THỐNG KÊ ỨNG DỤNG

ĐỖ LÂN

dolan@tlu.edu.vn
Đại học Thủy Lợi

Ngày 30 tháng 9 năm 2018

Nội dung môn học

1. Tổng quan về Thống kê
2. Thu thập dữ liệu
3. Tóm tắt và trình bày dữ liệu bằng bảng và đồ thị
4. Tóm tắt dữ liệu bằng các đại lượng thống kê mô tả
5. Xác suất căn bản và biến ngẫu nhiên
6. Phân phối của tham số mẫu và ước lượng tham số tổng thể
7. Kiểm định giả thuyết về tham số một tổng thể
8. Kiểm định giả thuyết về tham số hai tổng thể
9. **Phân tích phương sai**
10. Kiểm định phi tham số
11. Kiểm định chi - bình phương

Phần VIII

Phân tích phương sai

Mục lục

Đặt vấn đề

1. Đã tìm hiểu:
   - kiểm định so sánh trung bình của một tổng thể với một số.
   - kiểm định giả thiết so sánh trung bình hai tổng thể.
     $\longrightarrow$ Những điều kiện khác nhau thì thống kê được sử dụng là khác nhau.
2. Xét xem một nhân tố nào đó có ảnh hưởng lên một biến định lượng không?
3. Ta sẽ chỉ xét sự tác động của một nhân tố $\rightarrow$ so sánh trung bình của các nhóm chịu các tác động khác nhau của nhân tố đang xét.
   $\longrightarrow$ cần những giả thiết ngặt nghèo hơn so với những phần trước đây: các tổng thể phân phối chuẩn, có cùng phương sai, mẫu chọn ra phải độc lập.

Bài toán tình huống

Example

Gần đây, nhiều bạn sinh viên trong trường cho rằng: Thời gian học ở nhà không ảnh hưởng đến kết quả học tập! Tức là học nhiều hay ít thì điểm cũng vẫn vậy!

Bài toán tình huống

Example

Gần đây, nhiều bạn sinh viên trong trường cho rằng: Thời gian học ở nhà không ảnh hưởng đến kết quả học tập! Tức là học nhiều hay ít thì điểm cũng vẫn vậy!
Để kiểm tra ý kiến này, thầy Hiệu trưởng yêu cầu điều tra điểm thi của 3 nhóm sinh viên:
+ Nhóm I: Thời gian tự học ít
+ Nhóm II: Thời gian tự học Trung bình
+ Nhóm III: Thời gian tự học nhiều.

Bài toán tình huống 1

Nhóm I (TG tự học ít)	Nhóm II (TG tự học TB)	Nhóm III (TG tự học nhiều)
5.8	6.0	6.2
6.2	6.6	5.8
5.4	6.1	6.5
6.0	5.8	6.2
5.2	5.9	6.4
5.3	6.0	5.7
5.4	5.9	6.1
5.6	6.0	6.8
6.2	6.7	7.1
5.7	6.5	6.5
5.5	6.3	7.1
6.1	6.1	7.2
6.0	6.8	6.7
5.2	6.4	7.0
6.4	6.8	7.6
5.5	6.6	7.7
5.0	6.4	7.8
5.6	6.2	6.8
6.2	7.1	7.3
6.1	7.0	7.1
5.3	7.2	7.2

Bài toán tình huống

Câu hỏi

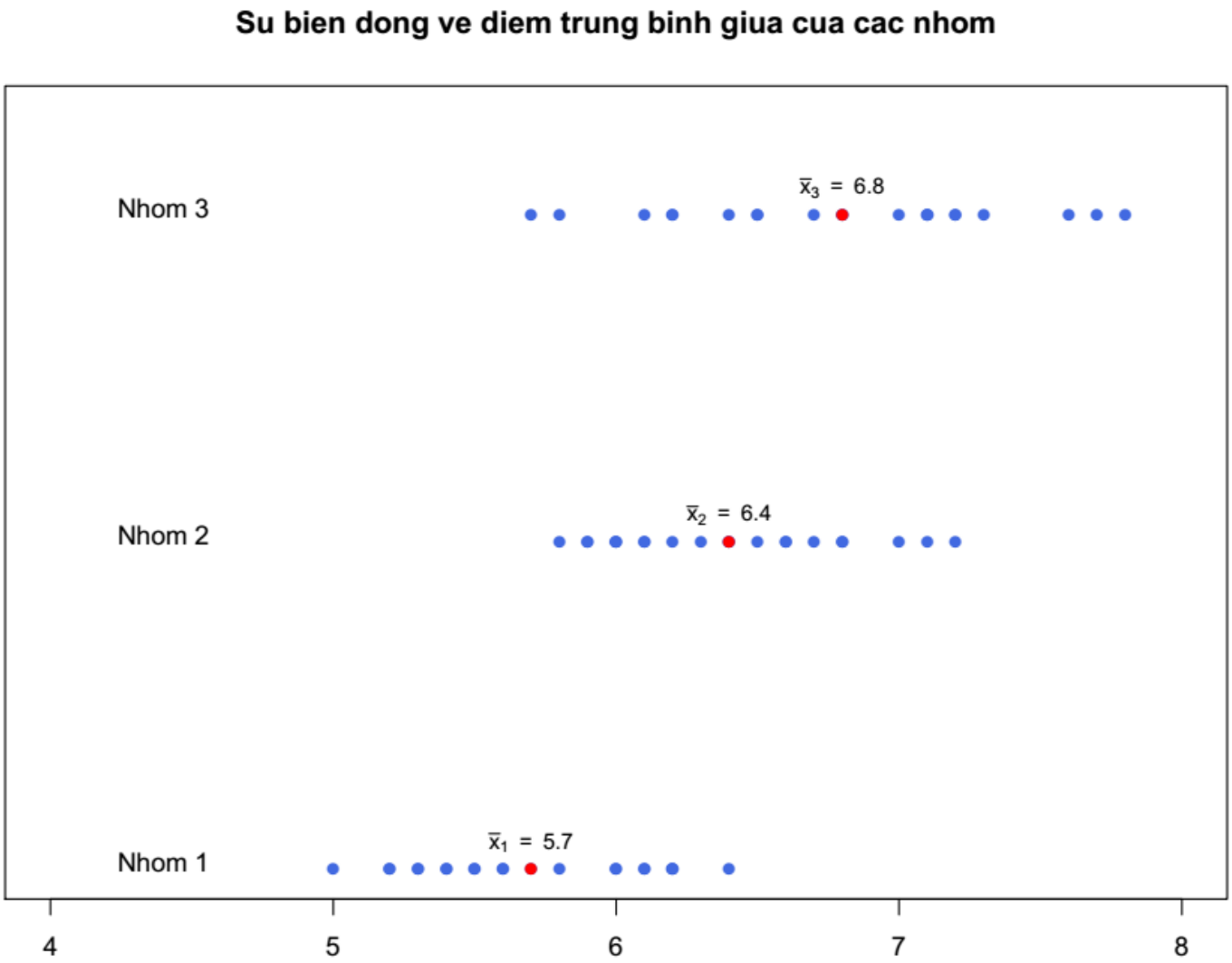
- Để xem xét thời gian tự học khác nhau ảnh hưởng đến kết quả học tập hay không, ta đi kiểm định bài toán nào?

Bài toán tình huống

Câu hỏi

- Để xem xét thời gian tự học khác nhau ảnh hưởng đến kết quả học tập hay không, ta đi kiểm định bài toán nào?
- Làm thế nào để kiểm định bài toán trên?

Minh họa sự biến động điểm Trung bình của 3 nhóm



Bài toán tình huống 2

Example

Để xem liệu điều kiện kinh tế khác nhau có ảnh hưởng đến số con trong một gia đình hay không, người ta thu phân loại ra 3 mức về điều kiện kinh tế: Trên mức khá giả, khá giả, dưới mức khá giả. Sau đó chọn ngẫu nhiên ở mỗi loại 4 gia đình và ghi lại số con của các gia đình như sau:

Trên mức khá giả	Khá giả	Dưới mức khá giả
2	1	3
3	2	4
3	1	2
2	2	3

Bài toán tình huống 2

Example

Để xem liệu điều kiện kinh tế khác nhau có ảnh hưởng đến số con trong một gia đình hay không, người ta thu phân loại ra 3 mức về điều kiện kinh tế: Trên mức khá giả, khá giả, dưới mức khá giả. Sau đó chọn ngẫu nhiên ở mỗi loại 4 gia đình và ghi lại số con của các gia đình như sau:

Trên mức khá giả	Khá giả	Dưới mức khá giả
2	1	3
3	2	4
3	1	2
2	2	3

- Liệu các con số đó có cho ta thấy số con trung bình của các hộ thuộc diện kinh tế khác nhau là như nhau không?

Bài toán tình huống 2

Example

Để xem liệu điều kiện kinh tế khác nhau có ảnh hưởng đến số con trong một gia đình hay không, người ta thu phân loại ra 3 mức về điều kiện kinh tế: Trên mức khá giả, khá giả, dưới mức khá giả. Sau đó chọn ngẫu nhiên ở mỗi loại 4 gia đình và ghi lại số con của các gia đình như sau:

Trên mức khá giả	Khá giả	Dưới mức khá giả
2	1	3
3	2	4
3	1	2
2	2	3

- Liệu các con số đó có cho ta thấy số con trung bình của các hộ thuộc diện kinh tế khác nhau là như nhau không?
- Điều kiện kinh tế ảnh hưởng như thế nào tới số con trong gia đình?

Khái niệm

Phân tích phương sai một nhân tố (One-way ANOVA) là phân tích ảnh hưởng của một yếu tố nguyên nhân (dạng biến định tính) đến một yếu tố kết quả (dạng biến định lượng) đang nghiên cứu.

Phân tích vấn đề

Ta gọi số con trung bình của các hộ thuộc mỗi nhóm tương ứng là m_1, m_2, m_3. Cặp giả thuyết sẽ là:

$$H_0 : m_1 = m_2 = m_3 \qquad H_1 : \exists i, j \in \{1, 2, 3\} : m_i \neq m_j;$$

1. Nếu H_0 đúng, tính trung bình mẫu của mỗi nhóm thì chúng có khác nhau không?
2. Nếu H_0 sai, tức là điều kiện kinh tế có ảnh hưởng đến quyết định sinh con ít hay nhiều, thì dấu hiệu nào đặc trưng cho điều này?
   - Số con trong mỗi mẫu của mỗi nhóm là biến động.
   - Nếu sự biến động đó là khác biệt nhiều giữa các nhóm khác nhau $\rightarrow$ sự ảnh hưởng của điều kiện kinh tế lên quyết định sinh con càng mạnh.
   - Nhưng bản thân nếu H_0 đúng thì sự biến động trong bản thân mỗi nhóm là vẫn có. Làm sao để nhận biết được sự biến động do sự khác biệt từ mỗi nhóm là đáng kể hay không?

Đo mức độ khác biệt giữa các nhóm

Người ta đo sự biến động giữa các nhóm như nào?

1. Lấy giá trị đại diện cho mỗi nhóm là trung bình: $\overline{x}_1 = 2.5, \overline{x}_2 = 1.5, \overline{x}_3 = 3$.
2. Tính trung bình chung: $\overline{x} = 7/3$.
3. Sự biến động đặc trưng cho sự biến động quanh trung bình chung:
   $SSG = 4(7/3 - 1.5)^2 + 4(7/3 - 2.5)^2 + 4(7/3 - 3)^2 = 14/3$.
   $\rightarrow$ Đại lượng này phản ánh sự chệnh lệch gây nên bởi điều kiện kinh tế khác nhau.

Đo mức độ biến động vốn có

Tính độ biến động trong mỗi nhóm như thế nào?
$\rightarrow$ Tính chênh lệch quanh trung bình trong bản thân mỗi nhóm:

- Nhóm kinh tế trên mức khá giả:
  $(2-2.5)^2+(3-2.5)^2+(3-2.5)^2+(2-2.5)^2=1.$
- Nhóm kinh tế khá giả:
  $(1-1.5)^2+(2-1.5)^2+(1-1.5)^2+(2-1.5)^2=1.$
- Nhóm dưới mức khá giả:
  $(3-3)^2+(4-3)^2+(2-3)^2+(3-3)^2=2$

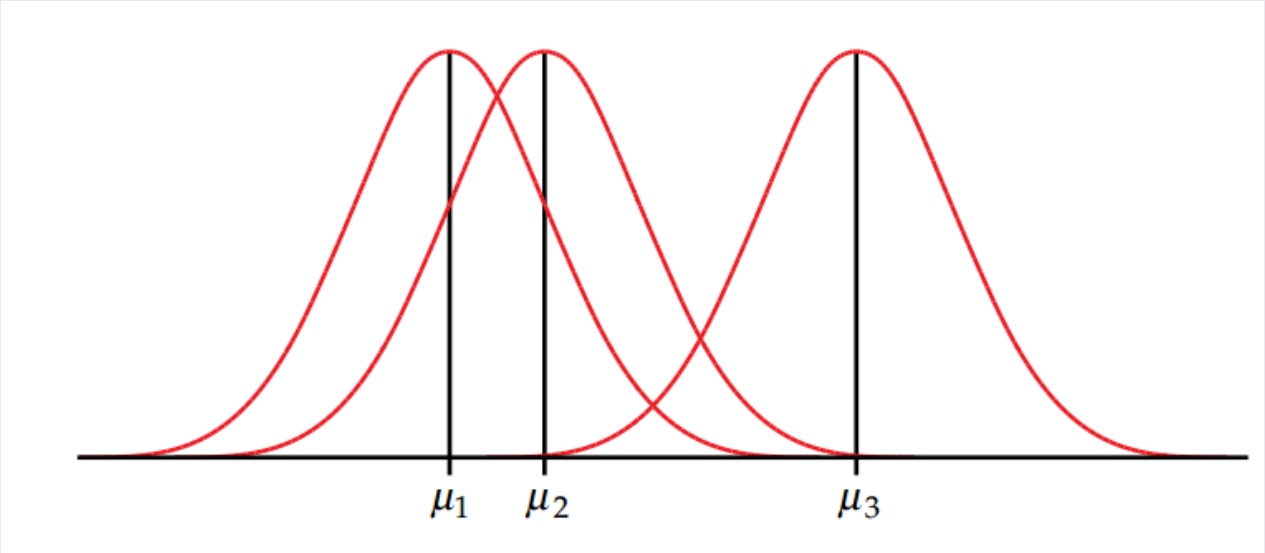
Tổng của sự biến động này là $SSW=4$
$\longrightarrow$ phản ánh tổng sự biến động vốn có trong nội bộ các nhóm, không phải do sự khác nhau về nhóm gây ra.

Mường tượng ý tưởng

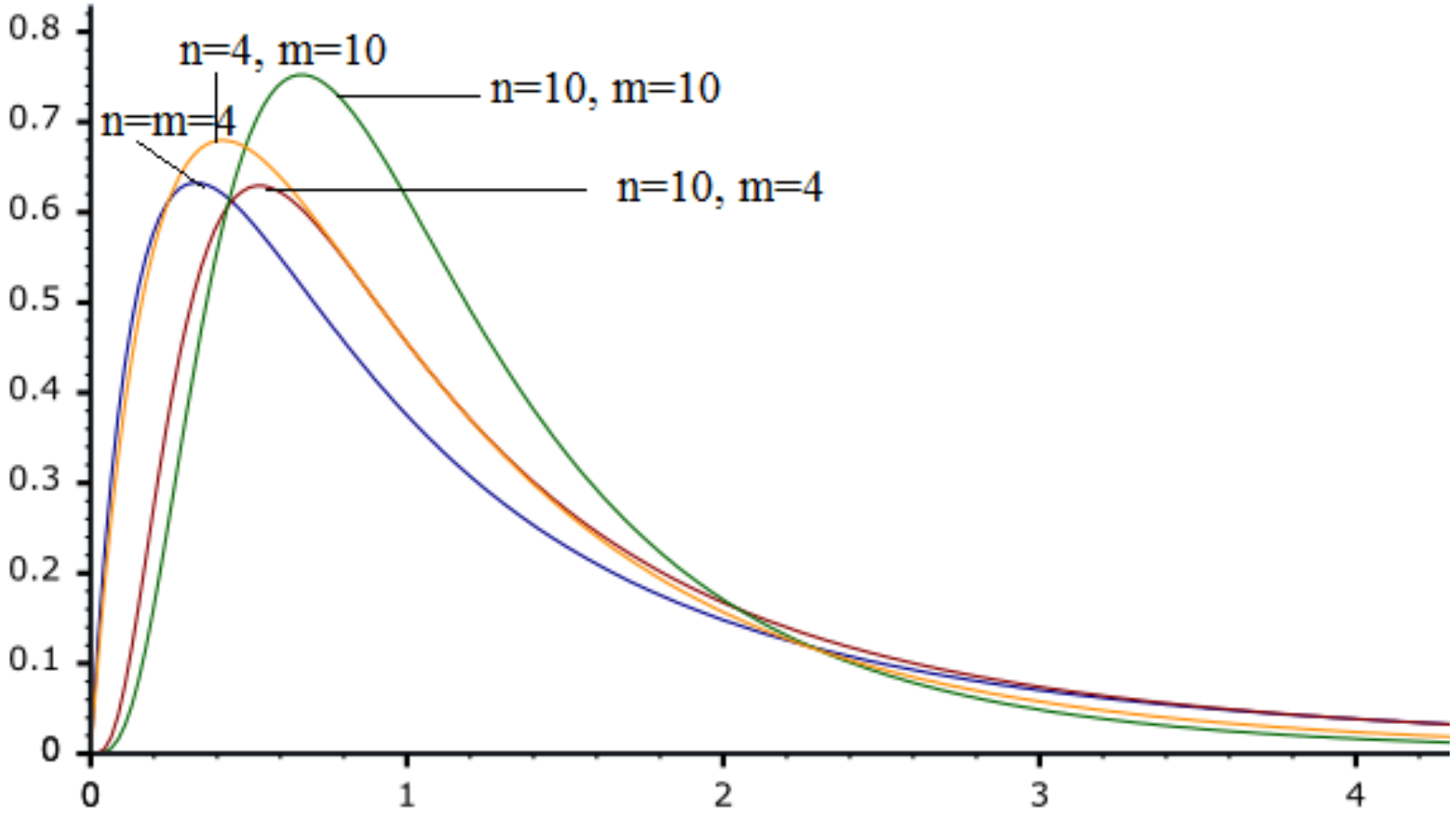
Tỉ lệ SSG/SSW càng lớn thì các yếu tố về điều kiện kinh tế sẽ tác động rõ rệt lên số con của các hộ gia đình.
Tuy nhiên, để có mốc so sánh thì ta cần xây dựng một thống kê có quy luật rõ ràng mà ta có thể xác định được phân bố của nó.

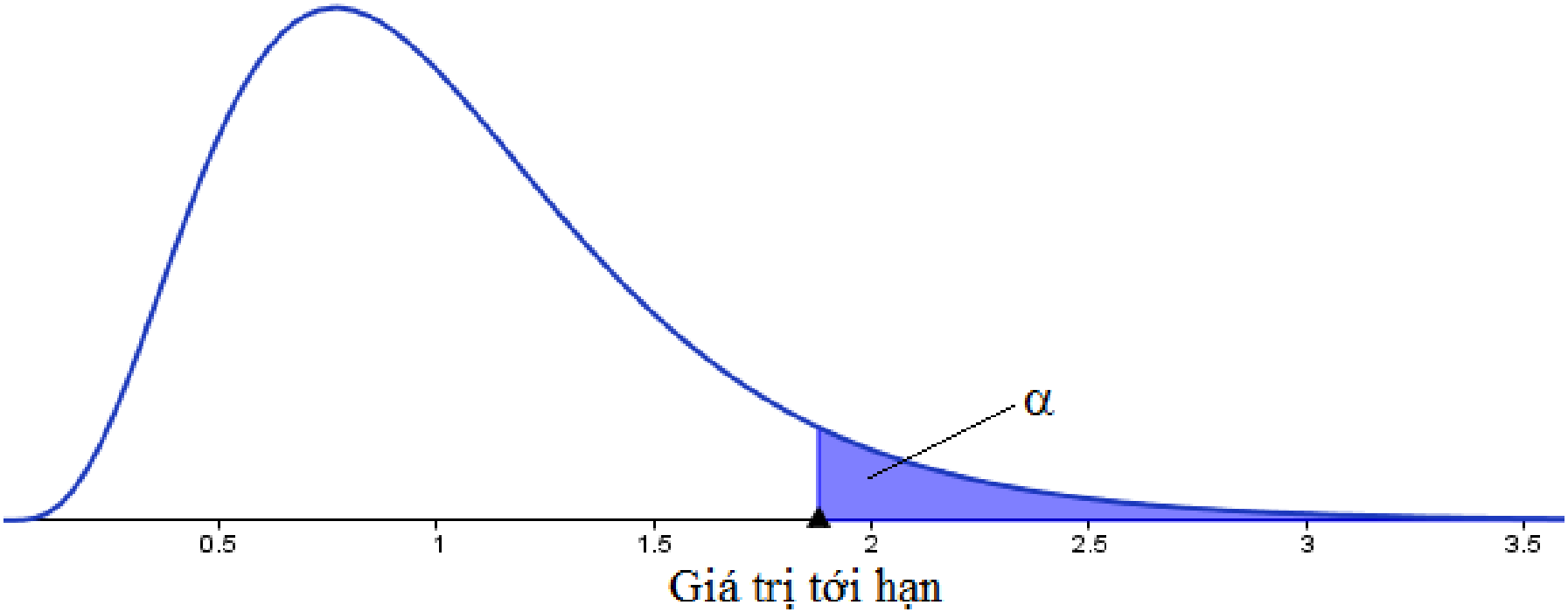
Theorem

Khi mỗi tổng thể đang xét có phân bố chuẩn với cùng phương sai và mẫu chọn ra là độc lập:

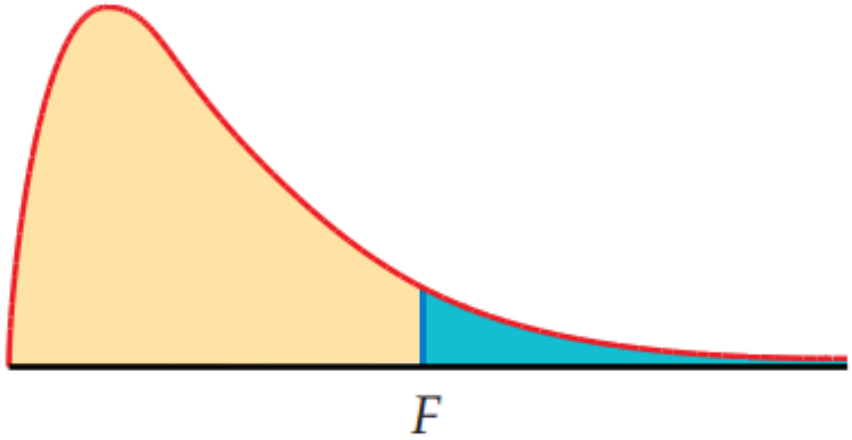


thì đại lượng: $F = \dfrac{MSG}{MSW} = \dfrac{SSG/(k-1)}{SSW/(n-k)}$ *tuân theo phân phối Fisher* $F_{k-1,n-k}$.


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
n=4, m=10
n=10, m=10
n=m=4
n=10, m=4


α
0.5
1
1.5
2
2.5
3
3.5
Giá trị tới hạn

$$F = \frac{\text{MSG}}{\text{MSE}}$$



Quy tắc bác bỏ

Vì thế khi tỉ số F tính ra lớn đến độ hiếm khi có, tức là rơi vào vùng từ $F_{k-1,n-k,\alpha} = qf(1-\alpha, k-1, n-k)$ trở lên thì ta bác bỏ H_0 (với α là mức ý nghĩa), vì nếu H_0 đúng thì khả năng F rơi vào vùng này chỉ là α. Nên khi bác bỏ H_0, sai lầm của ta là không quá α.
Như ta đã biết trong kiểm định so sánh phương sai hai tổng thể, ta cũng có thể tính P - giá trị $= P(F_{k-1,n-k} > F)$ và nếu nó nhỏ hơn α thì ta sẽ bác bỏ H_0.

Bài toán tổng quát

Bài toán

Giả sử có k tổng thể **tuân theo phân phối chuẩn, phương sai bằng nhau** *với trung bình lần lượt là* $\mu_1, \mu_2, \ldots, \mu_k$. *Ta cần so sánh trung bình của k tổng thể này dựa trên những mẫu* **ngẫu nhiên độc lập** *chọn ra từ k tổng thể này bằng cách kiểm định cặp giả thuyết*

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

$$H_1 : \exists i \neq j : \mu_i \neq \mu_j, i, j = \overline{1, k}.$$

Quy trình kiểm định

- Bước 1: Tính các trung bình mẫu.
  - Giả sử ta có k mẫu với số phần tử lần lượt là $n_1, n_2, \ldots, n_k$ chọn từ k tổng thể được cho ở bảng dưới đây:

1	2	3	k
x_{11}	x_{21}	$\ldots$	x_{k1}
x_{12}	x_{22}	$\ldots$	x_{k2}
$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_{1n_1}	x_{2n_2}	$\ldots$	x_{kn_k}

  - Trung bình mẫu của từng nhóm $x_1, x_2, \ldots, x_n$ theo công thức:

$$\overline{x}_i = \frac{x_{i1} + x_{i2} + \ldots + x_{in_i}}{n_i}.$$

  - Trung bình của k mẫu (mẫu gộp) $\overline{x}$ theo công thức:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \ldots + n_k\overline{x}_k}{n_1 + n_2 + \ldots + n_k}.$$

Quy trình kiểm định

- Bước 2: Tính tổng các chênh lệch bình phương
  - Tổng bình phương trong nội bộ nhóm SSW được tính bởi:

$$SSW = SS_1 + SS_2 + \ldots + SS_k,$$

trong đó, SS_i là tổng bình phương của từng nhóm được tính bởi công thức:

$$SS_i = (x_{i1} - \overline{x}_i)^2 + (x_{i2} - \overline{x}_i)^2 + \ldots + (x_{in_i} - \overline{x}_i)^2.$$

  - Tổng bình phương giữa các nhóm SSG được tính bởi công thức

$$SSG = n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \ldots + n_k(\overline{x}_k - \overline{x})^2.$$

  - Tổng bình phương toàn bộ SST được tính bởi công thức

$$SST = (x_{11} - \overline{x})^2 + \ldots + (x_{1n_1} - \overline{x})^2 + \ldots + (x_{k1} - \overline{x})^2 + \ldots + (x_{kn_1} - \overline{x})^2.$$

Ta có $SST = SSW + SSG$.

Quy trình kiểm định

- Bước 3: Tính các phương sai.
  - Phương sai trong nội bộ nhóm MSW được tính bởi công thức

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

  - Phương sai giữa các nhóm MSG được tính bởi công thức

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

- Bước 4: Kiểm định giả thuyết
  - Đặt $F = \frac{MSG}{MSW}$, khi đó F tuân theo phân phối Fisher với $k-1$ bậc tự do ở tử và $n-k$ bậc ở mẫu.
  - Bác bỏ giả thuyết H_0 tại mức ý nghĩa α nếu $F > F_{k-1,n-k,\alpha}$ hoặc P - giá trị $= P(F_{k-1,n-k} > F) < \alpha$.

Bảng phân tích phương sai

Tóm tắt qua bảng phân tích phương sai

Nguồn biến thiên	Tổng bình phương	Bậc tự do (df)	Phương sai (MS)	Tỉ số F	p - giá trị
Giữa các nhóm	SSG	k-1	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSW}$	$P(F_{k-1,n-k} > F)$
Nội bộ các nhóm	SSW	n-k	$MSW = \frac{SSW}{n-k}$		
Toàn bộ	SST	n-1			

Trở lại ví dụ

1. Cặp giả thuyết:

$$H_0 : m_1 = m_2 = m_3 \qquad H_1 : \exists i, j \in \{1, 2, 3\} \text{ để: } m_i \neq m_j;$$

2. Tính tổng các chênh lệch bình phương giữa các nhóm với trung bình có trọng số: $SSG = 14/3$.
   Tính tổng chênh lệch bình phương so với trung bình trong bản thân mỗi nhóm: $SSW = 4$.
3. Tính các phương sai:
$$MSG = \frac{SSG}{3-1} = \frac{14}{6}, MSW = \frac{SSW}{12-3} = \frac{4}{9}$$
4. Giá trị kiểm định $F = \frac{MSG}{MSW} = 5.25$.
   P - giá trị $= P(F_{2,9} > 5.25) = 1 - pf(5.25, 2, 9) = 0.03083$.
5. Do P - giá trị $< \alpha = 0.05$ nên ta bác bỏ H_0.
6. Vậy ở mức ý nghĩa 5%, số con trung bình của các hộ có điều kiện kinh tế khác nhau thì không như nhau.

Bảng phân tích phương sai

Nguồn biến thiên	Tổng BP	Bậc tự do	Phương sai	Tỉ số F	P - giá trị
Giữa các nhóm	14/3	2	14/6	5.25	0.03083
Nội bộ các nhóm	4	9	4/9		
Toàn bộ	26/3	11			

Bài toán

Trong bài toán so sánh nhiều trung bình, khi giả thuyết H_0 bị bác bỏ có nghĩa là kết luận trung bình của các tổng thể không bằng nhau. Ta cần phân tích sâu hơn (phân tích sâu ANOVA) để xác định trung bình của tổng thể nào khác tổng thể nào, trung bình của tổng thể nào lớn hơn hay nhỏ hơn.

Một số phương pháp phân tích sâu

Có nhiều phương pháp để tiếp tục phân tích sâu ANOVA khi bác bỏ giả thuyết H_0, chẳng hạn như phương pháp so sánh trực giao (Orthogonal comparison), phương pháp Student-Newman-Keuls, phương pháp Tukey, kiểm định đa khoảng Duncan (Duncans Multiple Range Test), kiểm định Scheffé (Scheffé Test) hay phương pháp khác biệt nhỏ nhất có ý nghĩa (Least-Significant Difference: LSD) ...Ở đây chúng ta sẽ tìm hiểu phương pháp khá thông dụng đó là phương pháp Tukey, phương pháp này còn được gọi là kiểm định HSD (Honestly Significant Differences).

Phương pháp TukeyHSD

Phương pháp Tukey dùng một thống kê tuân theo phân phối khoảng (còn gọi là phân phối q) trên cơ sở phân phối Student t với bậc tự do k và (n - k) để kiểm định (k là số tổng thể, n là tổng số quan sát).

Quy trình phân tích sâu

1. Giả sử cần so sánh trung bình của k tổng thể, khi đó ta cần so sánh trung bình của C_k^2 cặp tổng thể:

$$H_0 : \mu_i = \mu_j; H_1 : \mu_i \neq \mu_j, \forall i \neq j, i, j = \overline{1, k}.$$

2. Giá trị tới hạn Tukey được tính theo công thức:

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

   - n_{min} là số quan sát nhỏ nhất trong các mẫu chọn ra quan sát;
   - MSW là phương sai trong nội bộ nhóm;
   - $q_{\alpha,k,n-k}$ là giá trị của phân phối kiểm định Tukey tại mức ý nghĩa α, với bậc tự do k và n-k, n là tổng số quan sát $n = \sum n_i$.

3. Tiêu chuẩn quyết định là bác bỏ giả thuyết H_0 khi độ lệch tuyệt đối giữa các cặp trung bình mẫu lớn hơn hay bằng T giới hạn.
4. Đưa ra các kết luận.

Example

1. Các cặp giả thiết: $H_0 : m_i = m_j$, $H_1 : m_i \neq m_j$ với i, j lấy giá trị khác nhau trong tập $\{1, 2, 3\}$. Như vậy ta có 3 cặp giả thuyết.
2. Ta có $k = 3, \alpha = 5\%, n = 12$ và $MSW = 4/9$.
3. Giá trị q của phân phối Tukey:
   $q_{0.05,3,9} = qtukey(0.05, 3, 9, lower.tail = FALSE) \approx 3.95$.
4. Giá trị tới hạn: $T = 3.95\sqrt{\dfrac{4/9}{4}} \approx 1.32$.
5. Độ lệch tuyệt đối của các cặp trung bình mẫu tính được lần lượt như sau:
   - $|\overline{x}_1 - \overline{x}_2| = |2.5 - 1.5| = 1$;
   - $|\overline{x}_1 - \overline{x}_3| = |2.5 - 3| = 0.5$;
   - $|\overline{x}_2 - \overline{x}_3| = |1.5 - 3| = 1.5$.
6. Với T = 1.32, qui tắc bác bỏ H_0 cho ta các quyết định sau:
   - m_1 và m_2 như nhau vì $|\overline{x}_1 - \overline{x}_2| < T$;
   - m_1 và m_3 như nhau vì $|\overline{x}_1 - \overline{x}_3| < T$
   - m_2 và m_3 khác nhau vì $|\overline{x}_2 - \overline{x}_3| > T$.

Example

Do $\overline{x}_2 - \overline{x}_3 < 0$ nên có thể coi $m_2 < m_3$. Như vậy ở mức ý nghĩa 5%, theo tập dữ liệu này, trung bình số con của các hộ dưới mức khá giả lớn hơn các hộ khá giả, có nhỉnh hơn số con trung bình của các hộ trên mức giả nhưng chưa tới mức mang ý nghĩa thống kê.

Example

Một người bán hàng sử dụng một trang facebook để đăng quảng cáo. Anh ta muốn biết xem khung giờ nào là tốt nhất để đăng: sáng, trưa, chiều hay tối. Thử nghiệm của anh ta trong vài lần (cùng vào ngày thứ 7 với các loại hàng hóa đẹp tương đương nhau) cho thấy số lượt thích (đơn vị trăm) như sau:

Sáng	Trưa	Chiều	Tối
7	13	11	13
9	7	12	12
4	8	12	13
6	6	9	9

Giả sử số lượt thích của mỗi khoảng thời điểm trên là có phân bố chuẩn, có phương sai như nhau. Tại mức ý nghĩa 5%, đăng quảng cáo vào khoảng thời gian khác nhau trong ngày có dẫn tới lượt thích khác nhau không?

1. Phân tích phương sai một yếu tố
2. Phân tích sâu One-way ANOVA (Analysis of variance)
3. Phân tích anova với R

Sử dụng R trong phân tích phương sai

$$\text{Mẫu gộp} = \text{c(mẫu 1, mẫu 2, . . . , mẫu k)}$$
$$\text{Phân loại} = \text{factor(rep(c(1:k,c}(n_1, n_2 \ldots, n_k)))$$

Và tính toán kiểm định:

$$\text{anova(lm(Mẫu gộp} \sim \text{Phân loại))}$$

Kết quả sẽ cho ta bảng phân tích phương sai, kèm theo P - giá trị của bài toán.
Khi bác bỏ H_0, thực hiện phân tích sâu Tukey nhờ hàm:

$$\text{TukeyHSD(aov(Mẫu gộp} \sim \text{Phân loại))}$$

Kết quả sẽ cho ta bảng các P - giá trị của từng cặp dấu hiệu.

Example

Trong bài toán nói trên, ta có thể làm như sau:

```
> MauGop=c(2,3,3,2,1,2,1,2,3,4,2,3)
> PhanLoai=factor(rep(1:3,each=4))
> anova(lm(MauGop~PhanLoai))
Analysis of Variance Table

Response: MauGop
          Df Sum Sq Mean Sq F value  Pr(>F)
PhanLoai   2 4.6667 2.33333    5.25 0.03083 *
Residuals  9 4.0000 0.44444
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ta được P - giá trị là $0.0383 < 0.05$ nên bác bỏ H_0, tức là điều kiện kinh tế ảnh hưởng đến quyết định về số con trong các gia đình.

Example

Nếu muốn biết các cặp trung bình nào khác nhau ta dùng:

```
> TukeyHSD(aov(MauGop~PhanLoai))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = MauGop ~ PhanLoai)

$PhanLoai
    diff        lwr       upr     p adj
2-1 -1.0 -2.3161641 0.3161641 0.1402115
3-1  0.5 -0.8161641 1.8161641 0.5600698
3-2  1.5  0.1838359 2.8161641 0.0272373
```

Theo đó, chỉ có p - giá trị cho cặp μ_3 và μ_2 là < 0.05 nên chỉ có $\mu_3 \neq \mu_2$. Trong khi đó, với mức sai lầm hạn chế ở 5% ta chấp nhận $\mu_1 = \mu_2; \mu_3 = \mu_1$.

Example

Giả sử chúng ta muốn đánh giá tuổi tác có tác động đến việc đầu tư hay không, một mẫu ngẫu nhiên được chọn từ những tổng thể các nhà đầu theo độ tuổi: trẻ (<35 tuổi), trung niên (35 - 49), xế trung niên (50 - 65), già (trên 65) và nghi lại tỉ lệ tài sản của người đó dùng để đầu tư như sau:

Trẻ	Trung niên	Xế trung niên	Già
24.8	28.9	81.5	66.8
35.5	7.3	0.0	77.4
68.7	61.8	61.3	32.9
42.2	53.6	0.0	74.0

Giả sử tỉ lệ tài sản dành cho đầu tư của tổng thể có phân bố chuẩn.
Hãy kiểm tra điều kiện về phương sai (theo quy tắc nhanh đã nêu) cho bài toán phân tích phương sai.
Nếu điều kiện về phương sai thỏa mãn, tại mức ý nghĩa 5%, tuổi tác có tác động đến tỉ lệ tài sản mà người đó dành cho đầu tư không? Nếu có, độ tuổi nào người ta đầu tư với tỉ lệ tài sản lớn nhất?

Example (Ví dụ giả tưởng)

Một thành viên cái bang muốn xem trang phục của anh ta có ảnh hưởng đến lượng tiền xin được trong ngày không. Anh ta mặc ngẫu nhiên những áo với màu chủ đạo lần lượt là: xanh, đỏ, tím, vàng vào các ngày đi hành nghề ở cùng một khu vực. Kết quả được bảng sau đây (đơn vị nghìn đồng):

Xanh	Đỏ	Tím	Vàng
200	250	150	180
150	300	160	150
110	100	140	220
100	190	110	160
180	150	130	150

Giả sử tổng thể tiền xin được trong ngày của mỗi màu áo chủ đạo là tuân theo phân phối chuẩn, có cùng phương sai. Tại mức ý nghĩa 5%, hãy cho biết trang phục của người ăn xin có ảnh hưởng đến lượng tiền xin được không? Nếu có, trang phục nào giúp xin được nhiều nhất?

Example (Does the type of cooking pot affect iron content?)

Iron-deficiency anemia is the most common form of malnutrition in developing countries, affecting about 50% of children and women and 25% of men. Iron pots for cooking foods had traditionally been used in many of these countries, but they have been largely replaced by aluminum pots, which are cheaper and lighter. Some research has suggested that food cooked in iron pots will contain more iron than food cooked in other types of pots. One study designed to investigate this issue compared the iron content of some Ethiopian foods cooked in aluminum, clay, and iron pots. One of the foods was yesiga wet', beef cut into small pieces and prepared with several Ethiopian spices. The iron content of four samples of yesiga wet' cooked in each of the three types of pots is given below. The units are milligrams of iron per 100 grams of cooked food.

Example

Type of pot	Iron (mg/100 g food)			
Aluminum	1.77	2.36	1.96	2.14
Clay	2.27	1.28	2.48	2.68
Iron	5.27	5.17	4.06	4.22

1. Make a table giving the sample size, mean, and standard deviation for each type of pot. Is it reasonable to pool the variances? Note that with the small sample sizes in this experiment, we expect a large amount of variability in the sample standard deviations.
2. Run the analysis of variance. Report the F statistic with its degrees of freedom and P-value. What do you conclude?

RULE FOR EXAMINING STANDARD DEVIATIONS IN ANOVA

If the largest standard deviation is less than twice the smallest standard deviation, we can use methods based on the assumption of equal standard deviations, and our results will still be approximately correct. When we assume that the population standard deviations are equal, each sample standard deviation is an estimate of σ. To combine these into a single estimate, we use a generalization of the pooling method.

POOLED ESTIMATOR OF σ

Suppose we have sample variances $s_1^2, s_2^2, \ldots, s_I^2$ from I independent SRSs of sizes $n_1, n_2, \ldots, n_I$ from populations with common variance σ^2. The **pooled sample variance**

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_I - 1)}$$