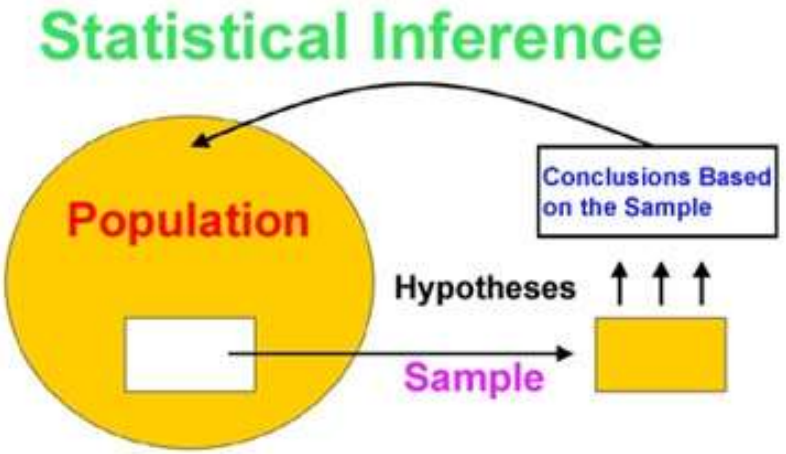
PHÂN TÍCH DỮ LIỆU KINH DOANH

LAB02. THỐNG KÊ SUY DIỄN

(Statistical inference)



CÔNG CỤ: R, PYTHON, EXCEL

Trình bày: Nguyễn Minh Nhựt

SĐT: 0939013911 - 09851734105

2.1. Kiểm định ANOVA – Kiểm định Levene

- ***Tiền xử lý kiểm định ANOVA***

Điều kiện để kiểm định ANOVA được hay không?

→ Cần một phép kiểm định trước là kiểm định ANOVA

- ***Phát biểu bài toán***

Giả thuyết: Phương sai giữa các nhóm bằng nhau

Đối thuyết: Phương sai giữa các nhóm không bằng nhau

→ *Nếu chấp nhận giả thuyết (Tức phương sai giữa các nhóm bằng nhau)* → *Có thể kiểm định ANOVA*

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

So sánh

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

Bác bỏ giả thuyết H0

2.1. Kiểm định ANOVA – Kiểm định Levene

- ***Tiền xử lý kiểm định ANOVA***

Điều kiện để kiểm định ANOVA được hay không?

→ Cần một phép kiểm định trước là kiểm định ANOVA

- ***Phát biểu bài toán***

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

$Z_{ij} = |Y_{ij} - median Y_i|$

$\bar{Z}_{..}$ = 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

2.2. Kiểm định ANOVA

- 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).
- Kiểm định **ANOVA** gồm 3 phương pháp: ***ANOVA một chiều*** *(One-way ANOVA)*, ***ANOVA hai chiều*** *(Two-way ANOVA)* và ***ANOVA đa biến*** *(MANOVA)*

→ *Kiểm định ANOVA so sánh các **giá trị trung bình** (tìm xem yếu tố này có ảnh hưởng yếu tố khác hay không).*

2.2. Kiểm định ANOVA

- ***Bài toán kiểm định ANOVA***

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

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

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

- ***Phát biểu bài toán***

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

Các giá trị lưu ý:

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

2.2. Kiểm định ANOVA

- ***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:***

$$\overline{N} = \frac{1}{n}\sum_{i=1}^{k} \overline{N_i}.n_i$$

- ***Tính các đại lượng biến thiên***

Biến thiên nội bộ trong nhóm i

$$SS_i = \sum_{i=1}^{n_i}\left(x_{ij} - \overline{N_i}\right)^2$$

Biến thiên trong nội bộ các nhóm

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

2.2. Kiểm định ANOVA

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

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

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***

$$SSG = \sum_{i=1}^{n_i} n_i \left(\overline{N_i} - \overline{N}\right)^2$$

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.*

2.2. Kiểm định ANOVA

- ***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 + Biên 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 H0**.

2.2. Kiểm định ANOVA

- ***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 → F lớn

So sánh $F > F_{\alpha}(k-1; n-k)$

Bác bỏ giả thuyết H0

2.2. Kiểm định ANOVA

- ***Bảng ANOVA một yếu tố***

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

2.2. Kiểm định ANOVA

Ví dụ 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 bộ. Kết quả ANOVA như sau:

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

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?

2.3. Kiểm định ANOVA – Kiểm định Turkey

- ***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

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

2.3. Kiểm định ANOVA – Kiểm định Turkey

- ***Các bước kiểm định Turkey***

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

$$D_{ij} = \left|\overline{N_i} - \overline{N_j}\right|$$

Bước 2: Tính chỉ số Turkey

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

Bước 3: Bác bỏ H_0 nếu $D_{ij} > T$

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Công cụ Excel***

  Tập dữ liệu ***Insurance Survey***

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?

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Công cụ Excel***

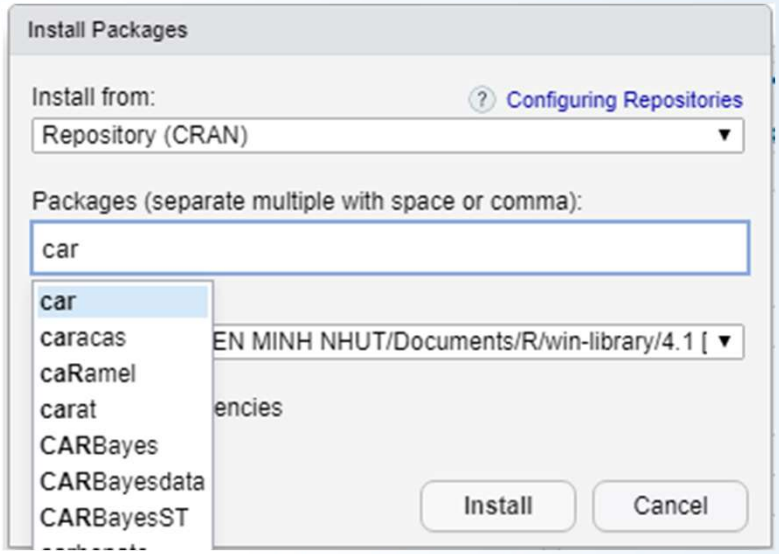
Chia nhóm về độ **Satisfaction*** của **Education**

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		

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ R***

1. Import dữ liệu **Insurance Survey** dạng CSV
2. Cài thêm **Package car** để kiểm Levene Test
3. Thực hiện kiểm định **Levene**



Hàm Levene Test

```
> leveneTest(value,group,center=mean)
```

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ R***

  Hàm Levene Test

```
> leveneTest(value,group,center=mean)
```

Kết quả

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  2  0.9434 0.4052
      21
```

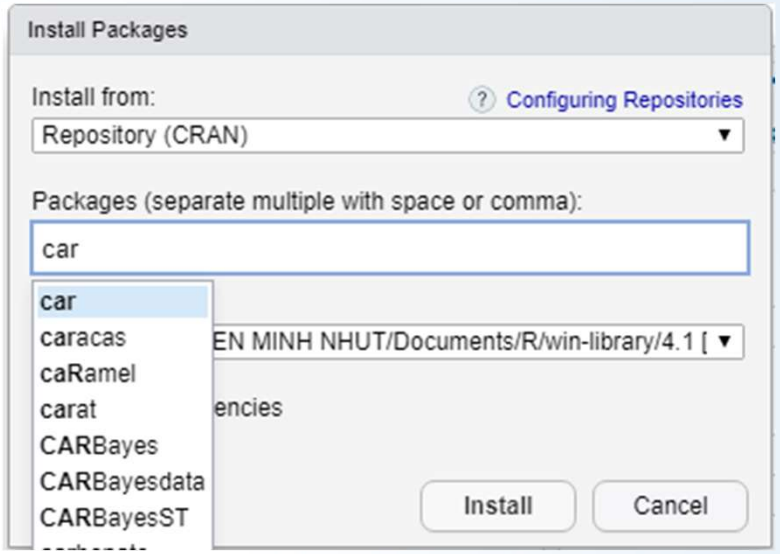
Kết luận

- Vì **`Pr(>F) > 0.05`** nên ta ***chấp nhận*** giả thuyết H_0
- Hoặc có thể giá trị **F-Value** để kiểm định

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ R***

1. Import dữ liệu **Insurance Survey** dạng CSV
2. Cài thêm **Package car** để kiểm Levene Test
3. Thực hiện kiểm định Levene
4. Kiểm định **ANOVA**



Hàm kiểm định ANOVA

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

Hàm tính F-Critical

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

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ R***

  Hàm kiểm định ANOVA

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

Kết quả

```
> rs=aov(Satisfaction.~Education,data=data)
> summary(rs)
             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
>
```

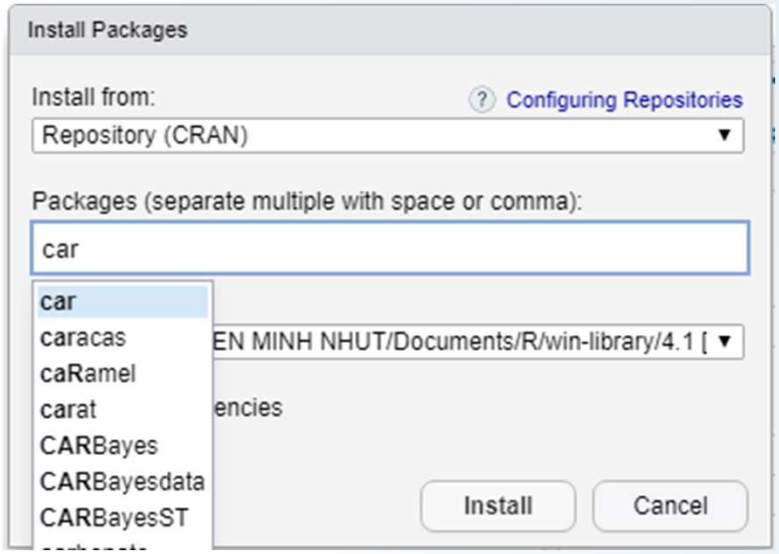
Kết luận

- Vì `Pr(>F) < 0.05` nên ta ***Bác bỏ*** giả thuyết H_0
- Hoặc có thể giá trị **F-Value** để kiểm định

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ R***

1. Import dữ liệu **Insurance Survey** dạng CSV
2. Cài thêm **Package car** để kiểm Levene Test
3. Thực hiện kiểm định Levene
4. Kiểm định **ANOVA**
5. Vì ta bác bỏ giả thuyết H_0 giữa các nhóm. Nên ta sẽ lấy lần lượt 2 nhóm ra so sánh. → Sử dụng ***kiểm định Turkey.***



2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ R***

  Hàm kiểm định ANOVA

  > `TukeyHSD(Result of ANOVA)`

Kết quả

```
> TukeyHSD(rs)
  Tukey multiple comparisons of means
    95% family-wise confidence level

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

$Education
                                       diff         lwr          upr     p adj
Graduate degree-College graduate  1.0555556 -0.1715336  2.28264475 0.1003252
Some college-College graduate    -0.3015873 -1.5742334  0.97105876 0.8230559
Some college-Graduate degree     -1.3571429 -2.6641246 -0.05016107 0.0409193
```

Đưa ra nhận xét?

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ Python***

1. Import dữ liệu **Insurance Survey** dạng CSV
2. Chia nhóm dữ liệu thành ***hàm chia nhóm***

Hàm chia nhóm

```
pythongroup = {}
def chiaGroup(dataframe, group):
    listGroup = dataframe[group].unique().tolist()
    i=1;
    for group_filter in listGroup:
        pythongroup[i] =
dataframe[dataframe[group]==group_filter]
        i = i+1
    return i-1
```

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ Python***

1. Import dữ liệu **Insurance Survey** dạng CSV
2. Chia nhóm dữ liệu thành ***hàm chia nhóm***
3. **Sử dụng hàm Levene của Python để kiểm định Levene**

Kiểm định Levene

```
stat, p = levene(group1, group2, group3,..., center = 'mean')
print(stat, p)
//stat: Giá trị kiểm định Levene W
//p: Giá trị p-value
```

Nhận xét:

- p-value > 0.05 chấp nhận giả thuyết H0 (Có thể kiểm định ANOVA)
- Ngược lại không chấp nhận không thể kiểm định ANOVA

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ Python***

1. Import dữ liệu Insurance Survey dạng CSV
2. Chia nhóm dữ liệu thành *hàm chia nhóm*
3. Sử dụng hàm Levene của Python để kiểm định Levene

4. Kiểm định ANOVA bằng Python thông qua hàm f_oneway

Kiểm định ANOVA

```
import scipy.stats as stats
fvalue, pvalue = stats.f_oneway(group1, group2, group3,...)
print(fvalue, pvalue)
```

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ Python***

1. Import dữ liệu Insurance Survey dạng CSV
2. Chia nhóm dữ liệu thành *hàm chia nhóm*
3. Sử dụng hàm Levene của Python để kiểm định Levene
4. Kiểm định ANOVA bằng Python thông qua hàm f_oneway
5. **Kiểm định Turkey bằng Python thông qua `pairwise_tukeyhsd`**

Kiểm định ANOVA

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

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ Python***

Cách làm khác

```
[32] li_group = [d["Satisfaction* "] for _, d in df.groupby("Education")]
```

```
stat_levene, p_value_levene = stats.levene(*li_group, center="mean")
```

```
[41] print(f"stats levene = {stat_levene}")
     print(f"p_value levene = {p_value_levene}")
```

```
stats levene = 0.9433580072525427
p_value levene = 0.40520616699352924
```

```
[42] fvalue_anova, pvalue_anova = stats.f_oneway(*li_group)
```

```
print(fvalue_anova, pvalue_anova)
```

```
3.9246517319277117 0.03563539756488997
```

2.4. Thực hành kiểm định ANOVA - Insurance Survey

- ***Ngôn Ngữ Python***

Cách làm khác

```
In [7]: df2 = data.groupby('Education')['Satisfaction* '].apply(list)
        print(df2)

        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 [8]: stat, p = levene(*df2,center='mean')

        print(stat, p)

        0.9433580072525427 0.40520616699352924
```

2.5. Kiểm định Chi-quare

- ***Đặt vấn đề bài toán Chi-quare***

1. Kiểm định ***sự độc lập/phụ thuộc*** của hai biến dạng phân loại.
2. Phát biểu bài toán

H_0: Hai biến phân loại là **độc lập**
H_1: Hai biến phân loại là **phụ thuộc**

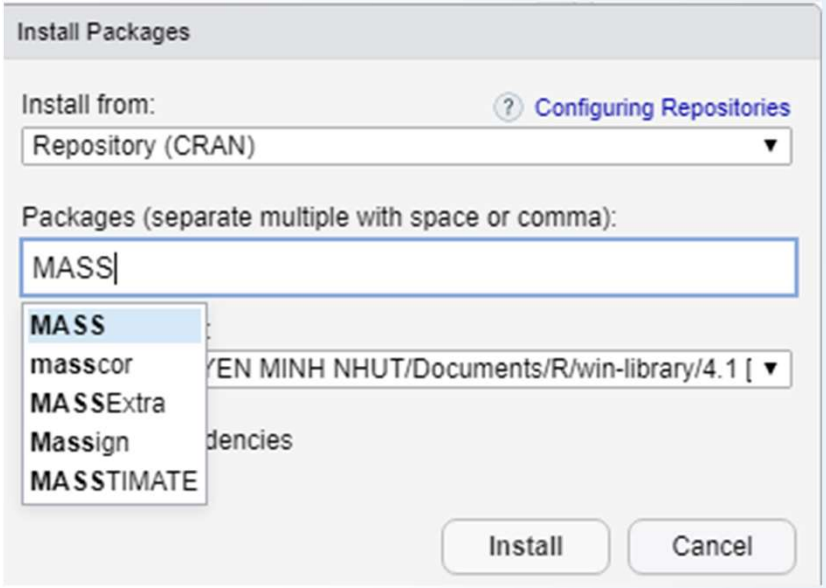
$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

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

2.6. Thực hành kiểm định Chi-quare - Energy Drink Survey

- ***Ngôn ngữ R***

1. Import dữ liệu **Energy Drink Survey** dạng CSV
2. Cài thêm **Package MASS** để kiểm Chi-quare
3. **Thực hiện kiểm định Chi-quare**



Thống kê bảng nhóm cần kiểm định chi-quare

```
> tb = table(group1 ,group2)
```

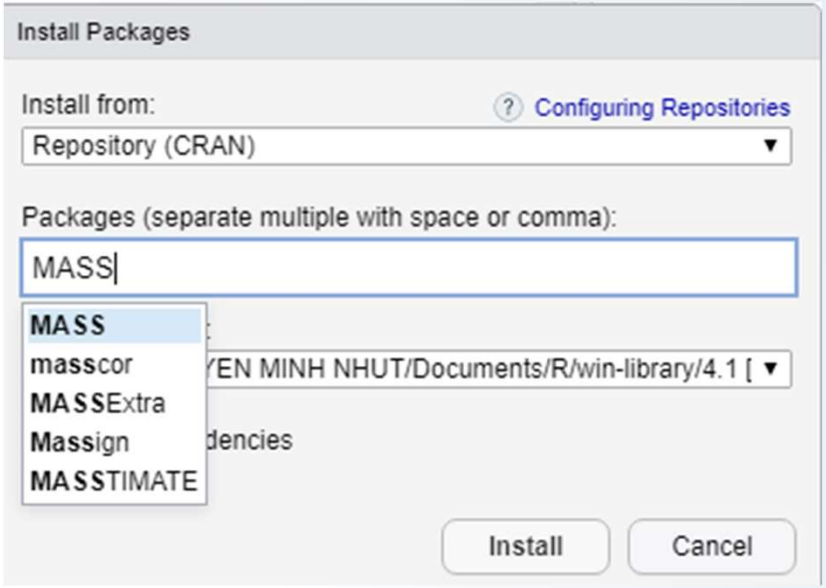
```
> attach(df)
> tb = table(Gender ,Brand.Preference)
> tb
        Brand.Preference
Gender   Brand 1 Brand 2 Brand 3
  Female       9       6      22
  Male        25      17      21
>
```

Tính Expect, Chiquare?

2.6. Thực hành kiểm định Chi-quare - Energy Drink Survey

- ***Ngôn ngữ R***

1. Import dữ liệu **Energy Drink Survey** dạng CSV
2. Cài thêm **Package MASS** để kiểm Chi-quare
3. **Thực hiện kiểm định Chi-quare**



Thống kê bảng nhóm cần kiểm định chi-quare

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

```
> chisq.test(tb)

        Pearson's Chi-squared test

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

>
```

Nhận xét kết quả?

2.6. Thực hành kiểm định Chi-quare - Energy Drink Survey

- ***Ngôn ngữ Python***

1. Import dữ liệu **Energy Drink Survey** dạng CSV
2. **Thực hiện kiểm định Chi-quare**

Thống kê bảng nhóm cần kiểm định chi-quare

```
> chisqt=pd.crosstab(group1, group2, margins = True)
print(chisqt)
```

Ví dụ ta được bảng sau đây:

```
In [118]: chisqt = pd.crosstab(df_chiq.Gender, df_chiq['Brand Preference'], margins=True)
          print(chisqt)

Brand Preference  Brand 1  Brand 2  Brand 3  All
Gender
Female                  9        6       22   37
Male                   25       17       21   63
All                    34       23       43  100
```

2.6. Thực hành kiểm định Chi-quare - Energy Drink Survey

- ***Ngôn ngữ Python***

1. Import dữ liệu **Energy Drink Survey** dạng CSV
2. Thực hiện kiểm định Chi-quare
3. Tiến hành kiểm định Chi-quare bằng

Thống kê bảng nhóm cần kiểm định chi-quare

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

Giải thích các giá trị:

- **c:** The test statistic
- **p:** The p-value of the test
- **dof:** Degrees of freedom
- **expected:** The expected frequencies, based on the marginal sums of the table