NGHIÊN CỨU MẪU NGẪU NHIÊN ĐƠN GIẢN VÀ MẪU NGẪU NHIÊN PHÂN TẦNG TRONG BÀI TOÁN CHỌN MẪU NGHIÊN CỨU

SIMPLE RANDOM SAMPLING AND STRATIFIED RANDOM SAMPLING

Trần Thị Kim Thanh

Trường Đại học Kinh tế - Kỹ thuật Công nghiệp; Email: ttkthanh@uneti.edu.vn

Tóm tắt - Ngày nay toán học thống kê được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, bởi những ưu điểm của phương pháp này là cho kết quả trung thực, khách quan với sai số tương đối nhỏ. Sử dụng phương pháp này bắt buộc phải lấy mẫu, các mẫu độc lập với nhau và đại diện cho một miền nào đó. Tồn tại một thực tế, không ít trường hợp mẫu được lấy, lại không đại diện trung thực và khách quan cho tổng thể nghiên cứu, dẫn đến các kết quả nghiên cứu không mong muốn, thậm chí trái với thực tiễn. Bài báo nghiên cứu hai phương pháp lấy mẫu ngẫu nhiên (Phương pháp lấy mẫu ngẫu nhiên đơn giản và Phương pháp lấy mẫu ngẫu nhiên phân tầng). Kết quả nghiên cứu cho thấy, mẫu ngẫu nhiên phân tầng tuy phức tạp, tốn nhiều thời gian và chi phí nhưng lại cho độ chính xác cao hơn mẫu ngẫu nhiên đơn giản.

Từ khóa - mẫu; ngẫu nhiên; mẫu ngẫu nhiên; mẫu ngẫu nhiên đơn giản; mẫu ngẫu nhiên phân tầng.

Abstract - Mathematical statistics has been used in various areas because of its accurate and objective results, and relatively small errors. Using statistics in research involves the collecting of samples, or a set of independent samples representing a whole group. There remain, however, cases where sample selection is not unbiased, the samples do not accurately represent the whole population, and then the results are undesirable and even contrary to the law of practice. In this paper, we present our study of two random sampling methods: simple random sampling and stratified random sampling. While stratified random sampling costs and is a complex and time-consuming process, its accuracy is higher than that of simple random sampling.

Key words - sample; random; random sampling; simple random sampling; stratified random sampling.

1. Đặt vấn đề

Trong thực tế, người ta thường phải nghiên cứu một đặc tính của một tập hợp nào đó như: mức độ hài lòng của khách hàng đối với sản phẩm của doanh nghiệp, kiểm tra an toàn thực phẩm của kho hoa quả, trình độ văn hóa của một khu dân cư,… Để xử lý và rút ra các kết luận cần thiết, đôi khi người ta sử dụng phương pháp nghiên cứu toàn bộ, tuy nhiên việc áp dụng phương pháp này gặp phải không ít khó khăn như:

- Nếu quy mô của tập nghiên cứu lớn thì việc nghiên cứu toàn bộ sẽ đòi hỏi nhiều chi phí vật chất và thời gian; có thể xảy ra trường hợp tính trùng hoặc bỏ sót một số phần tử trong vùng cần nghiên cứu. Do đó, đòi hỏi phải đưa ra được các giải pháp tối ưu, chi tiết, chặt chẽ và thật khoa học để hạn chế sai sót không mong muốn trong quá trình thu thập số liệu ban đầu.

- Trong nhiều trường hợp không thể nắm được toàn bộ các phần tử của tập cần nghiên cứu, do đó không thể tiến hành nghiên cứu toàn bộ được.

- Nếu các phần tử của tập hợp lại bị phá hủy trong quá trình nghiên cứu thì cũng không tiến hành nghiên cứu toàn bộ được.

Để kết quả phản ánh một cách trung thực khách quan, người ta thường nghiên cứu trên một tập nhỏ hơn gọi là mẫu, từ tập lớn gọi là tổng thể để phân tích, xử lý và đưa ra kết quả cần thiết. Vấn đề đặt ra cần chọn mẫu đại diện như thế nào, để mang đầy đủ các đặc tính của tổng thể, từ đó có thể đưa ra được các kết luận nhanh chóng, kịp thời mà giảm chi phí, nhưng vẫn đảm bảo độ chính xác cần thiết.

Bài báo này là kết quả nghiên cứu dựa trên cơ sở hai phương pháp lấy mẫu ngẫu nhiên đơn giản và lấy mẫu ngẫu nhiên phân tầng của lý thuyết xác suất - thống kê, để đưa ra những kết luận đánh giá về hai phương pháp chọn mẫu ngẫu nhiên phổ biến thường được sử dụng, từ đó giúp các nhà thống kê vận dụng linh hoạt khi xử lí thông tin cần thu thập.

2. Phương pháp nghiên cứu

2.1. Phương pháp lấy mẫu ngẫu nhiên đơn giản [3]

Lấy mẫu ngẫu nhiên đơn giản là phương pháp chọn ngẫu nhiên n phần tử trong số N phần tử đã cho. Từ đây ta có hai phương án lấy mẫu: lấy mẫu có hoàn lại và không hoàn lại.

● **Trường hợp:** Lấy mẫu ngẫu nhiên có hoàn lại

Ta rút ngẫu nhiên một phần tử, sau đó lại trả phần tử đó về tập hợp ban đầu. Cứ tiếp tục như vậy cho đến khi rút được n phần tử. Các phần tử rút ra trả lại cho tổng thể nên phương pháp này gọi là lấy mẫu ngẫu nhiên có hoàn lại.

● **Trường hợp:** Lấy mẫu ngẫu nhiên không hoàn lại

Ta rút ngẫu nhiên một phần tử, sau đó lại tiếp tục rút ngẫu nhiên phần tử thứ hai. Cứ tiếp tục như vậy cho đến khi rút được n phần tử. Các phần tử rút ra không trả lại cho tổng thể nên phương pháp này gọi là lấy mẫu ngẫu nhiên không hoàn lại.

2.2. Phương pháp lấy mẫu ngẫu nhiên phân tầng [2]

Tổng thể nghiên cứu của N phần tử được chia thành các tập con gồm N_1 , N_2 , …, N_L phần tử không trùng lặp sao cho:

$$N_1 + N_2 + \ldots + N_L = N$$

Các tập con gọi là các tầng. Mẫu được rút ra từ mỗi tầng và việc lấy mẫu là độc lập với nhau đối với các tầng. Cỡ mẫu trong các tầng ký hiệu bởi n_1 , n_2 , …, n_L tương ứng ($n_1 + n_2 + \ldots + n_L = n$).

Nếu mỗi tầng lấy ra một mẫu ngẫu nhiên thì tất cả các mẫu đó gọi là mẫu ngẫu nhiên phân tầng. Khi $\frac{n_h}{N_h} = \frac{n}{N}$ ký hiệu $f_h = f \ \forall h$ tức là tỷ suất lấy mẫu giống nhau trong tất cả các tầng. Sự phân tầng này gọi là sự phân tầng với số lượng n_h tỷ lệ.

3. Kết quả và thảo luận

3.1. Đánh giá điều kiện thực hiện mẫu ngẫu nhiên phân tầng và mẫu ngẫu nhiên đơn giản

Cả hai phương pháp đều lấy mẫu ngẫu nhiên nên xác suất của mỗi phần tử đã biết và có xác suất chọn như nhau, nghĩa là từ danh sách tất cả các cá thể trong quần thể định chọn mẫu, ta chọn đối tượng đến khi đủ mẫu. Tuy nhiên, mẫu ngẫu nhiên phân tầng đòi hỏi sự thay đổi trong tầng phải nhỏ, tức là các tầng phải có các đặc điểm chung như yếu tố vùng miền, giới tính, nhóm tuổi,… Nhưng sự thay đổi giữa các tầng phải đủ lớn để mỗi tầng được xét như một tổng thể riêng biệt, độc lập, từ đó trên mỗi tầng có thể lựa chọn phương pháp lấy mẫu phù hợp hoặc hiệu quả về giá nhất. Ví dụ: Một tòa soạn báo muốn tiến hành nghiên cứu trên một mẫu 1000 doanh nghiệp trong nước về sự quan tâm của họ với tờ báo nhằm tiếp thị việc đưa thông tin quảng cáo trên báo. Tòa soạn có thể căn cứ vào các tiêu chí: vùng địa lý (miền Bắc, miền Trung, miền Nam); hình thức sở hữu (quốc doanh, ngoài quốc doanh, công ty 100% vốn nước ngoài,..) để quyết định cơ cấu mẫu nghiên cứu. Số lượng mẫu trên từng tầng có thể thực hiện theo hai cách: có thể dựa vào tỉ lệ cỡ dân số tại vùng đó với tổng thể, chẳng hạn với mẫu hai tầng: thành thị 60% tổng thể và nông thôn 40% thì với cỡ mẫu 5000, ta lấy tầng thành thị 3000 và tầng nông thôn 2000 hoặc cỡ mẫu được chọn tương đương giữa các tầng. Vì vậy, mẫu ngẫu nhiên phân tầng phải lựa chọn được biến phân tầng hợp lí, do đó khó thực hiện hơn mẫu ngẫu nhiên đơn giản.

3.2. Đánh giá về thời gian và chi phí của mẫu ngẫu nhiên phân tầng và mẫu ngẫu nhiên đơn giản

Với bài toán lấy mẫu nghiên cứu, cỡ mẫu thường khá lớn với phạm vi điều tra rộng nên khi tiến hành phân tầng tổng thể, nhà thống kê phải điều tra để nắm rõ được các đặc điểm của vùng dân cư khảo sát như: yếu tố địa lý, trình độ văn hóa, tỉ lệ giới tính,… để tổng thể phân chia thành các nhóm nhỏ thực sự độc lập, phân biệt nhau. Do đó, khi tiến hành lấy mẫu ngẫu nhiên phân tầng sẽ tốn nhiều thời gian và chi phí hơn.

3.3. Đánh giá về độ chính xác tương đối giữa mẫu ngẫu nhiên phân tầng và mẫu ngẫu nhiên đơn giản

3.3.1. So sánh độ chính xác tương đối giữa hai mẫu ngẫu nhiên

Định lí sau cho ta kết quả mẫu ngẫu nhiên phân tầng chính xác hơn mẫu ngẫu nhiên đơn giản.

- **Định lí 1**

Ký hiệu: N_h là tổng số phần tử ở tầng h của tổng thể và V_{ran} , V_{prop} là phương sai của trung bình ước lượng của mẫu ngẫu nhiên đơn giản, mẫu ngẫu nhiên phân tầng với số lượng tỉ lệ

Nếu tỉ số $1/ N_h$ có thể bỏ qua được (tức là khá nhỏ so với 1) thì $V_{prop} \le V_{ran}$.

Chứng minh:

Theo định nghĩa

$$V_{ran} = (1-f)\frac{S^2}{n} \tag{1.1}$$

$$V_{prop} = \frac{1-f}{n}\sum W_h S_h^2 \tag{1.2}$$

Trong đó:

$S_h^2 = \dfrac{\sum_{i=1}^{N_h}\left(y_{h_i} - \overline{Y_h}\right)^2}{N_h - 1}$ là phương sai chân thực tầng h;

$W_h = \dfrac{N_h}{N}$ là trọng số tầng h;

$S^2 = \dfrac{\sum_h \sum_{i=1}^{N_h}\left(y_{h_i} - \overline{Y}\right)^2}{N-1}$ là phương sai của tổng thể.

Ta có:

$$(N-1)S^2 = \sum_h \sum_{i=1}^{N_h}\left(y_{h_i} - \overline{Y}\right)^2$$

$$\Rightarrow (N-1)S^2 = \sum_h \sum_{i=1}^{N_h}\left(y_{h_i} - \overline{Y_h}\right)^2 + \sum_h N_h\left(\overline{Y_h} - \overline{Y}\right)^2$$

$$\Rightarrow (N-1)S^2 = \sum_h (N_h - 1)S_h^2 + \sum_h N_h\left(\overline{Y_h} - \overline{Y}\right)^2 \tag{1.3}$$

Nếu số hạng $1/N_h$ bỏ qua được và do đó 1/N bỏ qua được thì (1.3) trở thành:

$$S^2 = \sum_h W_h S_h^2 + \sum_h W_h\left(\overline{Y_h} - \overline{Y}\right)^2 \tag{1.4}$$

Do đó, từ (1.1) và (1.4) ta có:

$$V_{ran} = (1-f)\frac{S^2}{n}$$

$$= \frac{1-f}{n}\sum_h W_h S_h^2 + \frac{1-f}{n}\sum_h W_h\left(\overline{Y_h} - \overline{Y}\right)^2$$

$$\Rightarrow V_{ran} = V_{prop} + \frac{1-f}{n}\sum_h W_h\left(\overline{Y_h} - \overline{Y}\right)^2 \tag{1.5}$$

Điều được chứng minh.

- **Ví dụ**

Số dân của 63 tỉnh, thành phố của nước ta năm 2012 được thể hiện trên Bảng 1 (số liệu lấy ở [5]). Các thành phố được sắp xếp theo hai tầng, tầng đầu tiên gồm 41 tỉnh, thành phố và tầng thứ hai gồm 22 tỉnh, thành phố còn lại. Tổng số dân trong tất cả các thành phố được ước lượng từ một cỡ mẫu 23.

Ta tính được tổng thể đầy đủ: $\overline{Y} = \frac{88772,9}{63} \approx 1409,09$

$$S^2 = \frac{217240908,2}{62} - \frac{63}{62}(1409,09)^2$$

$$\Rightarrow S^2 \approx 1\,486\,326,24$$

Bảng 1. *Dân số các tỉnh, thành phố của nước ta năm 2012 (đơn vị: nghìn người)*

Tầng			
h = 1		**h = 2**	
Tỉnh(TP)	**Số dân**	**Tỉnh(TP)**	**Số dân**
Hà Nội	6844,1	Hà Nam	790

Vĩnh Phúc	1020,6	Ninh Bình	915,9
Bắc Ninh	1079,9	Hà Giang	758
Quảng Ninh	1177,2	Cao Bằng	515,2
Hải Dương	1735,1	Bắc Kạn	301
Hải Phòng	1904,1	Tuyên Quang	738,9
Hưng Yên	1145,6	Lào Cai	646,8
Thái Bình	1787,3	Yên Bái	764,4
Nam Định	1836,9	Lạng Sơn	744,1
Thái Nguyên	1150,2	Điện Biên	519,3
Bắc Giang	1588,5	Lai Châu	397,5
Phú Thọ	1335,9	Hòa Bình	806,1
Sơn La	1134,3	Quảng Bình	857,9
Thanh Hóa	3426,6	Quảng Trị	608,1
Nghệ An	2952	Đà Nẵng	973,8
Hà Tĩnh	1230,5	Phú Yên	877,2
Thừa Thiên Huế	1114,5	Ninh Thuận	576,7
Quảng Nam	1450,1	Kon Tum	462,4
Quảng Ngãi	1227,9	Đắc Nông	543,2
Bình Định	1501,8	Bình Phước	912,7
Khánh Hòa	1183	Hậu Giang	769,7
Bình Thuận	1193,5	Bạc Liêu	873,4
Lào Cai	1342,7		
Đắc Lắc	1796,7		
Lâm Đồng	1234,6		
Tây Ninh	1089,9		
Bình Dương	1748		
Đồng Nai	2720,8		
Bà Rịa-Vũng Tàu	1039,2		
TP HCM	7681,7		
Long An	1458,2		
Tiền Giang	1692,5		
Bến Tre	1258,5		
Trà Vinh	1015,3		
Vĩnh Long	1033,6		
Đồng Tháp	1676,3		
An Giang	2153,7		
Kiên Giang	1726,2		
Cần Thơ	1214,1		
Sóc Trăng	1301,9		
Cà Mau	1217,1		

Bảng 2. *Tổng và tổng bình phương*

Tầng	Σy_{hi}	Σy_{hi}^2
h = 1	73420,6	205814253,6
h = 2	15352,3	11426654,55
Σ	88772,9	217240908,2

- Với mẫu ngẫu nhiên đơn giản:

$$V_{ran} = (1-f)\frac{S^2}{n} = \frac{N-n}{N}\frac{S^2}{n}$$

$$\Rightarrow V_{ran} = \frac{40}{63}.\frac{1486326,24}{23} = 41030,4$$

- Với mẫu phân tầng hai tầng với số lượng tỉ lệ:

$S^2_1 \approx 1\,858\,415,82$; $N_1 = 41$

$S^2_2 \approx 33\,968,18$; $N_2 = 22$

$$\Rightarrow V_{prop} = \frac{40}{23.63}[\frac{41}{63}.1858\,415,82 + \frac{22}{63}.33968,18]$$

$$\Rightarrow V_{prop} \approx 33714,48$$

Nhận xét: Trong ví dụ này, mẫu hai tầng được phân tầng tương đối hợp lí, tính đại diện và khái quát hóa cao (hai tầng có phương sai chênh lệch gần 55 lần). Kết quả mẫu hai tầng với số lượng tỉ lệ là chính xác hơn mẫu ngẫu nhiên đơn giản (độ chính xác tăng hơn 18,95%). **(I)**

3.3.2. Điều chỉnh độ chính xác trong mẫu ngẫu nhiên phân tầng

Trong mẫu ngẫu nhiên phân tầng, giá trị cỡ mẫu n_h ở tầng h tương ứng được lựa chọn có thể làm cực tiểu V_{prop} tức làm tăng độ chính xác. Điều này được thể hiện trong định lí về sự phân bổ Neymann.

- **Định lí 2 (Sự phân bổ Neymann)** [4]

Trong mẫu ngẫu nhiên phân tầng, V_{prop} nhỏ nhất với tổng cỡ mẫu n cố định nếu

$$n_h = n.\frac{N_h.S_h}{\sum N_h.S_h}$$

Khi đó, thay giá trị n_h vào công thức phương sai trung bình ước lượng của mẫu ngẫu nhiên phân tầng, ta được:

$$V^{\min}_{prop} = \frac{(\sum W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N}$$

Bây giờ, ta sẽ xây dựng công thức xác định mức chênh lệch cao nhất về độ chính xác có thể đạt được giữa việc chọn mẫu nghiên cứu là mẫu ngẫu nhiên đơn giản và mẫu ngẫu nhiên phân tầng.

Ta có: $V_{prop} \geq V_{prop}^{\min}$

$$V_{prop} - V^{\min}_{prop} = \frac{1}{n}\left[\sum W_h S_h^2 - (\sum W_h S_h)^2\right] \quad (2.1)$$

Từ (2.1) và (1.5), ta có:

$$V_{ran} - V^{\min}_{prop} = \frac{1-f}{n}\sum_h W_h\left(\overline{Y_h} - \overline{Y}\right)^2 + \frac{1}{n}\left[\sum W_h S_h^2 - (\sum W_h S_h)^2\right] \quad (2.2)$$

Hệ thức (2.2) biểu diễn độ chênh lệch giữa phương sai của mẫu ngẫu nhiên đơn giản và mẫu ngẫu nhiên phân tầng tối ưu nhất. Đặt vế phải của hệ thức (2.2) bằng A thì A gồm 2 thành phần: thành phần đầu tiên (số hạng sau dấu "=") thể hiện độ lệch giữa các trung bình tầng, số hạng còn lại là sự chênh lệch giữa mẫu phân tầng tỉ lệ và mẫu phân tầng tối ưu.

Sử dụng hệ thức:

$$\sqrt{a} - \sqrt{b} = \frac{a-b}{\sqrt{a}+\sqrt{b}} (a>0, b>0)$$

Ta có:

$$\sqrt{V_{ran}} - \sqrt{V^{\min}_{prop}} = \frac{A}{\sqrt{V_{ran}} + \sqrt{V_{ran}-A}}$$

Hay:

$$\sqrt{V_{ran}} - \sqrt{V^{\min}_{prop}} = \frac{A}{\sqrt{(1-f)\frac{S^2}{n}} + \sqrt{(1-f)\frac{S^2}{n} - A}} \quad (2.3)$$

với

$$A = \frac{1-f}{n}\sum_h W_h\left(\overline{Y_h} - \overline{Y}\right)^2 + \frac{1}{n}\left[\sum W_h S_h^2 - (\sum W_h S_h)^2\right]$$

Hệ thức (2.3) cho ta kết quả cần tìm. **(II)**

4. Kết luận

Bài báo nghiên cứu hai phương pháp lấy mẫu: Lấy mẫu ngẫu nhiên đơn giản và lấy mẫu ngẫu nhiên phân tầng dựa trên cơ sở Toán Lý thuyết Xác suất - Thống kê.

Từ định nghĩa, bài báo đưa ra kết quả, đánh giá, so sánh về thời gian, chi phí và độ chính xác của hai phương pháp lấy mẫu ngẫu nhiên khi tiến hành thu thập mẫu đại diện. Đánh giá này được kiểm chứng trong việc xử lý số liệu khi chọn mẫu nghiên cứu trên tổng thể là dân số các tỉnh, thành phố nước ta năm 2012. **(III)**

Từ những kết quả trên, nghiên cứu đã chỉ ra được lấy mẫu ngẫu nhiên phân tầng tuy phức tạp, tốn nhiều thời gian và chi phí nhưng cho kết quả chính xác hơn so với cách lấy mẫu ngẫu nhiên đơn giản. Hơn nữa, dựa vào định lí về sự phân bổ Neymann trong mẫu phân tầng thì sự chính xác của mẫu ngẫu nhiên phân tầng hoàn toàn có thể điều chỉnh tối ưu nhất (độ chính xác lớn nhất có thể). Tác giả cũng xây dựng được công thức (2.3) xác định giá trị mức chênh lệch về độ chính xác cao nhất có thể đạt được khi chọn mẫu nghiên cứu là mẫu ngẫu nhiên phân tầng lý tưởng (mẫu ngẫu nhiên phân tầng tối ưu) và chọn mẫu là mẫu ngẫu nhiên đơn giản. **(IV)**

Kết quả nghiên cứu là cơ sở khoa học cho việc ứng dụng vào thực tiễn để giải quyết các bài toán lấy mẫu có nhiều tham số đưa ra kết quả tối ưu.

TÀI LIỆU THAM KHẢO

[1] Đào Hữu Hồ (2008), *Xác suất thống kê"*, in lần thứ 11, Nhà xuất bản Đại học Quốc gia Hà Nội.

[2] Tống Đình Quỳ (2003), *Giáo trình xác suất thống kê*, trang 115, Nhà xuất bản Đại học Quốc gia Hà Nội.

[3] Đào Hữu Hồ, Nguyễn Văn Hữu, Hoàng Hữu Như (2004), *Thống kê toán học*, trang 1- 2, Nhà xuất bản Đại học Quốc gia Hà Nội.

[4] William G. Cochran, Sampling techniques (1977), third eddition, *JOHN WILLEY & SONS, INC, 94.*

[5] www.gso.gov.vn

(BBT nhận bài: 14/09/2014, phản biện xong: 26/09/2014)