MINI PROJECT

Managerial report for Par Inc.

**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.

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

**PROJECT OBJECTIVE**

The main aim of the project is to prove the hypothesis and obtain the results according to the problem statement with the help of R and excel.

* Importing the dataset into R
* Understanding the structure of the dataset and its contents
* Graphical exploration
* Descriptive statistics
* Insights from the statistics

**DATA**

The data provided in the above case that the number of samples is 40 i.e. there are 40 balls and the comparison is between the new and current model of the golf balls. The aim here is to increase the durability of the new golf balls. Here the driving distances are the main attributes to differentiate in design.

**ASSUMPTIONS**

The assumptions are made accordingly for hypothesis testing here it is apparent that the test sample size is 40. The important assumption is that the sampling distribution of the mean of the values in the case will be a normal distribution as iterated in the central limit theorem. Normal distribution is characterized by a bell curve.

**EXPLORATORY ANALYSIS STEP BY STEP APPROACH**

**Packages in R**

The necessary packages to be installed are “xlsx”. Use the command install.packages(“xlsx”). Since we need to read values of the excel file which are in xls format we need this package which is not present by default.

**Environment setup and data import**

To set up a working directory

A working directory is set up to make the import and export of files easier. The data relevant to the project are stored in the location which makes data handling simpler if all the data is stored in the same location retrieving them is not a difficult task. We use the following command

getwd(“file location”)

setwd to set the file location that’s chosen.

**Import and read the dataset**

The R-studio is used as a tool to explore the data and the cardio dataset is in .xls format which is extension or spreadsheet file format. The command used to import data is new<-read.xls(“filename”) or we can use a command to choose the dataset file directly from the directory using new<-read.xls(file.choose()).

**Variable identification**

The variables are the parameters given in the function on which the analysis is carried out. The parameters are called features.

The R function used for variable identification are as given below**:**

**Dim-** this function is used to find out the dimensions i.e. number of rows and columns in the dataset.

**Head**-this function is used to get the head values which are the starting values by default only 10 are displayed but it can be increased as per specification.

**Tail-** this function is used to read and focus only on the ending values or last few values by default only 10 are displayed but it can be increased as per specification.

Head and tail are used to check if the data provided in the particular dataset behaves similarly throughout the table or is random.

**View-**it gives the tabulated from of the data as shown in the dataset in the output window.

**Names-**this function gives only the names or the parameters specified in the dataset arranged as a row.

**Summary-**it is a generic function used to give summaries of the objects in the dataset to be analyzed with different parameters included such as minimum, maximum, median and quantiles depending on the class of the variable individually for all the features. Result of model fitting functions.

**Str-**the function str displays the structure of an arbitrary R object. It is a diagnostic function and an alternate to the summary function which displays the class of the object and its name.

**INSIGHTS**

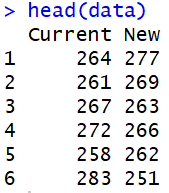
Dimensions (dim) specify that there are 2 columns and 40 rows of data in the given dataset. The distances are given in the rows of the dataset and columns specify the types of golf balls namely current and new.



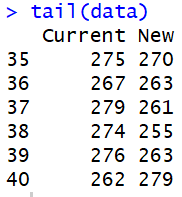
Names (names) prints the names of the parameters in the dataset for which the data points. Here both the datatypes are same. The names obtained are current and new which are mentioned in double quotes.



**Head:** This function returns the first part of vector,matrix and table. This shows only few of the values and can assess the behaviour of data points rather on viewing large amount of data.Here in our case only six of the first few values are shown for both the current and new.

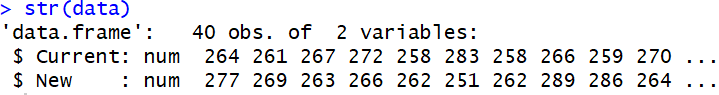


**Tail:** The function returns the last part of vector, matrix and table. In this case it displays only six values from the bottom of the table so that we can analyze the behaviour of data points rather on looking into large amount of data.

****

**Str**

This displays internal structure of R objects and used to diagnose the variables. One line for each basic structure is displayed.



**Summary**

It gives a summation of the key attributes such as the minimum value, maximum value, mean, median and quantile range for both the current as well as new values of the golf ball. From the observation it is clear that

**Current**

The least value of the driving distance is displayed as 255 and the maximum value of driving distance is 289. The mean value obtained from the table is 270.3 which is very necessary in the hypothesis testing and value at the center or the median is 270.

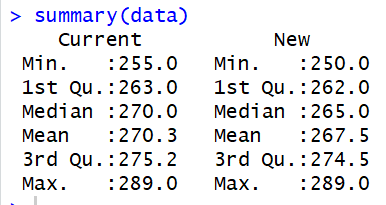
The first quartile is 263 and the third quartile is 275.2

**New**

The minimum value of the driving distance is displayed as 250 and the maximum value of driving distance is 289. Mean value obtained from the table is 267.5 which is essential in the hypothesis testing and value at the center or the median is 265.

The first quartile is 262 and the third quartile is 274.5

In comparison the values of new and current ball mean driving distance is higher for current ball and maximum value is same for both whereas the minimum value is lesser in new golf balls compared to the current ball.



**Table**

This command gives the frequency of the values of each feature in the dataset. Just like other commands it is case sensitive.

**Current:**

The table gives the values of the particular column “current”. The frequency of each entry is specified.

**New:**

The table gives the values of the particular column “new”. The frequency of each entry is specified.

**To find mean, variance and standard deviation of the data objects**

**CURRENT**

Mean value: mean(Current) =**270.2**

Standard deviation: sd(Current) =**8.75**

Variance: var(Current) =**76.61**

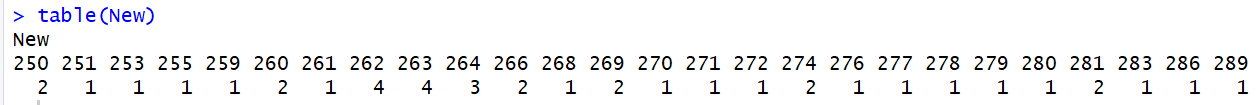
**NEW**

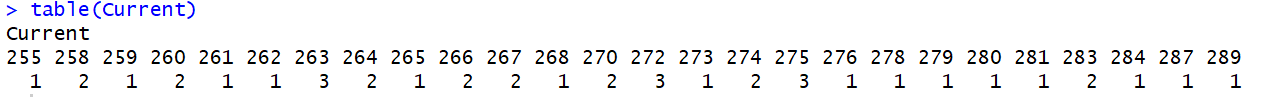
Mean value: mean(New) =**267.50**

Standard deviation: sd(New) =**9.89**

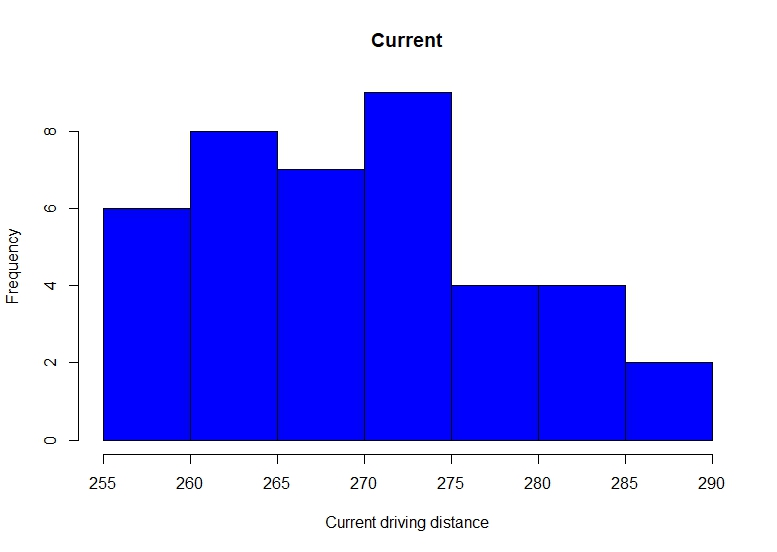
Variance: var(New) =**97.948**

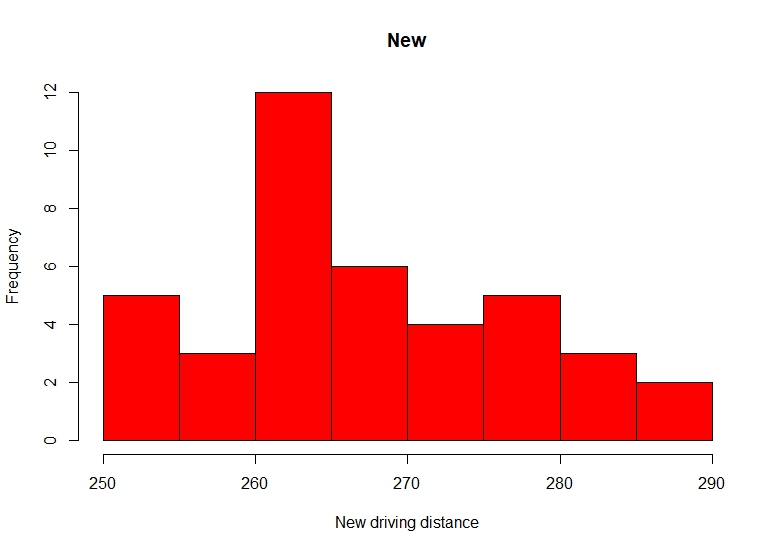
On observation it is evident that values are considerably higher for new golf ball here we compare the standard deviation from the mean and mean value also the variance is measured.

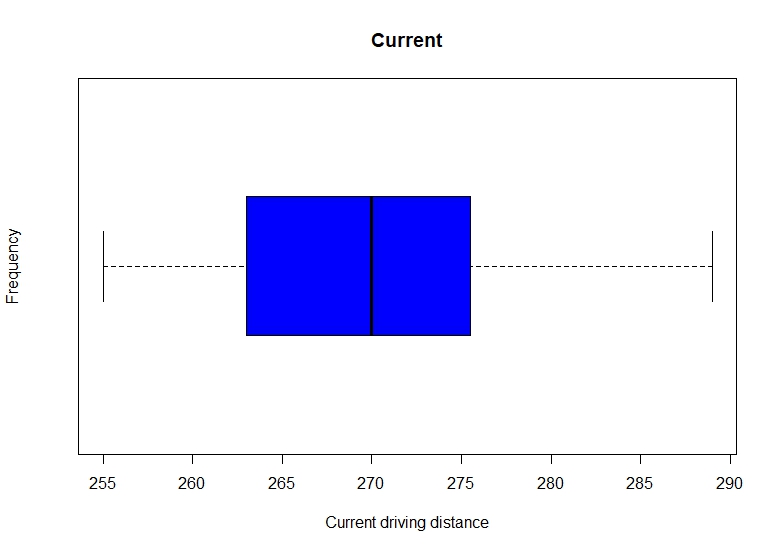


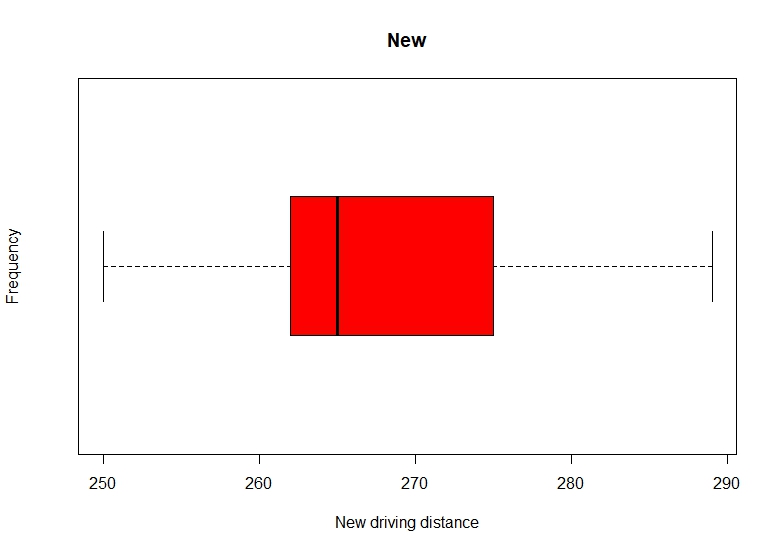


The visual representation of the dataset is given below in terms of Histogram and Boxplot which in marked in blue for current golf balls and in red for new golf balls. X axis gives the driving distance and Y axis gives the Frequency. From the values we obtained the values for standard deviation, variance and mean are higher for new golf balls. The boxplot shows the values as specified in summary for e.g. the mean value is marked with a dark line at the center and the quantile range in the box. There are no outliers which affect the values. The driving of new golf ball is more than the current golf ball. There is a difference of 270.2-267.5=2.7.

****

****

****

****

**HYPOTHESIS TESTING**

There two hypothesis to be formulated here which are the null hypothesis and the alternate hypothesis. The null hypothesis is denoted by H0 and the alternate hypothesis is denoted by Ha

**NULL HYPOTHESIS**

H0: mu1-mu2=0

Null hypothesis is denoted by H0 and mu1 is the mean driving distance of the current golf ball, mu2 is the mean driving distance of the new golf ball.

**ALTERNATE HYPOTHESIS**

Ha: mu1-mu2 != 0

Alternate hypothesis is denoted by Ha and mu1 is the mean driving distance of the current golf ball, mu2 is the mean driving distance of the new golf ball. This means that the mean driving distances are not the same for both the golf balls.

We must accept the alternate hypothesis and reject the null hypothesis.

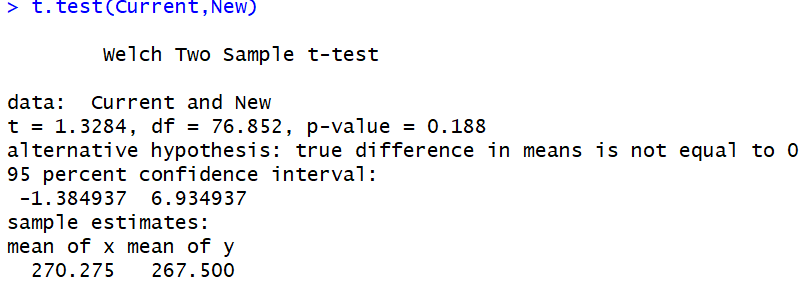
On framing the hypothesis we make an assumption are not dependent and are not very different from one another.

If the difference in the mean distances of the current ball and the new ball is same then the new design did not make any difference on the driving distance. If mu1 and mu2 are same then new design did not have any effect on the golf ball.

To determine the p value

We should use the command t.test in r studio for both the values since from the data set we have seen that the samples are independent and are non-overlapping. We use welch t test to test the hypothesis that two population have equal means.

And the result displayed is the welch two sample t test where values of t, df and p value is obtained along with the stats of 95% confidence interval



It is also called as welch two sample t test from the above values it can be stated that the distribution is two tailed and p value obtained is p=0.188 where p>α hence we fail to reject the null hypothesis. It is not necessary that the driving distance of new balls will be different than that of current golf balls. The design of the new golf ball can be carried out as planned.

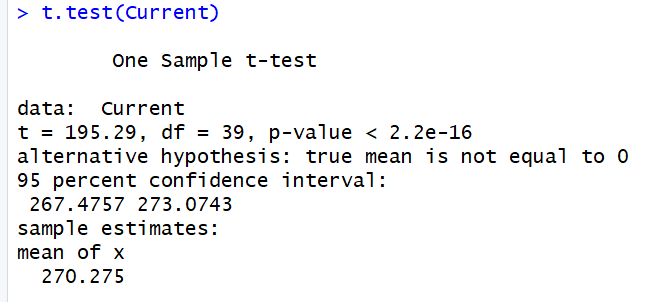
95% confidence interval for difference in means of the two population

The same command for the t test gives the difference in values as stated above the 95% confidence interval is in between -1.384937 and 6.934937. The difference in the values is difference of mean of x and y namely as mu1 and mu2 which 270.275-267.500=2.775.

Confidence interval

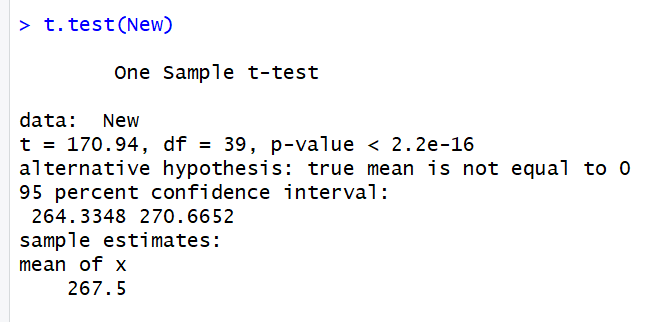
To get one sample t test to get 95% confidence interval we need to use the command t.test for both the columns of data Current and New individually.

t.test(Current)



From the above data we can draw a conclusion that value of p is less than 2.2e^(-16) and 95% confidence interval lies between values 267.4757 and 273.0743.

t.test(New)



From the above data we can draw a conclusion that value of p is less than 2.2e^(-16) and 95% confidence interval lies between values 264.3348 and 270.6652.

**T test Power**

To find the power of t test we use the command

power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = 0.05,

power = NULL,

type = c("two.sample", "one.sample", "paired"),

alternative = c("two.sided", "one.sided"),

strict = FALSE, tol = .Machine$double.eps^0.25)

n=40

standard deviation=9.342

Level of significance=5%=0.05

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

We get the value of the power to be 0.258536 which is 25.8% which means that the percentage of the rejection null hypothesis when its false is 25.8% so it is clear that the power of test is to be increased.

**Find sample size using the formula for power t test**

We assume the power of test to be 95% which will be 0.95 and level of significance to be 0.188

Now we calculate the power of t test using the same formula given above:

power.t.test(n=40,delta=2.775,sd=9.342,sig.level=0.188,type=”two sample”, alternative=”two sided”)

**CONCLUSION**

From the above statistics we derive a set of results

The new and current golf ball design follow a normal distribution and there are no outliers in the case.

* In hypothesis testing we decide that the new golf balls have a lower or higher mean driving distance.
* From the values we obtained the values for standard deviation, variance and mean are higher for new golf balls.
* Par Inc. must manufacture new golf balls as there is no drastic difference.
* To obtain a 95% power of test we need sample size of around 188 samples which can be estimated to be around 200.
* Power of test obtained is 25.8%
* 95% confidence interval lies between values 267.4757 and 273.0743 for Current golf balls.
* 95% confidence interval lies between values 264.3348 and 270.6652 for new golf balls.

**SOURCE CODE**

**#=======================================================================**

**#**

**# *Exploratory Data Analysis - Golf***

**#**

**#=======================================================================**

**#*Environment set up and Data Import***

**#set up working directory**

**>** setwd("C:/Users/Saurabh/Desktop")

> getwd()

[1] "C:/Users/Saurabh/Desktop"

> install.packages("xlsx")

Error in install.packages : Updating loaded packages

> install.packages("xlsx")

Installing package into ‘C:/Users/Saurabh/Documents/R/win-library/3.4’

(as ‘lib’ is unspecified)

Warning in install.packages :

package ‘xlsx’ is in use and will not be installed

> library("xlsx")

**#*to store information into a variable called data***

> data<-read.xlsx(file.choose())

Error in read.xlsx(file.choose()) :

> data

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

**#*to access variables as the names are in the dataset***

> attach(data)

> Current

[1] 264 261 267 272 258 283 258 266 259 270 263 264 284 263 260 283 255 272 266 268 270 287 289 280 272

[26] 275 265 260 278 275 281 274 273 263 275 267 279 274 276 262

> New

[1] 277 269 263 266 262 251 262 289 286 264 274 266 262 271 260 281 250 263 278 264 272 259 264 280 274

[26] 281 276 269 268 262 283 250 253 260 270 263 261 255 263 279

***#to find rows and columns***

> dim(data)

[1] 40  2

> names(data)

[1] "Current" "New"

> View(data)

> head(data)

Current New

1     264 277

2     261 269

3     267 263

4     272 266

5     258 262

6     283 251

> tail(data)

Current New

35     275 270

36     267 263

37     279 261

38     274 255

39     276 263

40     262 279

> str(data)

'data.frame': 40 obs. of  2 variables:

$ Current: num  264 261 267 272 258 283 258 266 259 270 ...

$ New    : num  277 269 263 266 262 251 262 289 286 264 ...

**#*to find all the values and their types***

> summary(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

> table(Current)

Current

255 258 259 260 261 262 263 264 265 266 267 268 270 272 273 274 275 276 278 279 280 281 283 284 287 289

1   2   1   2   1   1   3   2   1   2   2   1   2   3   1   2   3   1   1   1   1   1   2   1   1   1

> table(New)

New

250 251 253 255 259 260 261 262 263 264 266 268 269 270 271 272 274 276 277 278 279 280 281 283 286 289

2   1   1   1   1   2   1   4   4   3   2   1   2   1   1   1   2   1   1   1   1   1   2   1   1   1

**#*to plot tables for different parameters specified in the dataset*.**

> hist(Current,main="Current",xlab ="Current driving distance",ylab ="Frequency",col ="blue")

> hist(New,main="New",xlab ="New driving distance",ylab ="Frequency",col ="red")

> boxplot(Current,main="Current",xlab ="Current driving distance",ylab ="Frequency",col ="blue",horizontal=T)

> boxplot(New,main="New",xlab ="New driving distance",ylab ="Frequency",col ="red",horizontal=T )

> mean(Current)

[1] 270.275

> sd(Current)

[1] 8.752985

> var(Current)

[1] 76.61474

> mean(New)

[1] 267.5

> sd(New)

[1] 9.896904

> var(New)

[1] 97.94872

> obj1<-c(mean(Current),sd(Current),var(Current))

> obj1

[1] 270.275000   8.752985  76.614744

> obj2<-c(mean(New),sd(New),var(New))

> obj2

[1] 267.500000   9.896904  97.948718

> 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

> 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

> 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

> 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

> 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

***#end of the source code***