Mini Project 1: Golf Case Study

Project Report

Ravisekhar R

29.04.2019

PGP BABI March 2019

PROBLEM STATEMENT

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.

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. Identify the p-value for the test. Formulate a recommendation for Par Inc.

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

4. Identify the 95% confidence interval for the population mean of each model, and the 95% confidence interval for the difference between the means of the two population.

5. Discuss if there's need for larger sample sizes and more testing with the golf balls

Dataset used: See Appendix A

SOLUTION 1: HYPOTHESIS FORMULATION

Here the aim is to compare the two models (let's say Current and New) based on their distance tests. It is also known that the difference between the mean distances for the two models attributes to the difference in design.

So in order to say that the test is successful, there has to be a difference in the means of the two data samples (Current and New). In that case, when we formulate this problem for hypothesis testing, **the null hypothesis would claim that the means of the two samples are the same** and the **alternate hypothesis would claim that the means of**

the two samples are different. To formulate it in mathematical notation,

Mean of sample - Current: μ_C

Mean of sample - New: μ_N

Null hypothesis (H_0) : $\mu_C = \mu_N$

Alternate hypothesis (H_A) : $\mu_C \neq \mu_N$

Type 1 error rate (α) = 0.05 (default α value in most cases)

Decision rule:

If test statistic **(p-value)** < α, then **reject the null hypothesis**

If test statistic **(p-value)** > α, then **fail to reject the null hypothesis**

Here the sample size is not so huge (40 data points). Also, there is no information about the standard deviation of the population given. So, with the assumption that the data is normally distributed, we **apply the two sample t-test** to compare the means of the two samples. Assume that the variance is equal in both the samples.

SOLUTION 2: DATA ANALYSIS AND RECOMMENDATION

We perform t-test of the two models, current and new, using the t.test function in R programming language. The code snippet (Code snippet 1) is as shown below:

```
## Hypothesis testing
```{r}
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
```

(Code snippet 1)

From the R-code, we can see that the p-value of the test is 0.1879, which is greater than our critical value 0.05,

i.e: **p-value(0.1879)** > **α (0.05)**

Hence, according to our previous decision rule, we fail to reject the null hypothesis.

So the recommendation for Par Inc. would be that the means of driving distances of the current and new samples are the same.

SOLUTION 3: DESCRIPTIVE STATISTICAL SUMMARIES

The following R-code snippet (Code snippet 2) shows the stats calculation and plotting of boxplot.

```
## Descriptive Statistical Summaries
```{r}
mu_C=mean(Current)
var_C=var(Current)
sd_C=sd(Current)

mu_N=mean(New)
var_N=var(New)
sd_N=sd(New)

N=40

boxplot(Current,New,col=c("Red","Green"),horizontal = TRUE,main="Mean comparison",xlab="Driving distance",ylab="Models",names=c("Current","New"))
```

(Code snippet 2)

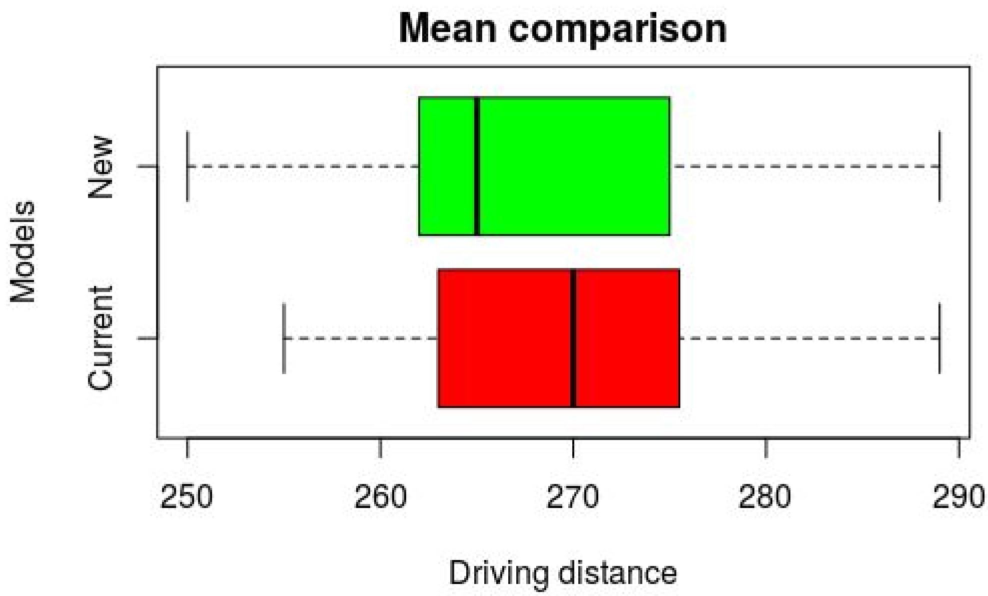
Number of instances in both samples = 40

The following table (Table 1) shows the comparison of mean, variance and standard deviation of the two models.

Model	Mean	Variance	Standard deviation
Current	270.28	76.61	8.75
New	267.5	97.95	9.90

(Table 1)

The following figure (1) shows a boxplot of the two models:
```



(Figure 1)

SOLUTION 4: 95% CONFIDENCE INTERVAL CALCULATION

The following R-code snippet (Code snippet 3) shows the calculation of 95% confidence interval for the two samples individually:

```
```{r}
t.test(Current,mu=mu_C)
t.test(New,mu=mu_N)
```
```

```
        One Sample t-test

data:  Current
t = 0, df = 39, p-value = 1
alternative hypothesis: true mean is not equal to 270.275
95 percent confidence interval:
 267.4757 273.0743
sample estimates:
mean of x
  270.275
```

```
        One Sample t-test

data:  New
t = 0, df = 39, p-value = 1
alternative hypothesis: true mean is not equal to 267.5
95 percent confidence interval:
 264.3348 270.6652
sample estimates:
mean of x
    267.5
```

(Code snippet 3)

From the Code snippet 3, we can see that the 95% confidence interval for :

Current Sample: **[267.4757,273.0743]**

New Sample: **[264.3348,270.6652]**

From the Code snippet 2, we can see that the 95% confidence interval for :

Population: **[-1.383958,6.933958]**

SOLUTION 5: DISCUSSION ON SAMPLE SIZE AND TESTING

The following R-code snippet shows the calculation of power of the t-test with sample size=40:

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

```
     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
```

From the output of the above code, we can see that with a sample size of just 40, the power that can be achieved is just 14.27%.

Now to increase the power of the test, we definitely need larger sample sizes and more testing with the golf balls.

The following R-code shows the calculation of sample size to achieve a power of 90%:

```
powerTest2=power.t.test(n=NULL,cohen.d,sig.level =
0.05,power=0.90,type="two.sample",alternative = "two.sided")
powerTest2
```
```

```
```
     Two-sample t test power calculation

              n = 516.4577
          delta = 0.2019067
             sd = 1
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

From the output of R code, we can see that to achieve a power of 90%, the sample size should be atleast 517.

APPENDIX - A

	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
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
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