

Statistical Inference

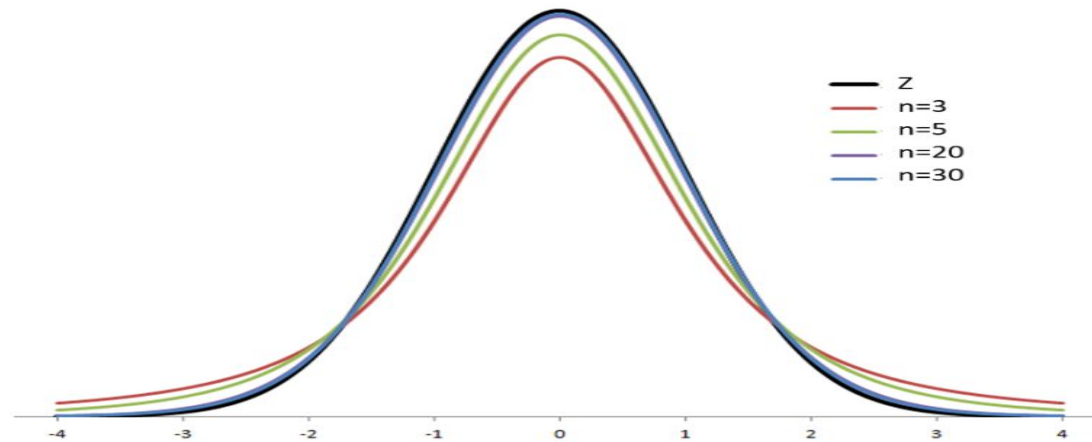
Parametric Tests I

Contents

1. Introduction to t-distribution
2. One Sample t-test
3. Independent samples t-test

t-distribution

- The t distribution is symmetric and its overall shape resembles the bell shape of a normally distributed variable with mean 0 and variance 1, except that it is a bit lower and wider.
- As the sample size increases so as the number of degrees of freedom grows, the t-distribution approaches the normal distribution with mean 0 and variance 1.



- In the above graph, z is normal distribution with mean 0 and variance 1.

A note on Degrees of Freedom (DF)

- Degrees of freedom (df) is defined as the number of independent terms.
- "Sum of the squared deviations about mean of n values" has $n-1$ degrees of freedom. Knowing $n-1$ values, we can find last value since sum of deviations about mean is always zero.
- Sampling distributions like t , F and chi square have shapes based on degrees of freedom.
- Example , Give 5 numbers such that sum is 20. You can use 4 numbers freely but fifth number should be such that sum is 20. Here $df = 4$

One sample t-test

- One sample t test is used to test the hypothesis about a single population mean.
- We use one-sample t-test when we collect data on a single sample drawn from a defined population.
- For this design, we have one group of subjects, collect data on these subjects and compare sample statistic to the hypothesized value of population parameter.
- Subjects in the study can be patients, customers, retail stores etc.

Case Study

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

Background

A large company is concerned about time taken by employees to complete weekly MIS report.

Objective

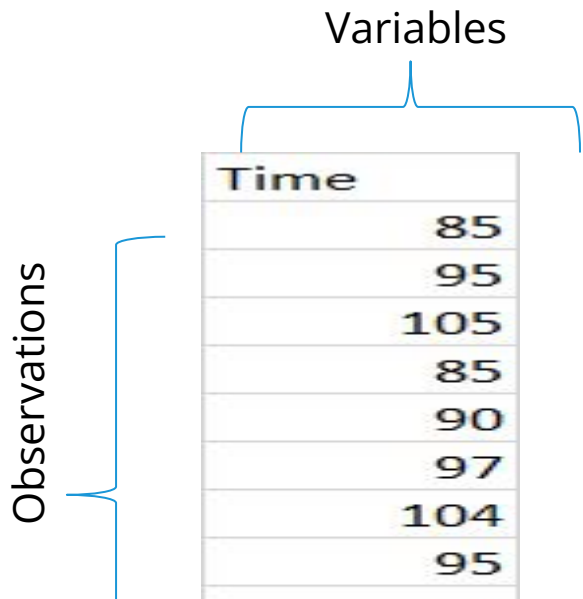
To check if average time taken to complete the MIS report is more than 90 minutes

Sample Size

Sample size: 12
Variables: Time

Data Snapshot

ONE SAMPLE t
TEST



Columns	Description	Type	Measurement	Possible values
Time	Time taken to complete MIS	Numeric	Minutes	Positive Values

Assumptions for one sample t-test

- The assumptions of the one-sample t-test are listed below:
 - Random sampling from a defined population
(employees are selected at random from the company)
 - Population is normally distributed
(Time taken to complete MIS report should be normally distributed).
 - Variable under study should be continuous.
- Normality test can be performed by any of the methods explained earlier.
- The validity of the test is not seriously affected by moderate deviations from 'Normality' assumption.

One sample t-test

Testing whether mean is equal to a test value.

Objective	To test the average time taken to complete MIS is more than 90 minutes
------------------	--

Null Hypothesis (H_0): $\mu = 90$

Alternate Hypothesis (H_1): $\mu > 90$

Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Notation	Value
Sample Size	n	12
Mean		93.5833
Standard Deviation	S	6.4731
Standard Error	s/ \sqrt{n}	1.8686
Difference		93.5833-90=3.5833
t	$\frac{\bar{x} - \mu_0}{S.E}$	1.9176

One sample t-test in Python

Import data

```
data2=pd.read_csv('ONE SAMPLE t TEST.csv')
```

t-test for one sample

```
from scipy.stats import ttest_1samp  
ttest_1samp(data2.Time, popmean=90,alternative='greater')
```

- ❑ **ttest_1samp()** from scipy package, returns two tailed t and p-value.
- ❑ **data.time** is the variable under study.
- ❑ **popmean=90** is the value to be tested.

Output:

```
Ttest_1sampResult(statistic=1.9176218472595046, pvalue=0.04074043079962237)
```

Interpretation :

- ❑ **scipy always gives the test statistic as signed.** This means that given p and t values from a two-tailed test, you would reject the null hypothesis of a greater-than test when $p/2 < \alpha$ and $t > 0$, and of a less-than test when $p/2 < \alpha$ and $t < 0$.
- ❑ Since $p/2$ is < 0.05 , reject H_0 . Average time taken to complete the MIS report is more than 90 minutes '

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

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

Assumptions for independent samples t-test

- The assumptions for independent samples t-test are listed below :
 - The samples drawn are random samples.
(Employees are selected at random from the company)
 - The populations from which samples are drawn have equal & unknown variances.
(F-test is used to validate this assumption which will be covered in next presentation)
 - The populations follow normal distribution.
(**Time taken to complete MIS report should be normally distributed for both groups**)

Normality assumption can be validated using method explained earlier).

Independent sample t-test

Testing whether means of two groups are equal.

Null Hypothesis (H_0): $\mu_1 = \mu_2$

Alternate Hypothesis (H_1): $\mu_1 \neq \mu_2$

μ_1 = average time taken by group1 to complete MIS

μ_2 = average time taken by group2 to complete MIS .

Objective	To test the average time taken to complete MIS by both the groups is same.
Test Statistic	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Group I	Group II
Sample Size	n1=12	n2=14
Mean		
Variance	$S_1^2=41.9015$	$S_2^2=27.1483$
Pooled Variance	$S_p^2=33.9102$	
Difference		
t	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 0.22345$	

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.

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 H_0 . There is no significant difference in average time taken to complete the MIS between both the group of employees.

Independent samples t-test when variances are not equal

- Welch's t test is used to test the equality of two means if variances of two groups can not be assumed equal.
- Welch's t-test defines the statistic t by the following formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- The denominator is not based on a pooled variance estimate.
- If 2 variance are not equal, t test syntax in Python is given below:

```
data=pd.read_csv('INDEPENDENT SAMPLES t TEST.csv')  
  
stats.ttest_ind(data['time_g1'],data['time_g2'], equal_var=False,  
nan_policy='omit')
```

Independent samples t-test when variances are not equal

Output:

```
Ttest_indResult(statistic=0.21965992515741178,pvalue=0.8282468548302413)
```

Interpretation :

- Since p-value is >0.05 , do not reject H_0 . There is no significant difference in average time taken to complete the MIS between both the group of employees.

Quick Recap

In this session, we continued to learn various parametric tests . Here is a quick recap :

Independent sample t test

- It compares the means of two independent groups on the same continuous variable.
- $H_0: \mu_1 = \mu_2$