BANA 7051-002

Statistical Methods

Final Project
on
Professional Golfers Association Data

Submitted by:

Rashmi Prathigadapa

UC ID: M12853369

Introduction

Two avid golfers were having a debate over whether scores were lower on Sundays. One of the golfers believed that courses were made easier on Sunday for viewers who wanted to see low scores. The other golfer countered that all the pressure on the golfers on Sunday would surely raise their scores and, in fact, the television coverage would make things worse.

McDougall and Higgins discussed how they would use the data if they were accessible. Questions they raised and discussed included the following:

- Are scores different from the first to the last day?
- Are scores different across the four rounds?
- Are younger people doing better than who are older?
- Do long hitters have lower scores?
- How important is driving accuracy in determining one’s score?
- Do people putt for ‘Dough’ and drive for ‘show’?

Using a dataset from the 2011 season, various questions about golfers and golf tournament can be addressed. The dataset is available from the case author. It contained over 1000 responses made over four generations.

This data was modified to show 270 unique golfers and a summary of what they accomplished in all tournaments where they played four rounds.

Information on the dataset

PlayerNumber	Unique number identifying the player
Age	Player age
FedExCupPoints	FedEx Cup Points
Money	Average money won per tournament
Round1Score	Average strokes in Round 1
Round2Score	Average strokes in Round 2
Round3Score	Average strokes in Round 3
Round4Score	Average strokes in Round 4
Total Strokes	Average strokes per tournament
Average Drive	Average driving distance
Drive_Rank	Driving Rank
Percent_Birdie_when_GIR	Percent birdies made when green is hit in regulation
Percent_Fairways	Percent of drives in fairway
Percent_GIR	Percent greens hit in regulation
Putt_Round	Average Putts per round
Percent_10foot	Percent of Putts inside 10' made
Percent_Outside10	Percent of Putts outside 10' made

1. Are scores different from the first to the last day?

In order to say if the scores are different from the first to last day;

Checking assumptions:

- The two groups of data are dependent.
- The differences between round-1 and round-4 follow normal distribution.

Since these assumptions are satisfied, we will have to perform a Paired T-test on the round 1 and round 4 scores for each player. So, let's assume,

Ho: Scores are same on the first day and last day.

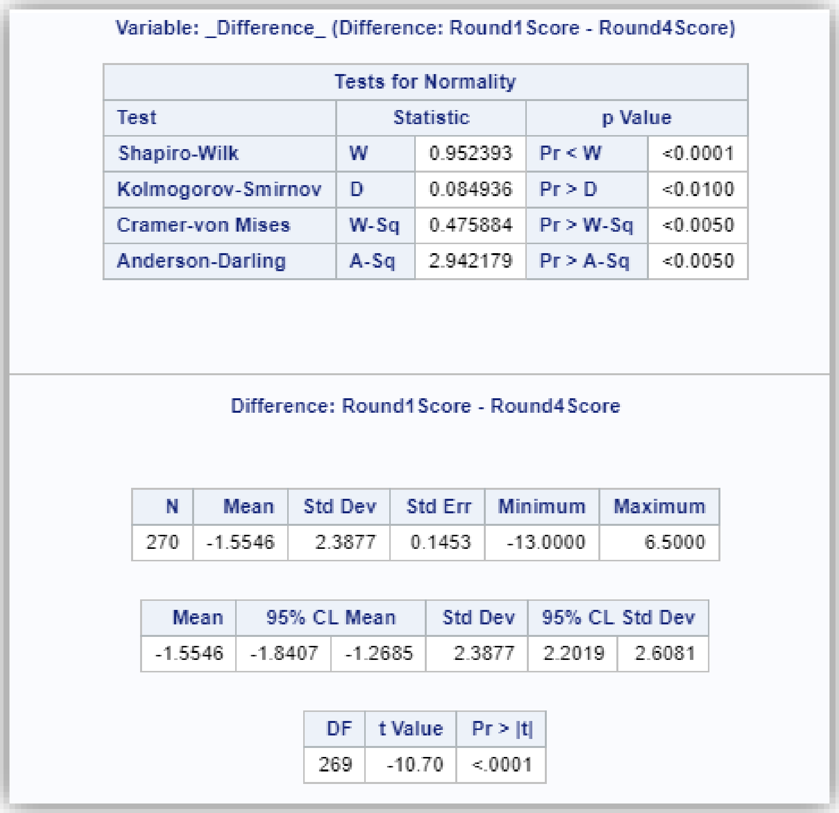
Ha: Scores are different on the first day and last day.

SAS Code:

```
data Work.Paired_diffs_;
      set WORK.PGA;
      _Difference_=Round1Score - Round4Score;
      label _Difference_="Difference: Round1Score - Round4Score";
run;

/* Test for normality */
proc univariate data=Work.Paired_diffs_ normal mu0=0;
      ods select TestsForNormality;
      var _Difference_;
run;

/* t test */
proc ttest data=WORK.PGA sides=2 h0=0 plots(showh0);
      paired Round1Score*Round4Score;
run;

/* Clean up */
proc delete data=work._paired_diffs_;
run;
```

Output:

From the output of the above code, we can say that the p value <0.0001, less than alpha value. Since the p value is significant, we fail to accept the null hypothesis. Therefore, there is difference in the scores of round-1 and round-4



Variable: _Difference_ (Difference: Round1Score - Round4Score)

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.952393	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.084936	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.475884	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.942179	Pr > A-Sq	<0.0050

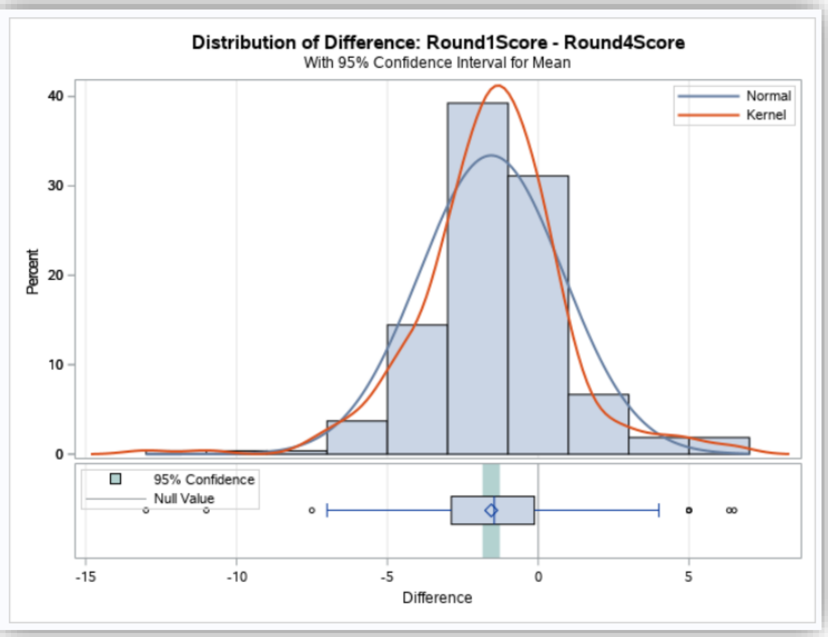
Difference: Round1Score - Round4Score

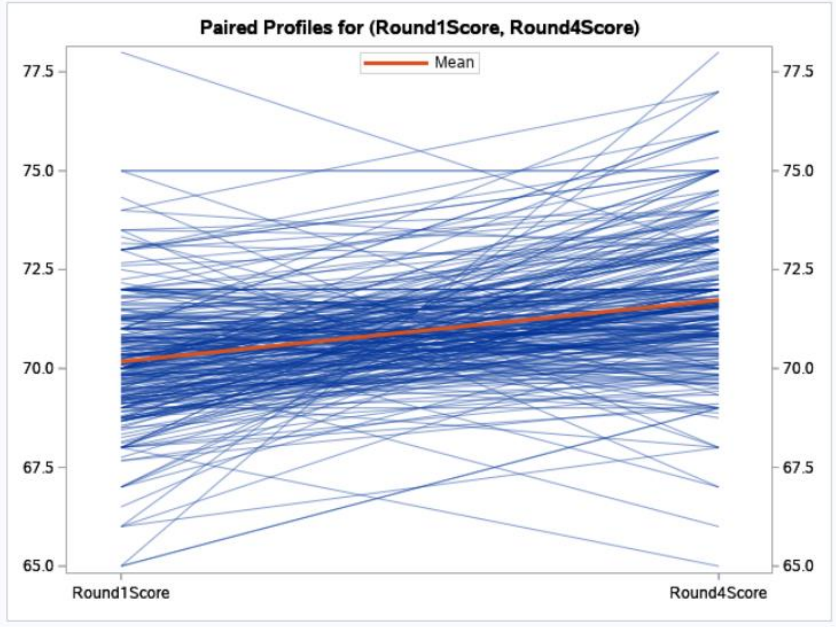
N	Mean	Std Dev	Std Err	Minimum	Maximum
270	-1.5546	2.3877	0.1453	-13.0000	6.5000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-1.5546	-1.8407	-1.2685	2.3877	2.2019	2.6081

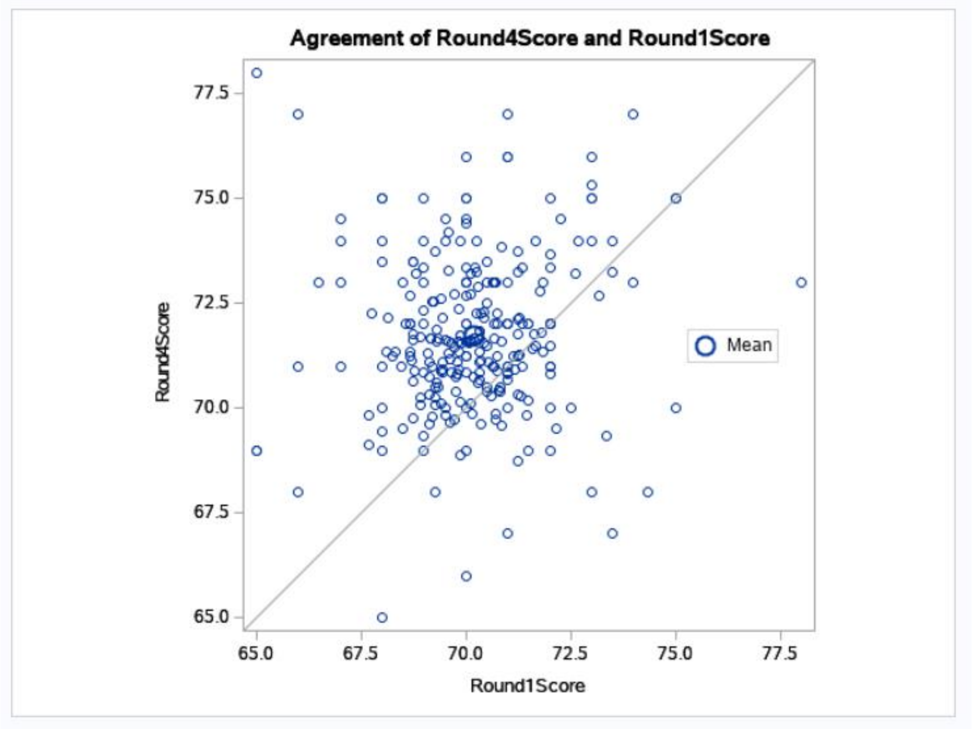
DF	t Value	Pr > |t|
269	-10.70	<.0001

From the below graph, we can say that the difference in the scores of first and last rounds follow normal distribution.

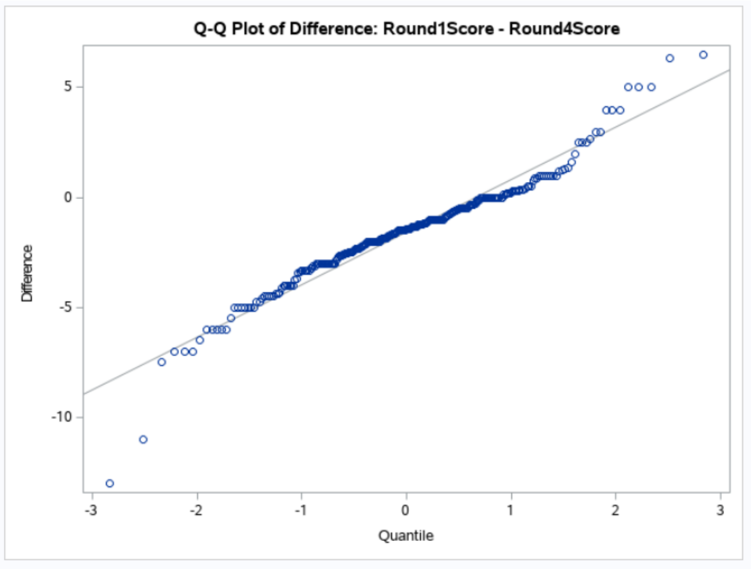




The agreement plot below reveals that only very few players have higher round1scores than round4 scores.



The below QQ plot assesses the normality assumption.



2. **Are young people doing better than those who are older?**

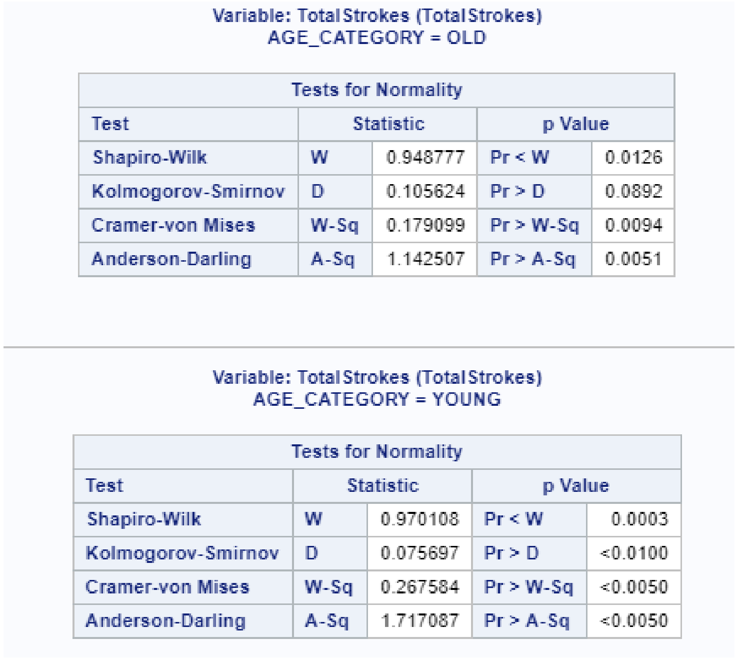
Ho: Young people and old people have same scores.

Ha: Young people and old people have different scores.

SAS Code:

```
/* Test for normality */
proc univariate data=WORK.PGA_2 normal mu0=0;
      ods select TestsForNormality;
      class AGE_CATEGORY;
      var FedExCupPoints;
run;

/* t test */
proc ttest data=WORK.PGA_2 sides=2 h0=0 plots(showh0);
      class AGE_CATEGORY;
      var FedExCupPoints;
run;
```

Output:



Variable: TotalStrokes (TotalStrokes)
AGE_CATEGORY = OLD

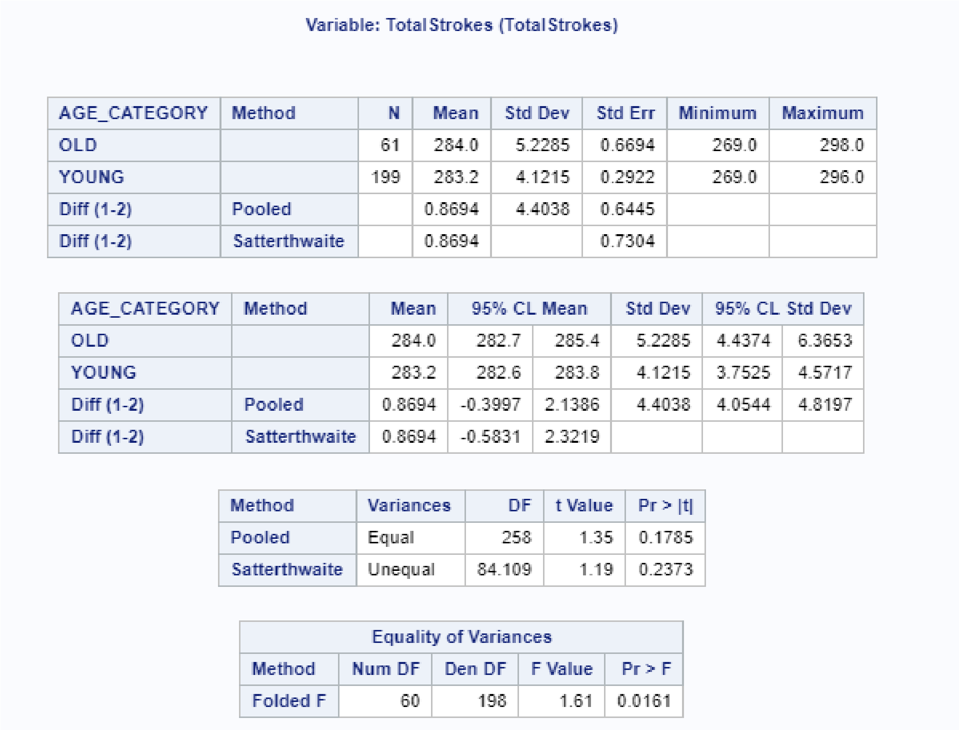
Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.948777	Pr < W	0.0126
Kolmogorov-Smirnov	D	0.105624	Pr > D	0.0892
Cramer-von Mises	W-Sq	0.179099	Pr > W-Sq	0.0094
Anderson-Darling	A-Sq	1.142507	Pr > A-Sq	0.0051

Variable: TotalStrokes (TotalStrokes)
AGE_CATEGORY = YOUNG

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.970108	Pr < W	0.0003
Kolmogorov-Smirnov	D	0.075697	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.267584	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.717087	Pr > A-Sq	<0.0050



Variable: Total Strokes (Total Strokes)

AGE_CATEGORY	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
OLD		61	284.0	5.2285	0.6694	269.0	298.0
YOUNG		199	283.2	4.1215	0.2922	269.0	296.0
Diff (1-2)	Pooled		0.8694	4.4038	0.6445		
Diff (1-2)	Satterthwaite		0.8694		0.7304		

AGE_CATEGORY	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
OLD		284.0	282.7	285.4	5.2285	4.4374	6.3653
YOUNG		283.2	282.6	283.8	4.1215	3.7525	4.5717
Diff (1-2)	Pooled	0.8694	-0.3997	2.1386	4.4038	4.0544	4.8197
Diff (1-2)	Satterthwaite	0.8694	-0.5831	2.3219			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	258	1.35	0.1785
Satterthwaite	Unequal	84.109	1.19	0.2373

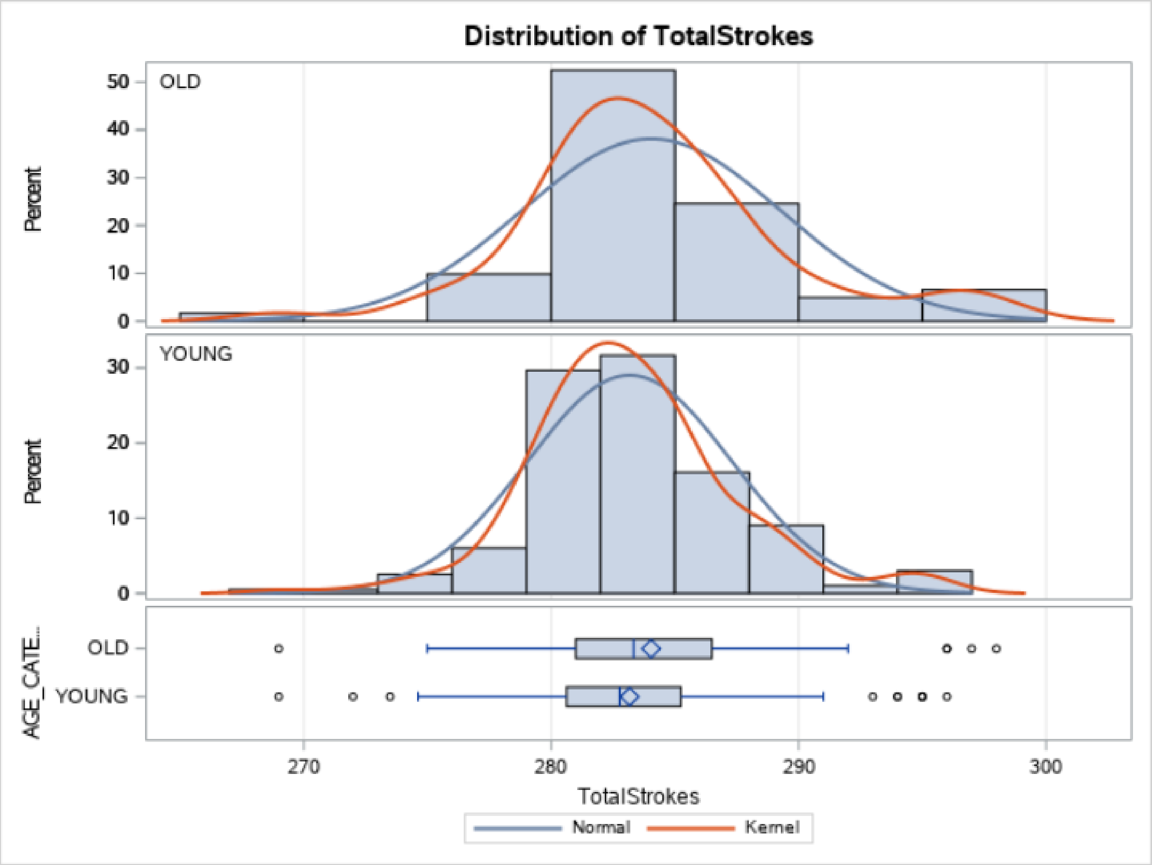
Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	60	198	1.61	0.0161

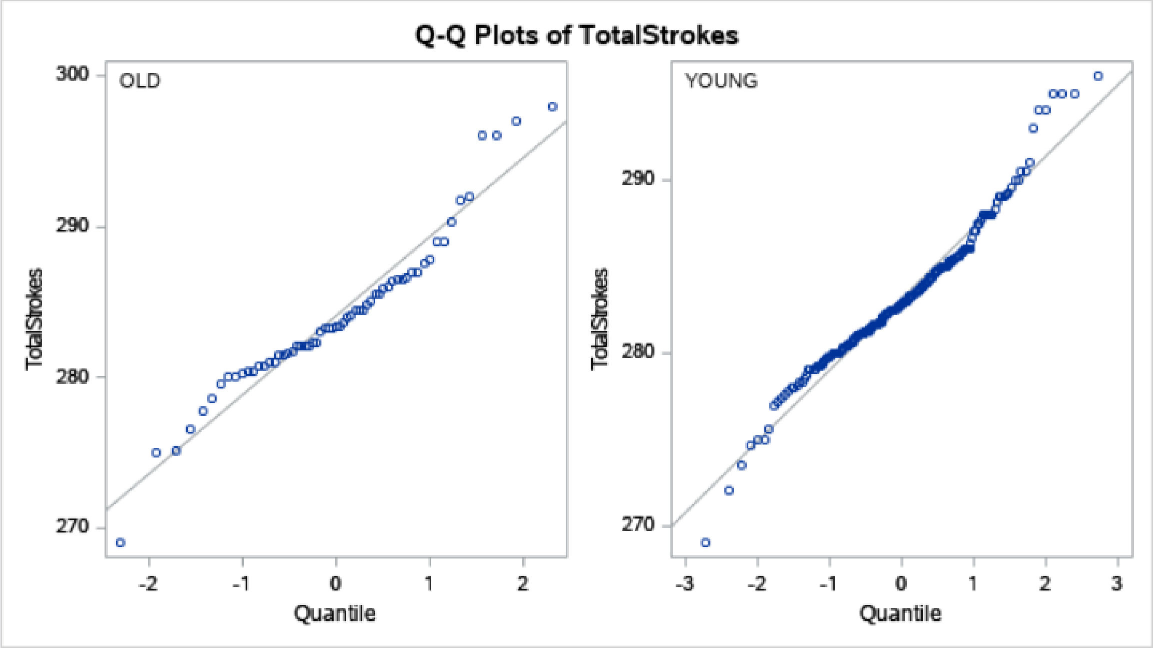
Conclusion:

From the output of the above code, we can say that the p value 0.01 <0.05, less than alpha value so we refer to the unequal variances table. The probability of the unequal variances is 0.23 i.e., > 0.05. Therefore, we fail to reject the null hypothesis.

Therefore, the young people and old people have same scores.



Distribution of TotalStrokes



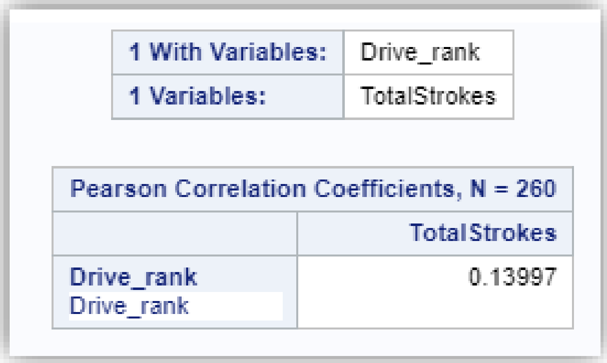
Q-Q Plots of TotalStrokes

3. Do Long hitters have low scores?

SAS Code:

```
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.PGA_2;
	scatter x=TotalStrokes y=Drive_rank /;
	xaxis grid;
	yaxis grid;
run;

ods graphics / reset;
```

Output:

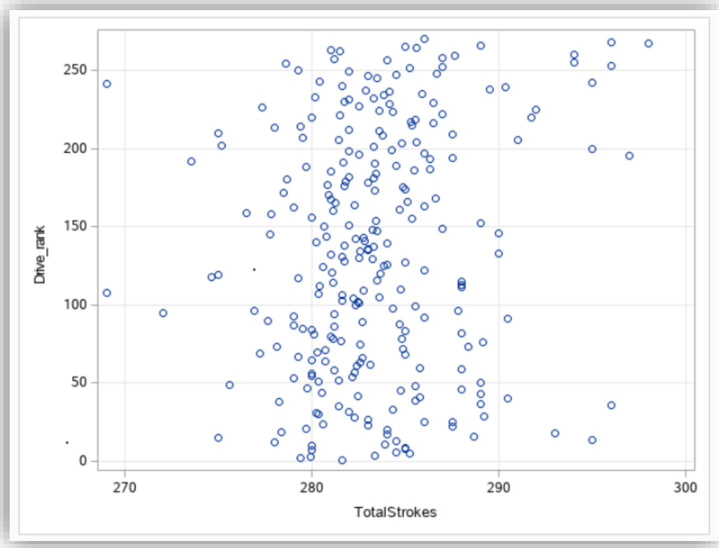


1 With Variables:	Drive_rank
1 Variables:	TotalStrokes

Pearson Correlation Coefficients, N = 260

	TotalStrokes
Drive_rank Drive_rank	0.13997

There is only 13% correlation between Total scores and Drive rank. Therefore, we cannot say that long hitters have low score.



Let us also perform a t-test to validate the above results.

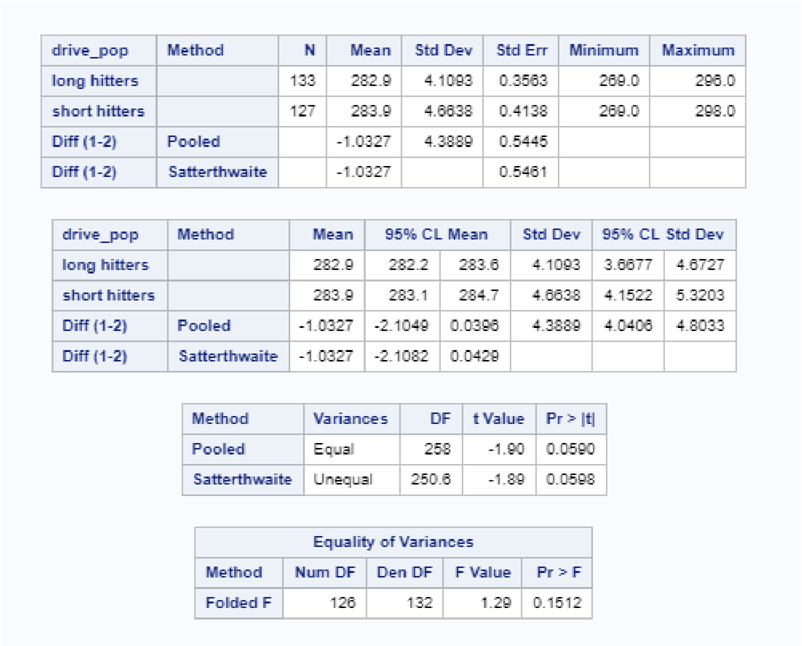
Ho: Long hitters have low scores.

Ha: Long hitters do not have low scores.

SAS Code:

```
/* Test for normality */
proc univariate data=WORK.PGA_1 normal mu0=0;
      ods select TestsForNormality;
      class drive_pop;
      var TotalStrokes;
run;

/* t test */
proc ttest data=WORK.PGA_1 sides=2 h0=0 plots(showh0);
      class drive_pop;
      var TotalStrokes;
run;
```

Output:



drive_pop	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
long hitters		133	282.9	4.1093	0.3563	269.0	296.0
short hitters		127	283.9	4.6638	0.4138	269.0	298.0
Diff (1-2)	Pooled		-1.0327	4.3889	0.5445		
Diff (1-2)	Satterthwaite		-1.0327		0.5461		

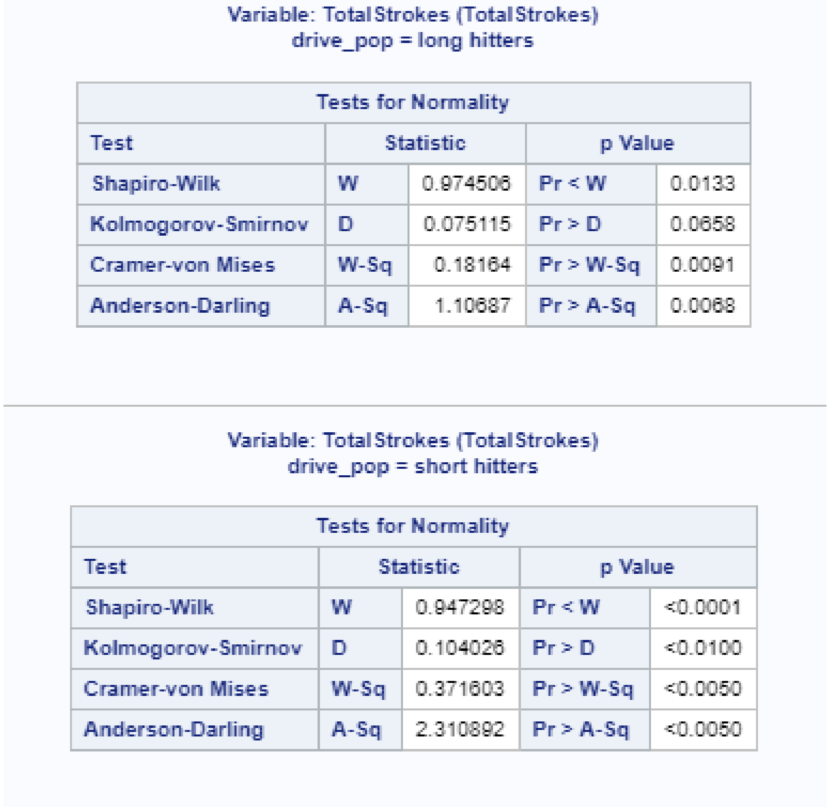
drive_pop	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
long hitters		282.9	282.2	283.6	4.1093	3.6677	4.6727
short hitters		283.9	283.1	284.7	4.6638	4.1522	5.3203
Diff (1-2)	Pooled	-1.0327	-2.1049	0.0396	4.3889	4.0406	4.8033
Diff (1-2)	Satterthwaite	-1.0327	-2.1082	0.0429			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	258	-1.90	0.0590
Satterthwaite	Unequal	250.6	-1.89	0.0598

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	126	132	1.29	0.1512

Conclusion:

The probability is 0.15 > 0.05 therefore, we refer to the equal variances. The equal variances value is 0.059>0.05. Therefore we can say that we accept null hypothesis and long hitters have low scores.

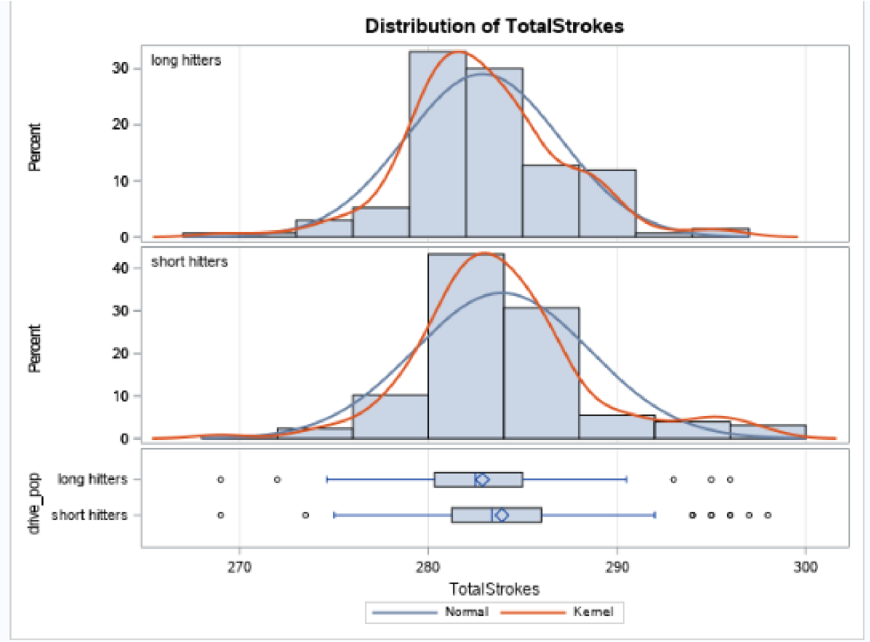


Variable: Total Strokes (Total Strokes)
drive_pop = long hitters

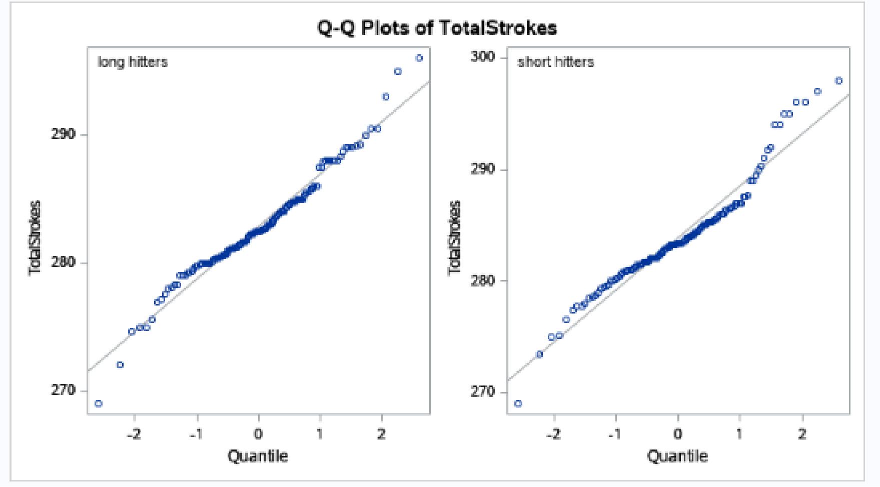
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.974506	Pr < W	0.0133
Kolmogorov-Smirnov	D	0.075115	Pr > D	0.0658
Cramer-von Mises	W-Sq	0.18164	Pr > W-Sq	0.0091
Anderson-Darling	A-Sq	1.10687	Pr > A-Sq	0.0068

Variable: Total Strokes (Total Strokes)
drive_pop = short hitters

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.947298	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.104026	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.371603	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.310892	Pr > A-Sq	<0.0050

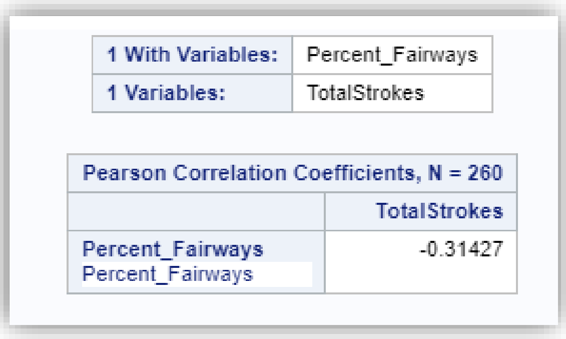


Distribution of TotalStrokes



Q-Q Plots of TotalStrokes

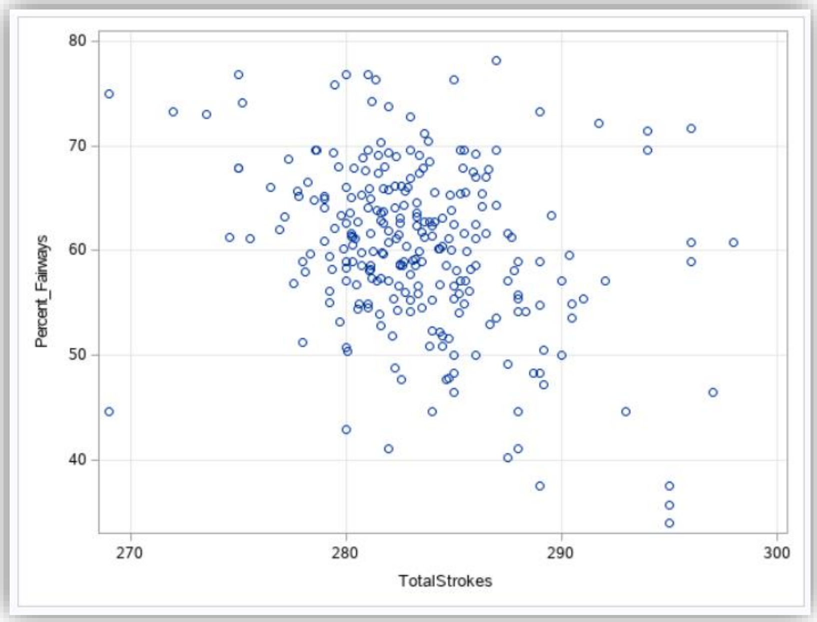
4. How important is driving accuracy in determining one's score?



1 With Variables:	Percent_Fairways
1 Variables:	TotalStrokes

Pearson Correlation Coefficients, N = 260	
	TotalStrokes
Percent_Fairways Percent_Fairways	-0.31427

There is 30% correlation between the variables Percent_Fairways and TotalStrokes.



Conclusion:

From the graph we can say that the population is concentrated at higher percentages of percent_fairways and at scores greater than 280. Therefore, higher driving accuracy leads to scores between 280-290 i.e., lower scores.

Appendix:

1. **Are scores different from the first to the last day?**

Code:

```
data Work.Paired_diffs_;
      set WORK.PGA;
      _Difference_=Round1Score - Round4Score;
      label _Difference_="Difference: Round1Score - Round4Score";
run;

/* Test for normality */
proc univariate data=Work.Paired_diffs_ normal mu0=0;
      ods select TestsForNormality;
      var _Difference_;
run;

/* t test */
proc ttest data=WORK.PGA sides=2 h0=0 plots(showh0);
      paired Round1Score*Round4Score;
run;

/* Clean up */
proc delete data=work._paired_diffs_;
run;
```

2. **Are young people doing better than old?**

Sas code:

```
proc univariate data=WORK.PGA_2 normal mu0=0;
      ods select TestsForNormality;
      class AGE_CATEGORY;
      var FedExCupPoints;
run;

/* t test */
proc ttest data=WORK.PGA_2 sides=2 h0=0 plots(showh0);
      class AGE_CATEGORY;
      var FedExCupPoints;
run;
```

3. **Do Long hitters have low scores?**

Sas code:

```
/* Test for normality */
proc univariate data=WORK.PGA_1 normal mu0=0;
      ods select TestsForNormality;
      class drive_pop;
      var TotalStrokes;
run;

/* t test */
proc ttest data=WORK.PGA_1 sides=2 h0=0 plots(showh0);
      class drive_pop;
      var TotalStrokes;
run;
```

Coclusion:

From our analysis on various questions, we have come to multiple conclusions and they are mentioned at the end of the each question.