

LAB02. THỐNG KÊ SUY DIỄN

IS403 – PHÂN TÍCH DỮ LIỆU KINH DOANH




TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

Người trình bày: **Nguyễn Minh Nhựt**
Số điện thoại: **0939013911 - 0981734105**

TÓM TẮT NỘI DUNG LAB2

- Sinh viên hiểu được các loại kiểm định ANOVA: **Levene Test**, **ANOVA Test**, **Turkey Test** trên bài toán kiểm định 1 phía
- Sinh viên nắm được bài toán kiểm định **Chi-Square**
- Thực hành trên các công cụ: **Python**, **R**, **Excel**



CHỦ ĐỀ 1 CÁC LẠI KIỂM ĐỊNH ANOVA VÀ ỨNG DỤNG (ANOVA TEST)

- *Leven, ANOVA, Turkey Test*
- *Thực hành trên R, Python và Excel*

1 - KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

HYPOTHESIS

Giả thuyết là
một nhận
định, một ý
kiến

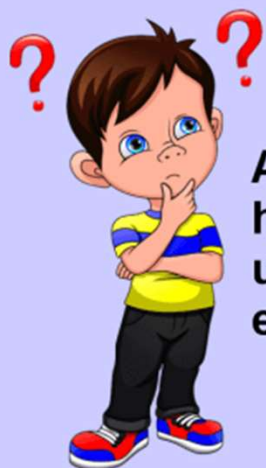


Giả thuyết khoa
học là một nhận
định, một ý kiến
yêu cầu có một
phương pháp
kiểm định

1 - KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

NULL HYPOTHESIS

Null hypothesis



H_0

???

A null hypothesis is a form of hypothesis that is deemed "true" until proven wrong based on experimental data.

=

• Equal

\geq


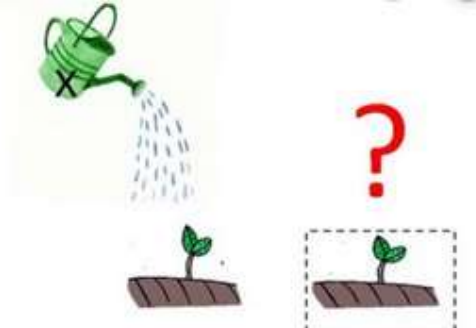
• Greater than or equal to

\leq

• less than or equal to

1 - KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

ALTERNATIVE HYPOTHESIS

	
H_1 : Application of bio-fertilizer 'x' increase plant growth.	H_0 : Application of bio-fertilizer 'x' <u>do not</u> increase plant growth.
Alternative hypothesis	Null hypothesis
✓ The alternative hypothesis is a hypothesis which the researcher tries to prove.	✓ The null hypothesis is a hypothesis which the researcher tries to disprove, or nullify.

1 - KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

DATASET DEMO

INSURANCE SURVEY DATASET

Insurance Survey						
Age	Gender	Education	Marital Status	Years Employed	Satisfaction*	Premium/Deductible**
36	F	Some college	Divorced	4	4	N
55	F	Some college	Divorced	2	1	N
61	M	Graduate degree	Widowed	26	3	N
65	F	Some college	Married	9	4	N
53	F	Graduate degree	Married	6	4	N
50	F	Graduate degree	Married	10	5	N
28	F	College graduate	Married	4	5	N
62	F	College graduate	Divorced	9	3	N
48	M	Graduate degree	Married	6	5	N
31	M	Graduate degree	Married	1	5	N
57	F	College graduate	Married	4	5	N
44	M	College graduate	Married	2	3	N
38	M	Some college	Married	3	2	N
27	M	Some college	Married	2	3	N
56	M	Graduate degree	Married	4	4	Y
43	F	College graduate	Married	5	3	Y
45	M	College graduate	Married	15	3	Y
42	F	College graduate	Married	12	3	Y
29	M	Graduate degree	Single	10	5	N
28	F	Some college	Married	3	4	Y
36	M	Some college	Divorced	15	4	Y
49	F	Graduate degree	Married	2	5	N
46	F	College graduate	Divorced	20	4	N
52	F	College graduate	Married	18	2	N

*Measured from 1-5 with 5 being highly satisfied.

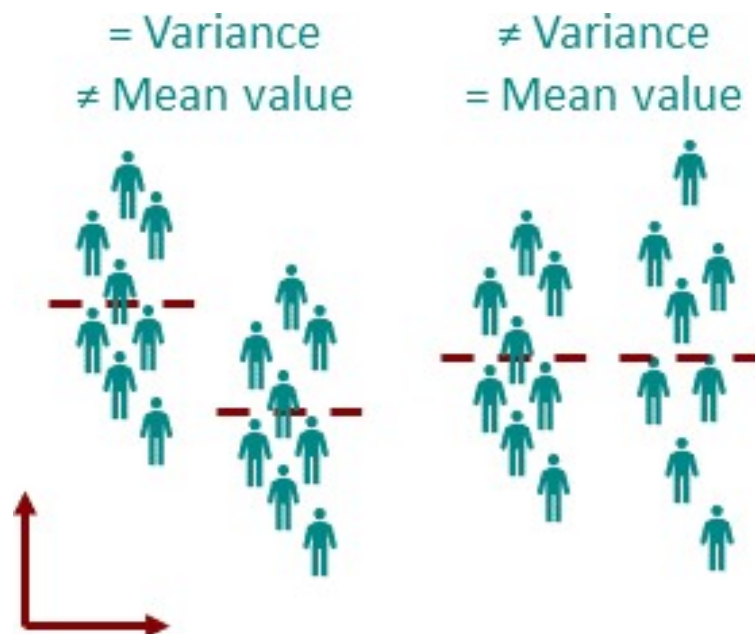
**Would you be willing to pay a lower premium for a higher deductible?

F	G	H	I	J
College graduate	Graduate degree	Some college		
5	3	4		
3	4	1		
5	5	4		
3	5	2		
3	5	3		
3	4	4		
3	5	4		
4	5			
2				

2- Kiểm định Levene (Levene's Testing)

LEVENE'S TEST

Phương sai
các nhóm là
bằng nhau
hay không
bằng nhau?



Kiểm tra tính
đồng nhất
của phương
sai.

Là bước tiền
điều kiện
kiểm định
ANOVA

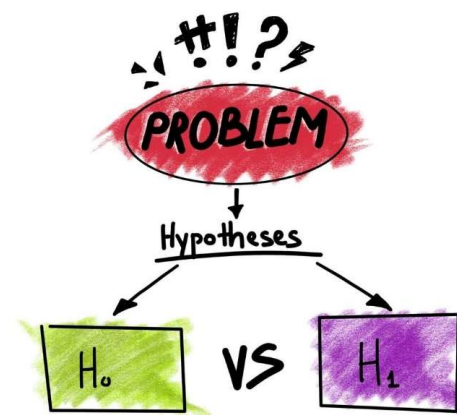
2 - Kiểm định Levene (Levene's Testing)

THEORY LEVENE'S TEST

H0: PHƯƠNG SAI GIỮA CÁC NHÓM LÀ BẰNG NHAU

H1: PHƯƠNG SAI GIỮA CÁC NHÓM LÀ KHÁC NHAU

→ Nếu chấp nhận H0 (Giả thuyết) thì ta có thể nói rằng phương sai các nhóm là bằng nhau → Có thể kiểm định ANOVA



2- Kiểm định Levene (Levene's Testing)

FORMULA LEVENE'S TEST

$$W = \frac{(N - k)}{(k - 1)} \frac{\sum_{i=1}^k N_i (\bar{Z}_{i.} - \bar{Z}_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_{i.})^2}$$

$$1. Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$$

where $\bar{Y}_{i.}$ is the mean of the i -th subgroup.

$\bar{Z}_{..}$ = Mean of all Z_{ij} data

$\bar{Z}_{i.}$ = Mean Z_{ij} group i

N = total number of samples

N_i = number of samples in group i

k = number of groups

$$W > F_{1-\alpha}(k-1; n-k)$$

**Bác bỏ giả
thuyết H_0**

2- Kiểm định Levene (Levene's Testing)

THỰC HÀNH KIỂM ĐỊNH LEVENE TRÊN R

```
warning message:
In leveneTest.default(y = y, group = group, ...) : group co
> leveneTest(Satisfaction., Education, center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  2  0.9434 0.4052
      21
warning message:
In leveneTest.default(Satisfaction., Education, center = me
  Education coerced to factor.
> |
```

P-value	Decision
Less than 0.05*	Reject Null (H_0) Hypothesis Statistical difference between groups
Greater than 0.05*	Fail to Reject Null (H_0) Hypothesis No statistical difference between groups, or not enough evidence (data) to find a difference

* Assuming $\alpha = 0.05$

- Cài thư viện R **car**
`install.packages("car")`
- Sử dụng thư viện
`require(car)`
- Kiểm định Levene trong R
`leveneTest(value, group, center=mean)`
- Fisher: `qf(p=.05, k-1, n-k, lower.tail=FALSE)`
- Tham khảo chọn **center = mean**
<https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm>

2- Kiểm định Levene (Levene's Testing)

THỰC HÀNH KIỂM ĐỊNH LEVENE TRÊN PYTHON

```
In [3]: sep=ex02.groupby('Education')['Satisfaction'].apply(list)
print(sep)
```

```
Education
College graduate    [5, 3, 5, 3, 3, 3, 3, 4, 2]
Graduate degree     [3, 4, 5, 5, 5, 4, 5, 5]
Some college        [4, 1, 4, 2, 3, 4, 4]
Name: Satisfaction, dtype: object
```

```
In [4]: from scipy.stats import levene
stat, p = levene(*sep, center='mean')
print(stat,p)

0.9433580072525427 0.40520616699352924
```

○ **Groupby** Education lấy Satisfaction đưa dạng list

```
sep = df.groupby(group)[value].apply(list)
```

○ **Sử dụng thư viện**

```
from scipy.stats import levene
```

○ **Kiểm định Levene trong Python**

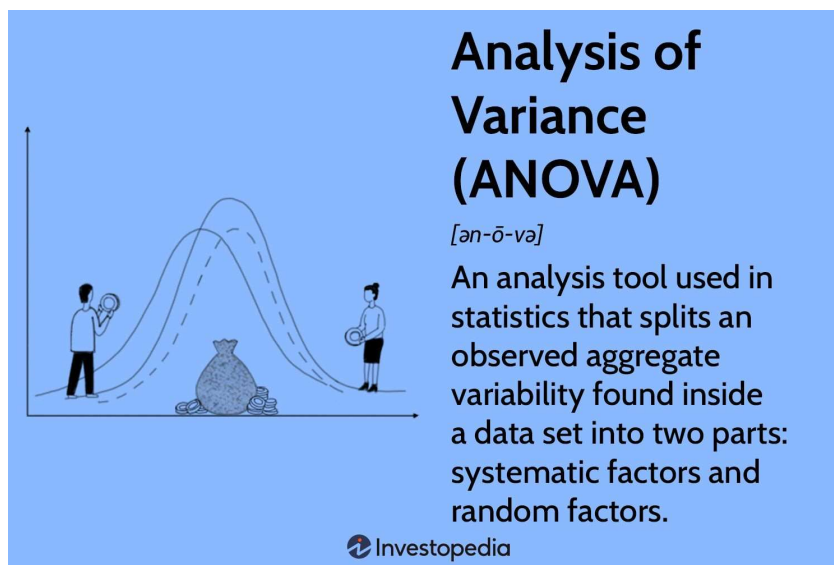
```
stat, p = levene(*sep, center = 'mean')
```

3- Kiểm định ANOVA (ANOVA's Testing)

ANOVA'S one TEST

Trung bình các nhóm là bằng nhau hay không bằng nhau?

Loại ANOVA:
One-way, Two-way, Multiple-way (MANOVA)



Kiểm định **ANOVA** hay tên gọi khác là phân tích phương sai (Analysis of Variance).

Là một **kỹ thuật thống kê** tham số được sử dụng để phân tích sự khác nhau giữa **giá trị trung bình** của các **biến phụ thuộc** với nhau (Ronald Fisher, 1918).

Mục tiêu: Tìm xem yếu tố này có **ảnh hưởng** yếu tố khác hay không?

3- Kiểm định ANOVA (ANOVA's Testing)

THEORY ANOVA'S one TEST

H0: TRUNG BÌNH CÁC NHÓM LÀ BẰNG NHAU

H1: TRUNG BÌNH CÁC NHÓM LÀ KHÁC NHAU

N1	N2	N3
x1	y1	z1
x2	y2	z2
x3	y3	z3
x4	y4	z4
x6	y5	z5

(Giả thuyết) $H_0 : \mu_1 = \mu_2 = \dots = \mu_n$

(Đôi thuyết) H_1 : ít nhất 1 cái khác nhau

Các giá trị lưu ý:

- k: số nhóm khảo sát
- n: Là số lượng tổng thể
- ni: Là số lượng phần tử thứ i

Phát biểu bài toán

Có thể cho rằng trung bình giữa các nhóm N1, N2 và N3 bằng nhau được hay không.

3- Kiểm định ANOVA (ANOVA's Testing)

THEORY ANOVA'S one TEST

- *Tính trung bình từng nhóm*

$N1$	$N2$	$N3$
x1	y1	z1
x2	y2	z2
x3	y3	z3
x4	y4	z4
x6	y5	z5
\overline{N}_1	\overline{N}_2	\overline{N}_3

- *Trung bình mỗi nhóm:*

$$\overline{N}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

- *Trung bình tổng thể:*

$$\overline{N} = \frac{1}{n} \sum_{i=1}^k \overline{N}_i \cdot n_i$$

- *Tính các đại lượng biến thiên*

Biến thiên nội bộ trong nhóm i

$$SS_i = \sum_{j=1}^{n_i} (x_{ij} - \overline{N}_i)^2$$

Biến thiên trong nội bộ các nhóm

$$SSW = SS_1 + SS_2 + \dots + SS_K$$

3- Kiểm định ANOVA (ANOVA's Testing)

THEORY ANOVA'S one TEST

- **Biến thiên trong nội bộ các nhóm**

$$SSW = SS_1 + SS_2 + \dots + SS_K$$

SSW (Within groups sum of square): là những biến thiên **không do yếu tố kiểm soát** (yếu tố dùng để phân tích nhóm) gây ra.

- **Tổng bình phương độ lệch giữa các nhóm SSG**

$$SSG = \sum_{i=1}^{n_i} n_i (\bar{N}_i - \bar{N})^2$$

SSG (Between groups sum of square): là những **biến thiên khác nhau giữa các nhóm** tức là biến thiên do yếu tố nghiên cứu gây ra.

- **Tổng biến thiên của 1 quan sát bất kỳ so với trung bình**

$$SST = SSG + SSW$$

SST (Total sum of square): là tổng bình phương các độ lệch giữa từng quan sát với trung bình của tất cả quan sát. **Biến thiên tổng = Biến thiên nghiên cứu + Biến thiên do các yếu tố khác.**

Nhận xét:

- Nếu phần biến thiên do các yếu tố tạo ra $SSG >$ biến thiên do các yếu tố khác SSW . Vậy yếu tố đang nghiên cứu thật sự ảnh hưởng đến yếu tố kết quả.

→ **Tăng khả năng bác bỏ H_0 .**

3- Kiểm định ANOVA (ANOVA's Testing)

THEORY ANOVA'S one TEST

- *Tính các phương sai*

Phương sai do các yếu tố khác tạo ra

$$MSW = \frac{SSW}{n - k}$$

Phương sai do yếu tố nghiên cứu tạo ra

$$MSG = \frac{SSG}{k - 1}$$

- *Kiểm định phương sai*

$$F = \frac{MSG}{MSW}$$

Nếu MSG lớn, MSW nhỏ \rightarrow F lớn

So sánh $F > F_{\alpha}(k - 1; n - k)$

Bác bỏ giả thuyết H_0

3- Kiểm định ANOVA (ANOVA's Testing)

THEORY ANOVA'S one TEST

- Bảng ANOVA một yếu tố*

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Within	$SSW = \sum_{j=1}^k \sum_{i=1}^l (X_{ij} - \bar{X}_j)^2$	$df_w = k - 1$	$MSW = \frac{SSW}{df_w}$	$F = \frac{MSB}{MSW}$
Between	$SSB = \sum_{j=1}^k (\bar{X}_j - \bar{X})^2$	$df_b = n - k$	$MSB = \frac{SSB}{df_b}$	
Total	$SST = \sum_{j=1}^n (\bar{X}_j - \bar{X})^2$	$df_t = n - 1$		

3- Kiểm định ANOVA (ANOVA's Testing)

THỰC HÀNH KIỂM ĐỊNH ANOVA TRÊN R

```
> tt=aov(Satisfaction~Education,data=ex02)
> summary(tt)
            Df Sum Sq Mean Sq F value Pr(>F)
Education    2  7.879    3.939   3.925 0.0356 *
Residuals   21 21.079    1.004
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

- Kiểm định Levene trong R

```
rs = aov(value~group,data=data_source)
```

```
Summary(rs)
```

- Fisher: `qf(p=.05,k-1,n-k, lower.tail=FALSE)`

THỰC HÀNH KIỂM ĐỊNH ANOVA TRÊN PYTHON

```
In [5]: import scipy.stats as stats
fvalue, pvalue=stats.f_oneway(*sep)
print(fvalue,pvalue)
3.9246517319277117 0.03563539756488997
```

- Kiểm định Levene trong Python

```
from scipy.stats import stats
```

```
fvalue, pvalue= stats.f_oneway(*sep)
```

```
Print(fvalue,pvalue)
```

- Fisher: `qf(p=.05,k-1,n-k, lower.tail=FALSE)`

4- Kiểm định ANOVA sâu (Turkey's Testing)

TURKEY TEST

- **Đặt vấn đề về kiểm định Turkey**

Kiểm định Turkey: Trong trường hợp **bác bỏ giả thuyết** H_0 ta muốn kết luận về **sự hơn kém giữa** các trung bình thì ta cần phân tích sâu hơn.

→ Được gọi là **phân tích ANOVA sâu** (Kiểm định Turkey)

- **Cách giải quyết bài toán kiểm định Turkey**

Với cùng mức ý nghĩa α , ta so sánh từng cặp trung bình để phát hiện các nhóm khác nhau.

Ví dụ 2: Trường hợp có 3 nhóm trung bình sánh

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \quad \begin{cases} H_0 : \mu_1 = \mu_3 \\ H_1 : \mu_1 \neq \mu_3 \end{cases} \quad \begin{cases} H_0 : \mu_2 = \mu_3 \\ H_1 : \mu_2 \neq \mu_3 \end{cases}$$

- **Các bước kiểm định Turkey**

Bước 1: Tính khoảng biến thiên trung bình giữa hai nhóm:

$$D_{ij} = |\bar{N}_i - \bar{N}_j|$$

Bước 2: Tính chỉ số Turkey

$$T = q_{\alpha}(k, n - k) \sqrt{\frac{MSW}{n_{\min}}}$$

Bước 3: Bác bỏ H_0 nếu $D_{ij} > T$

4- Kiểm định ANOVA sâu (Turkey's Testing)

THỰC HÀNH KIỂM ĐỊNH TURKEY TRÊN R

```
> TukeyHSD(tt)
```

```
Tukey multiple comparisons of means  
95% family-wise confidence level
```

```
Fit: aov(formula = Satisfaction ~ Education, data = ex02)
```

```
$Education
```

	diff	lwr
Graduate degree-College graduate	1.0555556	-0.1715336
Some college-College graduate	-0.3015873	-1.5742334
Some college-Graduate degree	-1.3571429	-2.6641246

	upr	p adj
Graduate degree-College graduate	2.28264475	0.1003252
Some college-College graduate	0.97105876	0.8230559
Some college-Graduate degree	-0.05016107	0.0409193

```
> |
```

- Kiểm định Turkey trong R
TukeyHSD(Result of ANOVA)

KẾT LUẬN

4- Kiểm định ANOVA sâu (Turkey's Testing)

THỰC HÀNH KIỂM ĐỊNH TURKEY TRÊN PYTHON

```
In [8]: from statsmodels.stats.multicomp import pairwise_tukeyhsd  
tukey = pairwise_tukeyhsd(endog=ex02['Satisfaction'], groups=ex02.Education, alpha=0.05)  
print(tukey)
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05  
=====
```

group1	group2	meandiff	p-adj	lower	upper	reject
College graduate	Graduate degree	1.0556	0.1003	-0.1715	2.2826	False
College graduate	Some college	-0.3016	0.8231	-1.5742	0.9711	False
Graduate degree	Some college	-1.3571	0.0409	-2.6641	-0.0502	True

```
=====
```

○ Kiểm định Turkey trong Python

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd  
tukey = pairwise_tukeyhsd(endog=value, groups=group, alpha=0.05)  
print(tukey)
```

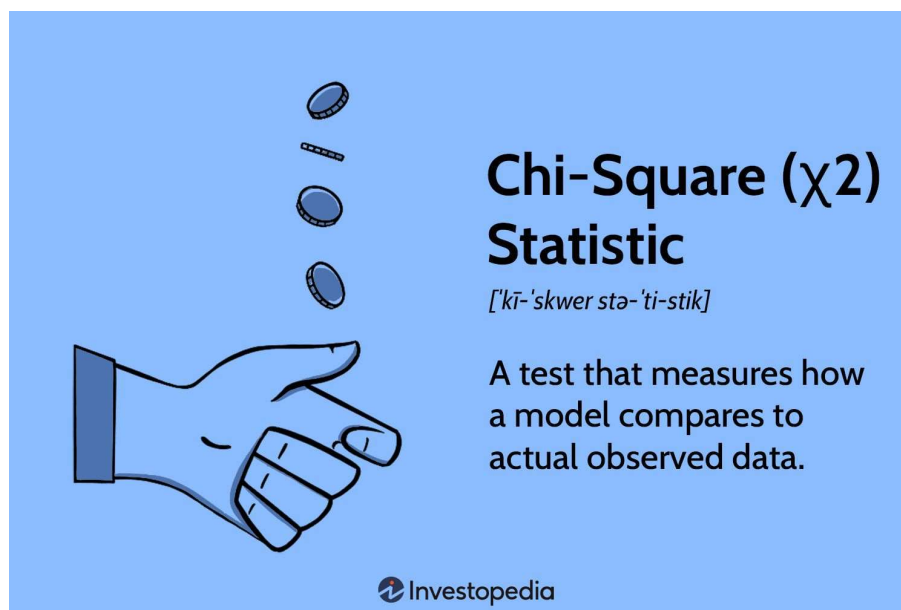


CHỦ ĐỀ 2 KIỂM ĐỊNH CHI-SQUARE (CHI-SQUARE TEST)

- *Chi-Square Test*
- *Thực hành trên R, Python và Excel*

1 - KIỂM ĐỊNH CHI-SQUARE TEST

CHI-SQUARE TEST



Kiểm định sự
độc lập/phụ
thuộc của hai
biến dạng phân
loại.

1 - KIỂM ĐỊNH CHI-SQUARE TEST

THEORY CHI-SQUARE TEST

H_0 : Hai biến phân loại là độc lập

H_1 : Hai biến phân loại là phụ thuộc

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

$$f_e \text{ của dòng } i \text{ cột } j = \frac{(\text{tổng } i) * (\text{tổng } j)}{\text{tổng quan sát}}$$

1 - KIỂM ĐỊNH CHI-SQUARE TEST

THEORY CHI-SQUARE TEST

Observed Frequencies				
Count of Respondent	Count of Respondent			
Count of Respondent	Brand 1	Brand 2	Brand 3	Grand Total
Female	9	6	22	37
Male	25	17	21	63
Grand Total	34	23	43	100

expected frequency in row i and column $j = \frac{(\text{grand total row } i)(\text{grand total column } j)}{\text{total number of observations}}$

Expected Frequencies				
Count of Respondent	Count of Respondent			
Count of Respondent	Brand 1	Brand 2	Brand 3	Grand Total
Female	12.58	8.51	15.91	37
Male	21.42	14.49	27.09	63
Grand Total	34	23	43	100

$(43 \times 37) / 100$

1 - KIỂM ĐỊNH CHI-SQUARE TEST

THEORY CHI-SQUARE TEST

(Observed - Expected)^2/Expected				
Count of Respondent				
Count of Respondent	Brand 1	Brand 2	Brand 3	Grand Total
Female	1.018792	0.740317	2.331119	4.0902278
Male	0.598338	0.43479	1.36907	2.40219728
Grand Total	1.61713	1.175107	3.700189	6.49242508

$(O-E)^2/E$

chi-square value X^2

X^2	6.492425079
df	2
p-value	0.038921342
chi-square critical value	5.991464547

$CHISQ.DIST.RT(X^2, df)$

$CHISQ.INV.RT(0.05, df)$

2- KIỂM ĐỊNH CHI-SQUARE TEST TRÊN R VÀ PYTHON

THỰC HÀNH CHI-SQUARE TRÊN R

```
> drink <- read.csv(file.choose(), header = TRUE)
> drink
```

	Gender	Brand.Preference
1	Male	Brand 3
2	Female	Brand 3
3	Male	Brand 3
4	Male	Brand 1
5	Male	Brand 1
6	Female	Brand 2
7	Male	Brand 2
8	Female	Brand 2
9	Male	Brand 1
10	Female	Brand 3
11	Male	Brand 3
12	Male	Brand 2
13	Female	Brand 3
14	Male	Brand 3
15	Female	Brand 3

```
> tb = table(Gender, Brand.Preference)
> tb
```

	Brand.Preference		
Gender	Brand 1	Brand 2	Brand 3
Female	9	6	22
Male	25	17	21

- Cài thêm **Package MASS** để kiểm Chi-quare

- Kiểm định Chi-Square trong R

`chisq.test(tb)` //tb: Là bảng đã được định dạng

```
> chisq.test(tb)
```

Pearson's Chi-squared test

```
data: tb
X-squared = 6.4924, df = 2, p-value = 0.03892
```

```
> |
```

2- KIỂM ĐỊNH CHI-SQUARE TEST TRÊN R VÀ PYTHON

THỰC HÀNH CHI-SQUARE TRÊN PYTHON

```
In [5]: drink = pd.read_csv("D:\\UIT-Term3\\PTTK\\Lab2\\11. Energy Drink Survey.csv")  
print(drink)
```

	Gender	Brand Preference
0	Male	Brand 3
1	Female	Brand 3
2	Male	Brand 3
3	Male	Brand 1
4	Male	Brand 1
..
95	Male	Brand 1
96	Male	Brand 3
97	Female	Brand 3
98	Male	Brand 2
99	Female	Brand 1

[100 rows x 2 columns]

```
In [17]: chisqt = pd.crosstab(drink.Gender, drink['Brand Preference'])  
print(chisqt)
```

Brand Preference	Brand 1	Brand 2	Brand 3
Gender			
Female	9	6	22
Male	25	17	21

- Crosstab (Đưa dữ liệu ban đầu về dạng bảng)

- Kiểm định Chi-Square trong Python

```
c, p, dof, expected = stats.chi2_contingency(chisqt)  
print(p)
```

```
In [19]: print(p)
```

0.038921342064441915

```
In [20]: print(c)
```

6.4924250792329055

```
In [21]: print(dof)
```

2



CHỦ ĐỀ 3

BÀI TẬP LAB02 TẠI LỚP

3- BÀI TẬP TẠI LỚP

Bài 1. Nghiên cứu về thu nhập của các hộ gia đình ở ngoại thành, người ta chia ngoại thành 7 địa bàn dân cư khác nhau. Chọn ngẫu nhiên các hộ gia đình trong từng địa bàn và ghi nhận địa bàn. Địa bàn dân cư thứ 3 có 13 hộ được chọn, các địa bàn còn lại đều chọn 19 hộ. Kết quả ANOVA như sau:

Source of Variation	SS	df	MS	F
Between Groups	187,2649			
Within Groups				
Total	1269,6891			

1. Điền vào những phần cam để hoàn thành bảng
2. Ở mức ý nghĩa 1% có thể kết luận rằng thu nhập trung bình của các hộ gia đình ở các địa bàn dân cư khác nhau là như nhau được hay không?

3- BÀI TẬP TẠI LỚP

Education Level	Income Level	Job Field
High School	Low	Education
College	High	Technology
High School	Low	Healthcare
Graduate School	High	Business
High School	Low	Technology
College	Low	Healthcare
Graduate School	High	Technology
High School	High	Business
Graduate School	High	Healthcare
College	Low	Business

Bài 2: Với tập dữ liệu. Số thứ tự chẵn làm bài 1, Số thứ tự lẻ làm bài 2

1. Kiểm định **Chi-Square Education Level** và **Income Level**
2. Kiểm định **Chi-Square Job Field** và **Income Level**

TÀI LIỆU THAM KHẢO.

1. <https://www.investopedia.com/>.
2. Roxy Peck and el, “Introduction to Statistics and Data Analysis”, 6th Edition
3. Phil Simon, “The Hundred-Page Machine Learning Book”

THANK YOU



TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN