Mini Project – Par Inc. New Golf Ball

Model Report

GREAT LAKES
INSTITUTE OF MANAGEMENT

Table of Contents

1 Project Objective

Par Inc., is a major manufacturer of golf equipment. Management believes that Par's market share could be increased with the introduction of a cut-resistant, longer-lasting golf ball. Therefore, the research group at Par has been investigating a new golf ball coating designed to resist cuts and provide a more durable ball. The tests with the coating have been promising. One of the researchers voiced concern about the effect of the new coating on driving distances. Par would like the new cut-resistant ball to offer driving distances comparable to those of the current-model golf ball. To compare the driving distances for the two balls, 40 balls of both the new and current models were subjected to distance tests. The testing was performed with a mechanical hitting machine so that any difference between the mean distances for the two models could be attributed to a difference in the design.

The results of the tests, with distances measured to the nearest yard, are contained in the data set "Golf". **Prepare a Managerial Report**

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
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.?
3. Provide descriptive statistical summaries of the data for each model
4. What is the 95% confidence interval for the population mean of each model, and what is the 95% confidence interval for the difference between the means of the two population?
5. Do you see a need for larger sample sizes and more testing with the golf balls? Discuss.

The Golf Dataset:

Sr. No.	Current	New
1	264	277
2	261	269
3	267	263
4	272	266
5	258	262
6	283	251
7	258	262
8	266	289
9	259	286
10	270	264
11	263	274
12	264	266
13	284	262
14	263	271

Sr. No.	Current	New
15	260	260
16	283	281
17	255	250
18	272	263
19	266	278
20	268	264
21	270	272
22	287	259
23	289	264
24	280	280
25	272	274
26	275	281
27	265	276
28	260	269

Sr. No.	Current	New
29	278	268
30	275	262
31	281	283
32	274	250
33	273	253
34	263	260
35	275	270
36	267	263
37	279	261
38	274	255
39	276	263
40	262	279

2 Assumptions

The sample size of the data set is 40 (> 30) from each model of golf ball. Central Limit Theorem states that irrespective of the shape of the original population, the sampling distribution of the mean will approach a normal distribution as the size of the sample increases and becomes large (>30). We also assume that the sample estimate will be reflective of the reality.

3 Step by step approach

We shall follow step by step approach to arrive to the final conclusion a follows:

1. Exploratory Data Analysis
2. Descriptive Statistics
3. Data Visualization
4. Hypothesis formation
5. Selection of appropriate Hypothesis Testing method
6. 95% Confidence Intervals
7. Need of Larger Sample Size
8. Final Conclusion and Recommendation.

3.1 Exploratory Data Analysis

Please refer Appendix Section AAA for related code.

3.1.1 Number of Rows and Columns:

- The number of rows in the dataset is 40
- The number of columns (Features) in the dataset is 2

3.1.2 Features and their Types:

- Both the features, i.e. Current and New are continuous variables.

Feature Code	Type	Continuous/ Categorical
Current	Integer	Continuous
New	Integer	Continuous

3.1.3 Check for Missing Values:

- The data was checked for Missing Values using R function colsums(is.na), and found **no missing values**.

3.1.4 Dataset Summary:

- ***Current:*** The minimum value is 255 yards and the maximum value is 289 yards. The average value is 270.3 yards
- ***New:*** The minimum value is 250 yards and the maximum value is 289 yards. The average value is 267.5 yards

Feature Code	Minimum Value	Maximum Value	Average Value
Current	255 yards	289 yards	270.3 yards
New	250 yards	289 yards	267.5 yards

3.2 Descriptive Statistics

Please refer Appendix Section AAA for related code.

In this step, the features are explored in detail. The goal is to describe or summarize data in ways that are meaningful and useful for insights generation. It provides simple summaries about the sample and the measures. Together with simple graphics analysis, it forms the basis of virtually every quantitative analysis of data.

Given both the features – 'Current' and 'New' are continuous in nature; the following measures are relevant to understand the central tendency and spread of the variable.

Measures of Central Tendency	Measures of Dispersion	Visualization Method
Mean	Range	Histogram
Median	1st Quartile	Boxplot
Mode	3rd Quartile	
Minimum	Inter Quartile Range (IQR)	
Maximum	Variance	
	Standard Deviation	

3.2.1 Measures of Central Tendency:

Measures of Central Tendency	Current	New
Mean	270.3	267.5
Median	270.0	265.0
Mode	272.0	263.0
Minimum	255.0	250.0
Maximum	289.0	289.0

3.2.2 Measures of Dispersion:

Measures of Dispersion	Current	New
Range	34.0	39.0
1st Quartile	263.0	262.0
3rd Quartile	275.2	274.5
Inter Quartile Range (IQR)	12.2	12.5
Variance	76.6	97.9
Standard Deviation	8.8	9.9

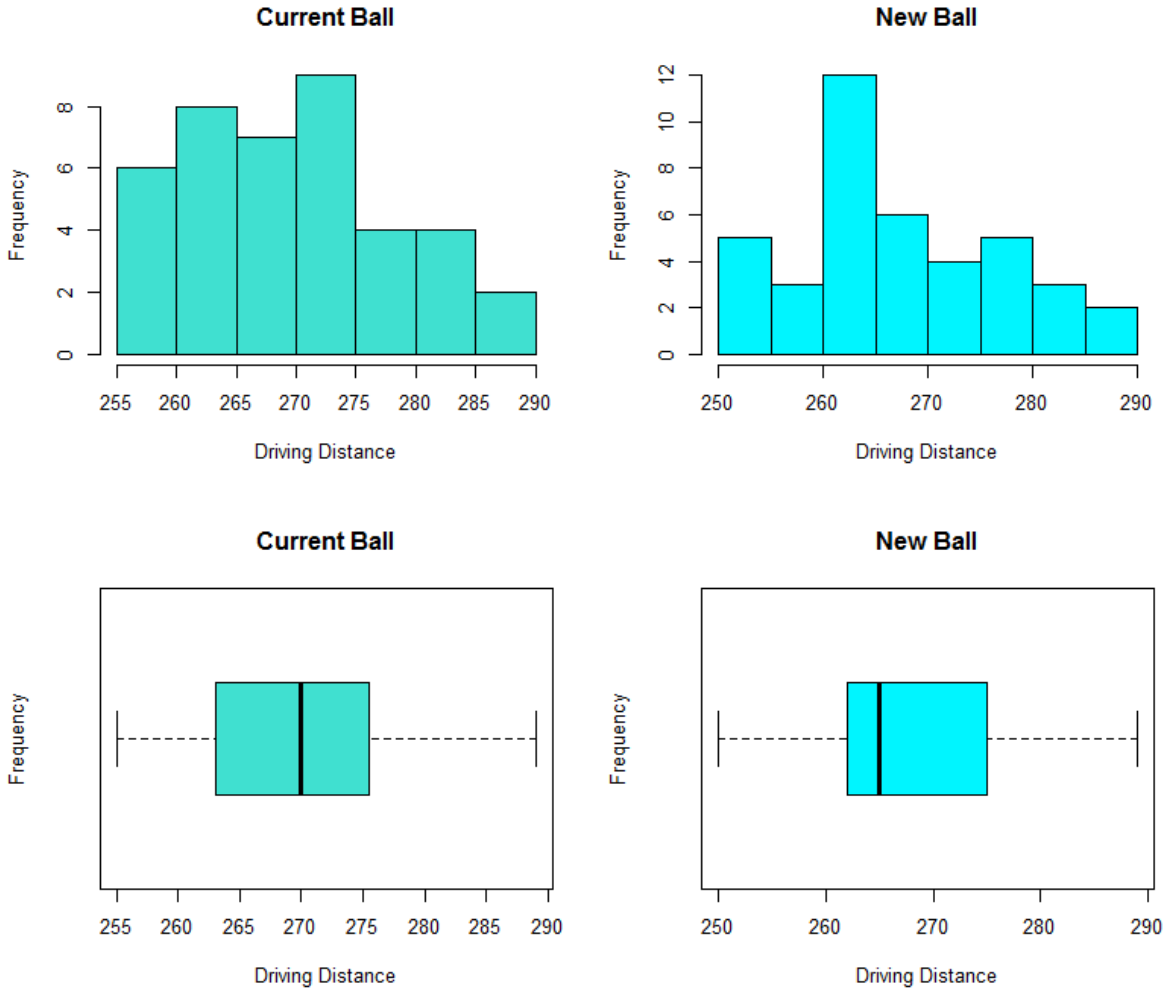
The data for 'Range', '1st Quartile' & '3rd Quartile' have been obtained using the R function 'summary' *(The output is provided in the section 9.1 Measures of Central Tendency).*

The data for 'Inter Quartile Range (IQR)' has been computed and is the difference in value between 3rd Quartile (75 percentile) & 1st Quartile (25 percentile).

The variances have been obtained using R function 'var':

The standard deviations have been obtained using R function 'sd':

3.3 Data Visualization – Histogram and Boxplot:



3.3.1 Key Observations:

- The average distance covered by 'New' golf ball (Mean = 267.5) is lower as compared to 'Current' golf ball (Mean = 270.3)
- 'New' golf ball observed to have relatively higher variation in the data distribution in comparison to 'Current' golf ball
- No outlier was observed in the data range for features – Current & New

3.4 Hypothesis formation

For the hypothesis formulation, we have to define the Null Hypothesis & Alternative Hypothesis.

- ***Null Hypothesis***
  - It is a hypothesis that says there is no statistical significance between the two variables. The null hypothesis is formulated such that the rejection of the null hypothesis proves the alternative hypothesis is true
- ***Alternative Hypothesis***

- It is one that states there is a statistically significant relationship between two variables. The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis

In Par Inc. case, the management would like to produce the new golf balls once it is comparable to the current golf balls. A sample of 40 balls of both the current and new models were tested with a mechanical hitting machine so that any difference between the mean distances for the two models could be attributed to a difference in the two models. Therefore, a hypothesis test that Par Inc. could use to compare the driving distance of the current and new golf balls. The Null Hypothesis & Alternative Hypothesis is formulated as follow:

Null Hypothesis (H0): μ1 - μ2 = 0 (i.e. they are the same)

Alternative Hypothesis (Ha): μ1 - μ2 ≠ 0 (i.e. they are not the same)

Where,

- ***μ1:*** Mean driving distance of current model golf ball
- ***μ2:*** Mean driving distance of new golf ball

By formulation of above hypotheses, we assume that the current and new golf balls show no significant difference to each other.

3.5 Hypothesis Testing Method

On the basis of the details shared for the Par Inc. project, we can assume the following:

- One machine
- Two populations
- No other influences considered
- Independently chosen

It seems to be an independent sample case. The two-tailed test will be applicable for the project.

Let us calculate the p-value using the R function 't.test'.

```
> t.test(Current,New)

        Welch Two Sample t-test

data:  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-tailed test, the p-value = 0.188 ÷ 2 = 0.094.

The p-value for the two-tailed test is 0.094, which is greater than level of significance α (0.05). Therefore, the Null Hypothesis (H0) will not be rejected. The conclusion is that this data does not provide statistical evidence that the new golf balls have either a lower mean driving distance or a higher mean driving distance. This implies that Par Inc. should take the new golf balls in production as the p-value indicate that there is no significant difference between estimated population mean of current as well as new golf balls.

3.6 Confidence Interval

Analysis of Current Model:

```
> t.test(Current)

        One Sample t-test

data:  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
```

Inference: The 95% confidence interval of population mean for Current model is between 267.4757 & 273.0743. This implies that, with 95% confidence, we can say that the sample mean driving distance of current balls will be within this range.

Analysis of New Model:

```
> t.test(New)

        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
```

Inference: The 95% confidence interval of population mean for New model is between 264.3348 & 270.6652. This implies that, with 95% confidence, we can say that the sample mean driving distance of New balls will be within this range.

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

```
> t.test(Current,New)

        Welch Two Sample t-test

data:  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
```

Inference: The 95% confidence interval of difference in population mean between both the models is between **–1.38 yards on the lower end 6.93 yards on the upper end**. This implies that, with 95% confidence, we can say that the difference in sample mean driving distance of both the models will be within the above range. For the 40 balls we have taken in this sample, the difference in mean driving distance is -2.775 yards (267.5 – 270.275) which falls in the range.

3.7 Need of Larger Sample Size:

Steps to follow:

- Get the difference between two sample means (2.775 as calculated earlier)
- Calculate pooled Standard Deviation using following formula:

$$SD*_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$$

- Execute the power T Test, with current parameters, and decide if larger size is needed.
- Calculate the samples number (in case Power of Test is insignificant)

Pooled Standard Deviation:

```
> delta=mean(Current)-mean(New)
> pooledSD <- (((40-1)*(8.75^2)+(40-1)*(9.9^2))/(40+40-2))^0.5
> delta
[1] 2.775
> pooledSD
[1] 9.342711
```

Power T Test:

```
> power.t.test(n=40, delta = 2.775, sd=9.342,
+              sig.level = 0.05,type = "two.sample",
+              alternative = "two.sided" )

     Two-sample t test power calculation

              n = 40
          delta = 2.775
             sd = 9.342
      sig.level = 0.05
          power = 0.258536
```

```
    alternative = two.sided

NOTE: n is number in *each* group
```

Inference: The Power of test is 0.258 or 25.8%, which means there are only 25% chances that the null hypothesis will not be rejected when it is false. Hence we need to revisit the number of samples to increase the power of test.

Recalculate Sample size using Power T Test:

Consider Power of test 95%, and significance level 0.188 (The P value calculated) and execute the Power T test once again.

```
> power.t.test(power=0.95, delta = 2.775,
+              sd=9.342,sig.level = 0.188,type = "two.sample",
+              alternative = "two.sided" )

     Two-sample t test power calculation

              n = 199.2145
          delta = 2.775
             sd = 9.342
      sig.level = 0.188
          power = 0.95
    alternative = two.sided

NOTE: n is number in *each* group
```

Inference: We can see that we need sample size of 200 (rounded up) in order to get 95% power of Test.

4 Conclusion and Recommendation

- From the Preliminary Data Analysis, we confirm that the mean driving distance of New Ball is less than Old Ball. (267.5 yards Vs 270.3 yards).
- When the Data Set explored further with the help of descriptive statistics and Visualization, we learnt that
  - The New Golf ball has relatively higher variation.
  - No outliers observed in both the samples.
  - Both the samples have nearly Normal distribution, however New Design is slightly more skewed towards right.
- After applying the Hypothesis Testing, we may conclude that Par Inc. should take the new golf balls in production as the p-value indicate that there is no significant difference between estimated population mean of current as well as new golf balls.
- However, with the current sample size of 40, the Power of test is 0.258 or 25.8%, which means there are only 25% chances that the null hypothesis will not be rejected when it is false. Hence it is recommended to have at least 200 sample size to have 95% power.