Statistical Inference

T Tests and Analysis of Variance Using Python

DATA SCIENCE
INSTITUTE

Independent samples t-test

- The independent-samples t-test compares the means of two independent groups on the same continuous variable.

- Following hypotheses are tested in independent samples t test

  - H0: Two population means are equal
  - H1: Two population means are not equal

DATA SCIENCE INSTITUTE

Case Study

To execute Parametric test in Python, we shall consider the below case as an example.

Background

The company is assessing the difference in time to complete MIS report between two groups of employees :
Group I: Experience(0-1 years)
Group II: Experience(1-2 years)

Objective

To test whether the average time taken to complete MIS by both the groups is same.

Sample Size

Sample size: 14
Variables: time_g1, time_g2

Data Snapshot

INDEPENDENT SAMPLES t TEST

Variables

servations

time_g1	time_g2
85	83
95	85
105	96
85	94

Columns	Description	Type	Measurement	Possible values
time_g1	Time to complete MIS report by group1	Numeric	Hours	Positive Values
time_g2	Time to complete MIS report by group2	Numeric	Hours	Positive Values

Independent samples t-test in Python

```
# Import data
```

```
import pandas as pd
data=pd.read_csv('INDEPENDENT SAMPLES t TEST.csv')
```

```
# t-test for independent samples
```

```
from scipy import stats
stats.ttest_ind(data['time_g1'],data['time_g2'],nan_policy='omit’,
,equal_var=True)
```

- ***ttest_ind()*** *from scipy, returns t & pvalue*
- ***nan_policy=‘omit’*** *Defines how to handle when input contains nan. 'propagate' returns nan, 'raise' throws an error, 'omit' performs the calculations ignoring nan values. Default is 'propagate'.*

* Before performing t test, normality test is done to ensure time variable is normally distributed in both the groups.

DATA SCIENCE INSTITUTE

Independent samples t-test in Python

```
# Output:
```

```
Ttest_indResult(statistic=0.22345590920212569,pvalue=0.8250717960964372)
```

Interpretation :

- *Since p-value is >0.05, do not reject H0. There is no significant difference in average time taken to complete the MIS between two group of employees.*

DATA SCIENCE INSTITUTE

Paired samples t-test

- The paired sample t-test is used to determine whether the mean difference between two sets of observations is zero ,where each subject or entity is measured twice resulting in pair of observations.

- Commonly used when observations are recorded ‘before’ and ‘after’ the treatment / training and objective is to test whether the treatment/training is effective.

DATA SCIENCE INSTITUTE

Case Study

To execute Parametric test in Python, we shall consider the below case as an example.

Background

The company organized a training program to improve efficiency. Time taken to complete MIS report before and after training are recorded for 15 employees.

Objective

To test whether the average time taken to complete MIS before and after training is not different.

Sample Size

Sample size: 15
Variables: time_before, time_after

DATA SCIENCE INSTITUTE

Data Snapshot

PAIRED t TEST

Variables

time_before	time_after
85	74
95	91
92	80
102	91

...servations

Columns	Description	Type	Measurement	Possible values
time_before	Time to complete MIS report before training	Numeric	Hours	Positive values
time_after	Time to complete MIS report after training	Numeric	Hours	Positive values

DATA SCIENCE INSTITUTE

Paired sample t-test

Testing whether means of two dependent groups are equal.

Objective	To test the average time taken to complete MIS before and after training is not different.

Null Hypothesis (H_0): There is no difference in average time before and after the training. i.e. D=0

Alternate Hypothesis (H_1):Average time is less after the training. (Training is effective.) D>0

D= μBefore – μAfter

Test Statistic	$t = \frac{\bar{d}}{s_d / \sqrt{n}}$	Where $\bar{d}$ is the sample mean of the difference i.e. before-a s_d , is the sample standard deviation of the difference, n is the sample size of difference. The quantity t follows a distribution called as 't distribution' with n-1 degrees of freedom.
Decision Criteria	Reject the null hypothesis if p-value < 0.05	

DATA SCIENCE INSTITUTE

Paired sample t-test in Python

```
# Import data
data=pd.read_csv('PAIRED t TEST.csv')
```

```
# t-test for paired samples
stats.ttest_rel(data['time_before'],data['time_after']
,alternative='greater')
```

- ***data['time_before']*** *and* ***data['time_after']*** *are the variables under study.*
- ***ttest_rel()*** *from scipy, returns t & pvalue*

* Before performing t test, normality test is done to ensure difference variable is normally distributed.

DATA SCIENCE INSTITUTE

Paired sample t-test in Python

```
# Output:
```

```
Ttest_relResult(statistic=8.22948711672449, pvalue=4.918935850301797e-07)
```

Interpretation :

- *Since p-value is <0.05, reject H0.*

DATA SCIENCE INSTITUTE

t-test for Correlation

- Correlation coefficient summarizes the strength of a linear relationship between two variables.
- t-test is used to check if there is significant correlation between two variables.
- Sample correlation coefficient (r) is calculated using bivariate data.
- Null hypothesis of this test is
  H0: there is no correlation between 2 variables under study ($\rho=0$)

DATA SCIENCE INSTITUTE

Case Study

To execute Parametric test in Python, we shall consider the below case as an example.

Background

A company with 25 employees has calculated job proficiency score & aptitude test score for its employees

Objective

To test if there is significant correlation between job proficiency and aptitude test score.

Sample Size

Sample size: 25
Variables: Empcode, Aptitude, Job_prof

Data Snapshot

Correlation test

Variables

Observations

Empcode	aptitude	job_prof
E101	86	88
E102	62	80
E103	110	96
E104	101	76
E105	100	80
E106	78	73

Columns	Description	Type	Measurement	Possible values
Empcode	Employee code	Numeric	-	
Aptitude	Score of aptitude test	Numeric	-	Positive values
Job_prof	Job proficiency score	Numeric	-	Positive values

DATA SCIENCE INSTITUTE

Correlation t-test

Testing for correlation coefficient value.

Objective	To test whether there exists significant correlation between job proficiency and aptitude score.

Null Hypothesis (H_0): There is no correlation between Job proficiency and Aptitude test

Alternate Hypothesis (H_1):There is correlation between Job proficiency and Aptitude test.

Test Statistic	$t = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$	where r is the sample correlation coefficient, n is the sample size. The quantity t follows a distribution called as 't distribution' with n-2 degrees of freedom.
Decision Criteria	Reject the null hypothesis if p-value < 0.05	

DATA SCIENCE INSTITUTE

Computation

	Notation	Value
Sample Size	n	25
Sample correlation coefficient	r	0.514411
t	$t = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$	2.8769

Correlation t-test in Python

```
# Import data
data=pd.read_csv('Correlation test.csv')

# t-test for correlation
stats.pearsonr(data['aptitude'], data['job_prof'])
```

- ***data['aptitude']*** *and* ***data['job_prof ']*** *are the variables under study.*
- ***pearsonr()*** *from scipy, returns t & pvalue*

Correlation t-test in Python

```
# Output:
```

```
(0.5144106946654772, 0.008517216152487137)
```

Interpretation :

- *Since p-value is <0.05, reject H0. There is statistically significant correlation between aptitude test and job proficiency.*

ANALYSIS OF VARIANCE

- Note that although the name is 'Analysis of Variance', the method is used to analyze the differences among group means.

- Variation in the variable is inherent in nature. In general, the observed variance in a particular variable is partitioned into components attributable to different sources of variation.

- The total variance in any variable is due to a number of causes which may be classified “assignable causes (which can be detected and measured)” and “chance causes (which is beyond control of human and cannot be traced separately)”.

- Hence, ANOVA is the separation of variance ascribable to one group of causes from the variance ascribable to other group.

Case Study

To execute analysis of Variance in Python, we shall consider the below case as an example.

Background

A large company is assessing the difference in 'Satisfaction Index' of employees in Finance, Marketing and Client-Servicing departments.

Objective

To test whether **mean satisfaction index** for employees in three departments (CS, Marketing, Finance) are equal.

Sample Size

Sample size: 37
Variables: satindex, dept

Data Snapshot

One way anova

Variables

Observations

satindex	dept
75	FINANCE
56	FINANCE
72	FINANCE
59	FINANCE
66	FINANCE
58	FINANCE
58	MARKETING
63	MARKETING

Columns	Description	Type	Measurement	Possible values
satindex	Satisfaction Index	Numeric		Positive Values
dept	Department	Character	MARKETING, CS, FINANCE	3

One Way ANOVA

Testing equality of means in one factor with more than two levels.

Objective	To test whether **mean satisfaction index** for employees in three departments (CS, Marketing, Finance) are equal.

Null Hypothesis (H_0): **Mean satisfaction index for 3 departments are equal** i.e. $\mu 1 = \mu 2 = \mu 3$
Alternate Hypothesis (H_1): Mean satisfaction index for 3 departments are not equal

Test Statistic	The test statistic is denoted as F and is based on F distribution.
Decision Criteria	Reject the null hypothesis if $p\text{-value} < 0.05$

DATA SCIENCE INSTITUTE

Calculation

Total SS = (75-65.59)^2+(56-65.59)^2+............+(65-65.59)^2+(76-65.59)^2
= 1840.92

Between Groups SS = 12*(64.42-65.59)^2+12*(63.25-65.59)^2+13*(68.85-65.59)^2
= 220.0599

Within Groups SS = Total SS – Between SS

Overall Mean	65.59	n=37
Mean for Finance	64.42	n1=12
Mean for Marketing	63.25	n2=12
Mean for CS	68.85	n3=13

DATA SCIENCE INSTITUTE

One Way ANOVA table

Sources of variation	Degrees of freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MS=SS/df)	F-Value
Between groups	K-1=3-1 =2	SSA=220.0599	MSA=110.03	F=2.3080
Within groups (error)	n-k=37-3 =34	SSE=1620.86	MSE=47.6724	
TOTAL	n-1=37-1 =36	TSS=1840.92		

DATA SCIENCE INSTITUTE

One Way ANOVA in Python

```
# Import data
import pandas as pd
data = pd.read_csv('One way anova.csv')
```

```
# ANOVA table
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols('satindex ~ C(dept)', data=data).fit()
aov_table = sm.stats.anova_lm(model, typ=2)
aov_table
```

```
# Output:
```

	sum_sq	df	F	PR(>F)
C(dept)	220.059945	2.0	2.308047	0.114836
Residual	1620.858974	34.0	NaN	NaN

- ***ols()*** *from statsmodels.formula.api is used to fit the model*
- *Independent variable to be specified as* ***C()***
- ***sm.stats.anova_lm()*** *from statsmodel.api is used to get ANOVA table*
- ***typ*** *= determines how the sum of squares is calculated &* ***typ = 2*** *if there is no significant interaction effect*

Interpretation :

- *Since p-value is >0.05, do not reject H0. There is no significant difference in satisfaction index among 3 different departments.*

Two Way ANOVA

- Two Way Anova is used when there are 2 factors under study.
- Each factor can have 2 or more levels . Example: Gender and Age can be 2 factors.
  Gender with 2 levels as Male and Female
  Age with 3 levels as 18-30,31-50 and >50
- Three hypothesis are tested.

Factor A H0: All group means are equal
H1: At least one mean is different from other means

Factor B H0: All group means are equal
H1: At least one mean is different from other means

Interaction H0: The interaction is not significant
H1: The interaction is significant

* For two-way ANOVA with interaction there has to be more than one observation per combination of the levels of factors.

Two Way ANOVA

- **Total variation is partitioned as below :**

Total SS= Between Groups SS due to factor A (SSA)

+ Between Groups SS due to factor B (SSB)

+ Interaction SS due to factor A and B (SSAB)

+ Error SS (SSE)

where, SS stands for sum of squares

* SS formulae for two-way ANOVA with interaction are not specified due to their complexity.

DATA SCIENCE INSTITUTE

Case Study

We will illustrate Two Way Anova in Python using following case study

Background

A large company is assessing the difference in 'Satisfaction Index' of employees in Finance, Marketing and Client-Servicing departments. Experience level is also considered in the study.(<=5 years and >5 years)

Objective

To test the equality of the satisfaction index among employees of three departments (CS, Marketing, Finance) and among different experience bands.

Sample Size

Sample size: 36
Variables: satindex, dept, exp

Data Snapshot

Two Way Anova

Variables

Observations

satindex	dept	exp
75	FINANCE	lt5
56	FINANCE	lt5
62	FINANCE	gt5
66	FINANCE	gt5
58	FINANCE	gt5
58	MARKETIN	lt5
63	MARKETIN	lt5
53	MARKETIN	lt5
74	MARKETIN	lt5
77	MARKETIN	lt5
69	MARKETIN	lt5
57	MARKETIN	gt5
70	MARKETIN	gt5
68	MARKETIN	gt5
77	CS	lt5

Columns	Description	Type	Measurement	Possible values
Satindex	Satisfaction Index	Numeric	-	Positive Values
Dept	Department	Character	MARKETING, CS, FINANCE	3
Exp	Years of Experience (grouped)	Character	lt5 = less than 5, gt5 = greater than 5	2

Two Way ANOVA

Testing equality of means in two factors.

Objective	To compare employee satisfaction index in three departments (CS, Marketing, Finance) and two experience level based groups.

Null Hypothesis

(H_{01}): Average satisfaction index is equal for 3 departments.

(H_{02}): Average satisfaction index is equal for 2 experience levels.

(H_{03}) Interaction effect(dept*exp) is not significant on satisfaction index.

The test statistic is computed for each of these null hypothesis.

Reject the null hypothesis if $p\text{-value} < 0.05$

DATA SCIENCE INSTITUTE

Two Way ANOVA in Python

```
# Import data
```

```
import pandas as pd
data = pd.read_csv('Two Way Anova.csv')
```

```
# ANOVA Table
```

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols('satindex ~ C(dept) + C(exp) + C(dept) : C(exp)',
data=data).fit()
sm.stats.anova_lm(model, typ=2)
```

- ❑ ***'sm.stats.anova_lm'*** *is the Python function for ANOVA .*
- ❑ ***formula*** *specifies 'satindex' as analysis (dependent) variable and 'dept' and 'exp' as factor (independent) variables.*
- ❑ ***C(dept) : C(exp)*** *specifies the interaction effect.*

DATA SCIENCE INSTITUTE

Two Way ANOVA in Python

```
# Output:
```

	sum_sq	df	F	PR(>F)
C(dept)	164.222222	2.0	1.678973	0.203624
C(exp)	78.027778	1.0	1.595479	0.216274
C(dept):C(exp)	20.222222	2.0	0.206748	0.814374
Residual	1467.166667	30.0	NaN	NaN

Interpretation :

- *Since p-value is >0.05 for all three (dept, exp and dept*exp), do not reject H0 for all three tests. There is no significant difference in satisfaction index among 3 different departments and 2 experience levels.*
- *Also interaction effect is not significant.*

DATA SCIENCE INSTITUTE

THANK YOU!

THANK YOU!

DATA SCIENCE
INSTITUTE

THANK YOU!

DATA SCIENCE
INSTITUTE