Golf dataset

Ajay

May 18, 2018

## R Markdown

# setting working directory

setwd(“C:/Users/majay/Desktop/Python/”) getwd()

# reading the data and assigning to variable gf

gf <- read.csv(“Golf.csv”)

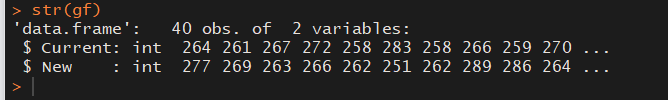
# using attach so that we can directly use the variables in the dataset

attach(gf)

# Lets do some exploratory data analysis of this data set

# lets check for the number of observations in our dataset

str(gf)

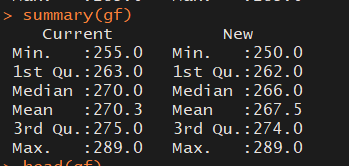


Observations:

We have 40 values and 2 columns. – Current and New

# we see ther are total 40 values and 2 variables.

summary(gf)



# Observations:

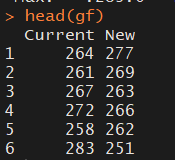
# we can see minimum value of 255 and 250 in the current and new columns

# Mean is 270.3 and 267.5 respectively

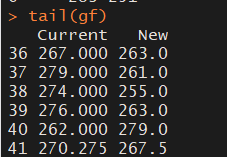
# both have 289 as max value

# Lets check the top 6 and bottom 6 values

head(gf)

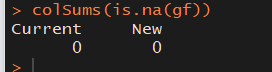


tail(gf)



# check for any missing values

colSums(is.na(gf))

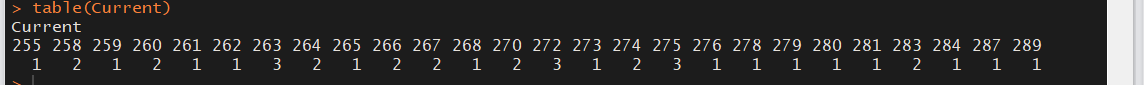


Observations:

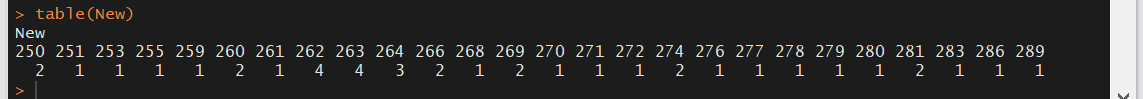
No Missing values.

# lets check the frequencies

table(Current)



table(New)

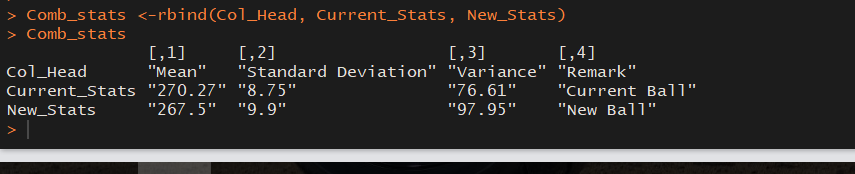


# Lets bind the mean, standard deviation and variance of the columns and see if we can find any observations

Col\_Head <- c(“Mean”, “Standard Deviation”, “Variance”, “Remark”) Current\_Stats <- c(round(mean(Current), digits = 2), round(sd(Current), digit = 2), round(var(Current), digits = 2), “Current Ball”)

New\_Stats <- c(round(mean(New), digits = 2), round(sd(New), digit = 2), round(var(New), digits = 2), “New Ball”)

Comb\_stats <-rbind(Col\_Head, Current\_Stats, New\_Stats) Comb\_stats



# Observations :

# As informed earlier mean of new golf balls is 267.5 whereas mean of current golf balls is 270.27

# New Golf balls have higher variance compared to current golf balls

# Coming to standard deviation, Current has 8.75 and new golf balls has 9.9

# Lets solve the problem by hypothesis :

# Lets µ1 be Mean driving distance of current model golf ball and

# µ2 be the Mean driving distance of new model golf ball

# Our Null hypothesis (H0) wil be : µ1 - µ2 = 0 (Meaning there is no difference in both new and current model)

# And our Alternate Hypothesis (Ha) will be : µ1 - µ2 != 0 (meaning New and Current are not the same)

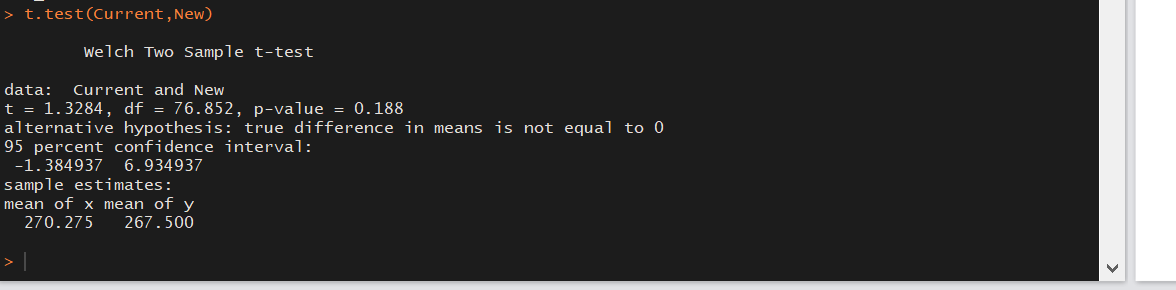
# Assumption :

# By formulation of above hypotheses, we assume that the current and new golf balls show no

# significant difference to each other. Its a case of independent sample, two-tailed test.

# Lets calculate the p value using Two sample t-test

t.test(Current,New)



# Observations:

# 1. In the two-tailed test,the p-value what we get is 0.188.

# 2. The p-value is 0.188, which is greater than level of significance 0.05.

# Hence, we cant reject the the Null Hypothesis (H0).

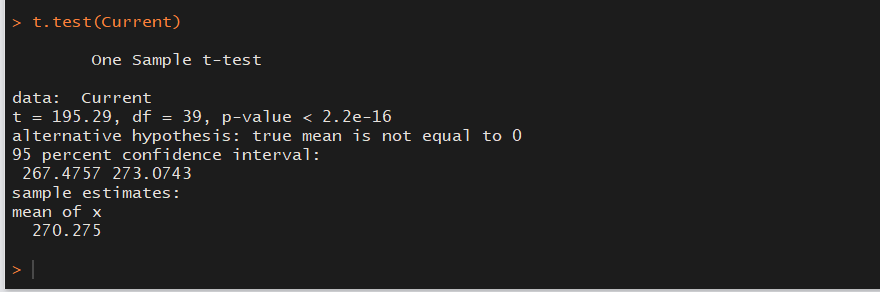
# 3. So the conclusion is that this data does not provide statistical evidence that the new golf balls have either a lower mean driving distance or a higher mean driving distance.

# 4. 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.

# Lets check for the Confidence interval - One sample t test to get 95% confidence interval

t.test(Current)

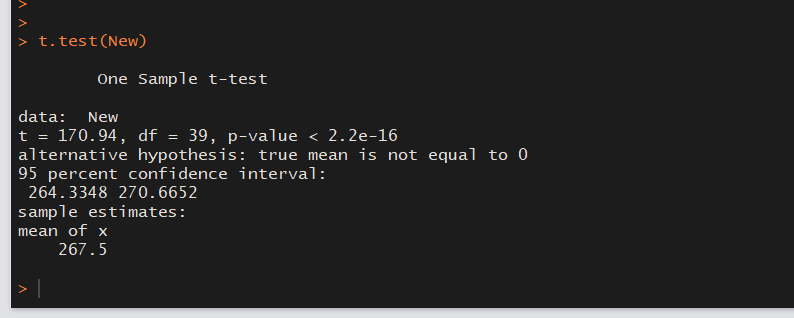


# Observations:

# The 95% confidence interval of population mean for Current model is between 267.4757 & 273.0743.

# This implies - with 95% confidence, we can say that the sample mean driving distance of current balls will be within this range.

t.test(New)

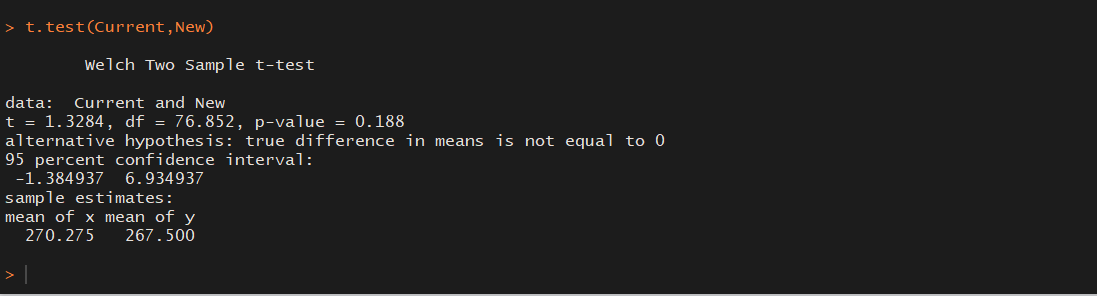


# Observations :

# 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.

# Now lets again see the 95% confidence interval for the difference in means of the two population

t.test(Current,New)



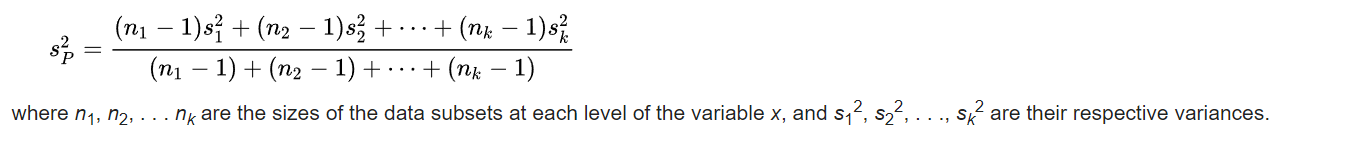
Observations:

**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 i.e |267.5 - 270.275| which falls in the range.**

# Now lets see if we need more samples :

# The difference between sample means is - 2.775

# The pooled standard deviation formula is –



Let’s calculate the value for pooled standard deviations –

Calculate the power T test and check if larger sample is required.

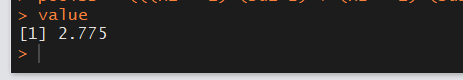
# We may need to find the samples number if in case Power of Test is insignificant.

# Pooled Standard Deviation -

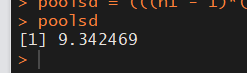
n1 = 40 n2 = 40 m1 = mean(Current) m2 = mean(New) sd1 = sd(Current) sd2 = sd(New) value = m1 - m2 poolsd = (((n1 - 1)*(sd1^2) + (n2 - 1)*(sd2^2))/(n1 + n2 - 2))^0.5

# Lets see the poolsd values -

Value

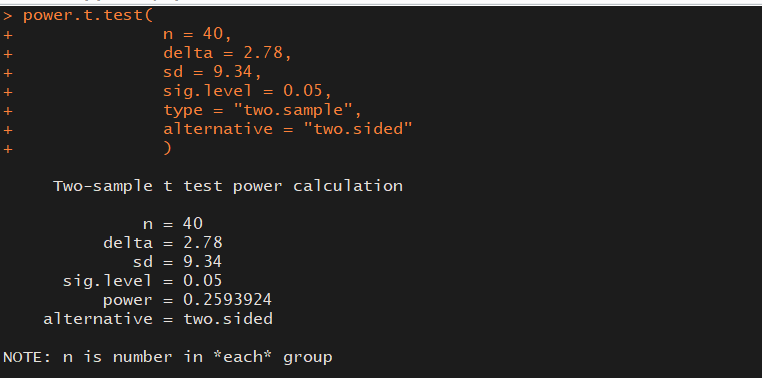


Poolsd



# Lets check the power T test -

power.t.test( n = 40, delta = 2.78, sd = 9.34, sig.level = 0.05, type = “two.sample”, alternative = “two.sided” )



# Obervations :

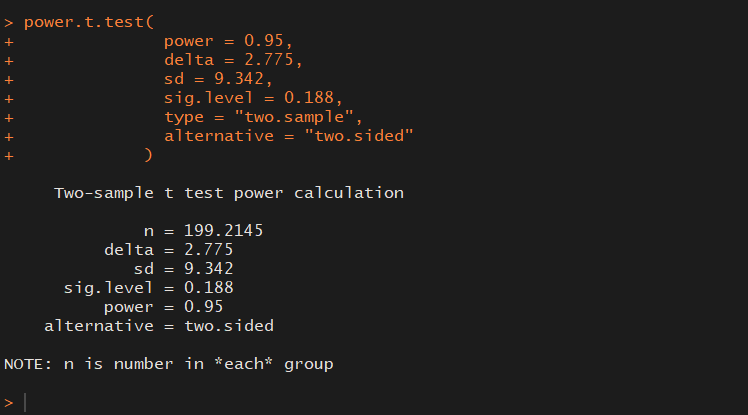
The Power of test is 0.2593 or 25.9%, which means there is only 26% chances that the null hypothesis will be rejected when it is false. Hence, we should revisit the number of samples to increase the power of test.

# Lets Recalculate the Sample Size using Power T Test :

# we will consider Power of test as 95% , significance level of 0.188 (The P value calculated using

# t.test(Current,New)) 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” )



# Observation :

# We can see that we need a sample size of 199-200 to get 95% power of Test.

# Now answering the below questions :

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

# Answers/Concusion fron the Test :

# 1. We can confirm that the mean driving distance of New Ball is less than Old Ball. (267.5 yards Vs 270.3 yards).

# 2. 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. - WHich 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. I have listed the descriptive statistical summaries earlier. Please refer it.

# 4. When it comes to confidence interval we can say :

# 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.

# 5. Talking about the power of test :

# Eventhough the 2 Tail Hypothesis Test recommends to launch the new ball design into Production, the # Power of Test is only around 26%.

# In order to have 95% Power of Test, it is recommended to have 199-200 samples for both the models.

Code used :

---

title: "Golf dataset"

author: "Ajay"

date: "May 18, 2018"

output: word\_document

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

## R Markdown

# setting working directory

setwd("C:/Users/majay/Desktop/Python/")

getwd()

# reading the data and assigning to variable gf

gf <- read.csv("Golf.csv")

# using attach so that we can directly use the variables in the dataset

attach(gf)

#Lets do some exploratory data analysis of this data set

# lets check for the number of observations in our dataset

str(gf)

# we see ther are total 40 values and 2 variables.

summary(gf)

# obersvations

# we can see minimum value of 255 and 250 in the current and new columns

# Mean is 270.3 and 267.5 respectively

# both have 289 as max value

# Lets check the top 6 and bottom 6 values

head(gf)

tail(gf)

# check for any missing values

colSums(is.na(gf))

# lts check if there are any repeating values. i.e frequencies

table(Current)

table(New)

# Lets bind the mean, standard deviation and variance of the coluumns and see if we can find any observations

Col\_Head <- c("Mean", "Standard Deviation", "Variance", "Remark")

Current\_Stats <- c(round(mean(Current), digits = 2),

round(sd(Current), digit = 2),

round(var(Current), digits = 2),

"Current Ball")

New\_Stats <- c(round(mean(New), digits = 2),

round(sd(New), digit = 2),

round(var(New), digits = 2),

"New Ball")

Comb\_stats <-rbind(Col\_Head, Current\_Stats, New\_Stats)

Comb\_stats

# Obrervations :

# As informed earlier mean of New gof balls is 267.5 whereas mean of current golf balls is 270.27

# New Golf balls have higher variance compared to current golf balls

# Coming to standard deviation, Current has 8.75 ad new golf balls has 9.9

# Lets solve the problem by hypothesis :

# Lets µ1 be Mean driving distance of current model golf ball and

# µ2 be the Mean driving distance of new model golf ball

# Our Null hypothesis (H0) wil be : µ1 - µ2 = 0 (Meaning there is no difference in both new and current)

# And our Alternate Hypothesis (Ha) will be : µ1 - µ2 != 0 (meaning New and Current are not the same)

# Assumption :

# By formulation of above hypotheses, we assume that the current and new golf balls show no

# significant difference to each other. Its a case of independent sample, two-tailed test.

# Lets calculate the p value using Two sample t-test

t.test(Current,New)

# Observations:

# 1. In the two-tailed test,the p-value what we get is 0.188.

# 2. The p-value is 0.188, which is greater than level of significance 0.05.

# Hence, we cant reject the the Null Hypothesis (H0).

# 3. So 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.

# 4. 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.

# Lets check for the Confidence interval - One sample t test to get 95% confidence interval

t.test(Current)

# Observations:

# The 95% confidence interval of population mean for Current model is between 267.4757 & 273.0743.

# This implies - with 95% confidence, we can say that the sample mean driving distance of current

# balls will be within this range.

t.test(New)

# Observations :

# 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.

# Now lets again see the 95% confidence interval for the difference in means of the two population

t.test(Current,New)

# Obseravtions :

# 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.

# Lets see if we need more samples.

# The difference between sample means is - 2.775

# The pooled standard deviation formula is -

# calculate the power T test and check if larger sample is required.

# We may need to find the samples number if in case Power of Test is insignificant.

# Pooled Standard Deviation -

n1 = 40

n2 = 40

m1 = mean(Current)

m2 = mean(New)

sd1 = sd(Current)

sd2 = sd(New)

value = m1 - m2

poolsd = (((n1 - 1)\*(sd1^2) + (n2 - 1)\*(sd2^2))/(n1 + n2 - 2))^0.5

# Lets see the value an poolsd values -

value

poolsd

# Lets check the power T test -

power.t.test(

n = 40,

delta = 2.78,

sd = 9.34,

sig.level = 0.05,

type = "two.sample",

alternative = "two.sided"

)

# Obervations :

The Power of test is 0.2593 or 25.9%, which means there is only 26% chances that the null hypothesis will be rejected when it is false. Hence, we should revisit the number of samples to increase the power of test.

# Lets Recalculate the Sample Size using Power T Test :

# we will consider Power of test as 95% , significance level of 0.188 (The P value calculated using

# t.test(Current,New)) 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"

)

# Observation :

# We can see that we need a sample size of 199-200 to get 95% power of Test.

# Now answering the below questions :

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

# Answers/Concusion fron the Test :

# 1. We can confirm that the mean driving distance of New Ball is less than Old Ball. (267.5 yards Vs

# 270.3 yards).

# 2. 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. - WHich 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. I have listed the descriptive statistical summaries earlier. Please refer it.

# 4. When it comes to confidence interval we can say :

# 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.

# 5. Talking about the power of test :

# Eventhough the 2 Tail Hypothesis Test recommends to launch the new ball design into Production, the # Power of Test is only around 26%.

# In order to have 95% Power of Test, it is recommended to have 199-200 samples for both the models.

############################################################################

T H E - E N D ############################################################################