INDUSTRIAL INTERNSHIP

**WEEKLY PERFORMANCE REPORT (WPR)**

# Student Name : Aman Srivastava

**Head-Coordinator Name : vatshal goshwami**

**Organization : Cureya Hours Worked : Monday 5 , Tuesday 5**

**Wednesday 5 , Thursday 5 , Friday 5 , Saturday 5 .**

On Monday a new task was given by team leader to Test the null hypothesis that whether the pay increases in 2016 And alternate hypothesis :the pay didn’t increase in 2016 And also check for type 1 and type 2 errors

**Monday**:- I started reseraching all the topic related to Hypothesis testing in Machine learning using Python

Here is the some glance:

*1. What is hypothesis testing ?*

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.

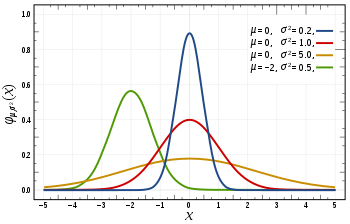
Ex : you say avg student in class is 40 or a boy is taller than girls.

all those example we assume need some statistic way to prove those. we need some mathematical conclusion what ever we are assuming is true.

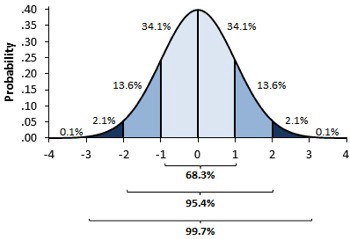
*2. why do we use it ?*

Hypothesis testing is an essential procedure in statistics. A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. When we say that a finding is statistically significant, it’s thanks to a hypothesis test.

1. *what are basic of hypothesis ?*



The basic of hypothesis is normalisation and standard normalisation. all our hypothesis is revolve around basic of these 2 terms. let’s see these.



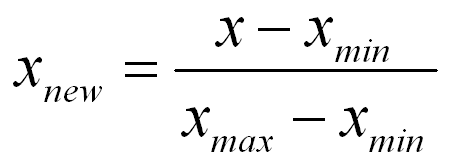
You must be wondering what’s difference between these two image, one might say i don’t find, while other will see some flatter graph compare to steep. well buddy this is not what i want to represent , in 1st first you can see there are different normal curve all those normal curve can have different mean’s and variances where as in 2nd image if you notice the graph is properly distributed and**mean =0 and variance =1 always**. concept of z-score comes in picture when we use **standardised normal data.**

**Tuesday:**

****Normal Distribution -****

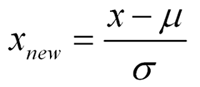
A variable is said to be normally distributed or have a **normal distribution** if **its distribution** has the shape of a **normal curve** — a special bell-shaped **curve**. … The graph of a **normal distribution** is called the **normal curve**, which has all of the following **properties**: 1. The mean, median, and mode are equal.

Normal distribution formula:-



****Standardised Normal Distribution —****

A standard normal distribution is a normal distribution with mean 0 and standard deviation 1



****Which are important parameter of hypothesis testing ?****

****Null hypothesis :-****In inferential statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured phenomena, or no association among groups

In other words it is a basic assumption or made based on domain or problem knowledge.

**Alternative hypothesis :-**

The alternativehypothesis is the hypothesis used in ****hypothesis**** testing that is contrary to the null hypothesis. It is usually taken to be that the observations are the result of a real effect (with some amount of chance variation superposed)

**Wednesday:**

****Level of significance:**** Refers to the degree of significance in which we accept or reject the null-hypothesis. 100% accuracy is not possible for accepting or rejecting a hypothesis, so we therefore select a level of significance that is usually 5%.

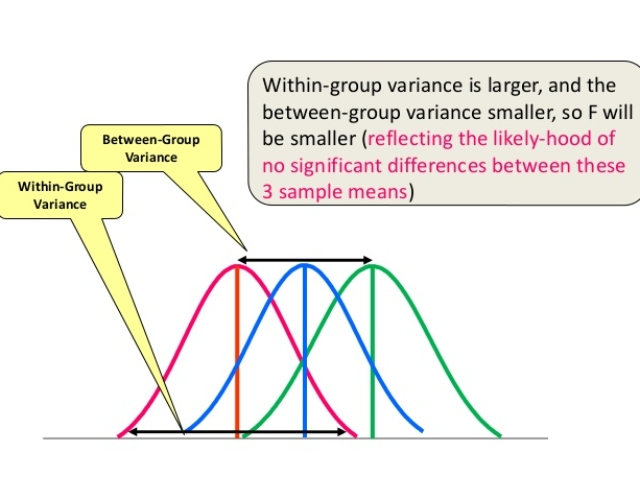
This is normally denoted with alpha(maths symbol ) and generally it is 0.05 or 5% , which means your output should be 95% confident to give similar kind of result in each sample.

****Type I error:**** When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha. In hypothesis testing, the normal curve that shows the critical region is called the alpha region

****Type II errors:**** When we accept the null hypothesis but it is false. Type II errors are denoted by beta. In Hypothesis testing, the normal curve that shows the acceptance region is called the beta region.

****ANOVA (F-TEST) :-**** The t-test works well when dealing with two groups, but sometimes we want to compare more than two groups at the same time. For example, if we wanted to test whether voter age differs based on some categorical variable like race, we have to compare the means of each level or group the variable. We could carry out a separate t-test for each pair of groups, but when you conduct many tests you increase the chances of false positives. The [analysis of variance](https://en.wikipedia.org/wiki/Analysis_of_variance) or ANOVA is a statistical inference test that lets you compare multiple groups at the same time.

****F = Between group variability / Within group variability****



Unlike the z and t-distributions, the F-distribution does not have any negative values because between and within-group variability are always positive due to squaring each deviation.

****One Way F-test(Anova) :-****It tell whether two or more groups are similar or not based on their mean similarity and f-score.

****Two Way F-test :-**** Two way F-test is extension of 1-way f-test, it is used when we have 2 independent variable and 2+ groups. 2-way F-test does not tell which variable is dominant. if we need to check individual significance then ****Post-hoc**** testing need to be performed.

**Thursday:**

# ****Import libraries****

# In [1]:

*# import package*

**%**matplotlib inline

**import** pandas **as** pd

**import** numpy **as** np

**import** statistics

**import** seaborn **as** sns

**import** matplotlib.pyplot **as** plt

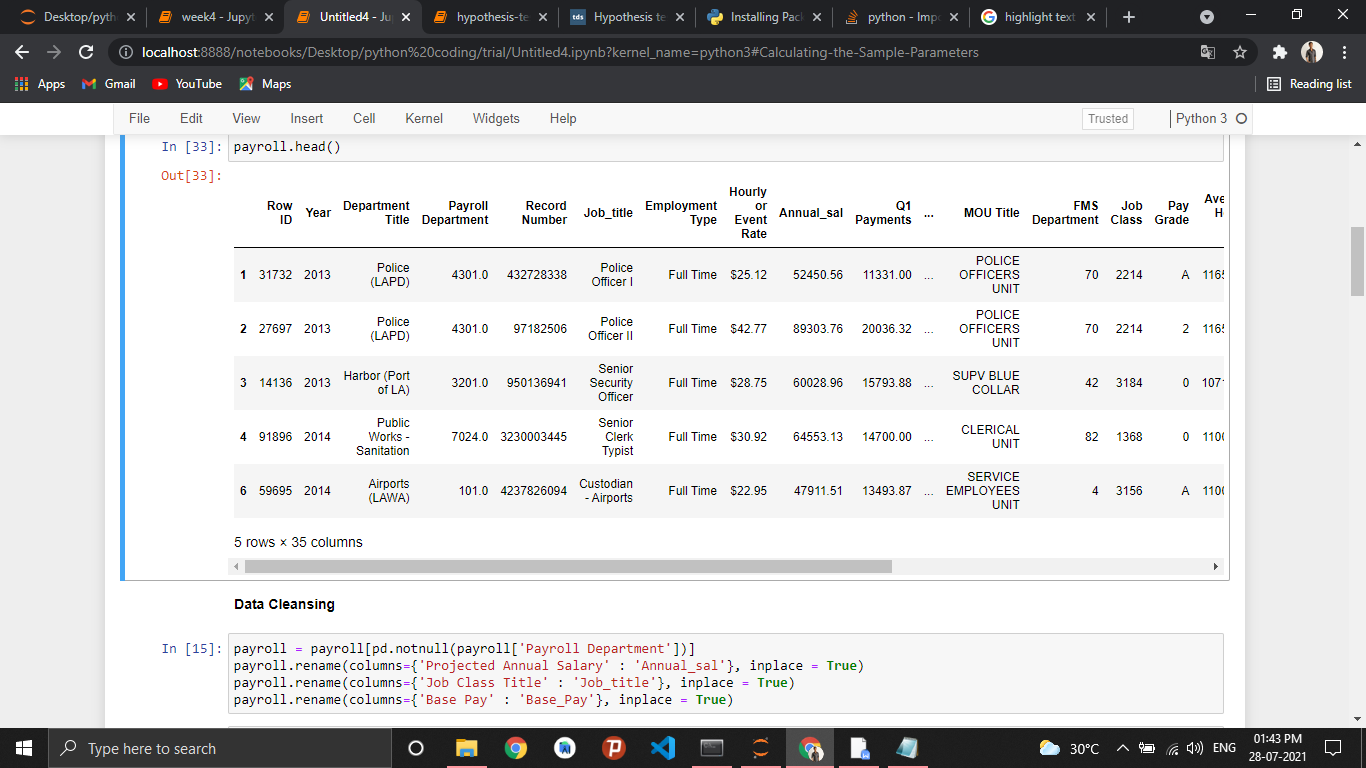
**from** scipy **import** stats

**Import Dataset:**

payroll = pd.read\_csv('data.csv')

**View top 5 dataset:**

data.head()



****Data Cleansing****

In [15]:

payroll **=** payroll[pd.notnull(payroll['Payroll Department'])]

payroll.rename(columns**=**{'Projected Annual Salary' : 'Annual\_sal'}, inplace **=** **True**)

payroll.rename(columns**=**{'Job Class Title' : 'Job\_title'}, inplace **=** **True**)

payroll.rename(columns**=**{'Base Pay' : 'Base\_Pay'}, inplace **=** **True**)

In [17]:

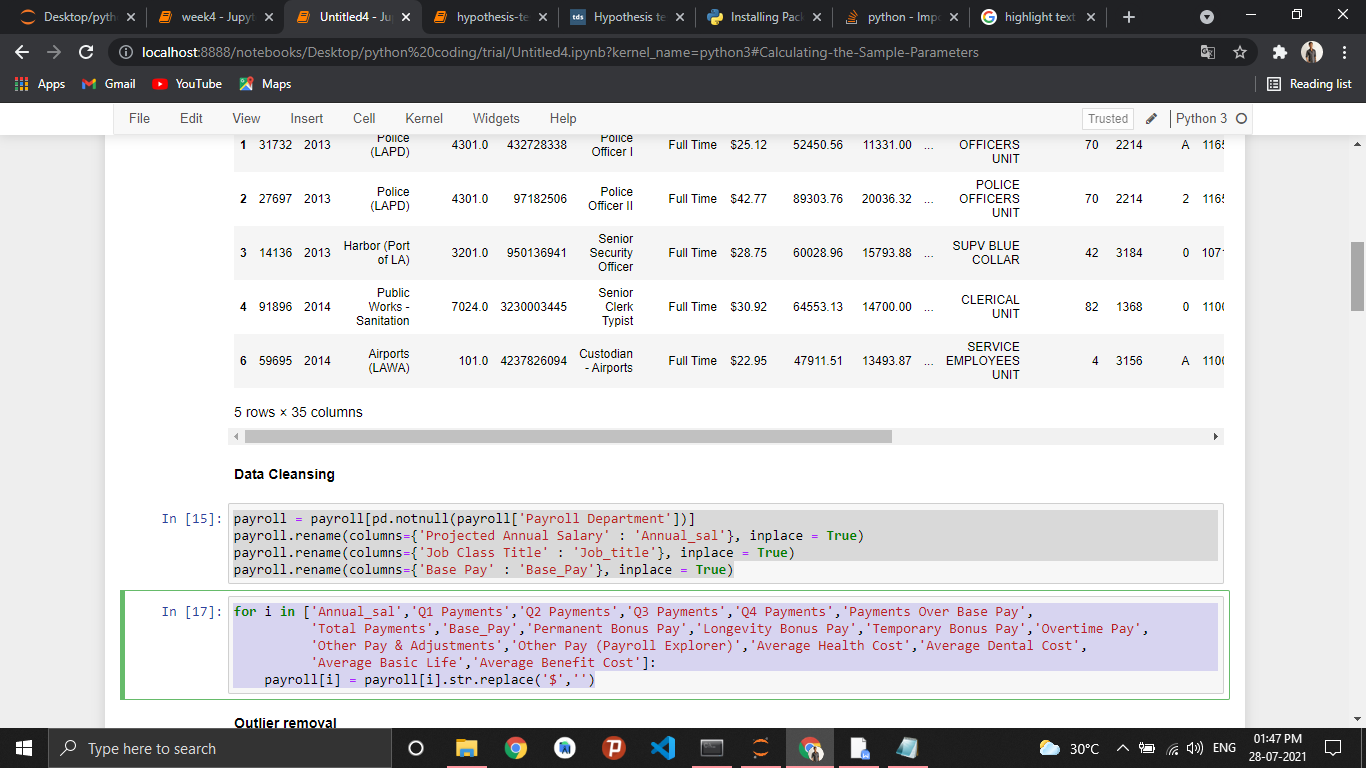
**for** i **in** ['Annual\_sal','Q1 Payments','Q2 Payments','Q3 Payments','Q4 Payments','Payments Over Base Pay',

'Total Payments','Base\_Pay','Permanent Bonus Pay','Longevity Bonus Pay','Temporary Bonus Pay','Overtime Pay',

'Other Pay & Adjustments','Other Pay (Payroll Explorer)','Average Health Cost','Average Dental Cost',

'Average Basic Life','Average Benefit Cost']:

payroll[i] **=** payroll[i].str.replace('$','')



****Outlier removal****

In [18]:

payroll **=** payroll[payroll.Annual\_sal **!=** 0]

payroll **=** payroll[payroll.Base\_Pay **!=** 0]

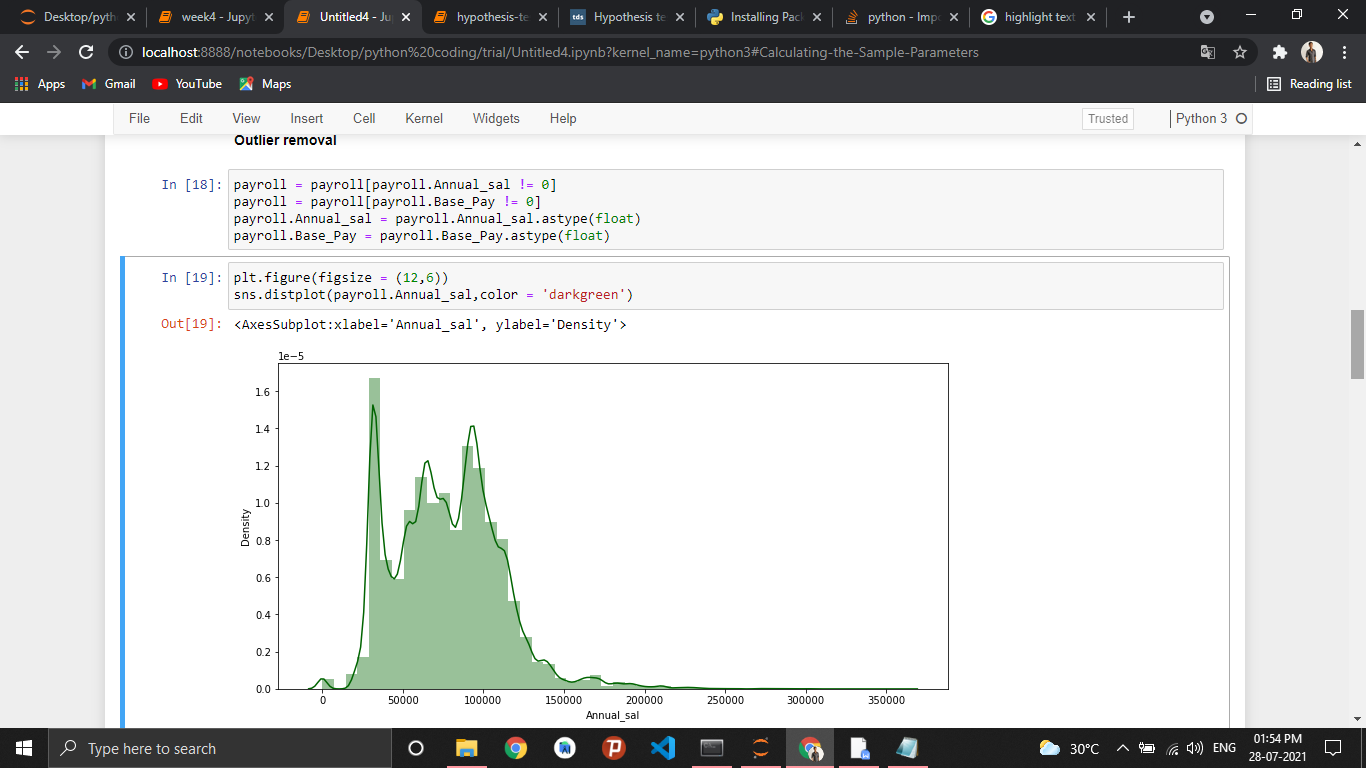
payroll.Annual\_sal **=** payroll.Annual\_sal.astype(float)

payroll.Base\_Pay **=** payroll.Base\_Pay.astype(float)

In [19]:

plt.figure(figsize **=** (12,6))

sns.distplot(payroll.Annual\_sal,color **=** 'darkgreen')



**Friday:**

****Creating different Sample from the population****

In [20]:

payroll\_2015 **=** payroll[payroll.Year **==**2015]

payroll\_2016 **=** payroll[payroll.Year **==**2016]

****Calculating the Population Parameters****

In [21]:

pop\_mean\_2015 **=** payroll\_2015['Annual\_sal'].mean()

pop\_std\_2015 **=** statistics.stdev(payroll\_2015.Annual\_sal)

print("Population Mean: "**+**str(pop\_mean\_2015))

print("Population Standard Deviation: "**+**str(pop\_std\_2015))

Population Mean: 76046.39152505301

Population Standard Deviation: 34232.25489729045

****Calculating the Sample Parameters****

In [23]:

print("Population Mean: "**+**str(payroll\_2016['Annual\_sal'].mean()))

payroll\_2016\_sample **=** payroll\_2016.sample(frac**=**0.10)

sample\_mean\_2016 **=** payroll\_2016\_sample['Annual\_sal'].mean()

print("Sample Mean: "**+**str(sample\_mean\_2016))

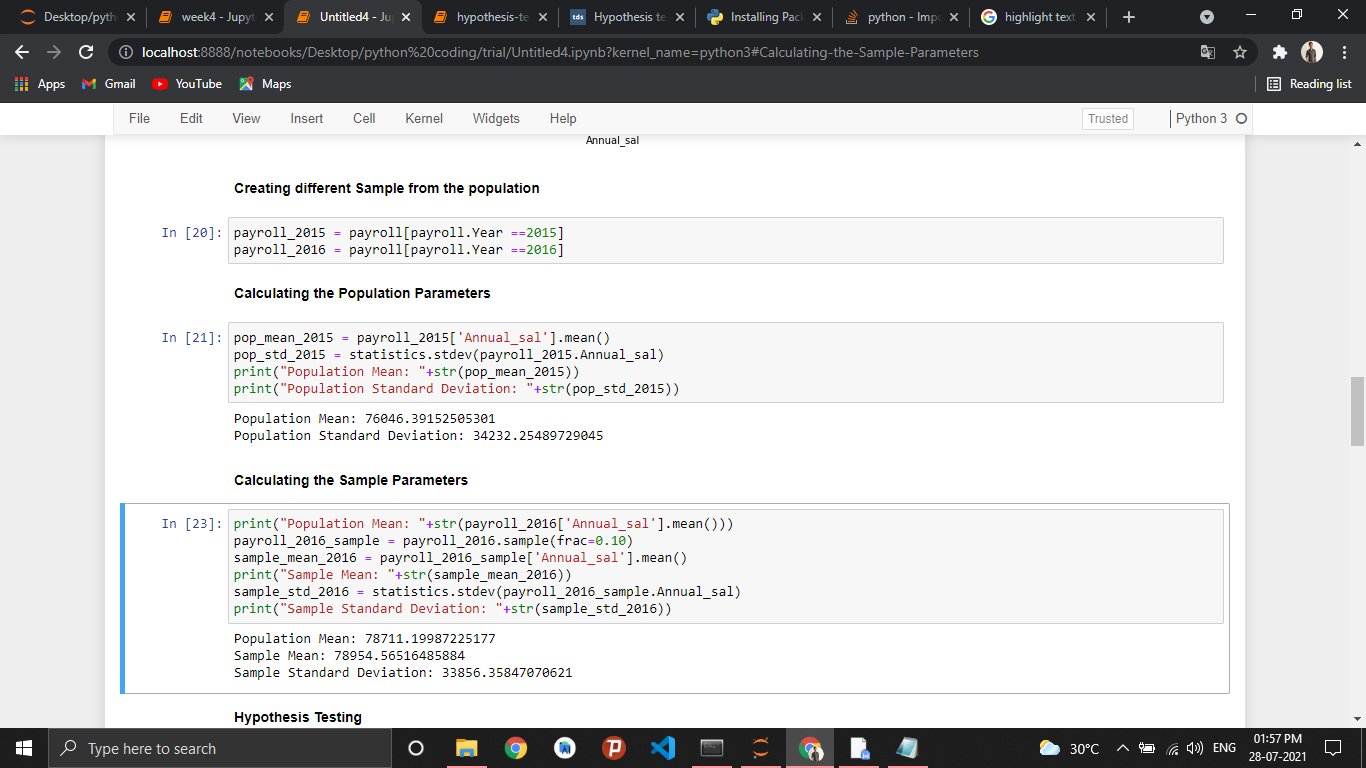
sample\_std\_2016 **=** statistics.stdev(payroll\_2016\_sample.Annual\_sal)

print("Sample Standard Deviation: "**+**str(sample\_std\_2016))

Population Mean: 78711.19987225177

Sample Mean: 78954.56516485884

Sample Standard Deviation: 33856.35847070621



**Hypothesis Testing**

**Normal distribution**

**Null Hypothesis :** It increases in 2016

**Alternate Hypothesis:**Pay does not increase in 2016

**Population parameters:** Mean = pop\_mean\_2015, standard deviation = pop\_std\_2015

**Sample parameters :** Mean = sample\_mean\_2016, standard deviation = sample\_std\_2016

In [25]:

**import** math

*# Confidence Level 95 % for one sided Normal curve*

zscore\_critical **=** 1.65

*# Calculate the test statistics*

zscore\_test\_stat **=** ((sample\_mean\_2016 **-** pop\_mean\_2015)**\***math.sqrt(8916))**/**sample\_std\_2016

print(zscore\_test\_stat)

8.110826644222016

**Conclusion:** As the test statistics fall into the rejection region the null hypothesis is rejected and it can be concluded that the Annual salaries increase in 2016. However, we will first check for the Type 1 and Type 2 Errors

# **Type 1 type 2 hypothesis error**

**Type I error:** When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha. In hypothesis testing, the normal curve that shows the critical region is called the alpha region

**Type II errors:** When we accept the null hypothesis but it is false. Type II errors are denoted by beta. In Hypothesis testing, the normal curve that shows the acceptance region is called the beta region.

In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (a "false positive"), while a type II error is incorrectly retaining a false null hypothesis (a "false negative").

In [26]:

*# we are basically checking the true value of the population characteristics*

pop\_mean\_2016 **=** payroll\_2016['Annual\_sal'].mean()

pop\_std\_2016 **=** statistics.stdev(payroll\_2016.Annual\_sal)

​

zscore\_error **=** ((pop\_mean\_2016 **-** pop\_mean\_2015)**/**pop\_std\_2016)

print(zscore\_error)

0.07937359898019851

We can clearly see that we encountered a Type 1 error as the population mean is well within acceptable region

**Student T-distribution**

**Null Hypothesis:** pay increase in 2016

**Alternate Hypothesis:**it did not increases in 2016

**Population parameters:** Mean = pop\_mean\_2015, standard deviation = pop\_std\_2015

**Sample parameters :** Mean = sample\_mean\_2016, standard deviation = sample\_std\_2016

In [28]:

*#Calculating the Sample Parameters\*\**

payroll\_2014 **=** payroll[payroll.Year **==**2014]

payroll\_2015 **=** payroll[payroll.Year **==**2015]

*# Creating Sample distribution for T statistics*

payroll\_t\_2015\_sample **=** payroll\_2015.sample(frac**=**0.00062)

In [29]:

payroll\_t\_2015\_sample **=** payroll\_2015.sample(frac**=**0.00062)

N **=** len(payroll\_t\_2015\_sample)

sample\_mean\_2015 **=** payroll\_t\_2015\_sample['Annual\_sal'].mean()

sample\_std\_2015 **=** statistics.stdev(payroll\_t\_2015\_sample.Annual\_sal)

pop\_std\_2014 **=** statistics.stdev(payroll\_2014.Annual\_sal)

pop\_mean\_2014 **=** payroll\_2014['Annual\_sal'].mean()

In [30]:

*# Confidence Level 95 % for one sided T curve*

t\_critical **=** 1.311

​

*# Calculate the test statistics*

tscore\_test\_stat **=** ((sample\_mean\_2015 **-** pop\_mean\_2014)**\***math.sqrt(N))**/**sample\_std\_2015

​

print(tscore\_test\_stat)

-0.06225377791500011

Depending on the Test score , we can accept/Reject the Null

**\*\* F Distribution and ANOVA**

The means from three different samples are compared using ANOVA It is similar to applying t-tests over multiple sample ANOVA : [https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php](https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php" \t "http://localhost:8888/notebooks/Desktop/python%20coding/trial/_blank)

In [31]:

payroll\_2014 **=** payroll[payroll.Year **==**2014]

payroll\_2014\_elec **=** payroll\_2014[payroll\_2014.Job\_title **==** 'Electrician']

payroll\_2015\_elec **=** payroll\_2015[payroll\_2015.Job\_title **==** 'Electrician']

payroll\_2016\_elec **=** payroll\_2016[payroll\_2016.Job\_title **==** 'Electrician']

sample\_elec\_2014 **=** payroll\_2014\_elec.sample(frac**=**0.47)

sample\_elec\_mean\_2014 **=** sample\_elec\_2014['Base\_Pay'].mean()

print("Sample Mean 2014 "**+**str(sample\_elec\_mean\_2014))

sample\_elec\_2015 **=** payroll\_2015\_elec.sample(frac**=**0.41)

sample\_elec\_mean\_2015 **=** sample\_elec\_2015['Base\_Pay'].mean()

print("Sample Mean 2015 "**+**str(sample\_elec\_mean\_2015))

sample\_elec\_2016 **=** payroll\_2016\_elec.sample(frac**=**0.22)

sample\_elec\_mean\_2016 **=** sample\_elec\_2016['Base\_Pay'].mean()

print("Sample Mean 2016 "**+**str(sample\_elec\_mean\_2016))

*# Creating the Samples of the base pays over three years*

sam\_1 **=** sample\_elec\_2014.Base\_Pay

sam\_2 **=** sample\_elec\_2015.Base\_Pay

sam\_3 **=** sample\_elec\_2016.Base\_Pay

Sample Mean 2014 67878.53171428571

Sample Mean 2015 65304.96400000001

Sample Mean 2016 34002.21742857143

In [32]:

f, p **=** stats.f\_oneway(sam\_1, sam\_2, sam\_3 )

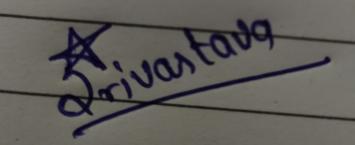
print ('F value:', f)

print ('P value:', p, '\n')

F value: 24.629006946921248

P value: 1.8748679645638236e-09

P value is very low, hence null hypothesis is rejected

Student Signature  Date 28/07/2021 Head- coordinator Signature \_Date

Instructions: After the completed report has been signed by both the student and Head-coordinator, the head-coordinator shall scan the form to a pdf format and email it to the Director-1 ([bpmishra435@gmail.com](mailto:bpmishra435@gmail.com)) of the company. Specific problems, concerns or suggestions from either the student/ head-coordinator should be emailed separately to the C.E.O.([info@cureya.in](mailto:info@cureya.in)) of the company.