BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỎ-ĐỊA CHẤT

BÁO CÁO HỌC THUẬT

CHỨNG MINH CÔNG THỨC VỀ PHÂN PHỐI CỦA CÁC ĐẶC TRƯNG MẪU

ThS Nguyễn Thu Hằng

Hà nội, tháng 6 năm 2021

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỎ-ĐỊA CHẤT

BÁO CÁO HỌC THUẬT

CHỨNG MINH CÔNG THỨC VỀ PHÂN PHỐI CỦA CÁC ĐẶC TRƯNG MẪU

Xác nhận của bộ môn

Hà nội, tháng 6 năm 2021

MỤC LỤC

LỜI GIỚI THIỆU

Từ thế kỷ thứ 17, phương pháp nghiên cứu mẫu đã ra đời, ngày càng được phát triển và được sử dụng rộng rãi trong nhiều lĩnh vực. Việc thu thập mẫu, trình bày mẫu, tính toán các đặc trưng mẫu là những bước hết sức quan trọng, có vai trò quyết định đến tính đúng, sai của bài toán phân tích thống kê toán học. Vì vậy, vấn đề chọn mẫu, biểu diễn số liệu cũng như tính toán các đặc trưng mẫu là không thể thiếu trong nghiên cứu cũng như trong nội dung giảng dạy ở các trường đại học, cao đẳng. Nhằm hoàn thiện giáo án giảng dạy cũng như cung cấp tài liệu tham khảo cho thầy cô và các em sinh viên, tôi chọn báo cáo "Chứng minh công thức về phân phối của các đặc trưng mẫu".

Báo cáo học thuật chia làm hai phần:

Phần 1: Trình bày một số khái niệm cơ bản về tổng thể và mẫu cũng như định nghĩa của một số đặc trưng mẫu cơ bản và tính chất của các đặc trưng mẫu này.

Phần 2: Trình bày chứng minh về phân phối của một số đặc trưng mẫu đã nêu trong phần 1.

1. Kiến thức cơ sở

1.1. Tổng thể và mẫu

Tổng thể là tập hợp tất cả các đối tượng có chung một tính chất nào đó mà ta đang quan tâm.

Mỗi phần tử trong tổng thể được gọi là một cá thể trong tổng thể đó.

Số lượng cá thể trong tổng thể được gọi là quy mô của tổng thể.

Biến ngẫu nhiên gốc X là đặc trưng cần nghiên cứu của tổng thể

Ví dụ: Để điều tra thời gian tự học của sinh viên Đại học Mỏ - Địa chất thì

- Tổng thể là toàn bộ sinh viên của trường Đại học Mỏ - Địa chất.
- Mỗi sinh viên của trường là một cá thể trong tổng thể đó.
- Biến ngẫu nhiên gốc X là biến ngẫu nhiên chỉ thời gian tự học của sinh viên.

Khi nghiên cứu về tổng thể, có 2 loại biến ngẫu nhiên

- Biến định lượng là biến chỉ các số đo của các cá thể. Ví dụ như chiều cao, cân nặng, thu nhập…
- Biến định tính là biến chỉ tính chất nào đó của cá thể. Ví dụ như giới tính, tôn giáo…

Ta cũng có thể gán các tính chất của biến định tính với một số nguyên tương ứng. Như vậy khi nghiên cứu tổng thể ta luôn có thể giả sử là các các phần tử của tổng thể có dấu hiệu định lượng.

Có tổng thể rất lớn cỡ hàng tỷ (ví dụ như tổng thể loài người...), cũng có tổng thể rất nhỏ (Tổng thể các con gấu trúc, hóa thạch khủng long...)

Nếu thu thập được số liệu của toàn bộ tổng thể thì kết quả nghiên cứu sẽ càng chính xác. Tuy nhiên, thông thường quy mô của tổng thể là rất lớn. Do đó,thu thập số liệu toàn bộ tổng thể sẽ có một số hạn chế sau:

- Tốn kém (Thời gian, tiền bạc…)
- Không thể thực hiện được trong nhiều trường hợp
- Phá hủy hiện vật

Mẫu là tập hợp n phần tử được lấy ra từ tổng thể, n được gọi là cỡ mẫu. Dựa trên thông tin trên mẫu, ta sẽ suy diễn ra các nhận định về toàn bộ tổng thể. Như vậy, ta làm việc trên mẫu nhưng mục tiêu lại là hiểu biết về toàn bộ tổng thể.

Kí hiệu $W = (X_1, X_2, \dots, X_n)$ là mẫu ngẫu nhiên

$w = (x_1, x_2, \dots, x_n)$ là mẫu cụ thể

1.2. Các số đặc trưng của mẫu

Khi nghiên cứu mẫu, người ta thường quan tâm đến một số đặc trưng của mẫu.

Định nghĩa 1. Cho mẫu ngẫu nhiên $W = (X_1, X_2, \dots, X_n)$. Trung bình mẫu ngẫu nhiên, kí hiệu là $\bar{X}$ và được tính bằng

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Nếu có mẫu cụ thể $w = (x_1, x_2, \dots, x_n)$ thì ta sẽ tính được giá trị của $\bar{X}$, kí hiệu là $\bar{x}$, gọi là trung bình mẫu cụ thể (trung bình mẫu). Như vậy

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Định lí 1. Nếu biến ngẫu nhiên gốc X có $EX = \mu$ và Var(X)= σ^2 thì $E\bar{X} = \mu$ và $Var(\bar{X}) = \frac{\sigma^2}{n}$.

<u>*Chứng minh.*</u>

Theo tính chất của kì vọng toán ta có

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n.\mu}{n} = \mu.$$

Để ý rằng các đại lượng X_i là độc lập, có cùng phân phối với biến ngẫu nhiên gốc X. Do đó, theo tính chất của phương sai, ta có

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}.n\sigma^2 = \frac{\sigma^2}{n}.$$

Định nghĩa 2. Cho mẫu ngẫu nhiên $W = (X_1, X_2, \dots, X_n)$. Phương sai mẫu ngẫu nhiên, kí hiệu là $\hat{S}^2$ và được tính bằng

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

Nếu có mẫu cụ thể $w = (x_1, x_2, \dots, x_n)$ thì ta sẽ tính được giá trị của $\hat{S}^2$, kí hiệu là $\hat{s}^2$, gọi là phương sai mẫu cụ thể (phương sai mẫu). Như vậy

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Định lí 2. Nếu biến ngẫu nhiên gốc X có $EX = \mu$ và Var(X)= σ^2 thì

$$E(\hat{S}^2) = \frac{n-1}{n}\sigma^2.$$

<u>*Chứng minh.*</u>

Ta có

$$(X_i - \bar{X})^2 = [(X_i - \mu) - (\bar{X} - \mu)]^2 = (X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2.$$

Do đó,

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu).\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) + (\bar{X} - \mu)^2.$$

Vì $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu = \bar{X} - \mu$ nên

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)^2 + (\bar{X} - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X} - \mu)^2.$$

Suy ra,

$$E(\hat{S}^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X} - \mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}E(X_i - \mu)^2 - E(\bar{X} - \mu)^2$$

Mặt khác,

Vì $E(X_i) = \mu$,nên $E(X_i - \mu)^2 = Var(X_i) = \sigma^2$

Vì $E(\bar{X}) = \mu$,nên $E(\bar{X} - \mu)^2 = Var(\bar{X}) = \frac{\sigma^2}{n}$

Ta được

$$E(\hat{S}^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Định nghĩa 3. Cho mẫu ngẫu nhiên $W = (X_1, X_2, \dots, X_n)$. Phương sai mẫu ngẫu nhiên hiệu chỉnh, kí hiệu là S^2 và được tính bằng

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

Nếu có mẫu cụ thể $w = (x_1, x_2, \dots, x_n)$ thì ta sẽ tính được giá trị của S^2, kí hiệu là s^2, gọi là phương sai mẫu cụ thể hiệu chỉnh (phương sai mẫu hiệu chỉnh). Như vậy

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Định lí 3. Nếu biến ngẫu nhiên gốc X có $EX = \mu$ và Var(X)= σ^2 thì $E(S^2) = \sigma^2$.

Chứng minh.

Từ định lí 2 suy ra

$$E(S^2) = \frac{n}{n-1}E(\hat{S}^2) = \frac{n}{n-1}.\frac{n-1}{n}\sigma^2 = \sigma^2.$$

2. Chứng minh về phân phối của một số đặc trưng mẫu

Bổ đề 1. Giả sử $X_i, i = 1,2,\dots,n$ là các biến ngẫu nhiên độc lập có cùng phân phối chuẩn $N(\mu, \sigma^2)$ với biến ngẫu nhiên gốc X. Khi đó, $\bar{X}$ và S^2 là các biến ngẫu nhiên độc lập.

Chứng minh.

Đầu tiên, ta chứng minh $\bar{X}$ độc lập với $X_i - \bar{X}, i = 1,2,..,n$. Thật vậy, vì $X_i, i = 1,2,\dots,n$ là độc lập nên hàm phân phối đồng thời là

$$f_{X_1,\dots,X_n}(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\}$$

Đổi biến

$$Y_1 = \bar{X} \qquad\qquad X_1 = Y_1 - (Y_2 + \cdots + Y_n)$$

$$Y_2 = X_2 - \bar{X} \qquad\qquad X_2 = Y_2 + Y_1$$

$$Y_3 = X_3 - \bar{X} \qquad \text{Khi đó} \qquad X_3 = Y_3 + Y_1$$

$$\ldots \qquad\qquad \ldots$$

$$Y_n = X_n - \bar{X} \qquad\qquad X_n = Y_n + Y_1$$

Định thức Jacobi khi đó là

$$\begin{vmatrix} 1 & -1 & -1 & \ldots & -1 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \ldots & 0 \end{vmatrix} = n$$

Suy ra

$$f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) = nf_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) = nf_{X_1,\ldots,X_n}(x_1, y_1 + y_2, \ldots, y_1 + y_n)$$

$= C.exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right\}$, trong đó, C là hằng số.

Mặt khác, tương tự biến đổi trong định lí 2, ta có

$$\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{\sigma^2}\left[(x_1 - \bar{x})^2 + \sum_{i=2}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right]$$

$$= \frac{1}{\sigma^2}\left[\left(\sum_{i=2}^{n}(x_i - \bar{x})\right)^2 + \sum_{i=2}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right]$$

$$= \frac{1}{\sigma^2}\left[\left(\sum_{i=2}^{n}y_i\right)^2 + \sum_{i=2}^{n}y_i^2 + n(y_1 - \mu)^2\right]$$

Hàm mật độ đồng thời của $Y_1, Y_2, \ldots, Y_n$ là

$$f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) = C\exp\left\{\frac{-1}{2\sigma^2}\left[\left(\sum_{i=2}^{n}y_i\right)^2 + \sum_{i=2}^{n}y_i^2 + n(y_1 - \mu)^2\right]\right\}$$

$$= C\exp\left\{\frac{-1}{2\sigma^2}\left[\left(\sum_{i=2}^{n}y_i\right)^2 + \sum_{i=2}^{n}y_i^2\right]\right\}\exp\left\{-\frac{n}{2\sigma^2}(y_1 - \mu)^2\right\}$$

$$= C.h(y_2, \ldots, y_n)g(y_1).$$

Suy ra, $Y_1 = \bar{X}$ độc lập với $Y_i = X_i - \bar{X}, i = 2, \ldots, n$.

Hơn nữa, $X_1 - \bar{X} = -\sum_{i=2}^{n}(X_i - \bar{X})$. Từ đó ta được $Y_1 = \bar{X}$ độc lập với $Y_i = X_i - \bar{X}, i = 1, 2, \dots, n$.
Ta lại có, S^2 là một hàm của $X_i - \bar{X}, i = 1,2, \dots, n$ nên S^2 độc lập với $\bar{X}$.

Định lí 4. Cho mẫu ngẫu nhiên $W = (X_1, X_2, \dots, X_n)$. Giả sử $X_i, i = 1,2, \dots, n$ là các biến ngẫu nhiên độc lập có cùng phân phối chuẩn $N(\mu, \sigma^2)$ với biến ngẫu nhiên gốc X. Khi đó

1. $\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$,
2. $V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$,
3. $\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim T_{n-1}$.

Chứng minh.

Từ tính chất của phân phối chuẩn (đã được trình bày trong báo cáo học thuật kì 1 năm học 2020-2021 của tác giả), dễ thấy $\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$.Tiếp theo ta sẽ chứng minh V có phân phối "chi – bình phương" với $n-1$ bậc tự do.

Trước hết, theo bổ đề 1, $\bar{X}$ và S^2 là các biến ngẫu nhiên độc lập.

Xét trường hợp $\mu = 0, \sigma = 1$.

Giả sử $S_n^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ nên $V = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$.

Mặt khác từ

$$\bar{X}_{n+1} = \frac{\sum_{i=1}^{n+1} X_1}{n+1} = \frac{X_{n+1} + n\bar{X}_n}{n+1} = \bar{X}_n + \frac{1}{n+1}\left(X_{n+1} - \bar{X}_n\right),$$

Suy ra

$$nS_{n+1}^2 = \sum_{i=1}^{n+1}\left(X_i - \bar{X}_{n+1}\right)^2 = \sum_{i=1}^{n+1}\left(X_i - \bar{X}_n - \frac{1}{n+1}\left(X_{n+1} - \bar{X}_n\right)\right)^2$$

$$= \sum_{i=1}^{n+1}\left[\left(X_i - \bar{X}_n\right)^2 - 2\left(X_i - \bar{X}_n\right)\frac{X_{n+1} - \bar{X}_n}{n+1} + \left(\frac{1}{n+1}\right)^2\left(X_{n+1} - \bar{X}_n\right)^2\right]$$

$$= \sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 + \left(X_{n+1} - \bar{X}_n\right)^2 - 2\frac{\left(X_{n+1} - \bar{X}_n\right)^2}{n+1} + \left(\frac{1}{n+1}\right)^2\left(X_{n+1} - \bar{X}_n\right)^2$$

$$= \sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 + \frac{n}{n+1}\left(X_{n+1} - \bar{X}_n\right)^2$$

$$= (n-1)S_n^2 + \frac{n}{n+1}\left(X_{n+1} - \bar{X}_n\right)^2.$$

Với $n = 1$ thì $S_2^2 = \frac{1}{2}(X_2 - X_1)^2$. Từ $\frac{X_2 - X_1}{\sqrt{2}} \sim N(0,1)$ suy ra $S_2^2 \sim \chi_1^2$.

Giả sử với $n = k$ thì $(k-1)S_k^2 \sim \chi_k^2$. Khi đó, với $n = k + 1$ ta có

$$kS_{k+1}^2 = (n-1)S_k^2 + \frac{k}{k+1}\left(X_{k+1} - \overline{X}_k\right)^2.$$

Từ $X_{k+1} - \overline{\mathrm{X}}_{\mathrm{k}} \sim N\left(0, \frac{k+1}{k}\right)$ ta suy ra $kS_{k+1}^2 \sim \chi_n^2$.

Trường hợp tổng quát làm tương tự. Chú ý rằng, $\frac{X_i - \mu}{\sigma} \sim N(0,1), i = 1,2 \dots, n$.

Như vậy, thống kê $V = \frac{(n-1)S^2}{\sigma^2}$ có phân bố “chi – bình phương” với $n - 1$ bậc tự do.

Cuối cùng, ta chứng minh phân bố của thống kê $\frac{\overline{X} - \mu}{\sigma}\sqrt{n}$.

Ta đã có, $\frac{Z_0}{\sqrt{\frac{1}{n}\sum_{i=1}^n Z_i^2}} \sim T_n$ trong đó, $Z_0, Z_1, \dots, Z_n$ là các biến ngẫu nhiên độc lập, có cùng phân phối chuẩn tắc $N(0,1)$. Suy ra $\frac{Z_0}{\sqrt{\frac{\chi_{n-1}^2}{n-1}}} = \frac{\sigma.Z_0}{S} \sim T_{n-1}$. Lấy $Z_0 = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$ ta thu được $\frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim T_{n-1}$.

KẾT LUẬN

Các đặc trưng mẫu có thể nói là phần không thể thiếu trong nghiên cứu thống kê cũng như trong nội dung giảng dạy thống kê của các trường đại học, cao đẳng. Báo cáo đã trình bày lại đầy đủ định nghĩa, tính chất và chứng minh tính chất của ba số đặc trưng mẫu cơ bản nhất là trung bình mẫu, phương sai mẫu và phương sai mẫu hiệu chỉnh. Báo cáo cũng chỉ ra và chứng minh về phân phối của ba đặc trưng mẫu nêu trên. Từ đó, báo cáo cung cấp tài liệu tham khảo cho các giảng viên giảng dạy thống kê toán học và các em sinh viên có nhu cầu tìm hiểu sâu kiến thức. Trong tương lai, với mục đích hoàn thiện giáo án, tạo nhiều bài giảng hấp dẫn, gần gũi cuộc sống và cung cấp nhiều tài liệu tham khảo bổ ích cho sinh viên, tôi sẽ tiếp tục nghiên cứu và làm rõ hơn những lý thuyết trong giáo trình cũng như ứng dụng của toán học trong thực tế đời sống cũng như ứng dung của toán học trong nhiều lĩnh vực khác.

TÀI LIỆU THAM KHẢO

1. Nguyễn Viết Phú, Nguyễn Duy Tiến, *Cơ sở Lý thuyết Xác suất*, NXB ĐHQG Hà Nội.

2. http://www2.stat.duke.edu/courses/Fall18/sta611.01/Lecture/lec12_mean_var_indep.pdf

3. Nguyễn Thu Hằng, *Tổng của các biến ngẫu nhiên độc lập cùng phân phối*, ERSD 2020.