0.53383	< 0.0001
Spearman's Rho Correlation Coefficient	-0.52945	< 0.0001
Kendall's Tau-b Correlation Coefficient	-0.37138	< 0.0001

7 Nonparametric Methods for Analysis

To ensure the results from nonparametric methods are accurate and reliable, the data used in these tests must not be normally distributed. Testing for normality therefore becomes crucial before beginning such tests. Using computations completed in SAS it was proven that the driving distance data is not normally distributed, therefore nonparametric methods are required. Nonparametric methods will also be used to analyze driving accuracy for comparison purposes.

7.1 Hypothesis Test

The Kruskal-Wallis test is used to determine whether or not there is a difference in the nine mean values for driving distance for the 2010 to 2018 seasons.

1. H_0: The average driving distance for all golfers who competed during a given year was the same over all nine years.
   H_1: At least one year had a different average driving distance for all players who competed during a given year.

2. $\alpha = 0.05$, N = 1678

3. Test statistic (since there are tires), under H_0:

$$\mathrm{T} = \frac{1}{S^2}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4} \sim \chi^2_{(\mathrm{k}-1)}, \text{ approximately}$$

$$\text{where } S^2 = \frac{1}{N-1}\left(\sum_{all} R\left(X_{ij}\right)^2 - \frac{N(N+1)^2}{4}\right)$$

4. Critical Region: Reject the null hypothesis when $T > \chi^2_{(k-1)}$. Since there are ties, we use the Chi-Square table, where DF = k - 1 $\Rightarrow$ DF = 9 - 1 $\Rightarrow$ DF = 8 and the probability p = 0.950.

5. Calculate T_{obs}: From SAS, $T_{obs} = 134.7227$. Since $T_{obs} = 134.7227 > \chi^2_{(8)} = 15.51$, we reject H_0 at the $\alpha = 0.05$ significance level.

6. We can conclude that at least one year had a different average driving distance for all golfers who competed during that year, at the $\alpha = 0.05$ significance level.

7. Calculate the p-value: From SAS $\hat{\alpha} < 0.001$, which is less than $\alpha = 0.05$, therefore we reject H_0 at the $\alpha = 0.05$ significance level.

7.2 Multiple Comparisons

To study the difference in the mean driving distance for players who played in each individual year, the Mann-Whitney and Kruskal-Wallis tests were performed. The Mann-Whitney test was used to compare consecutive pairs of years, since k = 2 in these cases, and the Kruskal-Wallis test was used to compare the "overall" difference in means.

Assumptions:

1. N = $\sum_{i=1}^{k} n_i$, combine k samples to one sample, all samples are independent.
2. The scale is at least ordinal
3. $R(X_{ij})$: The rank assigned to observations X_{ij} in the combined sample;

$$R_i = \sum_{j=1}^{n_i} \mathrm{R}\left(X_{ij}\right) : \text{ the sum of ranks in sample i, i} = 1, 2, \ldots, \mathrm{k}$$

These tests concluded that the mean driving distance for players who played during 2010 was significantly different from the mean value from those who played during 2011. From 2011 to 2012, the test showed there was no difference between these two mean driving distances. In more recent years, the p-values became even smaller, showing an even more significant difference between means for two consecutive years. To confirm this significant difference, the Mann-Whitney test was conducted to compare means from the starting year, 2010, and the ending year, 2018. This p-value was less than 0.0001, which further suggests the average driving distance was significantly different over these nine years.

Years	P-Value (Distance)
Overall	< 0.0001
2010, 2011	< 0.0001
2011, 2012	0.2029
2012, 2013	0.0275
2013, 2014	0.0302
2014, 2015	0.8809
2015, 2016	0.5062
2016, 2017	0.0777
2017, 2018	< 0.0001
2010, 2018	< 0.0001

7.3 Confidence Intervals

Assumptions:

1. $X_1, X_2, \ldots X_n$ and $Y_1, Y_2, \ldots, Y_m$ are independent random samples from respective populations.

2. Populations X & Y have identical distributions except for a possible difference in location parameter d. Ex. d = E(X) - E(Y).

For each of the nine years for which the data is being studied, lower and upper bounds were calculated to predict with 95% certainty where the true mean driving distance fell during each year. The true mean distance for players who competed during 2010 was between 286.4 yards and 289.2 yards, whereas that for 2018 is between 294.7 yards and 298.1 yards. This shows that the true mean driving distance has increased over time. The 95% confidence intervals for the mean driving distance for each year are summarized in the following table.

Year	N	Lower CL for Mean	Upper CL for Mean
2010	192	286.4	289.2
2011	186	289.9	292.1
2012	191	288.7	290.8
2013	180	286.3	289.6
2014	177	288.1	291.7
2015	184	288.4	291.8
2016	185	289.0	292.0
2017	190	291.0	294.3
2018	193	294.7	298.1

8 Parametric Methods for Analysis

As the driving accuracy data was proven to follow a normal distribution, parametric methods for analysis were used. The driving distance statistic does not follow a normal distribution, therefore these parametric methods will not give trustworthy results, but for comparison purposes they were still performed.

8.1 Hypothesis Test

An Analysis of Variance (ANOVA) Completely Randomized F test was performed to test the difference in driving accuracy for each player.

1. H_0: The average driving distance for all golfers who competed during a given year was the same over all nine years.
   H_1: At least one year had a different average driving distance for all golfers who competed during a given year.

2. $\alpha = 0.05$, N = 1678

3. Test statistic (since there are tires), under H_0:

$$F = \frac{T/(k-1)}{(N-1-T)/(N-K)} = \frac{SS_{\text{treat}}/(k-1)}{SS_{\text{error}}/(N-k)}$$

4. Critical Region: Reject the null hypothesis when $F > F_{(1-\alpha)}(k-1, N-k)$ where k - 1 $\Rightarrow$ 9 - 1 $\Rightarrow$ 8 and N - k $\Rightarrow$ 1678 - 9 $\Rightarrow$ 1669. $F_{(8,\infty)} = 1.94$.

5. Find F_{obs}: From SAS, $F_{obs} = 6.59$. Since $F_{obs} = 6.59 > F_{(8,\infty)} = 1.94$, we reject H_0 at the $\alpha = 0.05$ significance level.

6. We can conclude that at least one year has a different driving accuracy, at the $\alpha = 0.05$ significance level.

7. Calculate the p-value: From SAS $\hat{\alpha} < 0.001$, which is less than $\alpha = 0.05$, therefore we reject H_0 at the $\alpha = 0.05$ significance level.

8.2 Confidence Interval

Nine 95% confidence intervals were calculated for the true mean driving accuracy for each of the nine years in which the data was studied. The true driving accuracy for the year 2010 was between 62.64% and 64.09%, and that for 2018 is between 60.79% and 62.19%, which shows a decrease in driving accuracy over time. The 95% confidence intervals for the true mean driving accuracy during each of the nine years are summarized in the following table.

Year	N	Lower CL for Mean	Upper CL for Mean
2010	192	62.64	64.09
2011	186	61.05	62.53
2012	191	60.35	61.71
2013	180	60.59	61.99
2014	177	60.84	62.26
2015	184	61.12	62.61
2016	185	59.50	60.95
2017	190	59.57	61.09
2018	193	60.79	62.19

9 Comparison of Nonparametric and Parametric Methods

For the majority of the tests used in this study, the nonparametric and parametric results provided very similar results. Comparing the nonparametric and parametric confidence intervals

for each of the nine years for both variables, the results appear to be similar. The parametric and nonparametric confidence intervals for a given year and variable deviate only slightly, showing both methods give somewhat adequate results, but the normality assumption is what differentiates the two intervals for a given year and variable. Only the parametric intervals for accuracy give reliable results, since the accuracy data follows a normal distribution and the distance data does not. By the same logic, the nonparametric intervals for distance are the only nonparametric intervals which can be trusted, since nonparametric methods require data that is not normal. The Kruskal-Wallis and Mann-Whitney nonparametric tests for multiple comparisons were used to test the difference in the average distance over nine years, since the distance data does not follow a normal distribution. Parametric methods to study the differences in mean driving accuracy are more reliable than these nonparametric methods, but despite this fact, the nonparametric procedures still provide seemingly logical results. These results are summarized in the following tables:

9.1 Accuracy Confidence Intervals

Year	N	Type	Lower CL for Mean	Upper CL for Mean
2010	192	Nonparametric	62.93	64.31
2010	192	Parametric	62.64	64.09
2011	186	Nonparametric	60.98	62.93
2011	186	Parametric	61.05	62.53
2012	191	Nonparametric	60.36	61.91
2012	191	Parametric	60.35	61.71
2013	180	Nonparametric	60.51	62.31
2013	180	Parametric	60.59	61.99
2014	177	Nonparametric	60.47	62.07
2014	177	Parametric	60.84	62.26
2015	184	Nonparametric	60.73	63.03
2015	184	Parametric	61.12	62.61
2016	185	Nonparametric	59.25	60.99
2016	185	Parametric	59.50	60.95
2017	190	Nonparametric	59.58	60.94
2017	190	Parametric	59.57	61.09
2018	193	Nonparametric	60.47	62.82
2018	193	Parametric	60.79	62.19

9.2 Distance Confidence Intervals

Year	N	Type	Lower CL for Mean	Upper CL for Mean
2010	192	Nonparametric	286.40	289.20
2010	192	Parametric	286.34	288.68
2011	186	Nonparametric	289.90	292.10
2011	186	Parametric	289.88	292.30
2012	191	Nonparametric	288.70	290.80
2012	191	Parametric	288.86	291.25
2013	180	Nonparametric	286.30	289.60
2013	180	Parametric	286.74	289.10
2014	177	Nonparametric	288.10	291.70
2014	177	Parametric	288.73	291.31
2015	184	Nonparametric	288.40	291.80
2015	184	Parametric	288.97	291.63
2016	185	Nonparametric	289.00	292.00
2016	185	Parametric	289.64	292.15
2017	190	Nonparametric	291.00	294.30
2017	190	Parametric	291.23	293.91
2018	193	Nonparametric	294.70	298.10
2018	193	Parametric	295.48	297.79

The reason that both the parametric and nonparametric tests give similar results is partially due to the fact that a large sample size was used in this study. As the sample size increases for data that does not follow a normal distribution, this data becomes approximately normal. There were 1678 observations used in this study, and thus this sample size is large enough to assume an approximately normal distribution, which is an assumption for parametric tests. Normality was tested for the driving distance data, and although it was shown the data was likely not normal, the normal probability plot showed this distribution is not extremely far off from a normal distribution.

10 Conclusion

This study was conducted to test the difference in means over nine years of important factors that impact overall scoring in the game of golf, driving accuracy and driving distance. Throughout the time frame studied, average distance increased, meanwhile average accuracy decreased. This negative correlation between the variables shows that a given player may struggle to increase their accuracy and distance at the same time; in order to improve in one area, another tends to be sacrificed. In recent years, there has been an improvement in the technology of golf clubs, meaning golfers can now attain larger distances more easily than they could have nine years ago. This has

changed the way golfers and their coaches view the game. After analyzing the trends, players now know to focus their attention on improving distance rather than accuracy. This study showed that from 2010 to 2018, average driving distance increased by 9 yards, while average accuracy only decreased by 2%. This is expected to be a better trade off with respect to overall score rather than what improving accuracy could provide. Other studies have shown that when comparing scoring between players off of the tee shot, the strokes gained off the tee category was 60-65% reliant on driving distance, but only 30-35% reliant on driving accuracy. This further suggests the results obtained in this study are accurate. Average driving distance has improved in more recent years, while driving accuracy has decreased.