Mini – Project
Report on Golf Dataset

Report By

Chakradhar Varma Sangaraju

Table of Contents

1. Project Objective

The objective of the project is to perform Hypothesis testing, provide insights and test conclusions, provide Descriptive Statistical Summaries and Recommendations for the given Golf Dataset. The report would form a reference for the Management at Par Inc., to take a decision on whether to introduce the Newer model of cut-resistant, longer-lasting golf ball and whether there is any need for larger sample sizes and more testing with the golf balls. This report will address the following:

- Formulation for Hypothesis Test
- Hypothesis Test Conclusion
- Descriptive statistics
- Recommendations

2. Assumptions

The distribution of the given data is assumed to be Normally Distributed. The level of significance alpha is equal to 0.05.

3. Solutions

This Solutions section will answer the Questions asked in the following steps:

1. Preliminary Analysis
2. Hypothesis Testing
3. P-Value Approach and Conclusion
4. Descriptive Statistical Summaries
5. Analysis on Population Mean
6. Recommendations

3.1 Preliminary Analysis

To compare the driving distances for the two balls, 40 balls of both the new and current models were subjected to distance tests by par. So, the sample contains 40 observations.

- **Five-point Summary of the data set(gd_data)**

```
   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
```

3.2 Hypothesis Testing

The first thing to do is to formulate and present the rationale for a hypothesis test that Par Inc., could use to compare the driving distance of the current and new golf balls.

From the Question Statement we can infer that even though the tests with the coating have been promising on the durability of the improved product, another issue has been raised and that is the effect of the new coating on driving distances. 40 balls of both the new and current models were subjected to distance test. They are independent sample and test follows a large sample case according to central limit theorem.

By formulation of these hypothesis there is assumed that the new and current golf balls show no significant difference to each other. The Null hypothesis and Alternative hypothesis are formulated as follow:

Mean distance of current-model balls: μ1.

Mean distance of new cut-resistant balls: μ2.

H0: μ1 = μ2 (Mean distance of current balls equals mean distance of new balls).
H1: μ1 != μ2 (Mean distance of current balls is not equal mean distance of new balls)

Taking the level of significance. $\alpha = 0.05$ so $z = 1.96$

Since the population standard deviation is unknown and both the sample are independent, we will use classic Tstat test. As the Null hypothesis is having absolute equality, we should use two-tailed test.

The T-test result from R is as below:

```
        Two Sample t-test

data:  Current and 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 Current    mean of New
                        270.275            267.500
```

Figure 3.2: T-test Data for both Current and New Model

3.3 P-Value Approach

To test further hypothesis, we set α at .05 and our rejection criteria is Reject Ho and accept Ha if P < α (.05). Since hypothesis test done through statistic tools given p-value, we will use the P-value approach.

Looking at the preliminary Analysis for each model, we can initially conclude that Current model has a longer range of distance based on the 40 samples with a mean of 270.3 compare to 267.5 for new model.

P-value = 0.1879 > 0.05 = α.

P-value is greater than alpha so our decision rule for this problem is:

- Accept Null hypothesis H0.
- Mean distance of cut-resistant balls **equals** mean distance of current-model balls.
- The new cut-resistant balls have **no difference** in distance compared to the current-model one.

Therefore, we **recommend** for Par Inc., that they should introduce the new model.

3.4 Descriptive Statistics Summaries

Below is the descriptive data for the given Golf data set.

Measure	Current	New
Count	40	40
Mean	270.3	267.5
Minimum	255	250
Maximum	289	289
Sample variance	76.61	97.95
Sample Standard deviation	8.75	9.89
Median	270.0	265.0

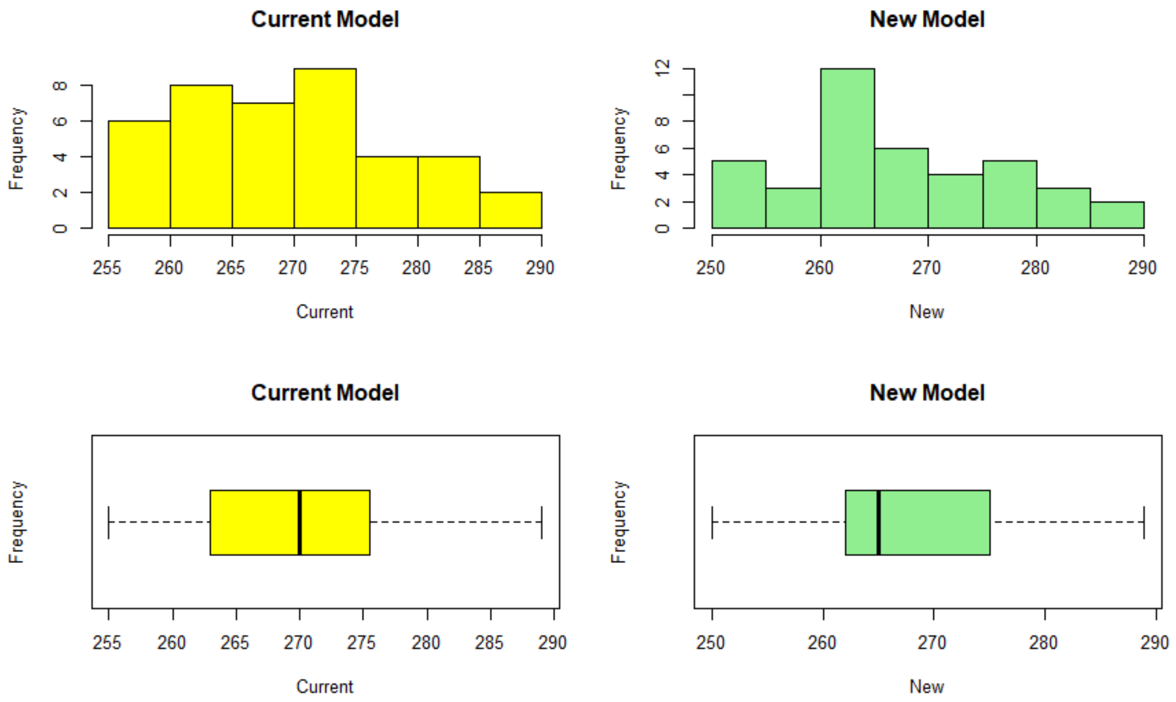

Figure.3.4: Histogram and Boxplot for the Current and New Models.

From the above table we can observe that the mean distance of Current Ball is a bit higher than that of New Ball. However, the standard deviation of Current Ball is lower than that of New Ball (8.75 compared to 9.90). Obviously, variance of Current Ball is lower than variance of New Ball (76.61 and 91.95). Since variance is a measure of group's dispersal, it means the data of New Ball fluctuates more than Current Ball's. It is almost indifferent between the interquartile of both sets of data.

The Mean value of current model is slightly more than Median and we can see some symmetry, where as in the case of new model median is closer to the first quartile and it has long right tail which indicates the right skewness.

There are no Outliers and extreme values present in the data. The range of box-and-whisker of New is taller, which indicates that it has larger variance than Current's. The Current Ball's distance is relatively more stable than the newer ones.

3.5 Analysis on Population Mean

At 95% Confidence Interval for Current Model:

Property	Value
Mean	270.275
Standard deviation	8.75298
N	40
Standard Error	1.383968
Upper confidence limit	273.0743
Lower confidence limit	267.4757

At 95% Confidence Interval for New Model:

Property	Value
Mean	267.5
Standard deviation	9.896904
N	40
Standard Error	1.564838
Upper confidence limit	270.6652
Lower confidence limit	264.3348

From the figure 3.2 above, the 95% confidence interval for the difference between the means of the two population is (-1.383958, 6.933958).

3.6 Recommendations

P-Value for this two tailed test is 0.1879, which is greater than level of significance α (0.05). Hence, H0 will not be rejected which shows that Par Inc., should introduce new ball as the P value indicates that there is no significant difference between estimated population mean of current as well new sample model. The 95% confidence interval for the population mean of the current model is 267.4757 to 273.0743 and of the new model is 264.3348 to 270.6652. It means that the estimated population mean for Par Inc., should lie within this range for consistent result. However, the 95% confidence interval for the difference between the means of the two populations is (-1.384, 6.934). The calculated test statistic value is far from the rejection area. The larger the sample size the smaller the standard deviations which means point estimator of mean will become more precise. Hence there is no need to take larger sample size

4. Conclusion

To conclude, Formulation for the Hypothesis test and the conclusion has been given on the given data set of "Golf". As part of analysis descriptive statistics has been applied, and graphs has been generated to visualize and to know the insights of the data. Recommendations has been provided based on the hypothesis test for the Par Inc.,

5. Appendix A – Source Code

```
#Installation of Packages and invoke libraries
install.packages('readxl')

library(readxl)

#set up of working directory
setwd("D:/BACP Program/R Directory")
getwd()

## [1] "D:/BACP Program/R Directory"

#create a data frame object and read the data file.
gd_data = read_excel("Golf.xls")

View(gd_data)

#Dimension of the data set
dim(gd_data)

## [1] 40  2

#Attach Data to the session
attach(gd_data)

#Find out summary of the data
summary(gd_data)

##     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

#Hypothesis Testing for Tstat under equal variance case
t.test(Current, New, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  Current and 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

#Descriptive Statistics
var(Current)

## [1] 76.61474

var(New)

## [1] 97.94872

sd(Current)

## [1] 8.752985

sd(New)

## [1] 9.896904

#Histogram and Boxplot for the features
par(mfrow=c(2,2))
hist(Current,main='Current Model',xlab='Current',ylab='Frequency',col='yel
low')
hist(New,main='New Model',xlab='New',ylab='Frequency',col='lightgreen')
boxplot(Current,main='Current Model',xlab='Current',ylab='Frequency',col='
yellow',horizontal = TRUE)
boxplot(New,main='New Model',xlab='New',ylab='Frequency',col='lightgreen',
horizontal = TRUE)
```


```
#confidence interval for the population mean of each 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

std_err_current = sd(Current)/sqrt(length(Current))
print(std_err_current)

## [1] 1.383968

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

std_err_new = sd(New)/sqrt(length(New))
print(std_err_new)
```

```
## [1] 1.564838
```