Mini Project – Par Inc (Business Statistics)

By Ramprasad Mohan

Hypothesis testing for New Product Development

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". **Managerial Report is to be prepared as below.**

2 Assumptions

The sample size of both the samples in 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 and also the sample size is sufficient for analysis.

3 Step by Step approach

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

- Exploratory Data Analysis
- Descriptive Statistics
- Data Visualization
- Hypothesis formation
- Selection of appropriate Hypothesis Testing method
- 95% Confidence Intervals
- Need of Larger Sample Size
- Conclusion and Recommendation.

4 Solution

4.1 Qn 1 : Provide descriptive statistical summaries of the data for each model.

Sample Size: 40, No. of samples: 2

Five Point Summary

```
> summary(golf)
    Current          New
 Min.   :255.0   Min.   :250.0
 1st Qu.:263.0   1st Qu.:262.0
 Median :270.0   Median :265.0
 Mean   :270.3   Mean   :267.5
 3rd Qu.:275.2   3rd Qu.:274.5
 Max.   :289.0   Max.   :289.0
```

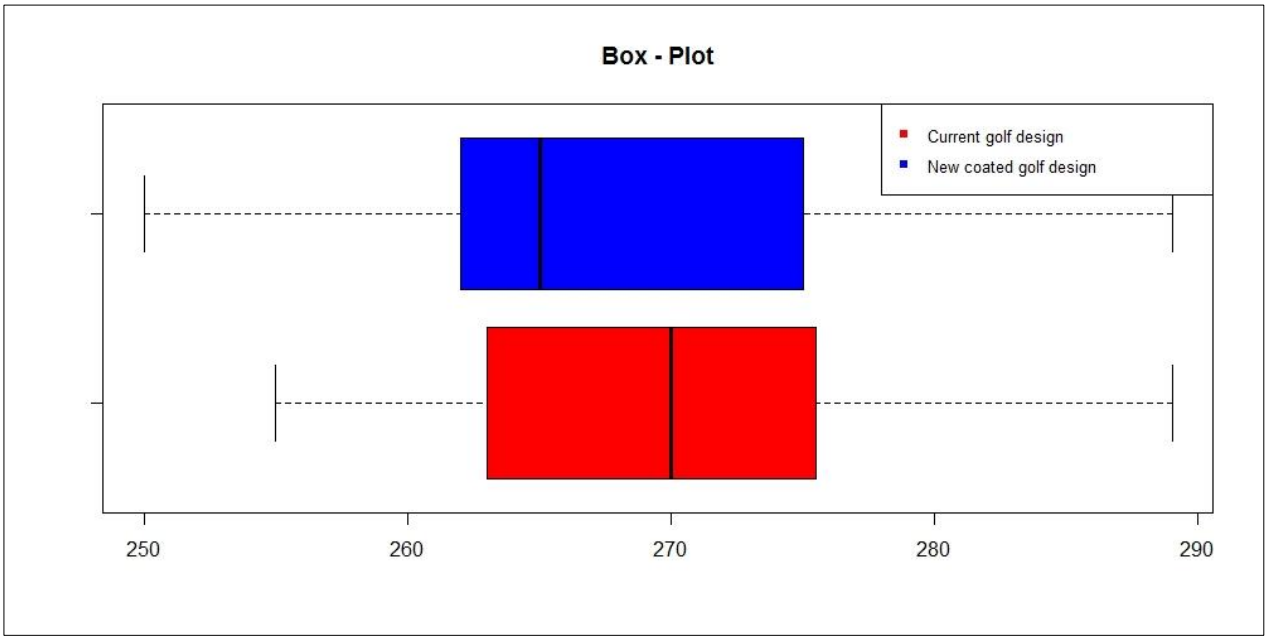
Observations:

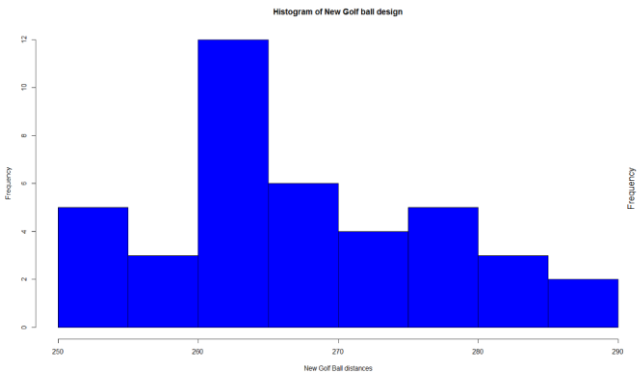
o Both the samples seems to be normally distributed.

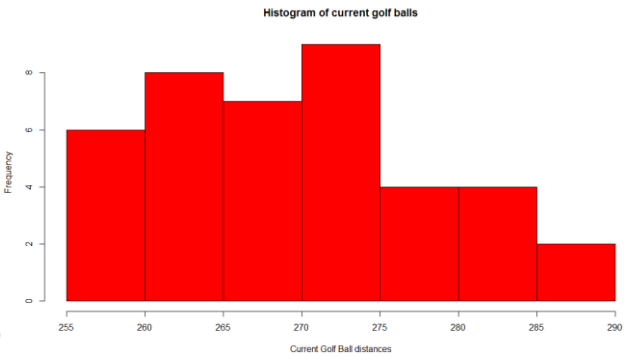
o Mean and Median Values have slight difference.

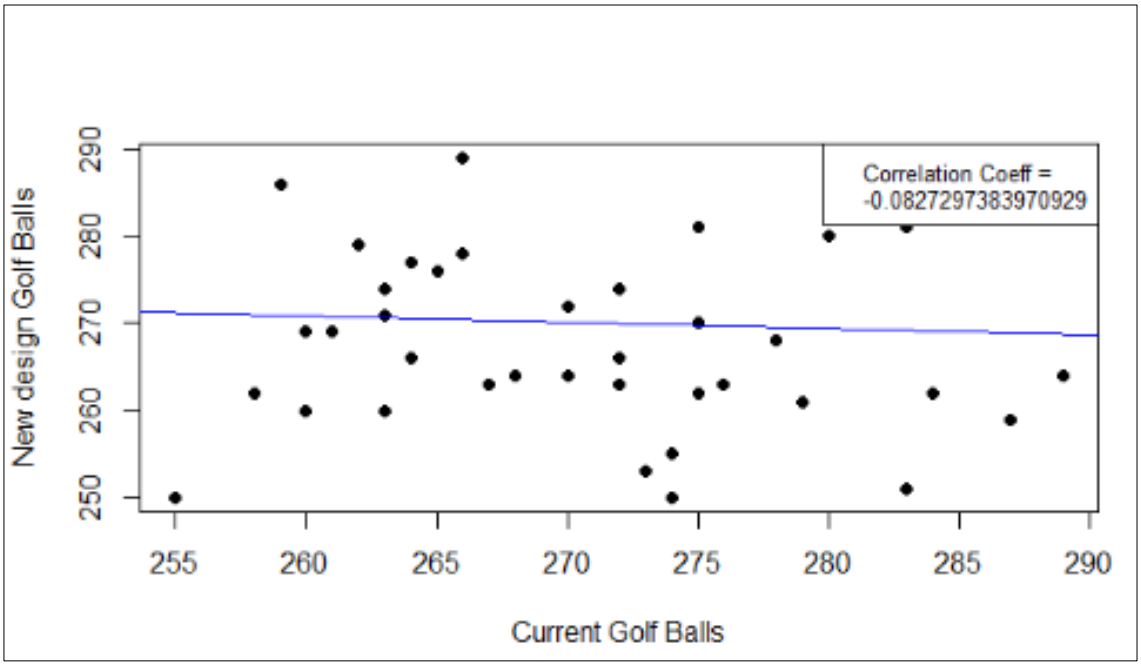
The new coated design data looks more normally distributed and left skewed, whereas the old design data looks right skewed.

A general scatter plot shows very weak correlation between the values which shows that each value/data is unbiased with different values from test results.









4.2 Qn 2 : 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

- The level of significance (Alpha) = 0.05.
- The sample size , N = 40 which is large for a Zstat Test for 1 sample.
- But since there are two independent samples, we have to use a Tstat test.
- Degree of Freedom: Since the sample is the same for both Sampling tests, we have (N-1)*2 degrees of freedom : 78

Null Hypothesis:

It is a hypothesis/status quo 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

Alternate 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.
A hypothesis test that Par Inc. could use to compare the driving distance of the current and new golf balls.

Hypothesis Formulation:

Null Hypothesis : No difference between the mean distances for the two models due to change in the design.

H0 : $\mu 1 = \mu 2$ i.e $\mu 1 - \mu 2 = 0$

Alternate Hypothesis : There is difference between the mean distances for the two models due to change in the design.

Ha : μ1 ≠ μ2 i.e μ1 - μ2 ≠ 0

T Test Equation:

$$t_{stat} = (\bar{d} - \mu D)/S_d\sqrt{n} \text{ Here } S_d = \sqrt{\Sigma(D - \bar{D})^2}_{n-1} \text{ and } d_f = n-1$$

4.3 Qn 3 : Analyze the data to provide the hypothesis testing conclusion. What is the p-value for your test?

Based on the details shared for the Par Inc. project, we can assume the following:

- Two populations
- No other influence factors considered
- Independently chosen

Since there are two independent sample case. The two-tailed test will be applicable for the project. The p-value is calculated using the R function 't.test'.

Code :

```
-------------------------------------------------------------------------
#### t test ###

x<-t.test(golf$Current,golf$New,var.equal = TRUE)
x
if(x$p.value<0.05)
{{
  print("Reject Null Hypothesis : accept that there is difference in mean
distances travelled due to design changes")
}
else{
  print("Accept Null Hypothesis : accept that There is no difference in mean
distances travelled due to design changes")
}}

-------------------------------------------------------------------------
```

Output

```
        Two Sample t-test

data:  golf$Current and golf$New
t = 1.3284, df = 78, p-value = 0.1879
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.383958  6.933958
sample estimates:
mean of x mean of y
  270.275   267.500
```

[1] "Accept Null Hypothesis: accept that there is no difference in mean distances travelled due to design changes"

P – Value : 0.1879

Since it is a two-tailed test, the p-value = 0.1879 ÷ 2 = 0.094 (approx.)

The calculated p-value is greater than level of significance α (0.05). Therefore, the Null Hypothesis (H0) will not be rejected.

4.4 Qn 4 : What is your recommendation for Par Inc.?

Recommendations to Par Inc : Though 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 The company can go ahead with the introduction of the newly coated, cut-resistant, longer-lasting designed golf ball to the market as the new features does not alter/change the driving distances compared to the existing golf balls. The p-value indicate that there is no significant difference between estimated population mean of current as well as new golf balls. However to further analyze the testing needs to be performed with more sample sizes.

4.5 Qn 5 : 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?

Code :

```
library(Rmisc)
library(lattice)
library(plyr)

### Function for CI estimate from Rmisc##

CI(golf$Current,ci = 0.95)
CI(golf$New,ci = 0.95)
```

95% C.I Interval for Current existing golf balls

Upper	mean	Lower
273.0743	270.2750	267.4757

95% C.I Interval for New design golf balls

Upper	mean	Lower
270.6652	267.5000	264.3348

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

```
    Two Sample t-test

data:  golf$Current and golf$New
t = 1.3284, df = 78, p-value = 0.1879
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.383958  6.933958
```

Inference: The 95% confidence interval for the difference between the means of the two sample size is -1.383958 and 6.933958

4.6 Qn 6 : Do you see a need for larger sample sizes and more testing with the golf balls? Discuss

To check need for larger sample size, the following is performed:

- 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)

Code :

Pooled Standard Deviation :

```
diff <- mean(golf$Current) - mean(golf$New)

# Need for larger sample size

df1 <- 40-1
df2 <- 40-1

sdcurrent <- sd(golf$Current)
sdnew <- sd(golf$New)

pooldsd <- sqrt((df1*(sdcurrent^2) + df2*(sdnew^2))/(df1+df2))

-----------------------------------------------------------------------------------------------------------

## power T test

power.t.test(n=length(golf$New),delta = diff,sd = pooldsd, sig.level =
0.05,type = "two.sample",alternative = "two.sided")
```

Output :

```
     Two-sample t test power calculation

              n = 40
          delta = 2.775
             sd = 9.342469
      sig.level = 0.05
          power = 0.2585147
    alternative = two.sided
```

Inference: The Power of test is 0.25851 or 25.8%, which means there are only 25% chances that the null hypothesis will not be rejected when it is false.(Type II Error) Hence, we should revisit the number of samples to increase the power of test.

Finding the Sample size for required Power of test value 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.

Code :

```
largersample<-power.t.test(power = 0.95,delta = diff,sd = pooldsd,sig.level =
0.188,type = "two.sample",alternative = "two.sided")
largersample$n
```

Output :

```
     Two-sample t test power calculation

              n = 199.2345
          delta = 2.775
             sd = 9.342469
      sig.level = 0.188
          power = 0.95
    alternative = two.sided
```

Inference: We can see that we require sample size of 200 (rounded up) to get 95% power of Test. Hence larger sample size would be helpful to reduce changes of hypothesis errors.

5 Conclusion

- From the Preliminary Data Analysis, we find that the mean driving distance of New Ball is less than Old Ball. (267.5 yards Vs 270.275 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.
- The Hypothesis Testing concludes that this data does not provide statistical evidence that the new golf balls have either a lower or 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.
- Confidence Interval:
    - The 95% confidence interval of population mean for Current model is between 267.4757 & 273.0743.
    - The 95% confidence interval of population mean for new model is between 264.3348 & 270.6652.
- Power of Test:
    - Although the 2 Tail Hypothesis Test recommends to launch the new ball design into Production, the Power of Test is only 25%.
    - In order to have 95% Power of Test, it is recommended to have the number of samples as 200 for both the designs, and then conclude the Hypothesis test recommendations.

6 Appendix A – Source Code

```
getwd()
setwd("D:/R_wd")

### Reading the Excel File ####
library(readxl)

golf <- read_excel("C:/Users/indway/Desktop/BACP/Fundementals of
Statistics/Mini Project/Golf.xls")
class(golf)

golf<-as.data.frame(golf)

#-------------------------------------------------------------------------

### Qn 3 , Descriptive statistics ####

library(ggplot2)
summary(golf)

qplot(golf$Current,golf$New,col=c("red","blue"),main = "General Scatter
Plot",xlab = "Current golf ball dist",ylab = "New golf ball dist" )

boxplot(golf$Current,golf$New,col= c("Red","Blue"),horizontal = TRUE,main =
"Box - Plot")

legend("topright",c("Current golf design","New coated golf
design"),cex=0.8,col=c("red","blue"),pch=15)
```

```r
hist(golf$Current,col="red",main = "Histogram of current golf
balls",xlab="Current Golf Ball distances")

hist(golf$New,col="blue",main = "Histogram of New Golf ball design",xlab="New
Golf Ball distances")


### Correlation ####

corrltion<-cor(golf$Current,golf$New,method = "pearson")
corrltion

plot(golf$Current,golf$New,xlab = "Current Golf Balls",ylab = "New design
Golf Balls",pch=16)

yyy<-lm(golf$Current~golf$New,data = golf)

abline(yyy,col="blue")

legend("topright",c("Correlation Coeff = ",corrltion),cex = 0.8 )

#------------------------------------------------------------------------
#### t test ###

x<-t.test(golf$Current,golf$New,var.equal = TRUE)
x
if(x$p.value<0.05)
{{
  print("Reject Null Hypothesis : accept that there is difference in mean
distances travelled due to design changes")
}
else{
  print("Accept Null Hypothesis : accept that There is no difference in mean
distances travelled due to design changes")
}}

#------------------------------------------------------------------------

###  Qn 5 Confidence Interval estimate ###

qplot(golf$Current,geom = "density",binwidth = 2.5)
qplot(golf$New,geom = "density",binwidth = 2.5)

mean(golf$Current)
sd(golf$Current)

library(Rmisc)
library(lattice)
library(plyr)

### Function for CI estimate from Rmisc##

CI(golf$Current,ci = 0.95)
CI(golf$New,ci = 0.95)

diff <- mean(golf$Current) - mean(golf$New)

# Need for larger sample size

df1 <- 40-1
df2 <- 40-1
```

```
sdcurrent <- sd(golf$Current)
sdnew <- sd(golf$New)

pooldsd <- sqrt((df1*(sdcurrent^2) + df2*(sdnew^2))/(df1+df2))

## power T test

power.t.test(n=length(golf$New),delta = diff,sd = pooldsd, sig.level =
0.05,type = "two.sample",alternative = "two.sided")

### to get the required sample size ###

largersample<-power.t.test(power = 0.95,delta = diff,sd = pooldsd,sig.level =
0.188,type = "two.sample",alternative = "two.sided")
largersample$n
```