Bài số 8

MẪU NGẪU NHIÊN VÀ PHÂN PHỐI CÁC THỐNG KÊ CƠ BẢN

Trong nhiều bài toán thực tế dẫn đến nghiên cứu một hay nhiều dấu hiệu định tính hoặc định lượng đặc trưng cho các phần tử của một tập hợp nào đó. Chẳng hạn, nếu muốn điều tra thu nhập bình quân của các gia đình ở Hà nội thì **tập hợp cần nghiên cứu** là các hộ gia đình, ***dấu hiệu nghiên cứu*** là thu nhập của từng gia đình.

Để xử lý dấu hiệu cần nghiên cứu đôi khi người ta sử dụng phương pháp nghiên cứu toàn bộ, đó là điều tra toàn bộ các phần tử của tập hợp theo dấu hiệu cần nghiên cứu để rút ra kết luận cần thiết. Tuy nhiên trong thực tế việc áp dụng phương pháp này gặp một số khó khăn sau:

+ Do quy mô nghiên cứu quá lớn nên việc nghiên cứu toàn bộ đòi hỏi nhiều chi phí về vật chất và thời gian, hơn nữa nếu không kiểm soát được sẽ dẫn đến cồng chéo hoặc bỏ sót.

+ Nhiều trường hợp không thể nắm được toàn bộ các phần tử của tập hợ cần nghiên cứu do đó không thể tiến hành toàn bộ được.

Vì thế trong thực tế, phương pháp nghiên cứu toàn bộ thường chỉ áp dụng với các tập hợp có quy mô nhỏ, còn chủ yếu người ta sử dụng phương pháp không toàn bộ mà đặc trưng là phương pháp nghiên cứu chọn mẫu.

I. TỔNG THỂ. MẪU. THỐNG KÊ

1. Tổng thể.

a. Định nghĩa. Khi nghiên cứu về một vấn đề người ta thường khảo sát trên một dấu hiệu nào đó, các dấu hiệu này thể hiện trên nhiều phần tử. Tập hợp các phân tử mang dấu hiệu được quan tâm được gọi là một ***tổng thể***.

Số lượng các phần tử trong một tổng thể được gọi là ***cỡ*** tổng thể.

Mỗi phần tử có mặt trong tổng thể được gọi là một ***cá thể*** của tổng thể đó.

Ví dụ 1. Nếu có 600 sinh viên trong trường mà chúng ta phân chia theo loại máu, thì chúng ta nói rằng chúng ta có một tổng thể 600 người. Các số trên các thẻ trong một tầng, chiều cao của khu dân cư trong thành phố nhất định và chiều dài của cá trong một hồ nhất định là các ví dụ về các tổng thể có cỡ hữu hạn.

Các quan sát thu được bằng cách đo áp suất khí quyển hàng ngày từ quá khứ đến tương lai hay tất cả các giá trị đo về độ sâu của hồ từ bất cứ vị trí nào có thể xác định được là các ví dụ về các tổng thể có cỡ vô hạn.

b. Đặc trưng của tổng thể.

Mỗi quan sát là một giá trị của một biến ngẫu nhiên X có phân phối xác suất $f(x)$. Khi chúng ta nhắc đến một "hàm phân phối nhị thức", một "phân phối chuẩn" hay, nói chung "hàm phân phối $f(x)$", chúng ta chỉ ra rằng một hàm phân phối có các quan sát là các giá trị của một biến ngẫu nhiên có phân phối nhị thức, phân phối chuẩn hay phân phối xác suất $f(x)$. Vì thế, trung bình mẫu và phương sai của một biến ngẫu nhiên hay phân phối xác suất cũng được xem là **trung bình mẫu** và ***phương sai của tổng thể*** tương ứng.

2. Mẫu.

a. Định nghĩa. Việc từ tổng thể ta lấy ra một tập con nào đó được gọi là ***phép lấy mẫu***.
Mỗi tập con được lấy ra gọi là một **mẫu.**
Số phần tử của mẫu được gọi là ***kích thước của mẫu.***

b.Chú ý.
+ Vì từ mẫu suy ra kết luận cho tổng thể nên mẫu phải đại diện cho tổng thể. Thường thì chúng ta cố gắng chọn một mẫu bằng cách chọn các thành phần và thuận lợi nhất của tổng thể. Cách thực hiện như thế có thể dẫn đến các kết luận sai lầm liên quan đến tổng thể. Bất cứ quy trình lấy mẫu nào đưa ra các kết luận luôn ước tính quá cao hay ước tính quá thấp một số đặc tính của tổng thể đó được gọi là **lấy mẫu chệch.** Để loại bỏ khả năng bị chệch trong lấy mẫu, tốt nhất nên lựa chọn **mẫu ngẫu nhiên** sao cho các quan sát được thực hiện độc lập và ngẫu nhiên.
+ Việc lấy mẫu được tiến hành theo hai phương thức: lấy mẫu có hoàn lại và lấy mẫu không hoàn lại.

c. Mẫu ngẫu nhiên. Giả sử $X_1, X_2, ..., X_n$ là n **biến ngẫu nhiên độc lập** và có cùng hàm phân phối xác suất $f(x)$. Khi đó chúng ta gọi $(X_1, X_2, ..., X_n)$ là một **mẫu ngẫu nhiên** có cỡ n từ tổng thể $f(x)$ và hàm phân phối xác suất đồng thời của chúng là:
$$f(x_1, x_2, ..., x_n) = f(x_1)f(x_2)...f(x_n).$$

Ví dụ 2. Kết quả điểm môn Toán của một lớp gồm 100 sinh viên được thông kê bởi bảng sau

Điểm	**3**	**4**	**5**	**6**	**7**
Số SV có điểm tương ứng	**25**	**20**	**40**	**10**	**5**

Gọi X là điểm môn Toán của một sinh viên được chọn ngẫu nhiên trong danh sách lớp thì X là một BNN có phân phối như sau

X	**3**	**4**	**5**	**6**	**7**
$P(X = x)$	**0,25**	**0,20**	**0,40**	**0,10**	**0,05**

Chọn ngẫu nhiên 5 sinh viên trong danh sách lớp để xem điểm. Gọi X_i là điểm của sinh viên thứ i. Khi đó ta có mẫu ngẫu nhiên kích thước $n = 5$ được xây dựng từ biến ngẫu nhiên X đó là:
$$(X_1, X_2, X_3, X_4, X_5).$$

3. Thống kê.

Mục đích chính của chúng ta trong lựa chọn các mẫu chính là việc tìm ra các thông tin về các tham số chưa biết trong phân phối của tổng thể. Ví dụ, giả sử rằng chúng ta đang rút ra kết luận liên quan đến tỷ lệ người uống cà phê tại Mỹ chỉ yêu thích một nhãn hiệu nhất định. Có thể tiến hành đặt câu hỏi với mọi người Mỹ uống cà phê để tính toán được giá trị thông số p đại diện cho tỷ lệ tổng thể. Hoặc chúng ta có thể lựa chọn một mẫu ngẫu nhiên lớn và tỷ lệ $\hat{p}$ người trong mẫu này yêu thích cùng một nhãn hiệu cà phê để tính toán. Giá trị $\hat{p}$ bây giờ được dùng để đưa ra kết luận liên quan đến tỷ lệ thực p. Khi đó $\hat{p}$ là một hàm của các giá trị quan sát được trong mẫu ngẫu nhiên; vì nhiều mẫu ngẫu nhiên có thể lấy từ cùng một tổng thể, chúng ta sẽ dự đoán $\hat{p}$ thay đổi theo từng mẫu. Điều đó có nghĩa là $\hat{p}$ là một giá trị của một hàm ký hiệu là $\hat{P}$. Một hàm của các quan sát được gọi là **một thống kê.**

a. Định nghĩa

Một hàm của ***biến ngẫu nhiên*** trong mẫu ngẫu nhiên được gọi là **một thống kê.**

b. Một số thống kê quan trọng.

i.Trung bình mẫu ngẫu nhiên.

Định nghĩa. Nếu $(X_1, X_2, ..., X_n)$ là một mẫu ngẫu nhiên có cỡ n, khi đó **trung bình mẫu** được xác định bằng thống kê:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Chú ý.

+ Vì $X_1, X_2, ..., X_n$ là các BNN nên $\overline{X}$ cũng là một BNN

+ Mẫu ngẫu nhiên $\overline{X}$ sẽ có giá trị $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ khi X_i có giá trị x_i, $i = 1, 2, ..., n$.

Ví dụ 3. Một thanh tra thực phẩm kiểm tra một mẫu ngẫu nhiên 7 hộp cá ngừ mang cùng nhãn hiệu để xác định phần trăm các tạp chất lạ. Các số liệu sau đây đã được ghi lại: 1,8; 2,1; 1,7; 1,6; 0,9; 2,7 và 1,8. Hãy tính trung bình mẫu mẫu.

Giải: + Giá trị $\overline{x}$ thu được của thống kê $\overline{X}$ là:

$$\overline{x} = \frac{1,8 + 2,1 + 1,7 + 1,6 + 0,9 + 2,7 + 1,8}{7} = 1,8\%$$

ii.Median mẫu (trung vị mẫu)

Định nghĩa. Nếu $(X_1, X_2, ..., X_n)$ là một mẫu ngẫu nhiên cỡ n, được sắp xếp theo ***thứ tự tăng dần*** của độ lớn, khi đó **median mẫu** được xác định bởi thống kê:

$$\widetilde{X} = \begin{cases} X_{\frac{n+1}{2}}, & \text{khi } n \text{ le} \\ \dfrac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}, & \text{khi } n \text{ chan} \end{cases}$$

Ví dụ 4. Số tàu nước ngoài đến cảng biển phía đông vào 7 ngày được lựa chọn ngẫu nhiên là 8, 3, 9, 5, 6, 8 và 5. Tìm median.

Giải: + Bố trí các quan sát theo trật tự tăng theo độ lớn, chúng ta có: 3 5 5 6 8 8 9

+ Và từ đó $\widetilde{X}$ = 6.

iii. Mode

Định nghĩa. Nếu $X_1, X_2, ..., X_n$ không nhất thiết khác nhau hoàn toàn, biểu diễn một mẫu ngẫu nhiên có cỡ n. Khi đó ***mode*** M là giá trị của mẫu mà xảy ra thường xuyên nhất hoặc có tần số lớn nhất.
Mode có thể không tồn tại và khi nó tồn tại không nhất thiết là giá trị duy nhất.

Ví dụ 5. Nếu các đồ tặng của một mẫu ngẫu nhiên các cư dân của Fairway Forest cho Hiệp hội phổi Virginia được ghi nhận là 9, 10, 5, 9, 9, 7, 8, 6, 10 và 11 đô la, khi đó mode là $m = 9$ *đô la,* đó là giá trị xảy ra với tần số cao nhất.

Ví dụ 6. Số lượng phim mà một tổng thể ngẫu nhiên 12 học sinh trung học đã tham gia đóng tháng vừa rồi được ghi nhận như sau: 2, 0, 3, 1, 2, 4, 2, 5, 4, 0, 1 và 4. Trong trường hợp này, có hai mode 2 và 4, vì cả 2 và 4 đều xảy ra với tần số cao nhất. Phân phối được xác định là **phân phối hai mốt.**

Nhận xét. + Trung bình mẫu là phương pháp đo được sử dụng phổ biến nhất để xác định vị trí trung tâm trong thống kê. Điểm yếu duy nhất đối với trung bình mẫu đó chính là nó có thể bị ảnh hưởng ngược lại do các giá trị cực trị.

+ Mode là tiêu chí thường được chú ý trong các bài toán kinh tế: Để bán được lượng hàng hóa lớn thì người bán hang nên quan tâm tới thị hiếu của số đông, chẳng hạn năm nay kiểu áo khoác nào được ưa chuộng.

3. Biên độ biến thiên của mẫu ngẫu nhiên.

Định nghĩa. ***Khoảng biến thiên*****(Biên độ mẫu)** của một mẫu ngẫu nhiên $X_1, X_2, ..., X_n$ được xác định bằng thống kê $X_{(n)} - X_{(1)}$ trong đó $X_{(n)}$ và $X_{(1)}$ tương ứng là các giá trị quan sát lớn nhất và nhỏ nhất trong mẫu.

Biên độ mẫu có thể là một đơn vị đo kém về độ biến đổi, nhất là nếu cỡ của mẫu hay tổng thể lớn. Phương pháp này chỉ xem xét các giá trị cực trị và không cho chúng ta biết về phân phối các giá trị ở giữa. Để khắc phục điều đó, chúng ta sẽ đi xem xét một giá trị đo độ biến đổi, đó là **phương sai mẫu:** xem xét vị trí của mỗi quan sát liên quan đến trung bình mẫu.

4. Phương sai mẫu.

Định nghĩa. Cho $(X_1, X_2, \ldots, X_n)$ là mẫu ngẫu nhiên cỡ n với trung bình mẫu là $\overline{X}$. Khi đó **phương sai mẫu** được xác định bởi thống kê:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Với mỗi mẫu cụ thể thì S^2 sẽ nhận giá trị $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$.

Ví dụ 7. So sánh giá cà phê ở 4 cửa hiệu tạp phẩm được lựa chọn ngẫu nhiên tại San Diego cho thấy các mức tăng từ tháng trước là 12, 15, 17 và 20 cent cho mỗi túi một pound. Tìm phương sai của mẫu ngẫu nhiên các mức tăng giá.

Giải: + Tính trung bình mẫu mẫu, chúng ta thu được:

$$\overline{x} = \frac{12+15+17+20}{4} = 16 \qquad cent$$

+ Vì thế,

$$s^2 = \frac{\sum_{i=1}^{4}(x_i - 16)^2}{3} = \frac{(12-16)^2 + (15-16)^2 + (17-16)^2 + (20-16)^2}{3}$$
$$= \frac{(-4)^2 + (-1)^2 + (1)^2 + (4)^2}{3} = \frac{3}{4}$$

Nếu $\overline{x}$ là một số thập phân được làm tròn, chúng ta sẽ tích lũy một sai số lớn khi sử dụng công thức phương sai mẫu của Định nghĩa trên. Để tránh điều này, chúng ta thường sử dụng công thức như trong định lý sau:

Định lý. Nếu S^2 là phương sai của một mẫu ngẫu nhiên cỡ n, khi đó:

$$S^2 = \frac{n\sum_{i=1}^{n}X_i^2 - \left(\sum_{i=1}^{n}X_i\right)^2}{n(n-1)}$$

Ví dụ 8. Tìm phương sai của số liệu 3, 4, 5, 6, 6 và 7 biểu diễn số cá hồi bắt được bằng một mẫu ngẫu nhiên của các ngư dân trong ngày 19 tháng 7 năm 1996 tại hồ Muskoka.

Giải. + Ta có: $\sum_{i=1}^{6} x_i^2 = 171, \quad \sum_{i=1}^{6} x_i = 31, \quad n = 6$.

+ Do đó: $s^2 = \dfrac{(6)(171)-(31)^2}{(6)(5)} = \dfrac{13}{6}$

5. <u>Độ lệch tiêu chuẩn mẫu</u>.

<u>Định nghĩa.</u>
Độ lệch tiêu chuẩn mẫu ký hiệu bằng S là căn số bậc hai dương của phương sai mẫu.

II. <u>PHÂN PHỐI XÁC SUẤT CỦA CÁC THỐNG KÊ CƠ BẢN</u>.

1.<u>Phân phối của trung bình mẫu</u>.

Phân phối mẫu quan trọng đầu tiên được xem xét là phân phối của trung bình mẫu $\overline{X}$. Giả sử một biến ngẫu nhiên của n các quan sát được lấy từ một tổng thể tuân theo phân phối chuẩn có kỳ vọng μ và phương sai σ^2. Mỗi quan sát $X_i,\ i = 1,2,\dots,n$ của mẫu ngẫu nhiên khi đó sẽ có cùng phân phối chuẩn như tổng thể đang xét. Vì thế, trung bình mẫu:

$$\overline{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

có kỳ vọng: $\mu_{\overline{X}} = \dfrac{\mu + \mu + \dots + \mu}{n} = \mu$

và phương sai: $\sigma_{\overline{x}}^2 = \dfrac{\sigma^2 + \sigma^2 + \dots + \sigma^2}{n^2} = \dfrac{\sigma^2}{n}$

Nếu lấy mẫu từ một tổng thể phân phối không xác định, cả hữu hạn và vô hạn, phân phối mẫu của $\overline{X}$ vẫn gần chuẩn với giá trị trung bình μ và phương sai $\dfrac{\sigma^2}{n}$ miễn là cỡ mẫu lớn. Kết quả này là một hệ quả trực tiếp của định lý sau đây, còn gọi là định lý giới hạn trung tâm.

<u>Định lý giới hạn trung tâm</u>. Nếu $\overline{X}$ là giá trị trung bình của một mẫu ngẫu nhiên có kích thước n lấy từ tổng thể có giá trị trung bình μ và phương sai hữu hạn σ^2, khi đó giới hạn phân phối dạng của

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \quad \text{là phân phối chuẩn } n(z;0,1) \text{ khi } n \to \infty.$$

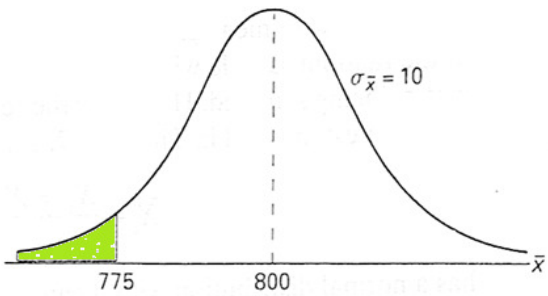
<u>Chú ý.</u> + Phép xấp xỉ chuẩn đối với $\overline{X}$ nói chung sẽ tốt nếu $n \geq 30$. Nếu $n < 30$ mức xấp xỉ sẽ chỉ tốt nếu tổng thể không quá khác so với phân phối chuẩn.

+ Nếu tổng thể có phân phối chuẩn, thì phân phối của trung bình mẫu $\overline{X}$ sẽ có phân phối chuẩn.

Ví dụ 9. Một công ty điện sản xuất các loại bóng điện có tuổi thọ được phân phối gần chuẩn, với giá trị trung bình gần bằng 800 giờ và độ lệch tiêu chuẩn 40 giờ. Tìm xác suất mà một biến ngẫu nhiên 16 bóng sẽ có tuổi thọ trung bình chưa đến 775 giờ.

Giải: + Phân phối của trung bình mẫu $\overline{X}$ sẽ xấp xỉ chuẩn với $\mu_{\overline{X}} = 800$ và $\sigma_{\overline{X}} = 40 / \sqrt{16} = 10$.

+ Xác suất cần tìm được cho trong miền gạch chéo trong hình vẽ dưới đây.



+ Tương ứng với $\overline{x} = 775$, chúng ta nhận thấy rằng

$$z = \frac{775 - 800}{10} = -2.5$$

+ Và vì thế: $P(\overline{X} < 775) = P(Z < -2.5) = 0.0062$

<u>Chú ý.</u> Một ứng dụng rất quan trọng của định lý giới hạn trung tâm là quá trình xác định các giá trị thích hợp của giá trị trung bình lý thuyết μ. Các chủ đề như kiểm tra giả thiết, ước lượng, kiểm tra chất lượng và các chủ đề khác tận dụng được định lý giới hạn trung tâm. Ví dụ sau đây biểu diễn cách sử dụng định lý giới hạn trung tâm xét về mối liên hệ đó.

Ví dụ 10. Một quá trình sản xuất quan trọng sản xuất ra các linh kiện xy lanh cho ngành sản xuất ô tô. Điều quan trọng là quá trình này sản xuất ra các linh kiện có giá trị trung bình của đường kính xi lanh bằng 5 milimét. Các Kỹ sư có các phỏng đoán giá trị trung bình lý thuyết là 5.0 milimét. Một thử nghiệm được tiến hành trong đó 100 linh kiện được sản xuất bằng quá trình được lựa chọn ngẫu nhiên và đường kính được đo trên mỗi linh kiện. Mọi người nhận thấy rằng độ lệch tiêu chuẩn tổng thể $\sigma = 0,1$. Thử nghiệm cho thấy đường kính trung bình mẫu $\overline{x} = 5,027$ mi-li-mét. Liệu thông tin mẫu này có bác bỏ lại phỏng đoán của kỹ sư hay không?

Giải. Thông tin có hỗ trợ hay bác bỏ lại phỏng đoán của các kỹ sư phụ thuộc vào việc sản phẩm thử nghiệm thu được ($\overline{x} = 5.027$) có lớn hơn so với giá trị được nhận định ($\mu = 5,0$) hay không. Nói cách khác, liệu có thường xuyên thu được $\overline{x} \geq 5.027$ với $n = 100$ nếu giá trị trung bình lý thuyết là $\mu = 5,0$?

+ Nếu xác suất này cho ta kết quả rằng $\overline{x} = 5.027$ không vô lý thì dự đoán không bị bác bỏ.
+ Nếu xác suất này rất thấp, thì dữ liệu không ủng hộ dự đoán rằng $\mu = 5,0$ là đúng.

+ Xác suất mà chúng ta chọn để tính toán xác định bởi:

$$P\left[(\bar{X}-5) > 0,027\right]$$

+ Ta có:

$$P\left[\left|\overline{X}-5\right| \geq 0,027\right] = P\left[(\overline{X}-5) \geq 0,027\right] + P\left[(\overline{X}-\mu) \leq -0,027\right] = 2P\left(\frac{\overline{X}-5}{0,1/\sqrt{100}} \geq 2,7\right)$$

Ở đây chúng ta tiêu chuẩn hóa đơn giản $\overline{X}$ theo dịnh lý giới hạn trung tâm.

+ Nếu dự đoán $\mu = 5,0$ là đúng, $\dfrac{\overline{X}-5}{0.1/\sqrt{100}}$ có phân phối $N(z;0,1)$. Vì thế

$$2P\left[\left(\frac{\overline{X}-5}{0,1/\sqrt{100}}\right) \geq 2,7\right] = 2P\left[Z \geq 2,7\right] = 2(0,0035) = 0,007$$

+ Như vậy sự kiện $\overline{x}$ lệch khỏi vị trí trung bình 0,027 mi-li-mét chỉ xảy ra 7 lần trong 1000 lần thử nghiệm.

+ Do vậy thử nghiệm này có $\overline{x}$=5.027 chắc chắn không ủng hộ cho dự đoán $\mu = 5,0$ của các kỹ sư.

2. Phân phối của hiệu hai trung bình mẫu.

Giả sử rằng chúng ta có hai tổng thể, tổng thể đầu tiên có giá trị trung bình μ_1 và phương sai σ_1^2 và tổng thể thứ hai có giá trị trung bình μ_2 và phương sai σ_2 .

Lấy $\overline{X_1}$ là trung bình của một mẫu ngẫu nhiên có cỡ n_1 được lựa chọn từ tổng thể đầu tiên, và thống kê $\overline{X_2}$ là trung bình của một mẫu ngẫu nhiên với cỡ n_2 được lựa chọn từ tổng thể thứ hai, độc lập với mẫu từ tổng thể đầu tiên. Khi đó ta có thể kết luận gì về phân phối lấy mẫu của độ chênh lệch $\overline{X_1} - \overline{X_2}$?

Theo Định lýgiá trị trung tâm, các biến $\overline{X_1}$ và $\overline{X_2}$ đều được phân phối gần chuẩn với các giá trị trung bình μ_1 và μ_2, các phương sai σ_1^2 / n_1 và σ_2^2 / n_2 tương ứng.

Và ta có thể kết luận rằng $\overline{X_1} - \overline{X_2}$ có phân phối có giá trị trung bình

$$\mu_{\overline{X_A}-\overline{X_B}} = \mu_A - \mu_B$$

và phương sai

$$\sigma^2_{\overline{X_1}-\overline{X_2}} = \frac{\sigma_1^1}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Từ đó ta có:

Định lý. Nếu các mẫu độc lập có kích thước n_1 và n_2 được lấy ngẫu nhiên từ hai tổng thể, rời rạc hoặc liên tục, có các giá trị trung bình μ_1 và μ_2, các phương sai σ_1^2 và σ_2^2, khi đó phân phối của các thống kê mẫu của các sự sai khác giữa hai giá trị trung bình: $\overline{X}_1 - \overline{X}_2$, được phân phối xấp xỉ chuẩn có giá trị trung bình và phương sai bằng:

$$\mu_{\overline{X}-\overline{X}} = \mu_1 - \mu_2 \quad \text{và} \quad \sigma^2_{\overline{X}_1-\overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Do đó

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}}$$

có phân phối xấp xỉ phân phối tiêu chuẩn $n(z;0,1)$.

Chú ý. + Nếu cả hai n_1 và n_2 lớn hơn hoặc bằng 30, thì phương pháp xấp xỉ chuẩn cho phân phối của $\overline{X}_1 - \overline{X}_2$ rất tốt bất kể hình dạng của hai tổng thể.

+ Tuy nhiên, thậm chí khi n_1 và n_2 nhỏ hơn 30, phép xấp xỉ chuẩn tương đối tốt ngoại trừ khi các tổng thể không hề bất thường.

+ Rõ ràng, nếu cả hai tổng thể là chuẩn, thì khi đó $\overline{X}_1 - \overline{X}_2$ có phân phối chuẩn bất kể cỡ n_1 và n_2.

3. Phân phối của phương sai mẫu.

a.Định nghĩa. Biến ngẫu nhiên liên tục X có ***phân phối*** χ^2 (***Khi bình phương***) với ν bậc tự do nếu hàm mật độ của nó được cho bởi

$$f(x) = \begin{cases} \dfrac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2} dx, & x > 0 \\ 0 & x \le 0 \end{cases}$$

ở đó v là một số nguyên dương.

b. Định lý. Nếu S^2 là phương sai của mẫu ngẫu nhiên có kích thước n được rút ra từ một tổng thể có phân phối chuẩn có phương sai σ^2, khi đó thống kê

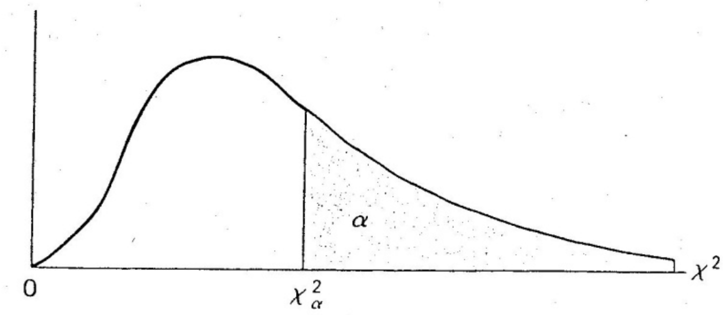
$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}$$

có phân phối χ^2 (Khi bình phương) với $v = n - 1$ bậc tự do.

+ Với mỗi mẫu cụ thể, giá trị của biến ngẫu nhiên χ^2 được tính theo công thức:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

+ Xác suất để một biến ngẫu nhiên sinh ra giá trị χ^2 - giá trị lớn hơn một số giá trị nhất định, bằng diện tích bên dưới biểu đồ sang phía bên phải giá trị này.



+ Bảng A.5 cho các giá trị của χ^2_α đối với các giá trị khác nhau của α và *v*.

4. <u>Phân phối t.</u>

<u>Định nghĩa.</u> Giả sử Z là một biến ngẫu nhiên tiêu chuẩn và V là biến ngẫu nhiên χ^2 có *v* bậc tự do. Nếu Z và V độc lập, khi đó phân phối của biến ngẫu nhiên T, trong đó

$$T = \frac{Z}{\sqrt{V / v}}$$

được xác định bởi:

$$h(t) = \frac{\Gamma\left[(v+1)/2\right]}{\Gamma(v/2)\sqrt{\pi v}}(1+\frac{t^2}{v})^{-(v+1)/2}, \ -\infty < 1 < \infty$$

Phân phối này được gọi là **phân phối *t*** có *v* bậc tự do.

<u>Tổng quát.</u> Giả sử X_1, X_2, …, X_n là các biến ngẫu nhiên độc lập, cùng phân phối chuẩn với giá trị trung bình μ và độ lệch chuẩn σ. Đặt

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} \quad \text{và } s^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$$

Khi đó biến ngẫu nhiên $T = \dfrac{\overline{X} - \mu}{S / \sqrt{n}}$ có một phân phối t với $v = n - 1$ bậc tự do.

5. <u>Phân bố *F*</u>

Phân bố F có nhiều ứng dụng rộng rãi trong so sánh các phương sai mẫu. Các ứng dụng của phân bố *F* có trong các bài toán gồm hai mẫu trở lên.

Định nghĩa. Giả sử U và V là hai biến ngẫu nhiên độc lập có các phân bố χ^2 với υ_1 và υ_2 các bậc tự do tương ứng. Khi đó phân bố của biến ngẫu nhiên $F = \dfrac{U/\upsilon_1}{V/\upsilon_2}$ được xác định bởi:

$$h(f) = \begin{cases} \dfrac{\Gamma\left[(v_1 - v_2)/2\right](v_1/v_2)^{v_1/2}}{\Gamma(v_2/2)\Gamma(v_2/2)} \dfrac{f^{v_1/2-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}} & ; 0 < f < \infty \\ 0 & \end{cases}$$

Phân bố này được gọi là ***phân bố F*** với υ_1 và υ_2 bậc tự do.

Định lý. Nếu S_1^2 và S_2^2 là các phương sai của các mẫu ngẫu nhiên độc lập có kích thước n_1 và n_2 được lấy từ các tổng thể chuẩn có các phương sai σ_1^2 và σ_2^2 tương ứng, khi đó

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

có phân bố F với $\upsilon_1 = n_1 - 1$ và $\upsilon_2 = n_2 - 1$ bậc tự do.

Về nhà:

Tự đọc: Mục 8.3

Bài tập: Tr. **231, 246, 260**

Đọc trước các Mục từ 9.1 đến 9.8 chuẩn bị cho **Bài số 9:**

Bài toán ước lượng một mẫu và hai mẫu