

# Bài giảng 16: Giới thiệu phân tích phương sai (analysis of variance)

**Nguyễn Văn Tuấn**

Garvan Institute of Medical Research, Australia  
Đại học Tôn Đức Thắng, Việt Nam

# Nồng độ hormone trong 4 nhóm bệnh nhân

A	B	C	D
8	7	28	26
9	17	21	16
11	10	26	13
4	14	11	12
7	12	24	9
8	24	19	10
5	11		11
	22		17
			15

**Có khác nhau giữa các nhóm về nồng độ hormone ?**

# Chi phí (\$1000) khởi đầu cơ sở buôn bán

pizza	bakers	shoes	gifts	pets
80	150	48	100	25
125	40	35	96	80
35	120	95	35	30
58	75	45	99	35
110	160	75	75	30
140	60	115	150	28
97	45	42	45	20
50	100	78	100	75
65	86	65	120	48
79	87	125	50	20
35	90			50
85				75
120				55
				60
				85
				110

**Chi phí có khác nhau giữa các cơ sở kinh doanh?**

# Nội dung

- Khái niệm biến thiên **giữa** các nhóm và **trong** mỗi nhóm
- Mô hình phân tích phương sai
- Tóm lược

**Between-group** variation và  
**Within-group** variation

# Nghiên cứu về glucose

- Đo nồng độ glucose trong máu của 100 đối tượng gồm 6 nhóm
- So sánh nồng độ glucose giữa các nhóm
- Có  $(5 \times 6)/2 = 15$  kiểm định giả thuyết !
- "Nguy cơ" khám phá dương tính giả (false +ve test)

# ANOVA – analysis of variance

**Ronald A. Fisher**, cha đẻ của  
thống kê hiện đại, nhà di truyền  
học, triết gia, 1920s

*"a genius who almost single-  
handedly created the  
foundations for modern  
statistical science"*



Ronald Fisher (1890 – 1962)

# Ý tưởng của ANOVA

- So sánh một biến liên tục giữa các nhóm (trên 2 nhóm)
- Giả thuyết vô hiệu (null hypothesis)  $H_0: \mu_1 = \mu_2 = \dots = \mu_k$
- Giả thuyết đảo (alternative hypothesis)  $H_a$ : tối thiểu một khác biệt



# Logic của ANOVA: khái niệm *variation*

- Cho một dãy gồm  $n$  giá trị  $X_i$  ( $X_1, X_2, X_3, \dots$ ) một **deviate** được định nghĩa như sau:

$$D = X_i - M$$

- Bình phương của  $D$ :

$$D^2 = (X_i - M)^2$$

- Tổng bình phương (variation):

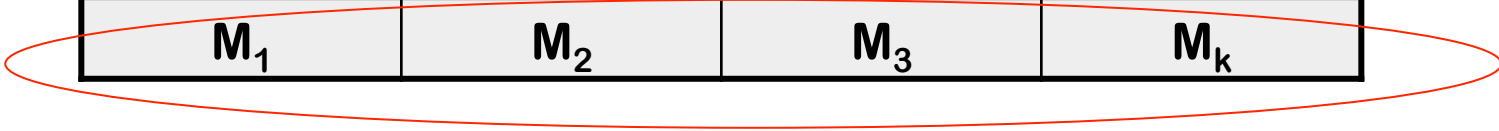
$$\begin{aligned} SS &= (X_1 - M)^2 + (X_2 - M)^2 + (X_3 - M)^2 + \dots + (X_n - M)^2 \\ &= \sum_{i=1}^n (X_i - M)^2 \end{aligned}$$

# Hai loại variation

- Between-group variation – biến thiên / khác biệt **giữa** các nhóm
- Within-group variation – biến thiên **trong** nhóm

# "Between-group" variation

Group 1	Group 2	Group 3	Group k
$X_{11}$	$X_{21}$	$X_{31}$	$X_{k1}$
$X_{12}$	$X_{22}$	$X_{32}$	$X_{k2}$
$X_{13}$	$X_{23}$	$X_{33}$	$X_{k3}$
$X_{14}$	$X_{24}$	$X_{34}$	$X_{k4}$
$X_{15}$	$X_{25}$	$X_{35}$	$X_{k5}$
$X_{16}$	$X_{26}$	$X_{36}$	$X_{k6}$
$M_1$	$M_2$	$M_3$	$M_k$



# "Within-group" variation

Group 1	Group 2	Group 3	Group k
$X_{11}$	$X_{21}$	$X_{31}$	$X_{k1}$
$X_{12}$	$X_{22}$	$X_{32}$	$X_{k2}$
$X_{13}$	$X_{23}$	$X_{33}$	$X_{k3}$
$X_{14}$	$X_{24}$	$X_{34}$	$X_{k4}$
$X_{15}$	$X_{25}$	$X_{35}$	$X_{k5}$
$X_{16}$	$X_{26}$	$X_{36}$	$X_{k6}$
$M_1$	$M_2$	$M_3$	$M_k$

# Logic của ANOVA

The diagram shows a table with 4 columns representing different groups and 6 rows representing individual observations. Blue ovals encircle the data points within each column, representing within-group variation. A red oval encircles the entire table, representing the total variation. A red arrow points to the mean row at the bottom.

Group 1	Group 2	Group 3	Group k
$X_{11}$	$X_{21}$	$X_{31}$	$X_{k1}$
$X_{12}$	$X_{22}$	$X_{32}$	$X_{k2}$
$X_{13}$	$X_{23}$	$X_{33}$	$X_{k3}$
$X_{14}$	$X_{24}$	$X_{34}$	$X_{k4}$
$X_{15}$	$X_{25}$	$X_{35}$	$X_{k5}$
$X_{16}$	$X_{26}$	$X_{36}$	$X_{k6}$
$M_1$	$M_2$	$M_3$	$M_k$

- So sánh **between variation (B)** với **within group variation (W)**
- Nếu  $B > W$ , đó là tín hiệu cho thấy có khác biệt giữa các nhóm.

# Ví dụ

Nồng độ một hormone trong máu của 4 nhóm bệnh nhân

A	B	C	D
8	7	28	26
9	17	21	16
11	10	26	13
4	14	11	12
7	12	24	9
8	24	19	10
5	11		11
	22		17
			15

# Ví dụ: biến thiên **giữa** 4 nhóm

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>
	8	7	28	26
	9	17	21	16
	11	10	26	13
	4	14	11	12
	7	12	24	9
	8	24	19	10
	5	11		11
		22		17
				15
Mean	<b>7.4</b>	<b>14.6</b>	<b>21.5</b>	<b>14.3</b>

**Overall mean = 14.2**

# Ví dụ: biến thiên **giữa** 4 nhóm

	A	B	C	D
Mean	<b>7.4</b>	<b>14.6</b>	<b>21.5</b>	<b>14.3</b>
N	<b>7</b>	<b>8</b>	<b>6</b>	<b>9</b>

**Overall mean = 14.2**

Tổng bình phương giữa các nhóm:

$$\text{SSB} = 7*(7.4 - \mathbf{14.2})^2 + 8*(14.6 - \mathbf{14.2})^2 + 6*(21.5 - \mathbf{14.2})^2 + 9*(14.3 - \mathbf{14.2})^2 = \mathbf{643.9}$$



# Ví dụ: biến thiên **trong** mỗi nhóm

	A	B	C	D
	8	7	28	26
	9	17	21	16
	11	10	26	13
	4	14	11	12
	7	12	24	9
	8	24	19	10
	5	11		11
		22		17
				15
Mean	7.4	14.6	21.5	14.3

Tổng bình phương trong nhóm A:

$$SSW_A = (8 - 7.4)^2 + (9 - 7.4)^2 + \dots + (5 - 7.4)^2 = 33.7$$

# Ví dụ: biến thiên **trong** mỗi nhóm

	A	B	C	D
	8	7	28	26
	9	17	21	16
	11	10	26	13
	4	14	11	12
	7	12	24	9
	8	24	19	10
	5	11		11
		22		17
				15
Mean	7.4	14.6	21.5	14.3

$$SSW_A = (8 - 7.4)^2 + (9 - 7.4)^2 + \dots + (5 - 7.4)^2 = 33.7$$

$$SSW_B = 247.9$$

$$SSW_C = 185.5$$

$$SSW_D = 214.6$$

# Ví dụ: biến thiên **trong** mỗi nhóm

$$SSW_A = (8 - 7.4)^2 + (9 - 7.4)^2 + \dots + (5 - 7.4)^2 = 33.7$$

$$SSW_B = 247.9$$

$$SSW_C = 185.5$$

$$SSW_D = 214.6$$

$$SSW = 33.7 + 247.9 + 185.5 + 214.6 = 681.6$$

# Bảng phân tích phương sai

Nguồn	Degrees of freedom	Sum of squares (SS)	Mean square (MS)
Giữa 4 nhóm		643.9	
Trong các nhóm		681.6	

# Bảng phân tích phương sai

Nguồn	Degrees of freedom	Sum of squares (SS)	Mean square (MS)
Giữa 4 nhóm	3	643.9	214.6
Trong các nhóm	26	681.6	26.2
Tổng số	29	1325.5	

$$\text{F-test} = 214.6 / 26.2 = 8.2$$

# Phân tích bằng R

```
A = c(8, 9, 11, 4, 7, 8, 5)
B = c(7, 17, 10, 14, 12, 24, 11, 22)
C = c(28, 21, 26, 11, 24, 19)
D = c(26, 16, 13, 12, 9, 10, 11, 17, 15)
```

```
x = c(A, B, C, D)
group = c(rep("A", 7), rep("B", 8),
rep("C", 6), rep("D", 9))
```

```
data = data.frame(x, group)
data
```

```
av = aov(x ~ group)
summary(av)
```

# Result

```
> av=aov(x ~ group)
> summary(av)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
group	3	642.3	214.09	8.197	0.000528	***
Residuals	26	679.1	26.12			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05  
'.' 0.1 ' ' 1

**Có sự khác biệt giữa các nhóm**

# Tóm lược

- Phân tích phương sai (ANOVA) kiểm định khác biệt giữa nhiều nhóm ( $>2$ )
- Hàm R cho ANOVA: **aov**

```
analysis = aov(y ~ group)
```



# **Phân tích hậu định (posthoc analysis)**

# Galactose ở 3 nhóm bệnh nhân

9 bệnh nhân Crohn, 11 bệnh nhân viêm đại tràng, và 19 người trong nhóm chứng

Crohn disease	Colitis	Control
1.343	1.264	1.809 2.850
1.393	1.314	1.926 2.964
1.420	1.399	2.283 2.973
1.641	1.605	2.384 3.171
1.897	2.385	2.447 3.257
2.160	2.511	2.479 3.271
2.169	2.514	2.495 3.288
2.279	2.767	2.525 3.358
2.890	2.827	2.541 3.643
	2.895	2.769 3.657
	3.011	

# Câu hỏi nghiên cứu

Có sự khác biệt nào về galactose giữa 3 nhóm bệnh nhân ?

Nếu có khác biệt, nhóm nào khác với nhóm nào?

# Nhập dữ liệu trực tiếp

Crohn disease	Colitis	Control
1.343	1.264	1.809 2.850
1.393	1.314	1.926 2.964
1.420	1.399	2.283 2.973
1.641	1.605	2.384 3.171
1.897	2.385	2.447 3.257
2.160	2.511	2.479 3.271
2.169	2.514	2.495 3.288
2.279	2.767	2.525 3.358
2.890	2.827	2.541 3.643
	2.895	2.769 3.657
	3.011	

```
crohn = c(1.343, 1.393, 1.420, 1.641, 1.897, 2.160, 2.169,  
2.279, 2.890)
```

```
colitis = c(1.264, 1.314, 1.399, 1.605, 2.385, 2.511, 2.514,  
2.767, 2.827, 2.895, 3.011)
```

```
control = c(1.809, 1.926, 2.283, 2.447, 2.479, 2.495, 2.525,  
2.541, 2.769, 2.850, 2.964, 2.973, 3.171, 3.257, 3.271,  
3.288, 3.358, 3.643, 3.657)
```

# Nhập dữ liệu trực tiếp

```
crohn = c(1.343, 1.393, 1.420, 1.641, 1.897, 2.160, 2.169,  
2.279, 2.890)  
colitis = c(1.264, 1.314, 1.399, 1.605, 2.385, 2.511, 2.514,  
2.767, 2.827, 2.895, 3.011)  
control = c(1.809, 1.926, 2.283, 2.447, 2.479, 2.495, 2.525,  
2.541, 2.769, 2.850, 2.964, 2.973, 3.171, 3.257, 3.271,  
3.288, 3.358, 3.643, 3.657)
```

```
gal = c(crohn, colitis, control)  
group=c(rep("Crohn", 9), rep("Colitis",11), rep("Control",  
19))
```

```
dat = data.frame(group, gal)
```

```
dat
```

# Thẩm định số liệu

```
boxplot(gal ~ group, col="Blue")
```

```
require(psych)
```

```
describe.by(gal, group, skew=F)
```

```
qqnorm(gal) ; qqline(gal)
```

# Phân tích ANOVA

```
model = aov(gal ~ group)
```

```
summary(model)
```

```
> summary(model)
Analysis of Variance Table

Response: gal
          Df Sum Sq Mean Sq F value    Pr(>F)
group      2  5.866  2.93301   8.8375 0.0007543 ***
Residuals 36 11.948  0.33188
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

# **Post-hoc analysis**

## **(Phân tích hậu định)**



# Phương pháp phân tích hậu định

- LSD (least significance difference) or Fisher's method
- Bonferroni's method
- Duncan's multiple range test
- Scheffé
- Tukey's Honest Significant Difference
- Dunnett's test

# Phương pháp Tukey' s HSD

- HSD = Honestly Significant Difference

$$Q = \frac{\bar{X}_j - \bar{X}_k}{\sqrt{MSW / \bar{n}}}$$

$\bar{n}$  là số đối tượng (trung bình) cho mỗi nhóm

- Nếu  $Q$  lớn hơn trị số  $Q$  lí thuyết (theoretical Tukey' s Studentized critical value) thì sự khác biệt có ý nghĩa thống kê

# Phương pháp Tukey' s studentized

- Studentized range statistic

$$Q_{k,n-k,\alpha} = \frac{\max \bar{X}_i - \min \bar{X}_i}{\sqrt{WMS}} \sqrt{N}$$

- Khác biệt giữa  $X_1$  và  $X_2$  có ý nghĩa nếu:

$$Q_{ij} = \frac{|\bar{X}_i - \bar{X}_j| \sqrt{N}}{\sqrt{WMS}} > Q_{k,n-k,\alpha}$$

- Khi cỡ mẫu không bằng nhau thì

$$N = 2n_i n_j / (n_i + n_j)$$

# Phương pháp nào thích hợp?

Phương pháp nào cho ra kết quả với **khoảng tin cậy ngắn nhất** là phương pháp tối ưu nhất

# R code - Tukey's Method

```
model = aov(gal ~ group)
```

```
TukeyHSD (model)
```

```
> TukeyHSD(model)
```

```
  Tukey multiple comparisons of means
```

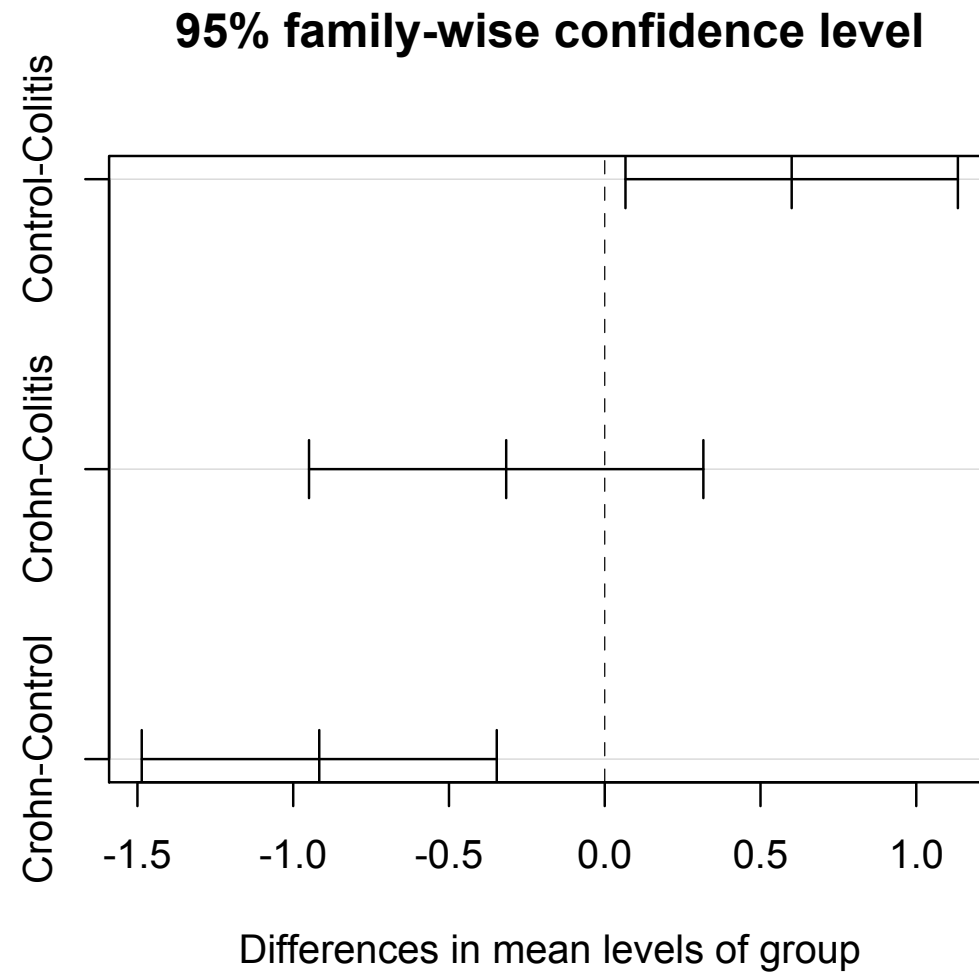
```
  95% family-wise confidence level
```

```
Fit: aov(formula = gal ~ group)
```

```
$group
```

	diff	lwr	upr	p adj
Control-Colitis	0.6000861	0.06658647	1.1335858	0.0245717
Crohn-Colitis	-0.3163232	-0.94923629	0.3165898	0.4483908
Crohn-Control	-0.9164094	-1.48621614	-0.3466026	0.0010497

```
plot(TukeyHSD(model) , ordered=T)
```



# Điều chỉnh cho nhiều so sánh

```
#Bonferroni
```

```
pairwise.t.test(gal, group,  
p.adjust="bonferroni", pool.sd=T)
```

```
# Benjamin-Hochberg
```

```
pairwise.t.test(gal, group, p.adjust="BH",  
pool.sd=T)
```

# Tóm lược

- Có nhiều phương pháp phân tích hậu định (posthoc)
- Phương pháp với khoảng tin cậy ngắn nhất là tối ưu nhất → TukeyHSD ?