Lec. VIII. T-test & Code

(B) Independent two sample test :

Used to compare the means of two-independent samples. The null hypoo states that there's no significant difference B|w the means of two samples while the alternate hypothesis states that there's a significant difference.

○ Assumptions for t-test :

(1) Independence of two samples, The two sample must be independent, meaning there's no relno. B|w the observations in one group & the observation in other group. The subjects in the two groups should be selected randomly & Independently.

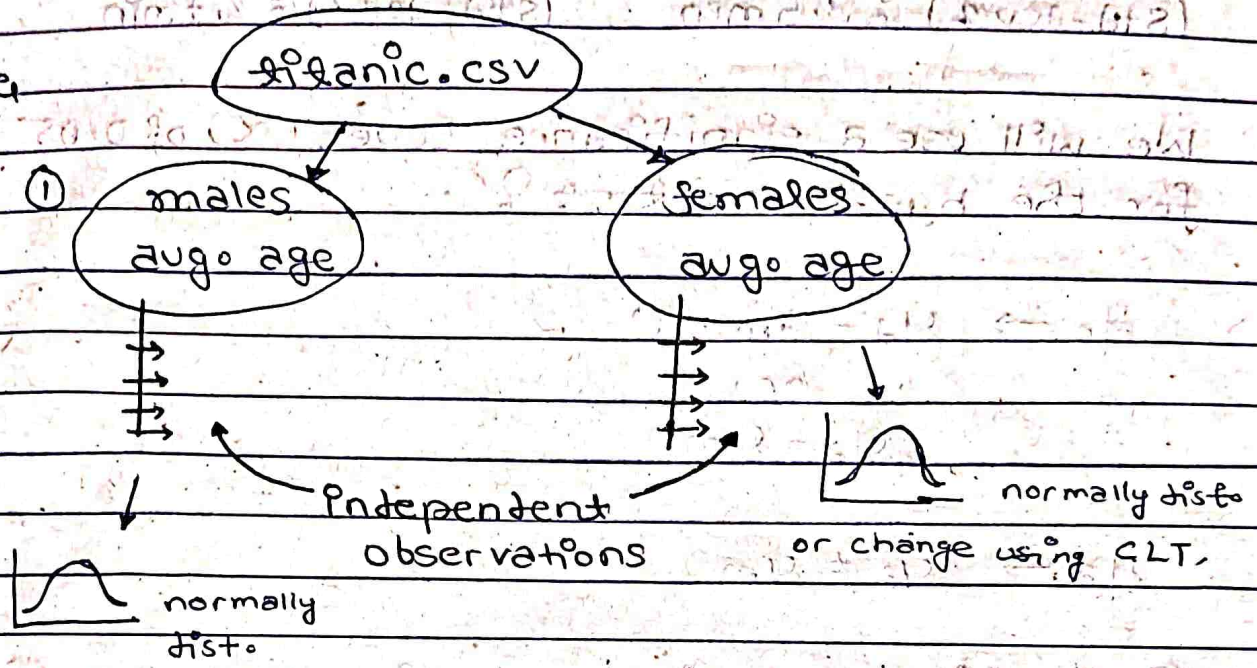
(2) Normality, The data in each of the two groups should be approximately normal dist°. The t-test is considered robust to mild violations of normality, especially when the sample sizes are large ($n \geq 30$) and the sample sizes of the two groups are similar.

(3) Equal Variances (Homoscedasticity), The variance of the two populations should be approximately equal. These assumptions can be checked using F-test, or if this assumption not met you can use Welch's t-test.

(4) Random Sampling, The data should be collected using a random sampling method from the respective populations. This ensures that the sample is representative of population & reduces the risk of selection bias.

(Independent two sample or Unpaired T-test) is a statistical method used to compare means of two independent groups to determine if there's a significant difference b/w them.

ie,



Q. Suppose a website owner claims that there's no difference in the avg. time spent on their website b/w desktop & mobile users. To test this claim, we collect data from 30 desktop users and 30 mobile users regards the time spent on the website in minutes. The sample statics are as follows —

desktop users = {12, 15, 10, 16, 20, 17, 14, 22, 19, 21, 23 18, 25, 17, 16, 24, 20, 19, 22, 18, 15, 14, 23, 16, 12, 21, 19, 17 20, 14}

mobile-users = {10, 12, 14, 13, 16, 15, 11, 17, 14, 16, 18, 14, 20, 15 14, 19, 16, 15, 17, 14, 12, 11, 18, 15, 10, 16, 15, 13, 16, 11}

desktop user ⇓ | mobile user ⇓

n1 = 30 | n2 = 30

mean1 = 10.5 min | mean2 = 14.3 min

(Std-dev 1) = 3.5 min | (Std-dev 2) = 2.7 min

We will use a significance level (α) of 0.05 for the hypothesis test.

$H_0 \rightarrow \mu_d - \mu_m = 0$

or

$\mu_d = \mu_m$

$H_a \rightarrow \mu_d \neq \mu_m$

(2) testing normality → shapiro test

(3) testing homoscedasticity → f-test / Welch's t-test

(discussed later)

	desktop	mobile
statistic →	0.97031141	0.97143560
pvalue →	0.779096	0.57916086

variances → using levene ⟶ ↓

if, p-value < 0.05 ⇒ $\sigma_A^2 \neq \sigma_B^2$ | statistic = 2.9439, p-value = 0.0153

0.0153 → 0.05

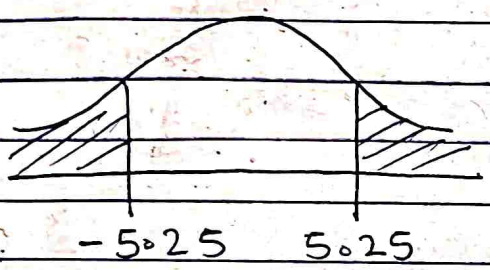
p-value > 0.05 ⇒ $\sigma_A^2 = \sigma_B^2$

both assumptions are true.

$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ replaces $\rightarrow$ $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

$\Rightarrow \frac{18.5 - 14.3}{\sqrt{\frac{(3.5)^2}{30} + \frac{(2.7)^2}{30}}} \Rightarrow \frac{4.2}{\sqrt{\frac{19.54}{30}}}$

$\Rightarrow 5.25$



$df = n1 + n2 - 2.$

$30 + 30 - 2 \Rightarrow 58$

p-value -> 0.00000--25

reject Null hypothesis.

Single Sample t-test code:

```
import seaborn as sns
import pandas as pd
import numpy as np

df = sns.load_dataset('titanic')
df.head()

pop = df['age'].dropna()
```

```
0    22.0
1    38.0
2    26.0
     ...
890  32.0
Name: age, Length: 714, dtype: float64
```

```
sample_age = pop.sample(25).values
```

```
sample_age
```

```
array([ 2. , 32. , 45. , 47. , ..... 30. , 34. ])
```

```
# Ho: The mean Age is 35
# Ha: The mean Age is less than 35.

# As the sample size is less than 30 (ie 25)

# Applying shapiro test to check whether
# It's Normally Distributed.

from scipy.stats import shapiro

shapiro_age = shapiro(sample_age)
shapiro_age
```

```
ShapiroResult(statistic=0.955, pvalue=0.328)
```

```
# pvalue > 0.05 --> Normally Distributed

pop_mean = sample_age.mean()
```

```
32.56
```

```
import scipy.stats as stats

t_statistic, p_value = stats.ttest_1samp
                       (sample_age, pop_mean)

print('t_statistic:', t_statistic)
print('p_value:', p_value/2)

alpha = 0.05

if p_value < alpha:
    print("Reject Null hypo")
else:
    print("fail to reject Null hypo")
```

Independent two-sample t-Test:

```
import seaborn as sns

df = sns.load_dataset('titanic')
df.head()

pop-male = df[df['sex'] == 'male']['age
pop-female = df[df['sex'] == 'female']['age'
                                .dropna(
                                .dropna()

sample-male = pop-male.sample(25)
sample-female = pop-female.sample(25)

alpha = 0.05

# H0 : mean age of male & female are same
# Ha : mean age of male is higher than female

from scipy.stats import levene

levens-test = levene(sample-male, sample
                                -female)

print (levens-test)
(statistic = 0.30, pvalue = 0.58)

0.58 > 0.05
```

```
import scipy.stats as stats

t-statistic, p-value = stats.ttest_ind
                    (sample_male, sample_female)

print ('t-statistic)
print (p-value 12):
```

(c) Paired two sample test:

Used to compare the means of two related or dependent groups.

Ie, (1) before-and-after studies — comparing the performance of ~~two~~ a groups before and after an intervention or treatment.

(2) Matched or correlated groups — Comparing the performance of two groups that are corelated or matched in some way, such as siblings.

Assumptions: (1) paired observations — The two set of observation must be related or paired in some way.

(2) Normality (3) Independence of pairs.

Q) Let's assume that a fitness centre is evaluating the effectiveness of a new 8 week weight loss program. They enroll 15 participants in program and measure their weight before & after program. The goal is to test whether the new weight loss program leads to a significant reduction in the participants weight?

significance level (α) = 0.05

name	wt before	wt after	difference	(difference)^2
1	80	73	7	49
2	92	98	-6	36
3	75	81	-6	36
4	68	67	1	1
5	85	88	-3	9
6	78	76	2	4
7	73	74	-1	1
8	90	91	-1	1
9	70	69	1	1
10	88	88	0	0
11	76	77	-1	1
12	84	81	3	9
13	82	80	2	4
14	77	79	-2	4
15	91	88	3	9

$\sum d \Rightarrow -1$ $\sum d^2 \Rightarrow 16$

$(\sum d)^2 \Rightarrow 1$

$n = 15$

H_0 : $\mu_{before} = \mu_{after}$

H_a : $\mu_{before} > \mu_{after}$

Check → difference → is normally distributed.

$$t = \frac{\Sigma d}{\sqrt{\frac{n(\Sigma d^2) - (\Sigma d)^2}{n-1}}} \Rightarrow \frac{-1}{\sqrt{\frac{15(165) - 1}{14}}}$$

$$\Rightarrow \frac{-1}{\sqrt{\frac{2474}{14}}} \Rightarrow \frac{-1}{\sqrt{176.71}} \Rightarrow \frac{-1}{13.29} \Rightarrow -0.075$$

★ Chi-Square Test

- A chi-square test (χ^2) is basically a comparison of two statistical data sets
- The chi-square test is used to estimate how likely the observation that are made would be, by considering the assumption of the Null hypo. as true.
- When we consider the null-speculation is true, the sampling distribution of the test-statistic is called chi-squared distribution
- The chi-square test helps us to determine whether there's a notable difference b/w the normal freq. and observed frequencies.

i.e. Applicable for categorical data such as men & women falling under categories of gender, age, height etc

- To calculate p-value, chi-square test is used
  - $p \leq 0.05 \rightarrow$ rejected
  - $p > 0.05 \rightarrow$ Accepted

properties: (1) mean = ν variance = 2ν
(ν: degree of freedom)

(2) $\nu_1 + \nu_2 = \chi_1^2 + \chi_2^2$

(3) chi-square distribution curve approaches the normal distribution when dof increases

$$\chi^2 = \sum \frac{(\text{observed value} - \text{expected value})^2}{\text{Expected value}}$$

or

$$\boxed{\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}}$$

Q) A survey on cars had conducted in 2011 and determined that 60% of cars owners have only one car, 28% have two cars and 12% have three or more. Supposing that you have decided to conducting your own survey and have collected the data below, determine whether your data supports the result of the day?

Use a significance level of 0.05 & given that out of 129 car owners, 73 had one car & 38 had two cars?

H_0 : The proportion of car owners with one car → 0.60, two car → 0.28, three or more → 0.12

H_a : The proportions doesn't match the proposed model.

A chi-square goodness of fit test is appropriate Because we are examining the distribution of a single categorical variable.

	Observed (O_i)	Expected (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$/E_i$
One car	73	$0.60 \times 129 \Rightarrow 77.4$	-4.4	19.36	0.250
Two car	38	$0.28 \times 129 \Rightarrow 36.1$	1.9	3.61	0.1
three or more car	18	$0.12 \times 129 \Rightarrow 15.5$	2.5	6.25	0.40 32
total	129				0.7533

$dof \Rightarrow 3 - 1 \Rightarrow 2$; χ^2

using table, the critical value of a 0.05 significance level with a df = 2 is 5.99

That means 95 times out of 100, a survey that agrees with a sample will have a χ^2 value of 5.99 or less

The chi-square statistic is only 0.7533, so we'll accept Null hypothesis

Code implementation :

The test is applied when you have two categorical variables from a single population. It's used to determine there's a significant association btw two variables.

```
import scipy.stats as stas
import seaborn as sns
       pandas as pd
       numpy as np

df = sns.load_dataset ('tips')

df.head()
```

↳ whether there's any association btw "sex" & "smoker".

```
df_table = pd.crosstab (df['sex'],
                        df['smoker'])
```

↳

smoker	yes	no
sex		
male	60	97
female	33	54

```
# Observed values

Obs = df_table.values
print ("Observed values :-\n", Obs)
```

↳ [[60 97]
[33 54]]

```
Val = stats.chi2_contigency (df_table)
```

↳ (0.0008 , 0.92 , 1 , array([[59.84, 97.15],
[33.15, 53.84]]),
chi-square statistics, pvalues, expected_values

```
Expected_values = Val[3]

nos_rows = len(df_table.iloc[0:2,0])
nos_cols = len(df_table.iloc[0,0:2])

ddof = (nos_rows - 1)(nos_colg - 1)
print ("Degree of freedom :", ddof)

alpha = 0.05

from scipy.stats import chi2

chi_square = sum([(o-e)**2./e for o,e
                  in zip(obs, expected_values)])

chi_square_statistic = chi_square[0] +
                       chi_square[1].
```

↳ (0.0019)

critical_value = chi2.ppf(q = 1 - alpha, df = ddof)

print("critical-value :", critical_value)

↳ 3.841

pvalue ↴

p-value = 1 - chi2.cdf(x = chi-square-statistic, df = ddof)

↓

0.964

If chi-square-statistical >= critical-value:
    print("Reject H0, there's a reln. b/w two categorical variables")

else:
    print("Retain H0, There's no reln. b/w two categorical variables")

if p-value <= alpha:

else: