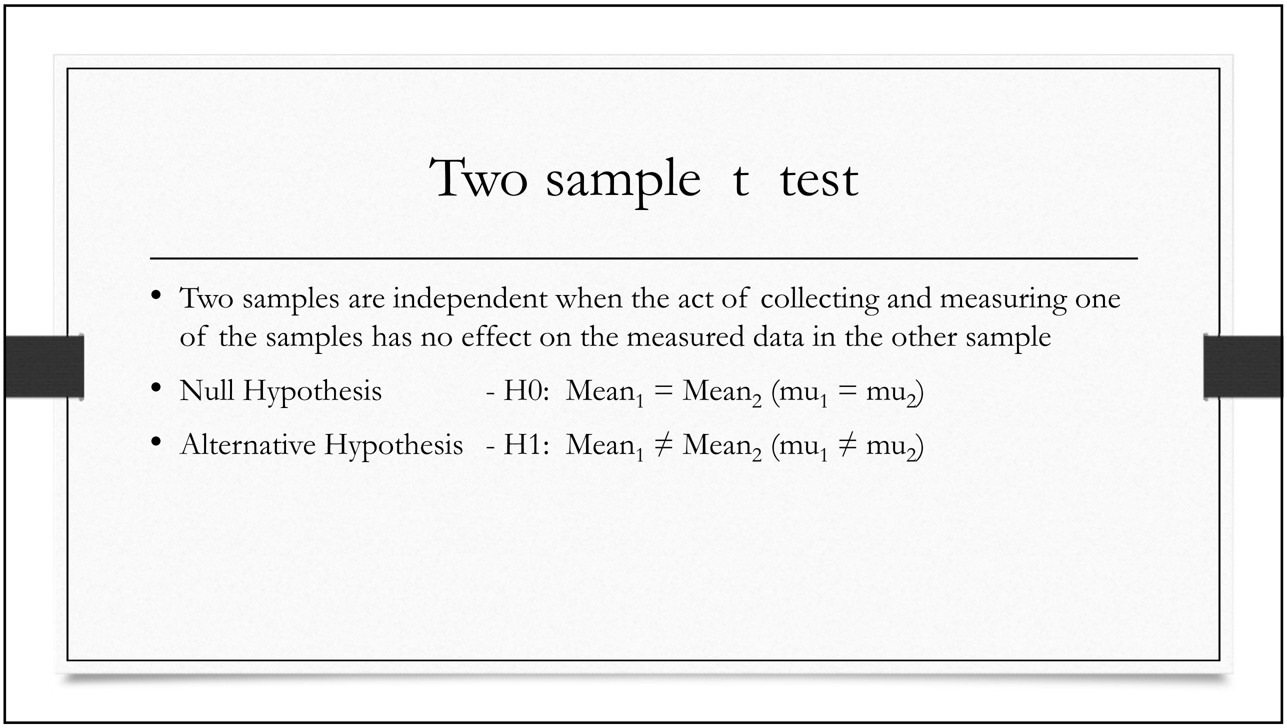
Test of hypothesis

Mohan M J

Hypothesis Testing

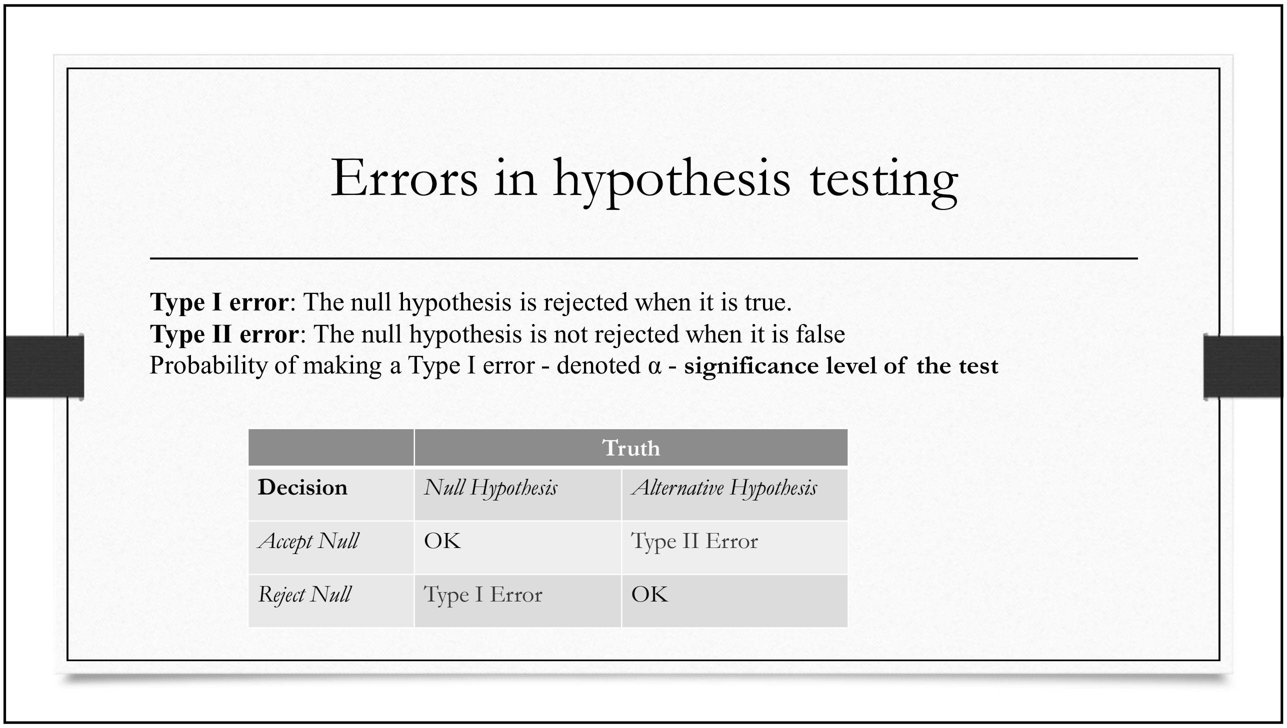
The general idea of hypothesis testing involves:

- Making an initial assumption
- Collecting evidence (data)
- Based on the available evidence (data), deciding whether to reject or not reject the initial assumption



Two sample t test

- Two samples are independent when the act of collecting and measuring one of the samples has no effect on the measured data in the other sample
- Null Hypothesis - H0: $\text{Mean}_1 = \text{Mean}_2$ ($\text{mu}_1 = \text{mu}_2$)
- Alternative Hypothesis - H1: $\text{Mean}_1 \neq \text{Mean}_2$ ($\text{mu}_1 \neq \text{mu}_2$)



Errors in hypothesis testing

Type I error: The null hypothesis is rejected when it is true.
Type II error: The null hypothesis is not rejected when it is false
Probability of making a Type I error - denoted α - **significance level of the test**

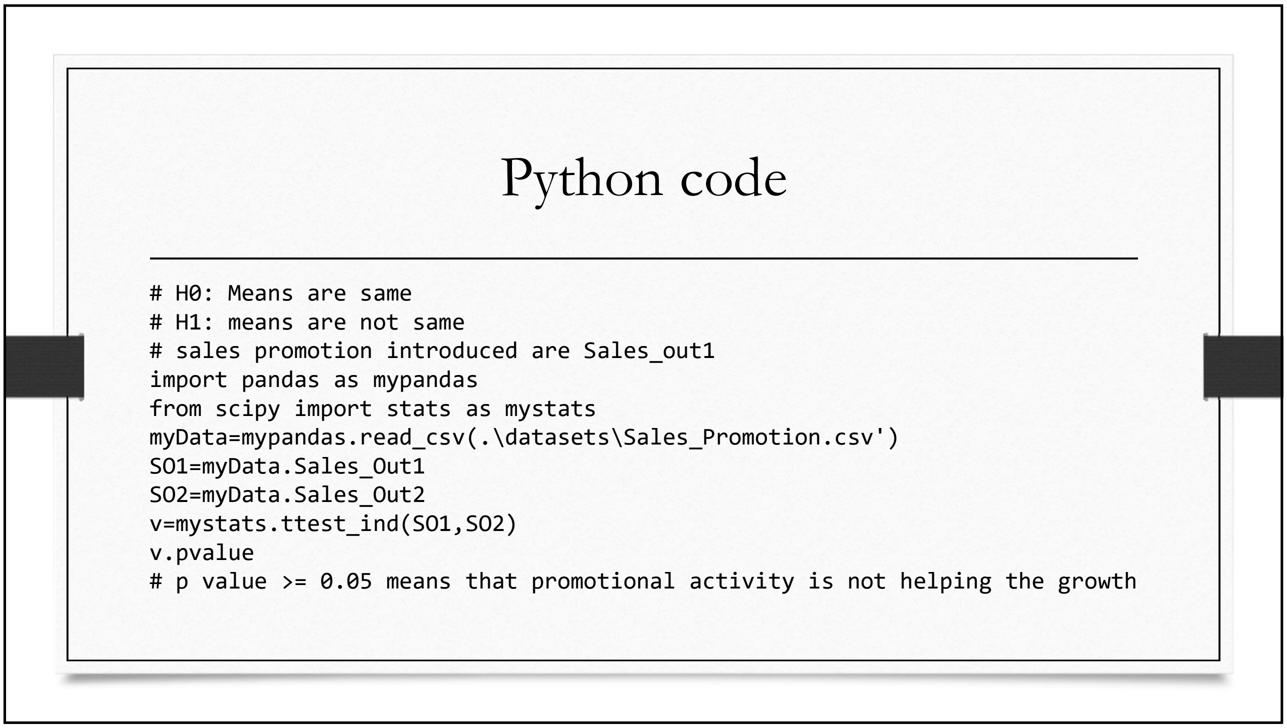
	Truth	
Decision	*Null Hypothesis*	*Alternative Hypothesis*
Accept Null	OK	Type II Error
Reject Null	Type I Error	OK

Methodology

- To test the two means are equal
- Calculate both the sample means xbar1 & xbar2
- Calculate SD1 & SD2
- Calculate test statistic t_0
- Test Statistc , t_0 = (xbar – xbar2)/[Sp/√{(1/n1)+(1/n2)}]
- Calculate p from t distribution
- If $p < 0.05$ then H0: $\text{Mean}_1 = \text{Mean}_2$ is rejected

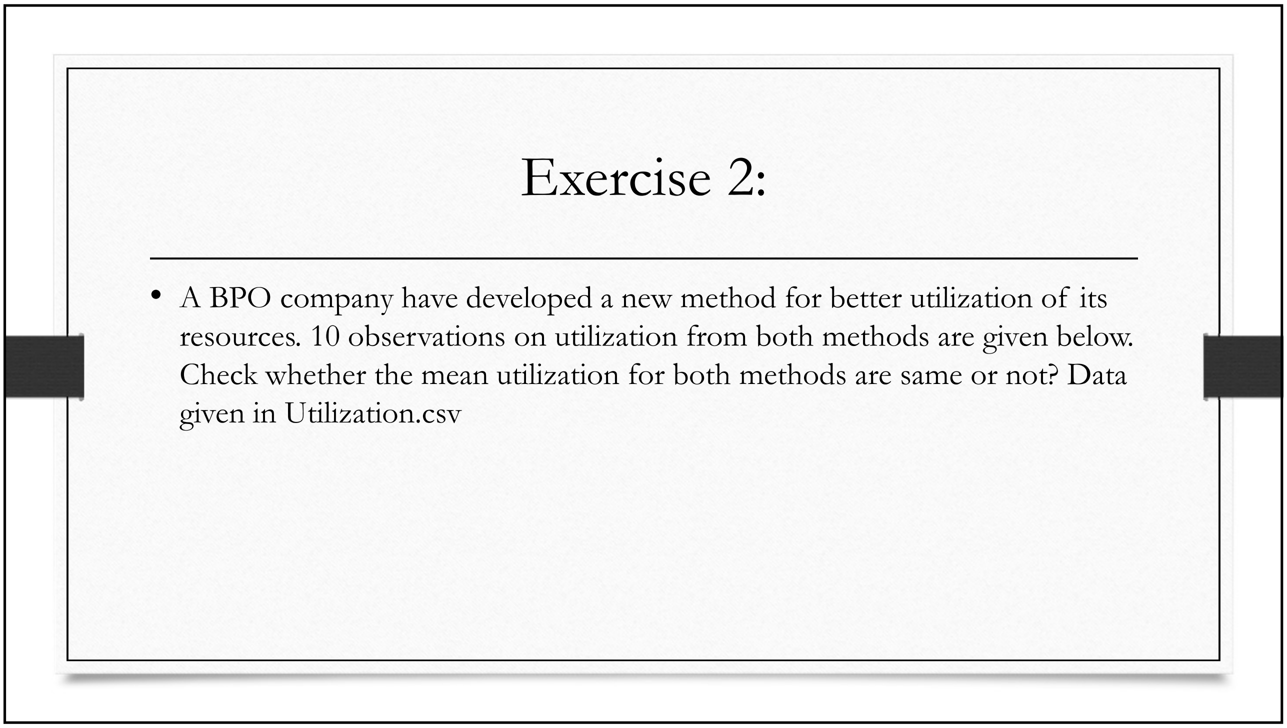
Exercise 1:

- A super market chain has introduced a promotional activity in its selected outlets in the city to increase the sales volume. Based on the data given below, check whether the promotional activity resulted in increasing the sales?
  - The outlets where promotional activity introduced are denoted by 1 and the others by 2
  - The data is given in Sales_Promotion.csv



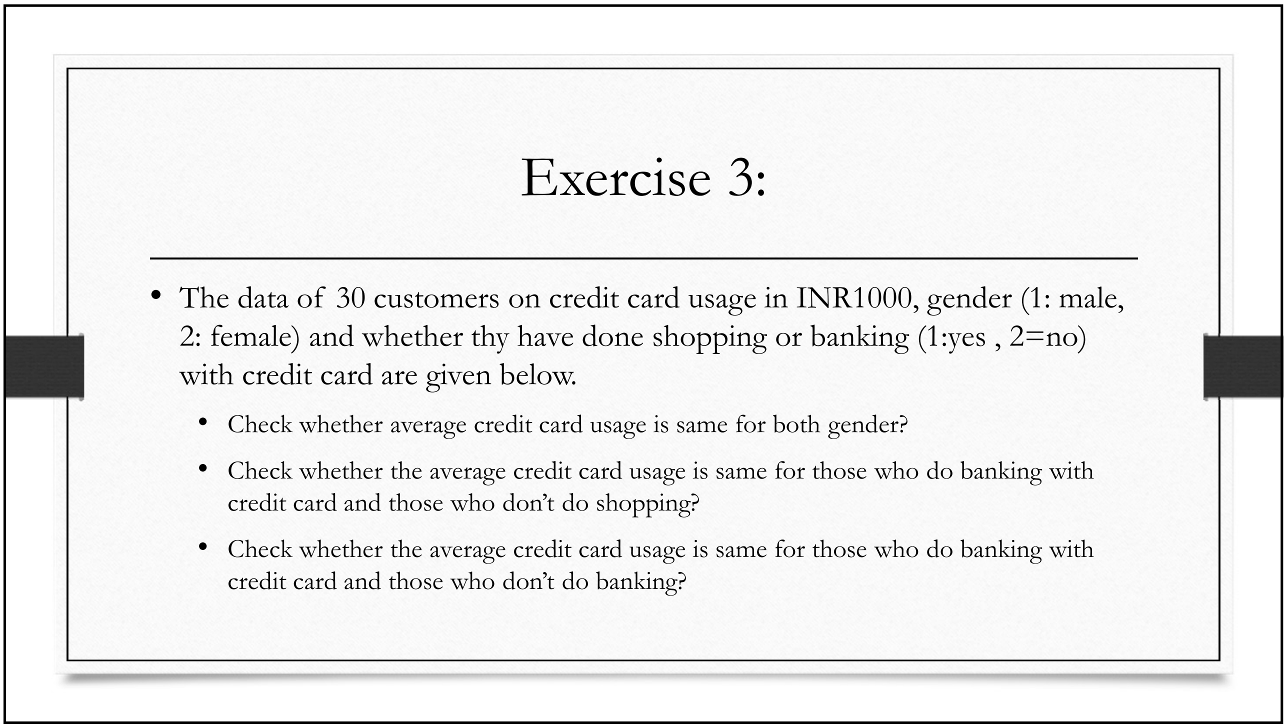
Python code

```
# H0: Means are same
# H1: means are not same
# sales promotion introduced are Sales_out1
import pandas as mypandas
from scipy import stats as mystats
myData=mypandas.read_csv(.\datasets\Sales_Promotion.csv')
SO1=myData.Sales_Out1
SO2=myData.Sales_Out2
v=mystats.ttest_ind(SO1,SO2)
v.pvalue
# p value >= 0.05 means that promotional activity is not helping the growth
```



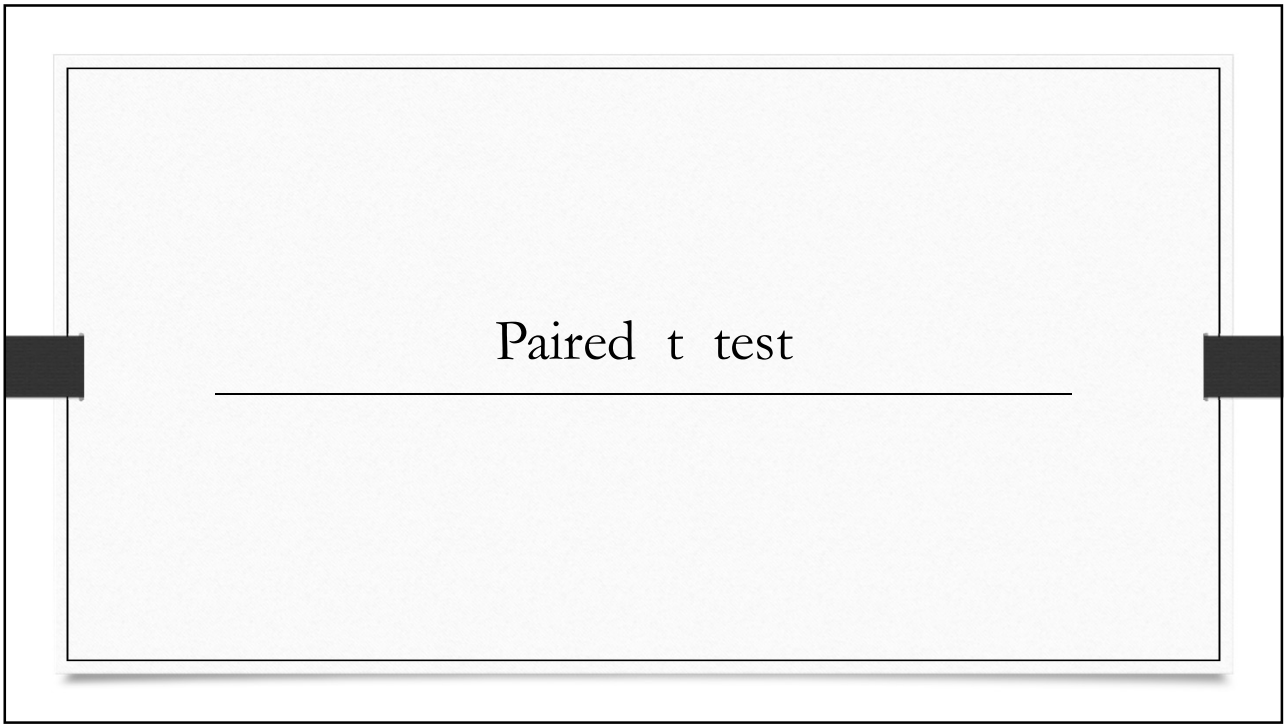
Exercise 2:

- A BPO company have developed a new method for better utilization of its resources. 10 observations on utilization from both methods are given below. Check whether the mean utilization for both methods are same or not? Data given in Utilization.csv



Exercise 3:

- The data of 30 customers on credit card usage in INR1000, gender (1: male, 2: female) and whether thy have done shopping or banking (1:yes , 2=no) with credit card are given below.
  - Check whether average credit card usage is same for both gender?
  - Check whether the average credit card usage is same for those who do banking with credit card and those who don’t do shopping?
  - Check whether the average credit card usage is same for those who do banking with credit card and those who don’t do banking?



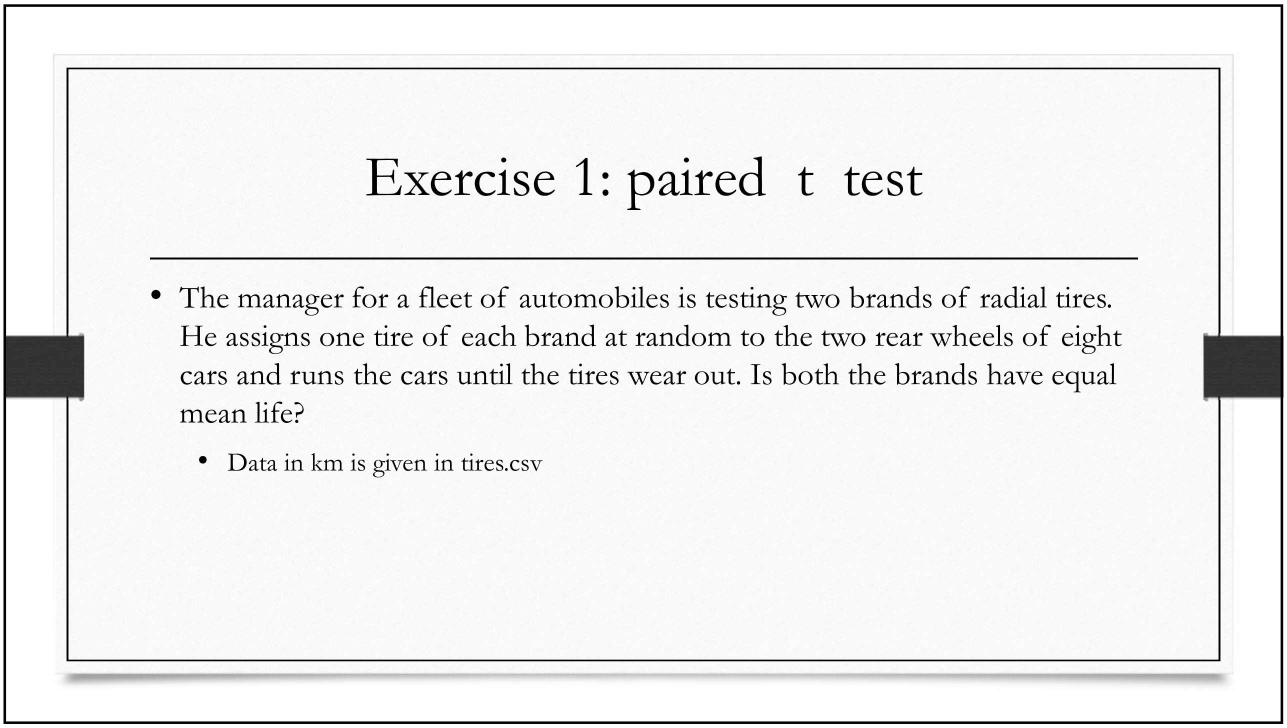
Paired t test

Paired t test

- A special case of two sample t test
- When the observations on two groups are collected in pairs
- Each pair of observation is taken under homogeneous conditions
- When conducting the two sample t test – a key assumption is that the data is independent

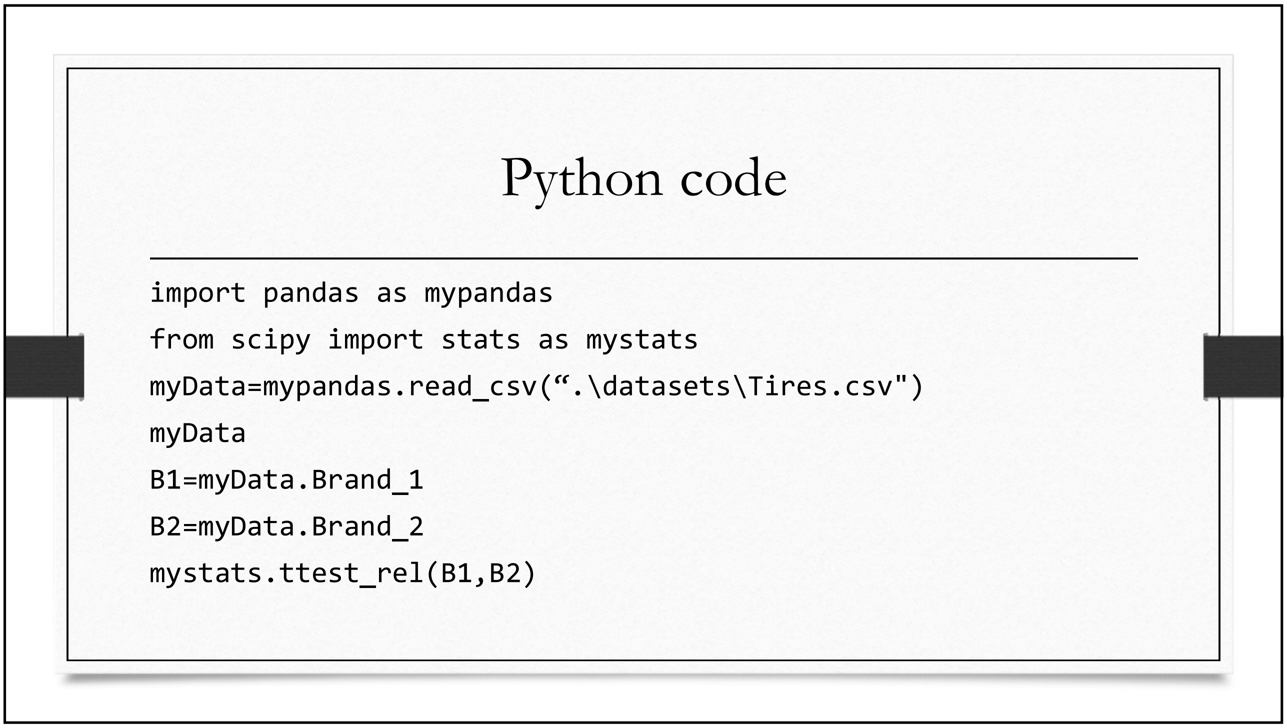
Procedure

- Compute d: difference in paired observations
- Let the difference in mean be $\mu_D = \mu_1 - \mu_2$
- Null Hypothesis : H0: $\mu_D = 0$
- Alternative Hypothesis H1: $\mu_D \neq 0$
- Test statistic, t0 = $D/(SD/\sqrt{n})$
- Reject H0 if p value < 0.05



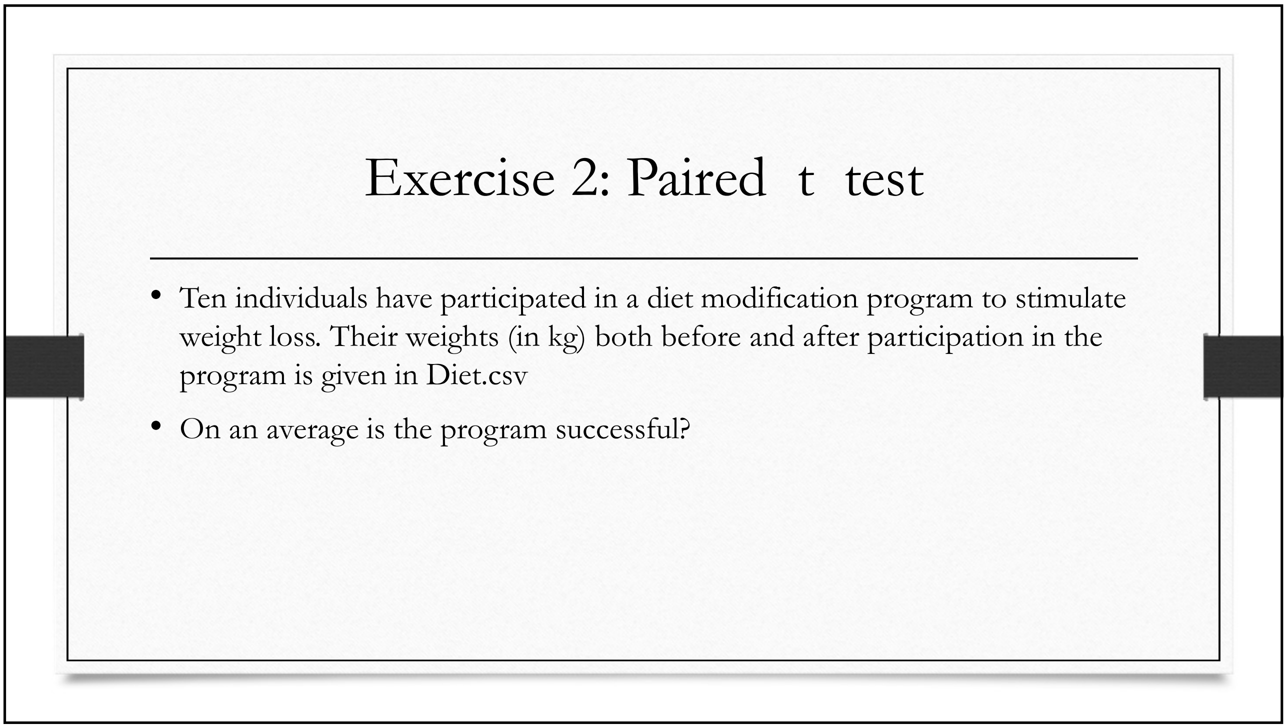
Exercise 1: paired t test

- The manager for a fleet of automobiles is testing two brands of radial tires. He assigns one tire of each brand at random to the two rear wheels of eight cars and runs the cars until the tires wear out. Is both the brands have equal mean life?
  - Data in km is given in tires.csv



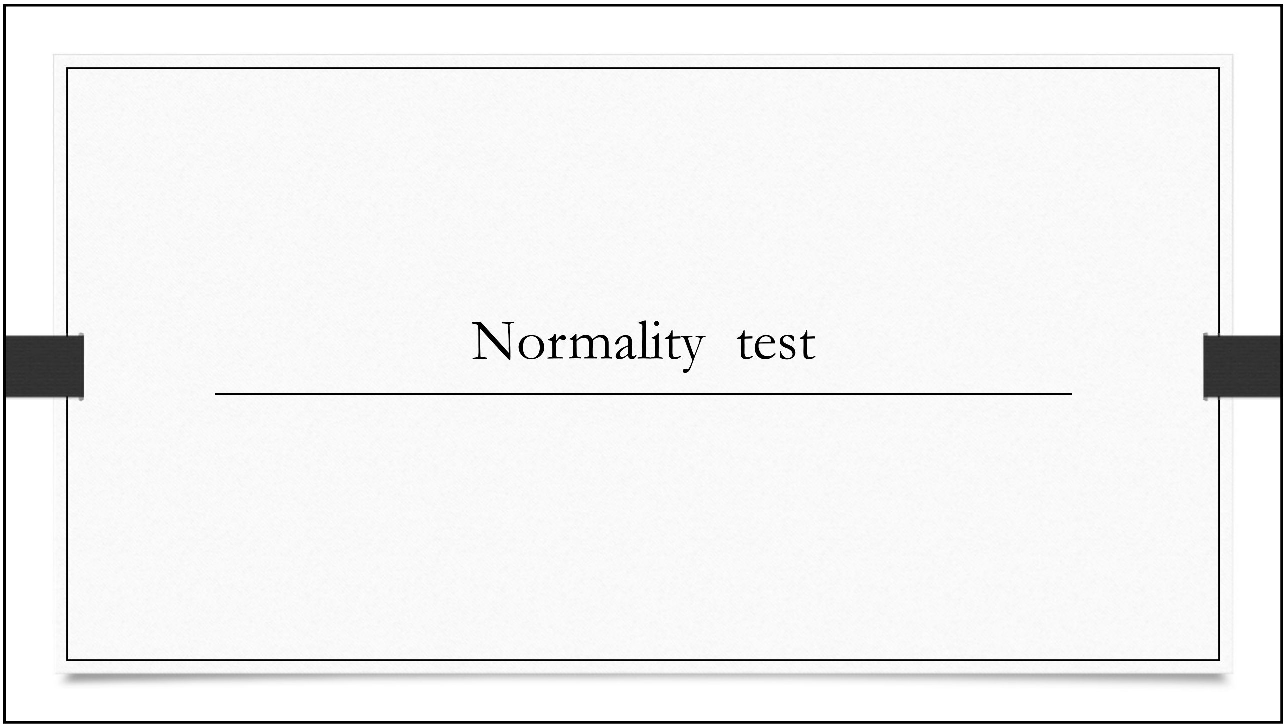
Python code

```
import pandas as mypandas
from scipy import stats as mystats
myData=mypandas.read_csv(“.\datasets\Tires.csv")
myData
B1=myData.Brand_1
B2=myData.Brand_2
mystats.ttest_rel(B1,B2)
```

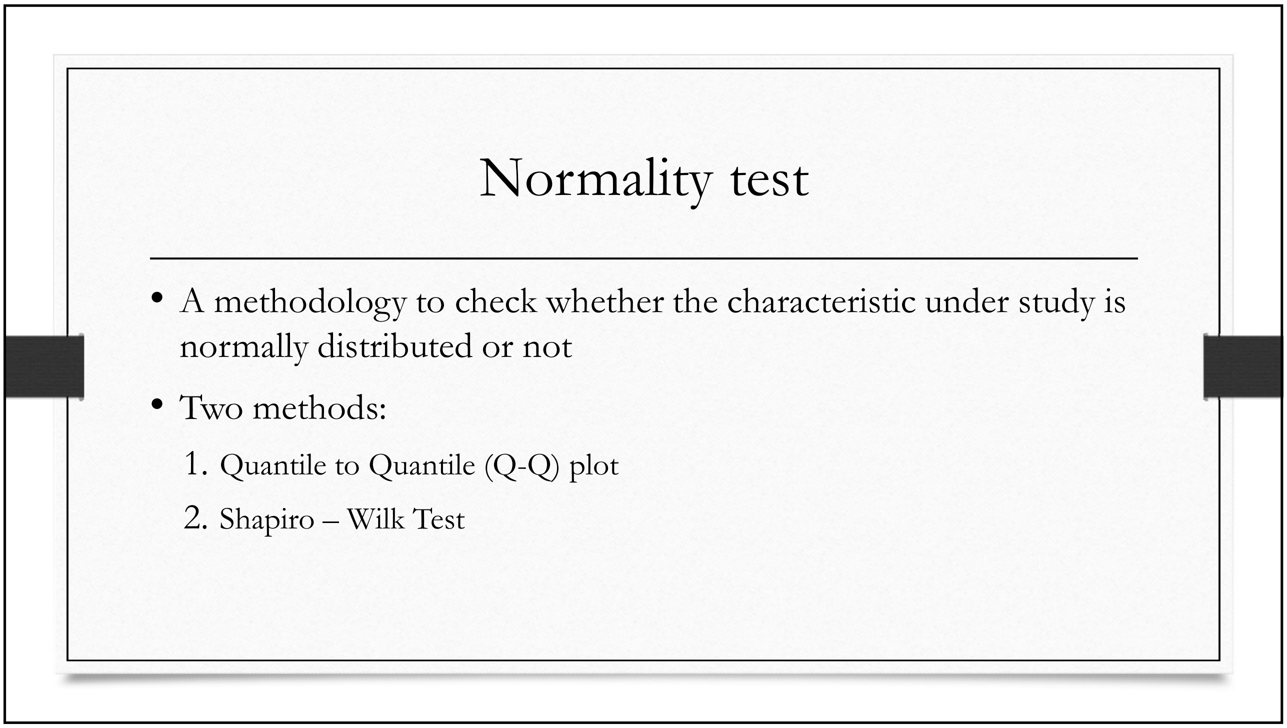


Exercise 2: Paired t test

- Ten individuals have participated in a diet modification program to stimulate weight loss. Their weights (in kg) both before and after participation in the program is given in Diet.csv
- On an average is the program successful?

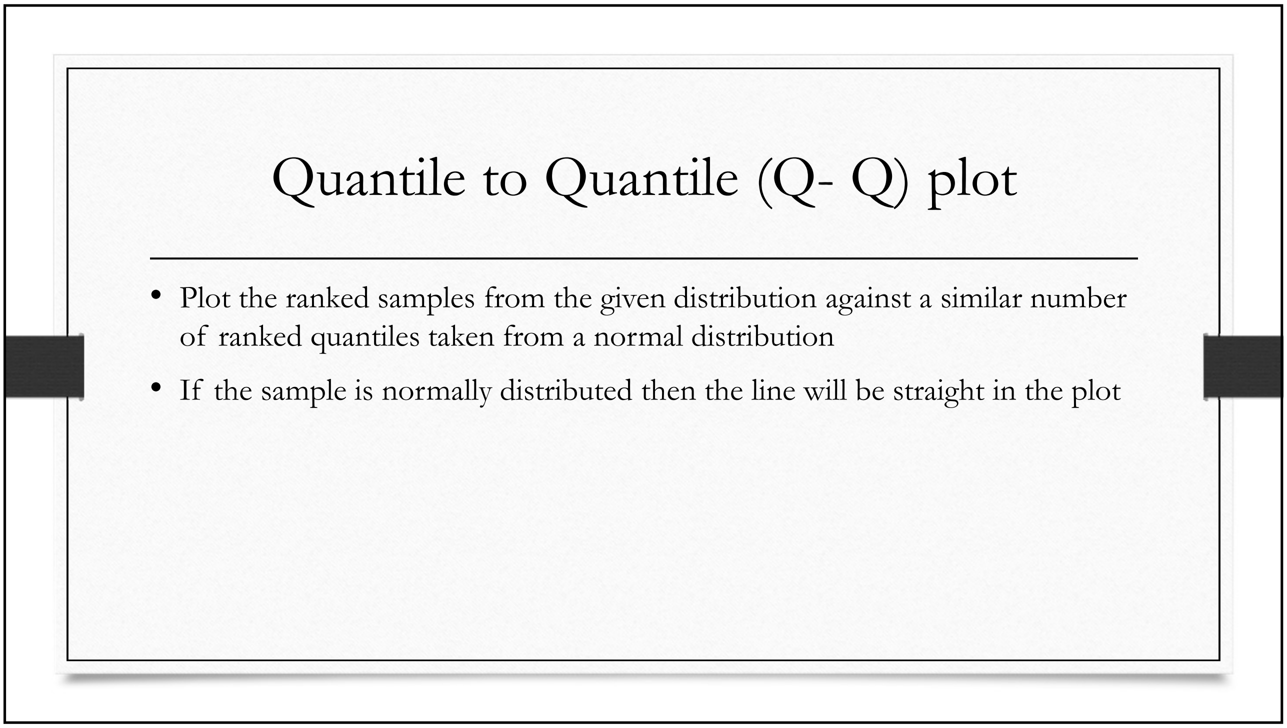


Normality test



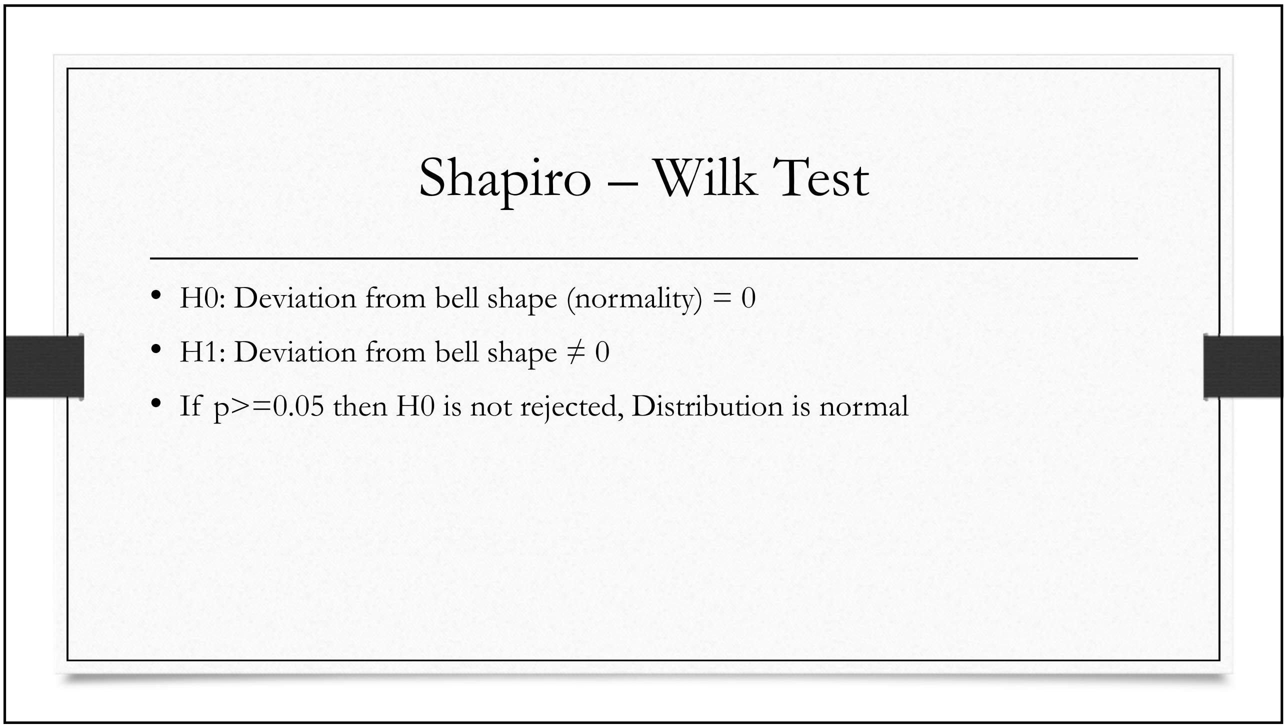
Normality test

- A methodology to check whether the characteristic under study is normally distributed or not
- Two methods:
  1. Quantile to Quantile (Q-Q) plot
  2. Shapiro – Wilk Test



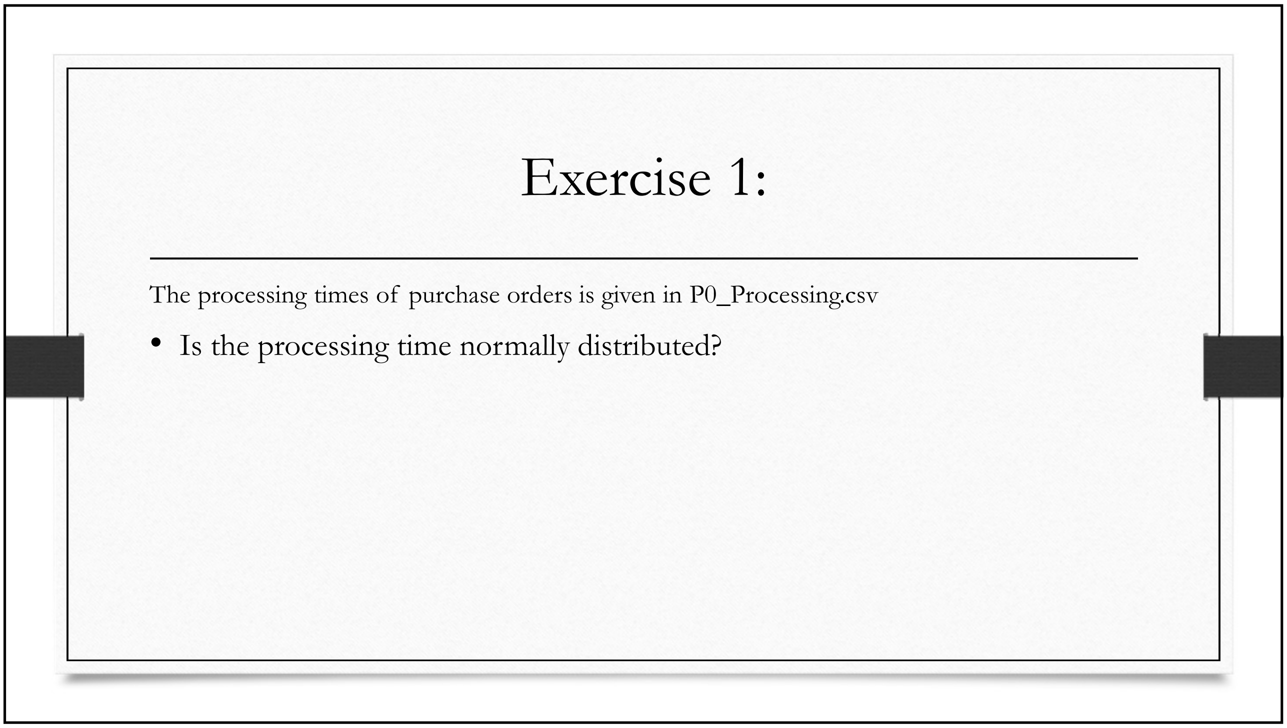
Quantile to Quantile (Q- Q) plot

- Plot the ranked samples from the given distribution against a similar number of ranked quantiles taken from a normal distribution
- If the sample is normally distributed then the line will be straight in the plot



Shapiro – Wilk Test

- H0: Deviation from bell shape (normality) = 0
- H1: Deviation from bell shape ≠ 0
- If p>=0.05 then H0 is not rejected, Distribution is normal



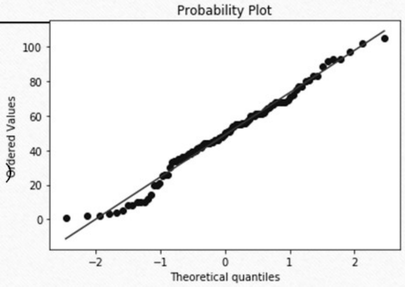
Exercise 1:

The processing times of purchase orders is given in P0_Processing.csv

- Is the processing time normally distributed?

Python code

```
import pandas as mypandas
from scipy import stats as mystats
import matplotlib.pyplot as myplot
myData=mypandas.read_csv(.\datasets\PO_Processing.csv'
myData
PT=myData.Processing_Time
mystats.probplot(PT,plot=myplot)
myplot.show()
mystats.mstats.normaltest(PT)
Out[] NormaltestResult(statistic=0.3396526182225921 8, pvalue=0.8438113662149449)
```



Exercise 2:

The impurity level (in ppm) is routinely measured in an intermediate chemical process. The data is given in Impurity.csv

- Check whether the impurity follows distribution?

THANKS