11. Phân tích phương sai

11.1 Phân tích phương sai đơn giản (one-way analysis of variance - ANOVA)

Ví dụ 17. Bảng dưới đây so sánh độ galactose trong 3 nhóm bệnh nhân: nhóm 1 gồm 9 bệnh nhân với bệnh Crohn; nhóm 2 gồm 11 bệnh nhân với bệnh viêm ruột kết (colitis); và nhóm 3 gồm 20 đối tượng không có bệnh (gọi là nhóm đối chứng). Câu hỏi đặt ra là độ galactose giữa 3 nhóm bệnh nhân có khác nhau hay không?

Độ galactose cho 3 nhóm bệnh nhân Crohn, viêm ruột kết và đối chứng

Nhóm 1: bệnh Crohn	Nhóm 2: bệnh viêm ruột kết	Nhóm 3: đối chứng (control)
1343	1264	1809 2850
1393	1314	1926 2964
1420	1399	2283 2973
1641	1605	2384 3171
1897	2385	2447 3257
2160	2511	2479 3271
2169	2514	2495 3288
2279	2767	2525 3358
2890	2827	2541 3643
	2895	2769 3657
	3011	
n=9 Trung bình: 1910 SD: 516	n=11 Trung bình: 2226 SD: 727	n=20 Trung bình: 2804 SD: 527

Chú thích: SD là độ lệch chuẩn (standard deviation).

Gọi giá trị trung bình của ba nhóm là μ_1, μ_2, và μ_3, và nói theo ngôn ngữ của kiểm định giả thiết thì giả thiết đảo là:

$$H_o: \mu_1 = \mu_2 = \mu_3$$

Và giả thiết chính là:

$$H_A: \text{có một khác biệt giữa 3 } \mu_j \ (j = 1,2,3)$$

Thoạt đầu có lẽ bạn đọc, sau khi đã học qua phương pháp so sánh hai nhóm bằng kiểm định t, sẽ nghĩ rằng chúng ta cần làm 3 so sánh bằng kiểm định t: giữa nhóm 1 và 2, nhóm 2 và 3, và nhóm 1 và 3. Nhưng phương pháp này không hợp lí, vì có ba phương sai khác nhau. Phương pháp thích hợp cho so sánh là phân tích phương sai. Phân tích phương sai có thể ứng dụng để so sánh nhiều nhóm cùng một lúc (simultaneous comparisons).

Để minh họa cho phương pháp phân tích phương sai, chúng ta phải dùng kí hiệu. Gọi độ galactose của bệnh nhân i thuộc nhóm j (j = 1, 2, 3) là x_{ij}. Mô hình phân tích phương sai phát biểu rằng:

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Hay cụ thể hơn:

$$x_{i1} = \mu + \alpha_1 + \varepsilon_{i1}$$
$$x_{i2} = \mu + \alpha_2 + \varepsilon_{i2}$$
$$x_{i3} = \mu + \alpha_3 + \varepsilon_{i3}$$

Trước hết, chúng ta cần phải nhập dữ liệu vào R. Bước thứ nhất là báo cho R biết rằng chúng ta có ba nhóm bệnh nhân (1, 2 vả), nhóm 1 gồm 9 người, nhóm 2 có 11 người, và nhóm 3 có 20 người:

```
> group <- c(1,1,1,1,1,1,1,1,1, 2,2,2,2,2,2,2,2,2,2,2,
3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3)
```

Để phân tích phương sai, chúng ta phải định nghĩa biến `group` là một yếu tố - factor.

```
> group <- as.factor(group)
```

Bước kế tiếp, chúng ta nạp số liệu galactose cho từng nhóm như định nghĩa trên (gọi object là `galactose`):

```
> galactose <- c(1343,1393,1420,1641,1897,2160,2169,2279,2890,
1264,1314,1399,1605,2385,2511,2514,2767,2827,2895,3011,
1809,2850,1926,2964,2283,2973,2384,3171,2447,3257,2479,3271,2495,3288,
2525,3358,2541,3643,2769,3657)
```

Đưa hai biến `group` và `galactose` vào một dataframe và gọi là `data`:

```
> data <- data.frame(group, galactose)
> attach(data)
```

Sau khi đã có dử liệu sẵn sàng, chúng ta dùng hàm `lm()` để phân tích phương sai như sau:

```
> analysis <- lm(galactose ~ group)
```

Trong hàm trên chúng ta cho R biết biến `galactose` là một hàm số của `group`. Gọi kết quả phân tích là `analysis`.

Kết quả phân tích phương sai. Bây giờ chúng ta dùng lệnh `anova` để biết kết quả phân tích:

```
> anova(analysis)
Analysis of Variance Table

Response: galactose
          Df   Sum Sq  Mean Sq F value    Pr(>F)
group      2  5683620  2841810  8.6655 0.0008191 ***
Residuals 37 12133923   327944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Trong kết quả trên, có ba cột: `Df` (degrees of freedom) là bậc tự do; `Sum Sq` là tổng bình phương (sum of squares), `Mean Sq` là trung bình bình phương (mean square); `F value` là giá trị *F*; và `Pr(>F)` là trị số P liên quan đến kiểm định *F*.

11.2 So sánh nhiều nhóm (multiple comparisons) và điều chỉnh trị số p

Cho *k* nhóm, chúng ta có ít nhất là *k(k-1)*/2 so sánh. Ví dụ trên có 3 nhóm, cho nên tổng số so sánh khả dĩ là 3 (giữa nhóm 1 và 2, nhóm 1 và 3, và nhóm 2 và 3). Khi *k*=10, số lần so sánh có thể lên rất cao. Như đã đề cập trong chương 7, khi có nhiều so sánh, trị số p tính toán từ các kiểm định thống kê không còn ý nghĩa ban đầu nữa, bởi vì các kiểm định này có thể cho ra kết quả dương tính giả (tức kết quả với p<0.05 nhưng trong thực tế không có khác nhau hay ảnh hưởng). Do đó, trong trường hợp có nhiều so sánh, chúng ta cần phải điều chỉnh trị số p sao cho hợp lí.

Có khá nhiều phương pháp điều chỉnh trị số p, và 4 phương pháp thông dụng nhất là: Bonferroni, Scheffé, Holm và Tukey (tên của 4 nhà thống kê học danh tiếng). Phương pháp nào thích hợp nhất? Không có câu trả lời dứt khoát cho câu hỏi này, nhưng hai điểm sau đây có thể giúp bạn đọc quyết định tốt hơn:

(a) Nếu $k < 10$, chúng ta có thể áp dụng bất cứ phương pháp nào để điều chỉnh trị số p. Riêng cá nhân tôi thì thấy phương pháp Tukey thường rất hữu ích trong so sánh.

(b) Nếu $k>10$, phương pháp Bonferroni có thể trở nên rất "bảo thủ". Bảo thủ ở đây có nghĩa là phương pháp này rất ít khi nào tuyên bố một so sánh có ý nghĩa thống kê, dù trong thực tế là có thật! Trong trường hợp này, hai phương pháp Tukey, Holm và Scheffé có thể áp dụng.

Quay lại ví dụ trên, các trị số p trên đây là những trị số chưa được điều chỉnh cho so sánh nhiều lần. Trong chương về trị số p, tôi đã nói các trị số này phóng đại ý nghĩa thống kê, không phản ánh trị số p lúc ban đầu (tức 0.05). Để điều chỉnh cho nhiều so sánh, chúng ta phải sử dụng đến phương pháp điều chỉnh Bonferroni.

Quay lại ví dụ trên, các trị số p trên đây là những trị số chưa được điều chỉnh cho so sánh nhiều lần. Trong chương về trị số p, tôi đã nói các trị số này phóng đại ý nghĩa thống kê, không phản ánh trị số p lúc ban đầu (tức 0.05). Để điều chỉnh cho nhiều so sánh, chúng ta phải sử dụng đến phương pháp điều chỉnh Bonferroni.

Chúng ta có thể dùng lệnh `pairwise.t.test` để có được tất cả các trị số p so sánh giữa ba nhóm như sau:

```
> pairwise.t.test(galactose, group, p.adj="bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  galactose and group
```

```
  1      2
2 0.6805 -
3 0.0012 0.0321

P value adjustment method: bonferroni
```

Kết quả trên cho thấy trị số p giữa nhóm 1 (Crohn) và viêm ruột kết là 0.6805 (tức không có ý nghĩa thống kê); giữa nhóm Crohn và đối chứng là 0.0012 (có ý nghĩa thống kê), và giữa nhóm viêm ruột kết và đối chứng là 0.0321 (tức cũng có ý nghĩa thống kê).

Một phương pháp điều chỉnh trị số p khác có tên là phương pháp Holm:

```
> pairwise.t.test(galactose, group)

        Pairwise comparisons using t tests with pooled SD

data:  galactose and group


  1      2
2 0.2268 -
3 0.0012 0.0214

P value adjustment method: holm
```

Kết quả này cũng không khác so với phương pháp Bonferroni.

Tất cả các phương pháp so sánh trên sử dụng một sai số chuẩn chung cho cả ba nhóm. Nếu chúng ta muốn sử dụng cho từng nhóm thì lệnh sau đây (`pool.sd=F`) sẽ đáp ứng yêu cầu đó:

```
> pairwise.t.test(galactose, group, pool.sd=FALSE)

        Pairwise comparisons using t tests with non-pooled SD

data:  galactose and group

  1      2
2 0.2557 -
3 0.0017 0.0544

P value adjustment method: holm
```

Một lần nữa, kết quả này cũng không làm thay đổi kết luận.

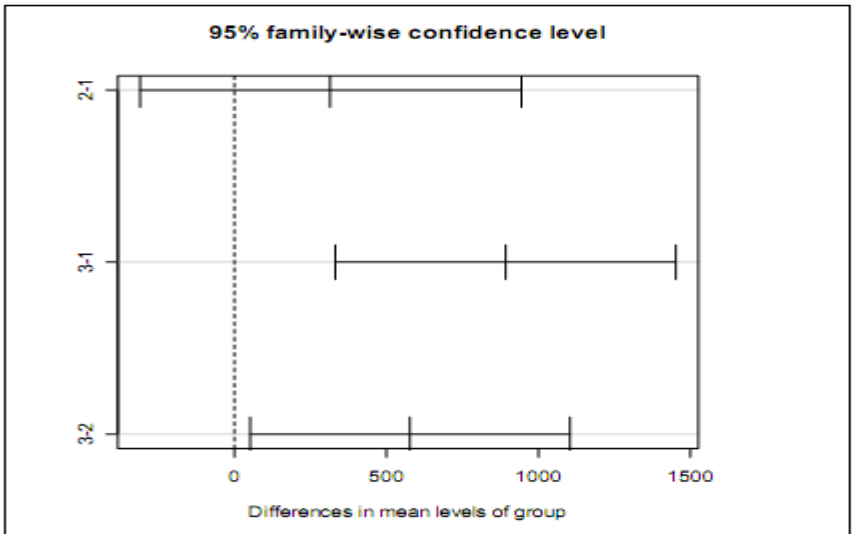
Trong các phương pháp trên, chúng ta chỉ biết trị số p so sánh giữa các nhóm, nhưng không biết mức độ khác biệt cũng như khoảng tin cậy 95% giữa các nhóm. Để có những ước số này, chúng ta cần đến một hàm khác có tên là `aov` (viết tắt từ analysis of variance) và hàm `TukeyHSD` (HSD là viết tắt từ Honest Significant Difference, tạm dịch nôm na là "Khác biệt có ý nghĩa thành thật") như sau:

```
> res <- aov(galactose ~ group)
> TukeyHSD (res)
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

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

$group
        diff         lwr      upr     p adj
2-1 316.3232 -312.09857  944.745 0.4439821
3-1 894.2778  333.07916 1455.476 0.0011445
3-2 577.9545   53.11886 1102.790 0.0281768
```

Kết quả trên cho chúng ta thấy nhóm 3 và 1 khác nhau khoảng 894 đơn vị, và khoảng tin cậy 95% từ 333 đến 1455 đơn vị. Tương tự, galactose trong nhóm bệnh nhân viêm ruột kết thấp hơn nhóm đối chứng (nhóm 3) khoảng 578 đơn vị, và khoảng tin cậy 95% từ 53 đến 1103.



Biểu đồ 24. Trung bình hiệu và khoảng tin cậy 95% giữa nhóm 1 và 2, 1 và 3, và 3 và 2. Trục hoành là độ galactose, trục tung là ba so sánh.

11.3 Phân tích bằng phương pháp phi tham số

Phương pháp so sánh nhiều nhóm phi tham số (non-parametric statistics) tương đương với phương pháp phân tích phương sai là Kruskal-Wallis. Cũng như phương pháp Wilcoxon so sánh hai nhóm theo phương pháp phi tham số, phương pháp Kruskal-Wallis cũng biến đổi số liệu thành thứ bậc (ranks) và phân tích độ khác biệt thứ bậc này giữa các nhóm. Hàm `kruskal.test` trong R có thể giúp chúng ta trong kiểm định này:

```
> kruskal.test(galactose ~ group)

        Kruskal-Wallis rank sum test

data:  galactose by group
Kruskal-Wallis chi-squared = 12.1381, df = 2, p-value = 0.002313
```

Trị số p từ kiểm định này khá thấp (p = 0.002313) cho thấy có sự khác biệt giữa ba nhóm như phân tích phương sai qua hàm `lm` trên đây. Tuy nhiên, một bất tiện của kiểm định phi tham số Kruskal-Wallis là phương pháp này không cho chúng ta biết hai nhóm nào khác nhau, mà chỉ cho một trị số p chung. Trong nhiều trường hợp, phân tích phi tham số như kiểm định Kruskal-Wallis thường không có hiệu quả như các phương pháp thống kê tham số (parametric statistics).

11.4 Phân tích phương sai hai chiều (two-way analysis of variance - ANOVA)

Phân tích phương sai đơn giản hay một chiều chỉ có một yếu tố (factor). Nhưng phân tích phương sai hai chiều (two-way ANOVA), như tên gọi, có hai yếu tố. Phương pháp phân tích phương sai hai chiều chỉ đơn giản khai triển từ phương pháp phân tích phương sai đơn giản. Thay vì ước tính phương sai của một yếu tố, phương pháp phân sai hai chiều ước tính phương sai của hai yếu tố.

Ví dụ 18. Trong ví dụ sau đây, để đánh giá hiệu quả của một kĩ thuật sơn mới, các nhà nghiên cứu áp dụng sơn trên 3 loại vật liệu (1, 2 và 3) trong hai điều kiện (1, 2). Mỗi điều kiện và loại vật liệu, nghiên cứu được lặp lại 3 lần. Độ bền được đo là chỉ số bền bĩ (tạm gọi là `score`). Tổng cộng, có 18 số liệu như sau:

Độ bền bĩ của sơn cho 2 điều kiện và 3 vật liệu

Điều kiện (*i*)	Vật liệu (*j*)		
	1	2	3
1	4.1, 3.9, 4.3	3.1, 2.8, 3.3	3.5, 3.2, 3.6
2	2.7, 3.1, 2.6	1.9, 2.2, 2.3	2.7, 2.3, 2.5

Gọi x_{ij} là `score` của điều kiện i ($i = 1, 2$) cho vật liệu j ($j = 1, 2, 3$). (Để đơn giản hóa vấn đề, chúng ta tạm thời bỏ qua k đối tượng). Mô hình phân tích phương sai hai chiều phát biểu rằng:

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

μ là số trung bình cho toàn quần thể, các hệ số α_i (ảnh hưởng của điều kiện i)và β_j (ảnh hưởng của vật liệu j) cần phải ước tính từ số liệu thực tế. ε_{ij} được giả định tuân theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2.

Để phân tích bằng R, chúng ta cần phải tổ chức dữ liệu sao cho có 4 biến như sau:

```
Condition    Material     Đối tượng   Score
(điều kiện) (vật liệu)
1            1            1           4.1
1            1            2           3.9
1            1            3           4.3
1            2            4           3.1
1            2            5           2.8
1            2            6           3.3
1            3            7           3.5
1            3            8           3.2
1            3            9           3.6
2            1            10          2.7
2            1            11          3.1
2            1            12          2.6
2            2            13          1.9
2            2            14          2.2
2            2            15          2.3
```

```
2           3           16          2.7
2           3           17          2.3
2           3           18          2.5
```

Chúng ta có thể tạo ra một dãy số bằng cách sử dụng hàm `gl` (generating levels).

```
> condition <- gl(2, 9, 18)
> material  <- gl(3, 3, 18)
```

Và tạo nên 18 mã số (từ 1 đến 18):

```
> id <- 1:18
```

Sau cùng là số liệu cho `score`:

```
> score <- c(4.1,3.9,4.3, 3.1,2.8,3.3, 3.5,3.2,3.6,
            2.7,3.1,2.6, 1.9,2.2,2.3, 2.7,2.3,2.5)
```

Tất cả cho vào một dataframe tên là `data`:

```
> data <- data.frame(condition, material, id, score)
> attach(data)
```

Bây giờ số liệu đã sẵn sàng cho phân tích. Để phân tích phương sai hai chiều, chúng ta vẫn sử dụng lệnh `lm` với các thông số như sau:

```
> twoway <- lm(score ~ condition + material)
> anova(twoway)
Analysis of Variance Table

Response: score
          Df Sum Sq Mean Sq F value    Pr(>F)
condition  1 5.0139  5.0139  95.575 1.235e-07 ***
material   2 2.1811  1.0906  20.788 6.437e-05 ***
Residuals 14 0.7344  0.0525
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ba nguồn dao động (variation) của `score` được phân tích trong bảng trên. Qua trung bình bình phương (mean square), chúng ta thấy ảnh hưởng của điều kiện có vẻ quan trọng hơn là ảnh hưởng của vật liệu thí nghiệm. Tuy nhiên, cả hai ảnh hưởng đều có ý nghĩa thống kê, vì trị số p rất thấp cho hai yếu tố. Chúng ta yêu cầu R tóm lược các ước số phân tích bằng lệnh `summary`:

```
> summary(twoway)

Call:
lm(formula = score ~ condition + material)
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.32778 -0.16389  0.03333  0.16111  0.32222


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9778     0.1080  36.841 2.43e-15 ***
condition2   -1.0556     0.1080  -9.776 1.24e-07 ***
material2    -0.8500     0.1322  -6.428 1.58e-05 ***
material3    -0.4833     0.1322  -3.655   0.0026 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.229 on 14 degrees of freedom
Multiple R-Squared: 0.9074,     Adjusted R-squared: 0.8875
F-statistic: 45.72 on 3 and 14 DF,  p-value: 1.761e-07
```

Kết quả trên cho thấy so với điều kiện 1, điều kiện 2 có `score` thấp hơn khoảng 1.056 và sai số chuẩn là 0.108, với trị số p = 1.24e-07, tức có ý nghĩa thống kê. Ngoài ra, so với vật liệu 1, `score` cho vật liệu 2 và 3 cũng thấp hơn đáng kể với độ thấp nhất ghi nhận ở vật liệu 2, và ảnh hưởng của vật liệu thí nghiệm cũng có ý nghĩa thống kê.

Giá trị có tên là "`Residual standard error`" được ước tính từ trung bình bình phương phần dư trong phần (a), tức là $\sqrt{0.0525} = 0.229$, tức là ước số của $\hat{\sigma}$.

Hệ số xác định bội (R^2) cho biết hai yếu tố điều kiện và vật liệu giải thích khoảng 91% độ dao động của toàn bộ mẫu. Hệ số này được tính từ tổng bình phương trong kết quả phần (a) như sau:

$$R^2 = \frac{5.0139 + 2.1811}{5.0139 + 2.1811 + 0.7344} = 0.9074$$

Và sau cùng, hệ số R^2 điều chỉnh phản ánh độ "cải tiến" của mô hình. Để hiểu hệ số này tốt hơn, chúng ta thấy phương sai của toàn bộ mẫu là s^2 = (5.0139 + 2.1811 + 0.7344) / 17 = 0.4644. Sau khi điều chỉnh cho ảnh hưởng của điều kiện và vật liệu, phương sai này còn 0.0525 (tức là residual mean square). Như vậy hai yếu tố này làm giảm phương sai khoảng 0.4644 – 0.0525 = 0.4119. Và hệ số R^2 điều chỉnh là:

$$\text{Adj } R^2 = 0.4119 / 0.4644 = 0.88$$

Tức là sau khi điều chỉnh cho hai yếu tố điều kiện và vật liệu phương sai của score giảm khoảng 88%.

Hiệu ứng tương tác

Để cho phân tích hoàn tất, chúng ta còn phải xem xét đến khả năng ảnh hưởng của hai yếu tố này có thể tương tác nhau (interactive effects). Tức là mô hình `score` trở thành:

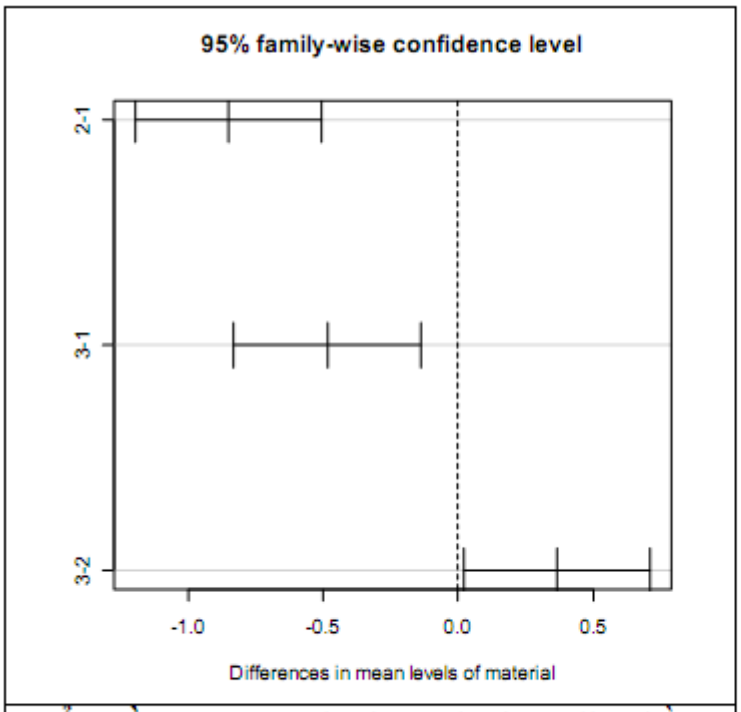
$$x_{ij} = \mu + \alpha_i + \beta_j + \left(\alpha_i \beta_j\right)_{ij} + \varepsilon_{ij}$$

Chú ý phương trình trên có phần $\left(\alpha_i \beta_j\right)_{ij}$ phản ánh sự tương tác giữa hai yếu tố. Và chúng ta chỉ đơn giản lệnh R như sau:

```
> anova(twoway <- lm(score ~ condition+ material+condition*material))
Analysis of Variance Table

Response: score
                   Df Sum Sq Mean Sq  F value    Pr(>F)
condition           1 5.0139  5.0139 100.2778 3.528e-07 ***
material            2 2.1811  1.0906  21.8111 0.0001008 ***
condition:material  2 0.1344  0.0672   1.3444 0.2972719
Residuals          12 0.6000  0.0500
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kết quả phân tích trên (p = 0.297 cho ảnh hưởng tương tác). Chúng ta có bằng chứng để kết luận rằng ảnh hưởng tương tác giữa vật liệu và điều kiện không có ý nghĩa thống kê, và chúng ta chấp nhận mô hình [4], tức không có tương tác.



Biểu đồ 25. So sánh giữa 3 loại vật liệu bằng phương pháp Tukey.