Mini Project - Par Inc

Table of Contents

1 Project Objective

The objective of the report is to explore the Golf data set ("Golf") in R and generate insights about the data set. This exploration report will consist of the following:

- Importing the dataset in R
- Understanding the structure of dataset
- Formulation of Hypothesis Test
- Calculation of P-Value & Recommendation
- Descriptive statistics of each model with Graphical representation
- 95% of Confidence interval for the Population Mean of each model
- 95% of Confidence interval for the difference between two Population Mean
- Insights from the dataset

2 Assumptions

- H0: $\mu 1 = \mu 2$
- Hα: $\mu 1$ is not equal to $\mu 2$

3 Exploratory Data Analysis – Step by step approach

A Typical Data exploration activity consists of the following steps:

1. Environment Set up and Data Import
2. Calculating the P Value
3. Reject/ Fail to reject – Null Hypothesis
4. Descriptive Statistics of each model
5. Graphical Visualization of each models

We shall follow these steps in exploring the provided dataset. A brief about these steps is given.

3.1 Environment Set up and Data Import

3.1.1 Install necessary Packages and Invoke Libraries

"RColorBrewer" Package has been used in the code to make the visuals appealing with colors.

```
install.packages("RColorBrewer")
library("RColorBrewer")
```

3.1.2 Set up working Directory

Working Directory had been set in the following path: "`D:/Par Inc Mini Project`"

```
setwd("D:/Par Inc Mini Project")
```

And to check, whether the directory had set correctly, used the getwd()

Please refer Appendix A for Source Code.

3.1.3 Import and Read the Dataset

The given dataset is in Golf.csv format. Hence, the command 'read.csv' is used for importing the file.

Please refer Appendix A for Source Code.

To use the available data in the Golf.csv, by calling the header name, used 'attach()' command in the code.

3.2 Variable Identification

3.2.1 Variable Identification – Inferences

Question 1 - Formulate and present the rationale for a hypothesis test that par could use to compare the driving distances of the current and new golf balls?

#Answer:

From the given information on the above question, had framed the Hypothesis Testing.
μ1 is the population mean driving distance for the current golf ball.
μ2 is the population means driving distance for the new golf ball.
Hypothesis Testing:
H0: μ1 = μ2
Hα: μ1 is not equal to μ2
Test statistics: Two sample t-test

Question 2 - Analyze the data to provide the hypothesis testing conclusion. What is the p-value for your test? What is your recommendation for Par Inc.?

#Answer:

I have used t.test() command, to analyze the data and to conclude the hypothesis testing between the current and new.

Code:

```
t.test(ï..Current,New)
```

The calculated p-value = 0.188. Since it is a two-tail test, the calculated p-value was divided by 2 and the final p_value is 0.094. Since, the p_value>0.05, Null Hypothesis has been accepted.
So, I conclude that population means driving distances of the current and new golf balls are equal.

Question 3 - Provide descriptive statistical summaries of the data for each model?

#Answer:

Following codes has been executed for the descriptive statistics & Graphical Visualization of each model.

For Current Ball Model:

```
summary(ï..Current)
sd(ï..Current)
var(ï..Current)
boxplot(ï..Current, horizontal = TRUE, col= "green", main="Current Golf Ball - Data Sets")
hist(ï..Current, col= brewer.pal(9,"Set3"), main = "Current Golf Ball - Histogram Pattern")
```

For New Ball Model:

```
summary(New)
sd(New)
var(New)
boxplot(New, horizontal = TRUE, col= "Red", main="New Golf Ball- Data Sets")
hist(New, col= brewer.pal(9,"Spectral"), main = "New Golf Ball - Histogram Pattern")
```

 # Question 4 - What is the 95% confidence interval for the population mean of each model? What is the 95% confidence interval for the difference between the means of the two population?

#Answer for 95% confidence interval for the population mean of each model:

I have executed the t.test() command to find out the confidence interval of the population mean for each model.

Current Model:

```
t.test(ï..Current)
```

The driving distance of population mean - "Current" Model will fall in between 267.4757 to 273.0743 with 95% Confidence Level.
And the mean of Current model is 270.275

New Model:

```
t.test(New)
```

The driving distance of population mean - "New" Model will fall in between 264.3348 to 270.6652 with 95% Confidence Level.
And the mean of Current model is 267.5

#Answer 95% confidence interval for the difference between the means of the two population:

```
t.test(ï..Current,New)
```

```
Welch Two Sample t-test

data:  ï..Current and New
t = 1.3284, df = 76.852, p-value = 0.188
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.384937  6.934937
sample estimates:
mean of x mean of y
270.275   267.500
```

95% confidence interval for the difference between the means of the two population of "Current" and "New" is falls in the between the range of -1.384937 and 6.934937.
-1.384937 on the lower end and 6.934937 on the upper end, this shows that with 95% of confidence the difference between both the models falls in this range.
The difference in the mean driving distance is 2.775 (270.275 - 267.500).

 # Question 5 - Do you see a need for larger sample sizes and more testing with the golf balls?

Answer: To check, whether we require a large sample size, required to perform a Power of Test.

```
Diff=ï..Current-New
Diff
MeanDiff=mean(Diff)
SDDiff=sd(Diff)
MeanDiff
SDDiff
cohen.d=MeanDiff/SDDiff
cohen.d
powertest=power.t.test(n=40,d=cohen.d,sig.level= 0.05,power=NULL,type = "two.sample",alternative = "two.sided")
powertest
```

The result of the power of test – Type 2 error is 86%, which is very high considering the current sample size. Since the type 2 error is higher, to find the right sample size that will decrease the type 2 error to lesser than or equal to 10%, required to assess what should be the required sample size. So, keeping the Power of test at 90%, again run the code to assess the required sample size.

```
powertest1=power.t.test(n=NULL,d=cohen.d,sig.level=0.05,power=0.90,type= "two.sample",alternative = "two.sided")
powertest1
```

This resulted, that we require minimum of 516.4577 as a sample. So, I would recommend the organization to provide more sample size to infer accurately.

3.3 Univariate Analysis

To visualize and see the data of each models, I have made the graphic visuals by using boxplot() & hist() command for each models.

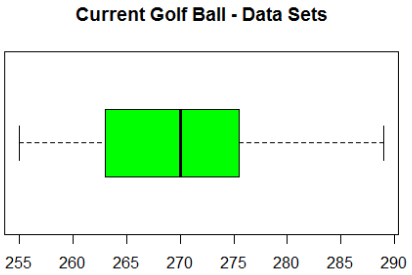
Descriptive Summary:

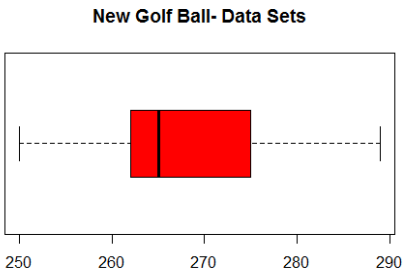
Summary									
Model	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Count**	**Sd**	**Variance**
Current Golf Ball	255.00	263.00	270.00	270.30	275.20	289.00	40.00	8.75	76.61
New Golf Ball	250.00	262.00	265.00	267.50	274.50	289.00	40.00	9.90	97.95

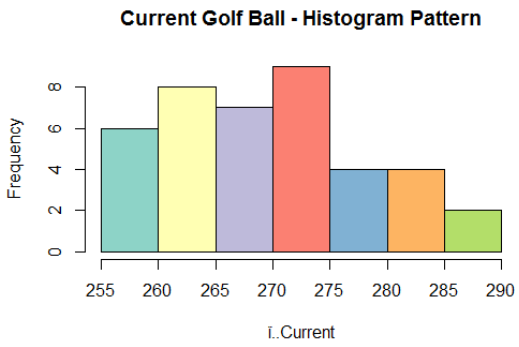
Minimum Value: Minimum driving distance Value is higher in the Current Golf Ball than compared to New Golf Ball.
1st Qu. Value: Driving Distance Value is slightly higher than by 1m in Current Golf Ball, compared to New Golf Ball.
Median Value: Minimum driving distance of Median Value is higher in the Current Golf Ball than compared to New Golf Ball.
Mean Value: Current Golf Ball, Mean Value of driving distance is higher compared to New Golf Ball.
3rd Qu. Value: Driving Distance Value is slightly higher than by 0.7m in Current Golf Ball, compared to New Golf Ball
Max Value: Maximum driving distance of both Current & New ball are same at 289 m.
SD Value: Current Ball, SD Value is lesser than New Golf Ball.

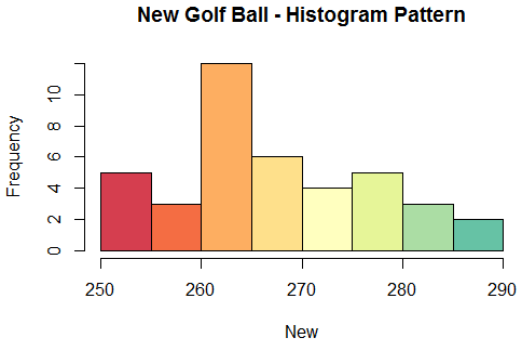
Variance Value: Variance Driving Distance Value is higher in New Ball compared to Current Golf Ball.

Visuals of Current & New Ball Data Sets:









4 Conclusion

In this Par Inc Golf Case Study, I have analyzed the given data sets of Current & New Ball by performing a T Test and calculated p-value. The pvalue of the test is 0.188. Since it is a two-tail test, the calculated p-value was divided by 2 and the final p_value is 0.094.
At 95% of confidence level, the calculate pvalue is >0.05, Null Hypothesis has been accepted.

So, I conclude that population means driving distances of the current and new golf balls are equal.

Also, the power of test – Type 2 error is at 86% with the sample size provided, which is higher. Hence, we need for larger sample size of 516 for the 90% Confidence Level.

5 Appendix A – Source Code

Here is the Source code of the Par Inc Golf Case study:

```
# Exploratory Data Analysis – Par Inc, Golf
# Environment Set up and Data Import
# Setup Working Directory

setwd("D:/Par Inc Mini Project")

getwd()


# Read Input File
Golf=read.csv("Golf.csv", header = TRUE)
Golf
attach(Golf)

# Question 1 - Formulate and present the rationale for a hypothesis test
that par could use to compare the driving distances of the current and
new golf balls.
#Answer: Hypothesis Testing: Ho: Mu1 = Mu2 & H1: Mu1 is not equal to Mu2 & it is Two Sample T
Test.

# Question 2 - Analyze the data to provide the hypothesis testing
conclusion.
What is the p-value for your test? What is your recommendation for Par
Inc.?

# Code:
t.test(ï..Current,New)

# Answer:
Welch Two Sample t-test
data:  ï..Current and New
t = 1.3284, df = 76.852, p-value = 0.188
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.384937  6.934937
sample estimates:
mean of x mean of y
270.275   267.500

# Since it is a two tail test, the actual PValue has been divided by 2.

Pvalue = 0.188/2
# Answer:
[1] 0.094

# As Pvalue>0.05 so we accept null hypothesis. So, we conclude that population means driving
distances of the current and new golf balls are equal.
```

#Question 3. Provide descriptive statistical summaries of the data for each model.

#Code: I wanted to use colors on the visuals, hence use this code.

```
install.packages("RColorBrewer")
library("RColorBrewer")
```

 # Code: "Current" Model – Summary Statistics
summary(ï..Current)
 # Answer:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
255.0	263.0	270.0	270.3	275.2	289.0

 # Code: Standard Deviation
sd(ï..Current)
 # Answer:
[1] 8.752985

 # Code: Variance
var(ï..Current)
 # Answer:
[1] 76.61474

 # Code: Boxplot & Histogram
boxplot(ï..Current, horizontal = TRUE, col= "green", main="Current Golf Ball - Data Sets")
hist(ï..Current, col= brewer.pal(9,"Set3"), main = "Current Golf Ball - Histogram Pattern")

……
……..

 # Code: "New" Model – Summary Statistics
summary(New)
 # Answer:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
250.0	262.0	265.0	267.5	274.5	289.0

 # Code: Standard Deviation
sd(New)
 # Answer:
[1] 9.896904

 # Code: Variance
var(New)
 # Answer:
[1] 97.94872

 # Code: Boxplot & Histogram
boxplot(New, horizontal = TRUE, col= "Red", main="New Golf Ball- Data Sets")
hist(New, col= brewer.pal(9,"Spectral"), main = "New Golf Ball - Histogram Pattern")

 # Question 4. What is the 95% confidence interval for the population mean of each model?

 # Code: "Current" Model
t.test(ï..Current)
 # Answer:
One Sample t-test
data: ï..Current
t = 195.29, df = 39, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
267.4757 273.0743
sample estimates:
mean of x
270.275

 # Code: "New" Model
t.test(New)
 # Answer:
One Sample t-test
data: New
t = 170.94, df = 39, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
264.3348 270.6652
sample estimates:
mean of x
267.5

 # Question 4.1 - what is the 95% confidence interval for the difference between the means of the two population?

 # Code: 95% Confidence Interval Difference between two population mean.
t.test(ï..Current,New)

 # Answer:
Welch Two Sample t-test
data: ï..Current and New
t = 1.3284, df = 76.852, p-value = 0.188
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.384937 6.934937
sample estimates:
mean of x mean of y
270.275 267.500

```
# Question 5 - Do you see a need for larger sample sizes and more testing
with the golf balls?

# Code:
Diff=ï..Current-New
Diff
# Answer:
 [1] -13  -8   4   6  -4  32  -4 -23 -27   6 -11  -2  22  -8   0   2   5   9 -12   4  -2  28  25
[24]   0  -2  -6 -11  -9  10  13  -2  24  20   3   5   4  18  19  13 -17
MeanDiff=mean(Diff)
MeanDiff
# Answer:
2.775
SDDiff=sd(Diff)
SDDiff
# Answer:
13.74397
cohen.d=MeanDiff/SDDiff
cohen.d
# Answer:
0.2019067
powertest=power.t.test(n=40,d=cohen.d,sig.level=0.05,power=NULL,type=
"two.sample",alternative = "two.sided")
powertest
# Answer:
Two-sample t test power calculation
          n = 40
      delta = 0.2019067
         sd = 1
  sig.level = 0.05
      power = 0.14274
alternative = two.sided
NOTE: n is number in *each* group

powertest1=power.t.test(n=NULL,d=cohen.d,sig.level=0.05,power=0.90,type=
"two.sample",alternative = "two.sided")
powertest1
# Answer:
Two-sample t test power calculation
          n = 516.4577
      delta = 0.2019067
         sd = 1
  sig.level = 0.05
      power = 0.9
alternative = two.sided
NOTE: n is number in *each* group
```