BÀI 5 CƠ SỞ LÝ THUYẾT MẪU

Hướng dẫn học

Bài này chuyển sang phần thứ hai của môn học, là phần Thống kê toán. Phần này không yêu cầu các suy luận logic nhiều như phần Lý thuyết xác suất, nhưng yêu cầu tính toán chính xác và thành thạo. Với các bài tập phần Thống kê toán, sinh viên cần nhận ra câu hỏi đặt vấn đề về điều gì trong các bài, về tham số, đại lượng nào, về loại bài tập nào để lựa chọn phương hướng và công thức đúng.

Với bài số 5, sinh viên tập trung vào khái niệm Mẫu và các đại lượng tính toán trong mẫu, phân biệt với tổng thể trong nghiên cứu. Các kiến thức của phần lý thuyết xác suất cần được nắm vững để áp dụng tại bài này, đặc biệt là các kiến thức về phân phối Chuẩn và phân phối Không – một. Kiến thức trong bài này là một bộ phận của bài toán thống kê hoàn chỉnh sẽ được giới thiệu tiếp trong bài 6 và 7. Để có thể làm được các bài tập trong chương sau buộc phải nắm được các ký hiệu, tên gọi, cách tính các đại lượng ở bài này. Những ứng dụng thực tế của bài này sinh viên có thể tham khảo thêm trong các tài liệu và kết hợp sử dụng các phần mềm máy tính.

Để học tốt bài này, sinh viên cần tham khảo các phương pháp học sau:

- Học đúng lịch trình của môn học theo tuần, làm các bài luyện tập đầy đủ và tham gia thảo luận trên diễn đàn.
- Đọc tài liệu: Giáo trình Lý thuyết xác suất và thống kê toán của NXB Đại học KTQD.
- Sinh viên làm việc theo nhóm và trao đổi với giảng viên trực tiếp tại lớp học hoặc qua email.
- Tham khảo các thông tin từ trang Web môn học.

Nội dung

- Khái niệm tổng thể;
- Các tham số của tổng thể;
- Khái niệm mẫu ngẫu nhiên, mẫu cụ thể;
- Thống kê đặc trưng mẫu: Trung bình, phương sai, độ lệch chuẩn, tỷ lệ;
- Các quy luật phân phối xác suất liên quan.

Mục tiêu

- Hiểu và phân biệt khái niệm Tổng thể và Mẫu;
- Hiểu và phân biệt khái niệm Tham số và Thống kê;
- Tính chính xác các thống kê đặc trưng mẫu bằng máy tính bấm tay;
- Nhớ được quy luật liên hệ để áp dụng tra bảng số.

Tình huống dẫn nhập

Khảo sát khách hàng

Trong các bài trước về xác suất, khi biết các thông tin cơ bản về kỳ vọng, phương sai của biến ngẫu nhiên, có thể tính được xác suất xảy ra một điều gì đó. Tuy nhiên trong thực tế để có được kỳ vọng, phương sai đó là điều không dễ. Liệu có cách thức nào tìm hiểu được những thông tin cơ bản đó chỉ thông qua một số thông tin điều tra được hay không?

Xét tình huống sau.

Có rất nhiều khách mua hàng ở cửa hàng, tuy nhiên người quản lý chỉ có trong tay hóa đơn thanh toán của 100 khách hàng tại một ngày (đơn vị: nghìn đồng) như sau:

169	210	160	196	203	221	208	174	260	164
119	177	248	208	321	214	283	234	197	221
234	118	191	141	60	197	182	195	287	311
299	219	174	179	165	237	156	225	68	138
181	294	116	173	234	211	223	256	338	175
204	220	159	171	258	174	184	189	182	301
130	152	210	157	297	195	65	101	216	227
221	304	174	106	198	160	252	198	153	139
176	234	281	232	196	165	301	211	222	170
175	184	127	215	227	258	195	160	219	231

Với số liệu trên, có thể nhận định thế nào về số tiền khách hàng chi tiêu nói chung (không phải chỉ 100 khách hàng này) tại cửa hàng?

1. Khách hàng nói chung có chi tiêu thế nào?
2. Số tiền khách chi là phân phối xác suất thế nào?
3. Số tiền trung bình tất cả các khách hàng chi là bao nhiêu?
4. Phương sai số tiền tất cả khách hàng chi là bao nhiêu?

Để trả lời được câu hỏi trên, cần hiểu rõ về số liệu đã cho:

- Số tiền chi của 100 khách hàng này phân phối thế nào?
- Số tiền chi trung bình của 100 khách hàng này là bao nhiêu?
- Phương sai số tiền chi của 100 khách hàng này là bao nhiêu?
- Tỷ lệ khách chi nhiều hơn 300 nghìn đồng bằng bao nhiêu?

5.1. Khái niệm cơ bản

Ta sẽ đề cập các khái niệm cơ bản của thống kê toán, những thuật ngữ, ký hiệu tại đây sẽ tiếp tục được sử dụng trong các mục sau.

5.1.1. Biến ngẫu nhiên gốc

Thống kê toán nghiên cứu các hiện tượng trong kinh tế – xã hội dựa trên các thông tin thu được từ các đối tượng nghiên cứu về một vấn đề nghiên cứu nào đó. Từ vấn đề nghiên cứu, ta sẽ có những khái niệm cơ bản là *đối tượng nghiên cứu*, *dấu hiệu nghiên cứu*, *đại lượng nghiên cứu*.

Ví dụ 5.1. Nghiên cứu sự hài lòng của sinh viên đang học Đại học Kinh tế Quốc dân (ĐHKTQD) với phương pháp giảng dạy của giảng viên của trường, đối tượng nghiên cứu sẽ là *các sinh viên đang học tại trường*. Dấu hiệu nghiên cứu là *sự hài lòng*. Tuy nhiên sự hài lòng là khái niệm trừu tượng, phải được thể hiện qua đại lượng đánh giá. Có hai cách thể hiện đánh giá:

- Cách 1: lấy ý kiến và ý kiến chỉ được có hai loại, Không hài lòng và Hài lòng, khi đó hai kiểu đánh giá đó có thể được thể hiện qua một đại lượng Không – một giá trị 0 ứng với trường hợp Không hài lòng và giá trị 1 ứng với trường hợp Hài lòng. Đại lượng 0 – 1 đó là đại lượng nghiên cứu.
- Cách 2: đặt một thang điểm từ 1 đến 5 với con số càng lớn thể hiện sự hài lòng càng nhiều. Như vậy cách đánh giá thể hiện qua một đại lượng với các giá trị rời rạc. Mức điểm là đại lượng nghiên cứu.

Ví dụ 5.2 (Ví dụ tình huống). Người quản lý cửa hàng quan tâm đến số tiền mà khách hàng chi tiêu tại cửa hàng của mình. Khi đó đối tượng nghiên cứu là các *khách hàng*, dấu hiệu nghiên cứu trong trường hợp này cũng là đại lượng nghiên cứu, là số tiền khách hàng chi. Số tiền của các khách hàng chi ra là không giống nhau, có thể coi như một đại lượng gần như liên tục.

Qua ví dụ trên có thể thấy các vấn đề nghiên cứu có thể quy về một hoặc một số đại lượng bằng số. Đại lượng đó có thể chỉ có hai giá trị như trường hợp Hài lòng hay Không hài lòng, có thể có một số hữu hạn giá trị, hoặc có vô hạn giá trị. Trong ngôn ngữ của Lý thuyết xác suất, đại lượng đó chính là các biến ngẫu nhiên, và gọi là *Biến ngẫu nhiên gốc*.

Định nghĩa 5.1 – Đại lượng nghiên cứu. *Với một vấn đề nghiên cứu, biến ngẫu nhiên gốc chính là đại lượng nghiên cứu, nhận các giá trị ngẫu nhiên tùy từng đối tượng nghiên cứu.*

Với ví dụ 5.1, theo cách đánh giá 1 thì biến ngẫu nhiên gốc $X = \{0 ; 1\}$; với cách đánh giá 2 thì biến ngẫu nhiên gốc là $X = \{1 ; 2 ; 3 ; 4 ; 5\}$. Với ví dụ 5.2, biến ngẫu nhiên gốc là số tiền khách hàng chi, do chỉ xét những khách hàng có chi tiền, nên $X = (0; +\infty)$.

5.1.2. Phương pháp nghiên cứu

Để có được thông tin về các đối tượng, có hai phương pháp nghiên cứu là nghiên cứu Tổng thể và nghiên cứu Mẫu.

Nghiên cứu tổng thể: là nghiên cứu toàn bộ các đối tượng theo dấu hiệu nghiên cứu đã xác định.

Ưu điểm của nghiên cứu tổng thể là thông tin sẽ đầy đủ, chính xác, trọn vẹn.

Tuy nhiên nghiên cứu tổng thể có hạn chế sau:

- Phải trả chi phí lớn về kinh tế và thời gian do số lượng các phần tử trong tập hợp toàn bộ có thể rất lớn.
- Có thể dẫn tới phá hủy toàn bộ tập hợp nghiên cứu. Chẳng hạn nghiên cứu thời gian hoạt động của các thiết bị điện tử hoặc các dây chuyền sản xuất đồ hộp. Khi áp dụng phương pháp này sẽ dẫn tới phá hủy toàn bộ các thiết bị điện tử và các sản phẩm đồ hộp.
- Có những tập hợp mà ta không thể nghiên cứu được toàn bộ vì không thể có đầy đủ thông tin. Chẳng hạn như nghiên cứu ô nhiễm nước ở một dòng sông mà muốn lấy thông tin toàn bộ nước ở dòng sông là không khả thi.

Với ví dụ 5.1 về sự hài lòng của sinh viên ĐHKTQD, thông tin của tổng thể có thể thu thập với sự hỗ trợ của phòng đào tạo và các đơn vị khác, tuy nhiên, có nhiều trường hợp sinh viên xin bảo lưu hoặc hiện không có mặt tại trường, hoặc không muốn trả lời, nên cũng khó có thể có toàn bộ thông tin từ toàn bộ sinh viên. Bên cạnh đó nhiều sinh viên còn không trả lời thật, nên thông tin thu được dù nhiều cũng chưa phải là thông tin tổng thể.

Với ví dụ 5.2 về mức chi của khách hàng, nếu chỉ quan tâm đến các khách hàng đã từng mua hàng thì với hệ thống thanh toán hiện đại, có thể có đầy đủ các hóa đơn của khách. Tuy nhiên nếu có giai đoạn chưa áp dụng thiết bị hiện đại thì thông tin có thể không được lưu trữ, hoặc thông tin vì một lý do nào đó đã được xóa. Tuy nhiên trong việc quan tâm đến mức chi của khách hàng thì tổng thể có thể không chỉ là cách khách hàng đã mua, mà còn là sẽ mua. Khi đó tổng thể là không thể điều tra được.

Trên thực tế, đa số các trường hợp nghiên cứu toàn bộ tổng thể là không khả thi. Khi đó ta sử dụng phương pháp nghiên cứu mẫu.

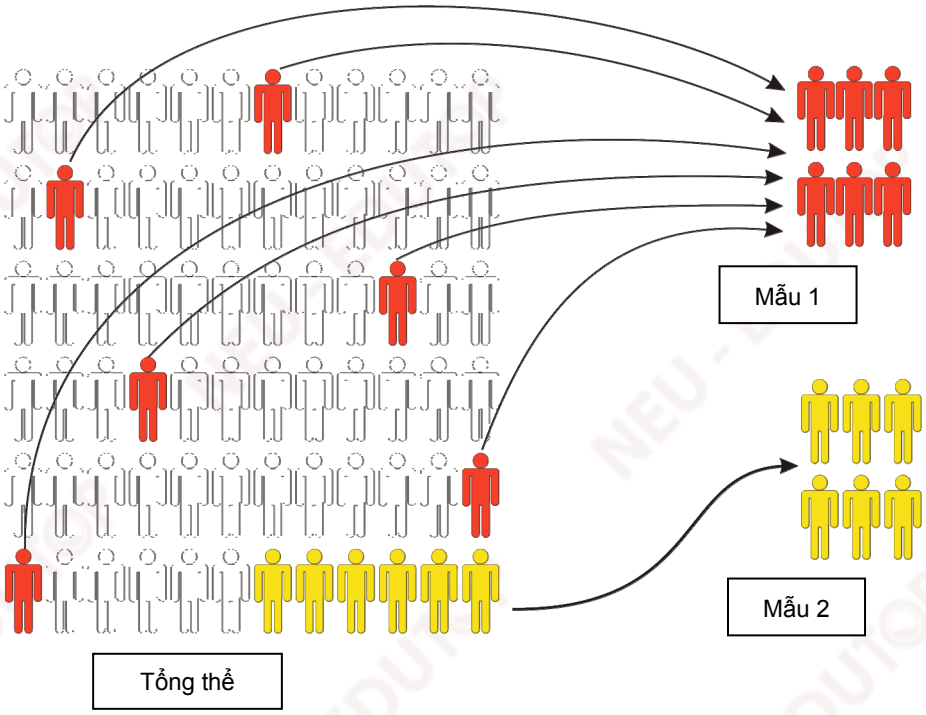
Nghiên cứu mẫu: là nghiên cứu bộ phận, từ tổng thể nghiên cứu ta lấy ra một tập con và nghiên cứu các phần tử trong tập con đó và từ đó ra kết luận cho toàn bộ các phần tử trong tập hợp nghiên cứu.

Phương pháp này thường được áp dụng trên thực tế vì các ưu điểm sau:

- Tính khả thi: khi tổng thể là không thể điều tra toàn bộ được thì phải chọn mẫu.
- Chi phí ít tốn kém hơn so với điều tra toàn bộ tổng thể.
- Khả năng bị trùng lặp thấp, và vì không phải điều tra toàn bộ nên có thể bỏ qua một số phần tử.
- Lượng thông tin thu thêm được trên các phần tử điều tra có tính giảm dần.
- Nếu mẫu được lấy ngẫu nhiên và khoa học thì các thông tin vẫn đảm bảo tính chính xác.

Với ví dụ 5.1 ta có thể xây dựng 1 mẫu điều tra một số sinh viên dựa trên các lớp chuyên ngành, các lớp tín chỉ hoặc cơ cấu các môn học để thu thập thông tin. Với ví dụ 5.2, thông tin từ ví dụ tình huống ở đầu bài học là một mẫu.

Hình 5.1 thể hiện mối quan hệ giữa tổng thể và mẫu điều tra.



Hình 5.1. ***Từ tổng thể rút ra hai mẫu. Mẫu nào tốt hơn, tại sao?***

Vấn đề của thống kê toán chính là làm sao có thể dùng thông tin từ Mẫu trả lời cho các câu hỏi đặt ra về Tổng thể. Tổng thể là điều ta cần biết, muốn biết nhưng trong hầu hết các trường hợp của kinh tế – xã hội, thông tin của tổng thể là không biết hoặc không thể biết. Thông tin của mẫu có thể có nhưng lại không đầy đủ và hoàn hảo. Nếu ta chờ có được thông tin của tổng thể thì các quyết định có thể sẽ không bao giờ thực hiện được. Khi đó buộc phải dùng thông tin từ mẫu để quyết định về tổng thể, sao cho chính xác nhất có thể có. Thống kê toán sẽ thực hiện yêu cầu này.

5.2. Tổng thể nghiên cứu

Để có thể thực hiện việc nghiên cứu tổng thể, trước hết ta cần xem xét các khái niệm về bản chất và các đại lượng tính toán trên tổng thể.

5.2.1. Định nghĩa

Định nghĩa 5.2 – Tổng thể: *Tổng thể là tập hợp các phần tử đồng nhất theo một dấu hiệu nghiên cứu định tính hoặc định lượng nào đó.*

Số phần tử trong tổng thể gọi là kích thước của tổng thể, ký hiệu là N; N có thể bằng vô cùng.

Ví dụ 5.3.

- Tổng thể về đánh giá của sinh viên đại học các hệ đang học tại ĐHKTQD, *N* bằng 20 nghìn (số liệu của Phòng Quản lý đào tạo).
- Tổng thể về mức chi của khách hàng đã và sẽ mua ở một cửa hàng, *N* có thể bằng vô cùng.
- Tổng thể về số vốn đăng ký của các doanh nghiệp thành lập mới trong năm 2013, *N* bằng 76955 (con số của Tổng cục Thống kê công bố).

- Tổng thể về giá vàng bán ra tại các cửa hàng trên địa bàn Hà Nội trong năm 2013, N rất lớn, có thể coi như vô cùng vì có rất nhiều cửa hàng, có nhiều ngày bán, trong mỗi ngày giá lại có thể thay đổi, giá bán cho người mua khác nhau có thể khác nhau, giá bán khi bán với tổng khối lượng khác nhau cũng khác nhau.

Khi nghiên cứu tổng thể thì dấu hiệu nghiên cứu trong tổng thể có thể là định lượng hoặc định tính, do đó cũng có hai loại biến tương ứng là biến định lượng và biến định tính.

Biến định lượng: là các biến số, thể hiện các số đo của phần tử trong tổng thể nghiên cứu.

Ví dụ: Cân nặng, chiều cao, tuổi, thu nhập…

Khi đó biến ngẫu nhiên gốc của tổng thể chính là đại lượng đo lường đó, và có đơn vị là đơn vị của đại lượng đo lường.

Biến định tính: là các biến chất lượng, thể hiện tính chất nào đó không lượng hóa được của phần tử trong tổng thể nghiên cứu.

Ví dụ: Giới tính người lao động (nam, nữ), loại tốt nghiệp của sinh viên (giỏi, khá, trung bình), Hình thức sở hữu của doanh nghiệp (nhà nước, tư nhân, nước ngoài…).

Biến định tính còn có thể phân làm hai loại nhỏ hơn là biến định danh và biến thứ bậc. Với yếu tố định tính có thể có nhiều trạng thái, thường đặt mã hóa để chuyển hóa thành con số. Chi tiết và phân loại và mã hóa có thể xem trong giáo trình.

Trong chương trình môn học này, ta chỉ xét loại biến định tính có hai trạng thái: Có và Không có một tính chất nào đó, ví dụ chỉ xét giới là Nam và Không phải nam, tốt nghiệp loại Giỏi và không phải loại Giỏi, sở hữu Tư nhân và Không phải sở hữu tư nhân. Do đó nếu gán 1 cho trường hợp Có và 0 cho trường hợp Không có thì biến ngẫu nhiên gốc có dạng Không – một.

Như vậy với tổng thể có kích thước N, biến ngẫu nhiên gốc X trong tổng thể có thể viết dưới dạng: $X = \{x_1, x_2,\ldots, x_N\}$ với x_i là các giá trị có thể có, $i = 1,2,\ldots, N$. Nếu dấu hiệu nghiên cứu định tính thì x_i chỉ có thể là 0 hoặc 1.

5.2.2. Mô tả tổng thể

Khi biến ngẫu nhiên gốc X gồm các phần tử $\{x_1, x_2,\ldots, x_N\}$, việc liệt kê tất cả các phần tử có thể rất dài khi số lượng phần tử là rất lớn. Nếu ta không quan tâm từng phần tử gắn với giá trị nào mà chỉ quan tâm đến độ lớn và sự phân bố của giá trị của X, thì việc liệt kê đủ N con số là không cần thiết. Khi đó ta chỉ cần xét trên các con số khác nhau.

Giả sử trong N con số của N phần tử, chỉ có k giá trị khác nhau: $x_1, x_2,\ldots, x_k$, số lượng tương ứng của mỗi giá trị là $N_1, N_2,\ldots, N_k$, ta có thể thu gọn thông tin của tổng thể nghiên cứu bằng cách gộp các giá trị giống nhau lại và biểu diễn dưới dạng:

X	x_1	x_2	…	x_k
Tần số	N_1	N_2	…	N_k

trong đó N_i $(i = 1,2,\ldots, k)$ là số lần giá trị x_i xuất hiện trong tổng thể, gọi là *tần số tổng thể*, như vậy:

$$0 \le N_i \le N\ (i = 1,2,\ldots, k) \text{ và } \sum_{i=1}^{k} N_i = N$$

Đặt $p_i = \frac{N_i}{N}$ được gọi là *tần suất tổng thể* hay *tỷ lệ tổng thể* của giá trị x_i và ta có bảng tần suất, hay tỷ lệ của tổng thể:

X	x_1	x_2	…	x_k
Tần suất/Tỷ lệ	p_1	p_2	…	p_k

Trong đó: $0 \le p_i \le 1$ (i = 1,2,…, k) và $\sum_{i=1}^{k} p_i = 1$

Bảng tần suất giống như một bảng phân phối xác xuất của biến ngẫu nhiên, và gọi là *phân phối gốc* của tổng thể.

Ví dụ 5.4. Giả sử điều tra được đánh giá về phương pháp giảng dạy của giảng viên từ tất cả 20 nghìn sinh viên đang học tại ĐH KTQD, thang điểm từ 1 đến 5, nếu liệt kê tất cả các giá trị ta sẽ có dạng: $X = \{1 ; 5 ; 3 ; \dots ; 2\}$ với 20 nghìn con số, nhưng chỉ có 5 giá trị khác nhau. Khi đó bảng tần số và tần suất tổng thể có dạng:

Điểm đánh giá (x_i)	1	2	3	4	5
Tần số (N_i)	1000	2000	4000	8000	5000
Tần suất/Tỷ lệ (p_i)	$\frac{1}{20}$ = 0,05	$\frac{2}{20}$ = 0,1	$\frac{4}{20}$ = 0,2	$\frac{8}{20}$ = 0,4	$\frac{5}{20}$ = 0,25

Khi đó ta có thể nói: Tỷ lệ tổng thể hay tần suất tổng thể sinh viên đánh giá 5 điểm là 0,25 hay 25%. Cũng có thể nói rằng: Khi chọn ngẫu nhiên một sinh viên trong tổng thể thì xác suất để sinh viên đó đánh giá 5 điểm là 0,25. Tỷ lệ tổng thể cũng chính là xác suất.

Tỷ lệ, hoặc xác suất sinh viên đánh giá điểm từ 4 trở lên là $\frac{8+5}{20} = 0,65$ hay 65%.

5.2.3. Các tham số đặc trưng của tổng thể

Cũng giống như nghiên cứu biến ngẫu nhiên, khi nghiên cứu tổng thể ta cũng xét một số giá trị đặc trưng cơ bản để có thể phán đoán, phân tích, nhận xét.

Định nghĩa 5.3 – Tham số tổng thể: *Các đại lượng tính trên các đại lượng nghiên cứu của tổng thể, hay trên biến ngẫu nhiên gốc, phản ánh về một khía cạnh của tổng thể, gọi là* ***tham số tổng thể****, gọi tắt là* ***tham số****.*

Có rất nhiều loại tham số, ta tập trung vào các tham số là Trung bình tổng thể, Phương sai tổng thể, Độ lệch chuẩn tổng thể, Tỷ lệ tổng thể.

Trung bình tổng thể

Định nghĩa 5.4 – Trung bình tổng thể: *Trung bình tổng thể, ký hiệu là* ***m****, là trung bình cộng tất cả các giá trị của biến ngẫu nhiên gốc trong tổng thể.*

Như vậy, công thức tính trung bình tổng thể là:

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Nếu chỉ có k giá trị khác nhau $x_1, x_2,\dots, x_k$ với các tần số tương ứng $N_1, N_2,\dots, N_k$ thì trung bình tổng thể có thể tính bằng công thức:

$$m = \frac{1}{N}\sum_{i=1}^{k} N_i x_i = \sum_{i=1}^{k} p_i x_i = E(X)$$

Ta thấy m chính là kỳ vọng của biến ngẫu nhiên gốc *X*. Trong nghiên cứu tổng thể, trung bình tổng thể m đại diện cho độ lớn của của đại lượng lượng hóa dấu hiệu nghiên cứu về mặt trung bình.

Trung bình tổng thể có đơn vị là đơn vị của *X*.

Ví dụ 5.5. Nếu khu vực A là một tổng thể, khu vực này có tổng cộng 1000 hộ gia đình, tổng thu nhập của cả khu vực là 1,8 triệu USD, thì trung bình thu tổng thể khu vực A là:

$$m_A = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{1000}\sum_{i=1}^{1000} x_i = \frac{1,8\times 10^6}{10^3} = 1800\,(USD)$$

Phương sai tổng thể

Nếu trung bình tổng thể cho biết giá trị bình quân của đại lượng trong tổng thể, thì khi cần đo sự biến động của các phần tử trong tổng thể, ta cần một đại lượng để đánh giá, là phương sai tổng thể.

Định nghĩa 5.5 – Phương sai tổng thể: *Phương sai tổng thể, ký hiệu là σ^2, được tính theo công thức*:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2$$

Giải thích ý nghĩa: giá trị $(x_i - m)$ cho biết sai lệch của một phần tử cá biệt so với trung bình tổng thể, thể hiện sự dao động của phần tử đó khỏi xu thế chung. Tuy nhiên giá trị này có thể mang dấu âm hoặc dương, do đó khi bình phương lên ta đã triệt tiêu dấu. Tổng các bình phương này thể hiện tổng biến động của các phần tử khỏi trung bình tổng thể, nên cần chia cho tổng số phần tử. Như vậy phương sai tổng thể chính là *trung bình của bình phương sai lệch của các phần tử khỏi giá trị trung tâm*.

Có thể chứng minh được: $\sigma^2 = V(X)$

Vậy phương sai tổng thể bằng phương sai biến ngẫu nhiên gốc *X*, có ý nghĩa hoàn toàn giống phương sai của biến ngẫu nhiên. Phương sai tổng thể có đơn vị là bình phương đơn vị của *X*.

Phương sai tổng thể σ^2 dùng để đo sự dao động, thay đổi, phân tán (hoặc đồng đều, ổn định, tập trung) của các giá trị phần tử trong tổng thể, hay các giá trị của biến ngẫu nhiên gốc *X*. Với cùng một biến ngẫu nhiên gốc, tổng thể nào có phương sai lớn hơn sẽ có sự dao động, thay đổi, phân tán nhiều hơn; ngược lại, tổng thể nào có phương sai nhỏ hơn sẽ có sự ổn định, đồng đều hơn đối với dấu hiệu nghiên cứu được đặc trưng bởi biến ngẫu nhiên *X*.

Độ lệch chuẩn tổng thể

Định nghĩa 5.6 – Độ lệch chuẩn tổng thể: *Độ lệch chuẩn tổng thể, ký hiệu là σ, là căn bậc hai của phương sai tổng thể*: $\sigma = \sqrt{\sigma^2}$

Độ lệch chuẩn có đơn vị là đơn vị của *X*.

Tương tự như phương sai, độ lệch chuẩn cũng là một thước đo sự phân tán, dao động, đồng đều, ổn định của biến ngẫu nhiên. Độ lệch chuẩn càng lớn thì tổng thể càng phân tán, độ lệch chuẩn càng nhỏ thì tổng thể càng đồng đều.

Ví dụ 5.6. Nếu nghiên cứu hai khu vực A và B, với cùng biến ngẫu nhiên gốc X là thu nhập hộ gia đình, m_A, m_B lần lượt là trung bình tổng thể của khu vực A và khu vực B. Nếu $m_A > m_B$ thì có thể nói rằng thu nhập trung bình ở khu vực A cao hơn khu vực B, hoặc ngắn gọn hơn nữa là khu vực A có thu nhập cao hơn khu vực B (bỏ bớt chữ trung bình).

Nếu σ_A^2 và σ_B^2 lần lượt là phương sai tổng thể của khu vực A và khu vực B, và $\sigma_A^2 > \sigma_B^2$ thì có thể nói rằng thu nhập ở khu vực B đồng đều hơn khu vực A, hay thu nhập của khu vực A là phân tán hơn khu vực B. Cũng có thể nói rằng xét về thu nhập thì khu vực B bình đẳng hơn khu vực A.

Tỷ lệ tổng thể

Định nghĩa 5.7 – Tỷ lệ tổng thể: *Tỷ lệ tổng thể (hay còn gọi là tần suất tổng thể) của một dấu hiệu A, ký hiệu là p, là tỉ số giữa số phần tử của tổng thể mang dấu hiệu đó và kích thước tổng thể.*

Nếu ký hiệu số phần tử chứa dấu hiệu A là M, thì tần suất tổng thể của dấu hiệu A, được tính theo công thức:

$$p = \frac{M}{N}$$

Ví dụ 5.7 (Tiếp theo ví dụ 5.4). Với số liệu trong ví dụ 5.4, hãy tính

(a) Trung bình tổng thể

(b) Phương sai và độ lệch chuẩn tổng thể

(c) Tỷ lệ tổng thể điểm số nhỏ hơn 4

Điểm đánh giá (x_i)	1	2	3	4	5
Tần số (N_i)	1000	2000	4000	8000	5000
Tần suất/Tỷ lệ (p_i)	0,05	0,1	0,2	0,4	0,25

Giải:

(a) Trung bình tổng thể là:

$$m = \frac{1}{20000}(1\times1000 + 2\times2000 + 3\times4000 + 4\times8000 + 5\times5000)$$
$$= 1\times0,05 + 2\times0,1 + 3\times0,2 + 4\times0,4 + 5\times0,25 = 3,7$$

(b) Phương sai tổng thể là:

$$\sigma^2 = \frac{(1-3,7)^2 1000 + (2-3,7)^2 2000 + (3-3,7)^2 4000 + (4-3,7)^2 8000 + (5-3,7)^2 5000}{20000}$$

$= 1,21$ (điểm^2)

Cách tính nhanh hơn là:

$\sigma^2 = 1^2\times0,05 + 2^2\times0,1 + 3^2\times0,2 + 4^2\times0,4 + 5^2\times0,25 - 3,7^2 = 1,21$ (điểm^2)

Độ lệch chuẩn tổng thể là:

$\sigma = \sqrt{1,21} = 1,1$ (điểm)

Như vậy nếu chọn ngẫu nhiên một sinh viên bất kì, thì điểm đánh giá về phương pháp giảng dạy của giảng viên của sinh viên đó có kỳ vọng là 3,7 (điểm), phương sai 1,21 (điểm^2), độ lệch chuẩn 1,1 (điểm).

(c) Tỷ lệ tổng thể có điểm nhỏ hơn 4 là:

$$p_{(X<4)} = \frac{M_{(X<4)}}{N} = \frac{1000+2000+4000}{20000} = 0,35 \text{ hay } 35\%.$$

Cũng có thể nói: Nếu chọn ngẫu nhiên một sinh viên thì xác suất sinh viên đó đánh giá điểm số về phương pháp giảng dạy nhỏ hơn 4 là 0,35.

Ví dụ 5.8. Nghiên cứu về giới của trẻ sơ sinh tại một bệnh viện trong năm 2013, đặt X là biến ngẫu nhiên nhận giá trị bằng 0 nếu trẻ sơ sinh là con gái, $X = 1$ nếu trẻ sơ sinh là con trai. Giả sử trong năm 2013 có 1200 trẻ được sinh ra tại bệnh viện đó, trong đó có 612 con trai và 588 con gái. Do đó X nhận giá trị $x_i = 1$ với 612 trường hợp, $x_i = 0$ với 588 trường hợp còn lại, khi đó trung bình tổng thể là:

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{1200}\sum_{i=1}^{1200} x_i = \frac{612}{1200} = 0,51$$

Dễ thấy con số $\frac{612}{1200}$ cũng chính là tỷ lệ tổng thể của số trẻ trai. Trong trường hợp biến ngẫu nhiên gốc có dạng Không – một thì trung bình tổng thể cũng là tỷ lệ tổng thể.

5.3. Mẫu ngẫu nhiên

Như phần trên đã trình bày, tổng thể là rất khó hoặc thậm chí không thể thu thập đủ thông tin, các ví dụ trên chỉ là giả định. Khi đó để nghiên cứu được các vấn đề trong tổng thể, cần điều tra một mẫu, là một bộ phận thông tin của tổng thể.

5.3.1. Phương pháp chọn mẫu

Để phản ánh về tổng thể một cách chính xác nhất, người nghiên cứu mong muốn mẫu phải có tính đại diện tốt nhất. Để có một mẫu đại diện tốt nhất cho tổng thể người ta thường phải tiến hành xây dựng mẫu theo một quy định chọn ngẫu nhiên các phần tử của mẫu. Một mẫu như vậy được gọi là *mẫu ngẫu nhiên*.

Có rất nhiều phương pháp chọn mẫu ngẫu nhiên để thỏa mãn tính đại diện tốt nhất cho tổng thể và phù hợp với mục tiêu nghiên cứu:

- Mẫu ngẫu nhiên đơn;
- Mẫu ngẫu nhiên hệ thống;
- Mẫu chùm;
- Mẫu phân tổ;
- Mẫu nhiều cấp.

Trong nội dung bài giảng, ta không đi sâu vào các phương pháp lấy mẫu. Sinh viên có thể đọc thêm trong giáo trình. Ta sẽ đi sâu vào khái niệm về mẫu ngẫu nhiên trong mục sau.

5.3.2. Mẫu ngẫu nhiên và mẫu cụ thể

Trong mục trên có đề cập khái niệm mẫu ngẫu nhiên. Hiểu một cách đơn giản, mẫu là một bộ phận nhỏ hơn tương đối so với tổng thể, được rút ra từ tổng thể để điều tra. Phương pháp lấy mẫu ngẫu nhiên tức là làm sao để mỗi phần tử trong tổng thể đều có khả năng được điều tra là như nhau, hay xác suất để mỗi phần tử bị chọn là như nhau trong mỗi lần chọn. Vì trong mỗi lần chọn mẫu lấy ra một phần tử, và các phần tử đó có khả năng bị chọn là như nhau, nên chúng là *độc lập* với nhau, và các phần tử trong mỗi lần chọn có *các đặc tính là như nhau*. Vì vậy kỳ vọng, phương sai của đại lượng nghiên cứu với mỗi phần tử được chọn đều giống nhau.

Để lấy một mẫu gồm *n* phần tử, hay còn gọi là mẫu kích thước *n*, cần thực hiện *n* lần chọn ngẫu nhiên. Nếu mỗi lần chọn được một phần tử, và đại lượng nghiên cứu của phần tử đó chính là *X*, thì *X* là ngẫu nhiên và giống nhau ở mọi lần. Từ đó ta có định nghĩa về mẫu ngẫu nhiên.

Định nghĩa 5.8 – Mẫu ngẫu nhiên: *Một mẫu ngẫu nhiên kích thước n là tập hợp n biến ngẫu nhiên độc lập X_1, X_2,..., X_n được thành lập từ biến ngẫu nhiên X trong tổng thể và có cùng phân phối với biến ngẫu nhiên gốc X.*

Ký hiệu mẫu ngẫu nhiên: $W = (X_1, X_2,\ldots,X_n)$

Do mỗi lần lấy phần tử cho mẫu, biến ngẫu nhiên *X* đều là như nhau, do đó kỳ vọng và phương sai của chúng đều bằng nhau.

$$E(X_1) = E(X_2) = \ldots = E(X_n) = E(X) = m$$

$$V(X_1) = V(X_2) = \ldots = V(X_n) = V(X) = \sigma^2$$

Mẫu ngẫu nhiên như vậy là mẫu lấy một cách trừu tượng, chưa thực hiện. Khi thực hiện chọn n phần tử một cách thực sự, được một bộ số. Nếu lần chọn đầu tiên được giá trị là $X_1 = x_1$; lần chọn thứ hai $X_2 = x_2$,..., cho đến $X_n = x_n$ với $x_1, x_2,\ldots, x_n$ là các con số, thì ta có một mẫu đã điều tra, gọi là mẫu cụ thể, gồm n con số, hay chính là một bộ số liệu.

Định nghĩa 5.9 – Mẫu cụ thể: *Mẫu cụ thể là một bộ n số thực $(x_1, x_2,\ldots, x_n)$, là kết quả khi thực hiện một phép thử của mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$.*

Ký hiệu mẫu cụ thể là $w = (x_1, x_2, \ldots, x_n)$.

Mỗi con số gọi là một quan sát. Do đó mẫu kích thước n sẽ có n quan sát.

Như vậy:

- Mẫu ngẫu nhiên là một bộ *n* biến ngẫu nhiên, ký hiệu viết hoa.
- Mẫu cụ thể là một bộ số liệu gồm *n* con số cụ thể, ký hiệu viết thường.

Ví dụ 5.9 (tiếp ví dụ 5.4). Từ tổng thể 20 nghìn sinh viên trong ví dụ 5.4, với đại lượng nghiên cứu là điểm đánh giá của sinh viên, có trung bình tổng thể là 3,7 và phương sai tổng thể là 1,21. Nếu ta "*sẽ điều tra*" ngẫu nhiên 3 sinh viên, mỗi lần điều tra là một trong số 20 nghìn sinh viên đó, thì ta có bộ ba đại lượng, ký hiệu là (X_1, X_2, X_3), trong đó X_1 là kết quả thu được khi "sẽ điều tra" người thứ nhất, do đó X_1 là ngẫu nhiên. Ta có:

- Kỳ vọng của X_1 chính là trung bình tổng thể: $E(X_1) = m = 3,7$.
- Phương sai của X_1 chính là phương sai tổng thể: $V(X_1) = \sigma^2 = 1,21$.

Tương tự, X_2, X_3 là kết quả "sẽ điều tra" lần thứ hai và thứ ba, thì $E(X_2) = E(X_3) = 3,7$ và $V(X_2) = V(X_3) = \sigma^2 = 1,21$.

Bộ ba $W = (X_1, X_2, X_3)$ là mẫu ngẫu nhiên kích thước là $n = 3$.

Nếu điều tra thực sự ba sinh viên, người thứ nhất trả lời là 3, người thứ hai trả lời là 5, người thứ ba trả lời là 2, ta được kết quả: (3; 5; 2) là một mẫu cụ thể gồm 3 quan sát. Cũng có thể có mẫu cụ thể khác, chẳng hạn (2; 4; 4). Nếu gộp hai mẫu trên lại ta sẽ được mẫu gồm 6 quan sát: (3; 5; 2; 2; 4; 4).

Tương tự, có vô số mẫu cụ thể có thể xảy ra.

5.3.3. Mô tả mẫu cụ thể

Tương tự như mô tả tổng thể, để thể hiện mẫu cụ thể có nhiều cách trình bày số liệu.

Cách thứ nhất: Liệt kê tất cả các giá trị của mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$.

Cách thứ hai: Nếu các giá trị của mẫu gồm k giá trị có thể có là $x_1, x_2, \ldots, x_k$ với tần số tương ứng $n_1, n_2, \ldots, n_k$ hoặc tần suất tương ứng $f_1, f_2, \ldots, f_k$ (còn gọi là tần suất thực nghiệm) với $f_i = \frac{n_i}{n}$ thì có thể mô tả mẫu cụ thể w như sau

Giá trị (x_i)	x_1	x_2	…	x_k
Tần số (n_i)	n_1	n_2	…	n_k
Tần suất/Tỷ lệ (f_i)	f_1	f_2	…	f_k

Trong đó n_i là số lần giá trị x_i xuất hiện trong mẫu. Ta có:

$$\sum_{i=1}^{n} n_i = n \text{ và } \sum_{i=1}^{n} f_i = 1$$

Cách thứ ba: Ta có thể gộp các giá trị thành các nhóm để có thể nhận ra sự phân bố một cách dễ dàng hơn. Các nhóm thường có dạng giá trị trong một khoảng hoặc một đoạn nào đó.

Trong các tính toán về sau, cách thứ ba sẽ được quy về cách thứ hai bằng việc lấy giá trị ở giữa của khoảng hoặc đoạn làm đại diện.

Ví dụ 5.10: Khảo sát về độ tuổi của 200 khách hàng, đây là một bộ số liệu, mẫu cụ thể, có thể mô tả mẫu như sau:

Cách 1: Lập một danh sách gồm 200 con số là tuổi của 200 khách hàng. Danh sách này có thể quản lý bằng phần mềm như Excel.

Cách 2: Tuổi của khách hàng rời rạc từ 20 đến 59, có tổng cộng 40 trường hợp khác nhau, có thể lập bảng, chẳng hạn

Tuổi (x_i)	20	21	22	…	58	59
Số người (n_i)	3	2	6	…	4	1

Cách 3: Có thể gộp thành các nhóm tuổi, chẳng hạn

Nhóm tuổi	20–29	30–39	40–49	50–59
Số người	25	60	80	35

Trong thực tế, nhiều khi số liệu thu thập được đã ở dạng nhóm tuổi chứ không phải dạng số rời rạc, vì người điều tra không cần biết tuổi một cách quá chi tiết.

Trong ví dụ 5.8, do tuổi là đại lượng rời rạc, nên cận trên của nhóm trước phải nhỏ hơn cận dưới của nhóm sau là 1 đơn vị, chẳng hạn cận trên nhóm trước là 29 thì cận dưới nhóm sau phải là 30. Tuy nhiên có những đại lượng là liên tục, cách thể hiện sẽ phải thay đổi.

Ví dụ 5.11. Kết quả khảo sát về thời gian chờ đợi ở một quầy dịch vụ (đơn vị: phút) như sau:

Thời gian chờ (phút)	0 đến dưới 5	5 đến dưới 10	10 đến dưới 15	15 đến dưới 20	20 đến dưới 25
Số người	6	12	18	15	9

Vì thời gian là liên tục, theo phần lý thuyết xác suất, không cần quan tâm đến cận nên ta có thể viết lại số liệu trên dưới dạng:

Thời gian chờ (phút)	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25
Số người	6	12	18	15	9

Khi cần tính toán, có thể quy về giá trị ở giữa làm đại diện, có bảng tần số:

x_i	2,5	7,5	12,5	17,5	22,5
n_i	6	12	18	15	9

Bên cạnh việc dùng bảng tần số, có thể mô tả số liệu bằng đồ thị. Có nhiều dạng đồ thị tương ứng với các số liệu khác nhau, sinh viên có thể đọc chi tiết trong giáo trình.

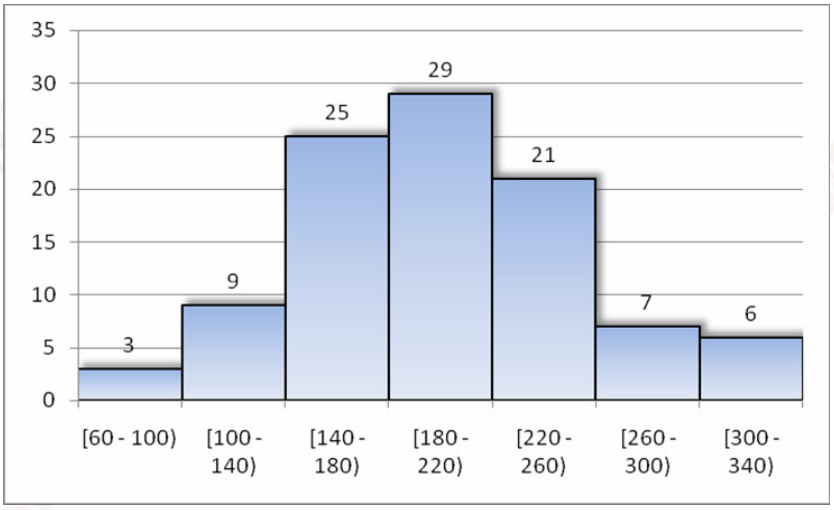
Ví dụ 5.12. Xét tình huống dẫn nhập đầu tiên, với số liệu gồm 100 quan sát của 100 khách hàng về chi tiêu tại cửa hàng. Cách liệt kê tất cả các con số có thể khó cho việc theo dõi phân phối giá trị, do đó có thể nhóm lại như sau, với X là chi tiêu của khách (nghìn đồng).

x_i	60 đến dưới 100	100 đến dưới 140	140 đến dưới 180	180 đến dưới 220	220 đến dưới 260	260 đến dưới 300	300 đến dưới 340
n_i	3	9	25	29	21	7	6

Vì coi thu nhập là biến ngẫu nhiên liên tục, nên ta có thể viết dưới dạng:

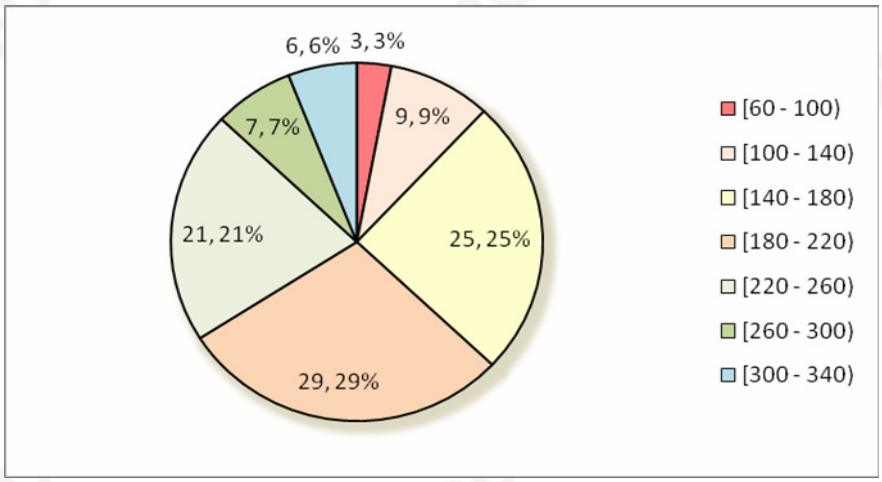
x_i	60–100	100–140	140–180	180–220	220–260	260–300	300–340
n_i	3	9	25	29	21	7	6

Mô tả bằng biểu đồ cột:



Hình 5.2. ***Biểu đồ cột mô tả mẫu***

Mô tả bằng biểu đồ tròn:



Hình 5.3. ***Biểu đồ tròn mô tả mẫu***

Ghi nhớ:

- Số liệu mẫu thường được cho dưới dạng bảng tần số hoặc tần suất.
- Nếu giá trị của mẫu là các khoảng thì lấy giá trị ở giữa làm đại diện.

5.4. Thống kê

Mẫu là bộ phận thông tin lấy được từ tổng thể để phản ánh về tổng thể. Trong tổng thể có các tham số để đặc trưng cho tổng thể, trong mẫu cũng có các đại lượng để đặc trưng cho mẫu, gọi là các *thống kê*. Các thống kê thông thường đều tương ứng với các tham số của tổng thể.

5.4.1. Định nghĩa về thống kê

Cho tổng thể với biến ngẫu nhiên gốc X, mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, \ldots, X_n).$$

Định nghĩa 5.10 – Thống kê: Thống kê là một hàm của thành phần của mẫu ngẫu nhiên.

Nếu hàm là f, ký hiệu là: $G = f(X_1, X_2, \ldots, X_n)$.

Do các đối số của hàm là các biến ngẫu nhiên, kết quả của G cũng là một biến ngẫu nhiên, có phân phối xác suất phụ thuộc vào phân phối xác suất của X, chứa đựng thông tin về X.

Với mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$, thống kê G là một giá trị cụ thể:

$$g = f(x_1, x_2, \ldots, x_n)$$

Ví dụ 5.13. Với mẫu ngẫu nhiên kích thước là 3: $W = (X_1, X_2, X_3)$, thống kê G là hàm "giá trị lớn nhất" của các đối số.

$G = \max\{X_1, X_2, X_3\}$ sẽ là một kết quả ngẫu nhiên tùy thuộc mẫu, do các giá trị X_1, X_2, X_3 là ngẫu nhiên.

Với mẫu cụ thể, chẳng hạn $w_1 = (3; 5; 2)$, thì giá trị cụ thể của thống kê là $g_1 = \max\{3; 5; 2\} = 5$; với mẫu $w_2 = (2; 4; 4)$ thì $g_2 = \max\{2; 4; 4\} = 4$.

Do mẫu là bộ phận của tổng thể, được rút ra nhằm tìm hiểu thông tin về tổng thể, nên các thống kê của mẫu thường có liên quan chặt chẽ với tham số của tổng thể. Các thống kê được sử dụng cần có các đặc điểm sau:

- Dùng để phản ánh về một tham số nào đó của tổng thể;
- Có kỳ vọng bằng chính tham số đó;
- Có phương sai nhỏ dần khi kích thước mẫu tăng lên;
- Có quy luật phân phối xác suất xác định.

Trong nội dung bài học, ta có ba tham số cơ bản của tổng thể: trung bình, phương sai (và độ lệch chuẩn), tỷ lệ. Trong mẫu cũng sẽ có ba thống kê tương ứng với ba tham số đó.

5.4.2. Trung bình mẫu

Định nghĩa 5.11 – Trung bình mẫu: Trung bình mẫu là trung bình cộng các giá trị của các thành phần mẫu.

Với mẫu ngẫu nhiên $W = (X_1, X_2, \ldots, X_n)$, trung bình mẫu ký hiệu là $\overline{X}$ được tính theo công thức:

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

Trung bình mẫu có đơn vị là đơn vị của biến ngẫu nhiên X. Trung bình mẫu ngẫu nhiên $\overline{X}$ là một biến ngẫu nhiên tùy thuộc vào các phần tử trong mẫu.

Chứng minh được các tính chất của $\overline{X}$ như sau:

- Kỳ vọng: $E(\overline{X}) = E(X) = m$
- Phương sai: $V(\overline{X}) = \frac{V(X)}{n} = \frac{\sigma^2}{n}$
- Độ lệch chuẩn: $\sigma(\overline{X}) = \sqrt{V(\overline{X})} = \frac{\sigma}{\sqrt{n}}$

Kỳ vọng của trung bình mẫu bằng trung bình tổng thể, có thể nói xét về mặt số lớn và về mặt xác suất thì trung bình mẫu phản ánh được giá trị của trung bình tổng thể, về cá biệt thì có sự sai lệch, sự sai lệch nhiều hay ít được đánh giá bằng phương sai tổng thể chia cho kích thước mẫu. Khi kích thước mẫu càng lớn thì sự sai lệch càng

giảm đi, hay nói khác đi là kích thước mẫu càng lớn, thì việc dùng trung bình mẫu để phản ánh về trung bình tổng thể là càng chính xác.

Với mẫu cụ thể $w=(x_1,x_2,...,x_n)$, mẫu có k giá trị có thể có $x_1,x_2,...,x_k$ với tần số tương ứng $n_1,n_2,...,n_k$, thì trung bình mẫu cụ thể được tính theo công thức:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{n}\sum_{i=1}^{k}n_ix_i$$

Trung bình mẫu cụ thể $\bar{x}$ là giá trị cụ thể tùy thuộc mẫu điều tra. Nếu có hai mẫu điều tra khác nhau trên cùng một tổng thể thì sẽ có hai trung bình mẫu cụ thể, và nếu ghép hai mẫu đó lại thành một mẫu mới, thì sẽ có trung bình mẫu cụ thể mới.

5.4.3. Phương sai mẫu

Nhắc lại phương sai tổng thể được tính theo công thức: $\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-m)^2$, do đó trong mẫu, ta cũng tính một đại lượng tương tự, gọi là trung bình của bình phương sai lệch.

Định nghĩa 5.12 – MS: *Trung bình của bình phương sai lệch, ký hiệu là MS, được tính theo công thức sau*:

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

Sau biến đổi, có thể chứng minh được:

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

Nếu xét trung bình mẫu $\overline{X}$ là thống kê phản ánh, thay thế cho tham số trung bình m của tổng thể, thì *MS* có phương thức tính khá giống với phương sai tổng thể. Tuy nhiên, cần xét kỳ vọng của *MS*. Chứng minh được:

$$E(MS)=\frac{n-1}{n}\sigma^2$$

Kỳ vọng của MS không bằng phương sai tổng thể, xét về mặt số lớn thì *MS* không phản ánh đúng giá trị phương sai tổng thể. Do $n-1<n$ nên xét về trung bình, *MS* sẽ luôn thấp hơn σ^2 một chút. Sự chênh lệch này không đáng kể nếu n rất lớn, nhưng với n nhỏ thì sự sai lệch có thể sẽ ảnh hưởng đến kết quả tính toán. Để điều chỉnh sự chênh lệch đó, cần triệu tiêu giá trị $\frac{n-1}{n}$ bằng cách nhân với $\frac{n}{n-1}$ được đại lượng *phương sai mẫu*.

Định nghĩa 5.13 – Phương sai mẫu: *Phương sai mẫu, ký hiệu là S^2, được tính bằng công thức*

$$S^2=\frac{n}{n-1}MS$$

Dễ thấy phương sai mẫu cũng có thể được tính bằng các công thức:

$$S^2 = \frac{n}{n-1}\left[\overline{X^2} - (\overline{X})^2\right]$$

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

Qua cách tính, ta thấy $E(S^2) = \sigma^2$: xét về mặt số lớn, kỳ vọng, phương sai mẫu S^2 phản ánh đúng giá trị của phương sai tổng thể.

Phương sai mẫu có đơn vị là bình phương đơn vị của biến ngẫu nhiên gốc. Do đó cần tính *độ lệch chuẩn*.

Định nghĩa 5.14 – Độ lệch chuẩn mẫu: *Độ lệch chuẩn mẫu, ký hiệu là S, là căn bậc hai của phương sai mẫu.*

$$S = \sqrt{S^2}$$

Phương sai và độ lệch chuẩn mẫu với các công thức trên đều là ngẫu nhiên. Trong các bài tập ta sẽ tính với mẫu cụ thể.

Với mẫu cụ thể

$$ms = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \overline{x^2} - (\overline{x})^2$$

Nếu trong mẫu được chia làm *k* nhóm giá trị có thể có thì:

$$ms = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2\right) - (\overline{x})^2 = \overline{x^2} - (\overline{x})^2$$

Phương sai mẫu cụ thể:

$$s^2 = \frac{n}{n-1}ms \text{ hoặc } s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2$$

Độ lệch chuẩn mẫu cụ thể $s = \sqrt{s^2}$

Như vậy ta nhận thấy cách ký hiệu:

- Thống kê với mẫu ngẫu nhiên là chữ cái in hoa, không có giá trị bằng số.
- Thống kê với mẫu cụ thể là chữ viết thường, có giá trị bằng số.

5.4.4. Tỷ lệ mẫu

Định nghĩa 5.15 – Tỷ lệ mẫu: *Tỷ lệ mẫu, ký hiệu là f, là tỉ số giữa số lần xuất hiện biến cố A trong mẫu và kích thước mẫu.*

Nếu trong mẫu ngẫu nhiên kích thước n, biến cố A xuất hiện X_A lần, X_A là biến ngẫu nhiên, $X_A = \{0, 1, \ldots, n\}$, thì tần suất mẫu của biến cố A:

$$f = \frac{X_A}{n}$$

Tần suất mẫu f là biến ngẫu nhiên $f = \left\{\frac{0}{n}, \frac{1}{n}, \ldots, \frac{n}{n}\right\}$

Đặt xác suất của biến cố A là p: $P(A) = p$, khi đó các tham số đặc trưng của tần suất mẫu:

$$E(f) = p \qquad V(f) = \frac{p(1-p)}{n} \qquad \sigma(f) = \sqrt{\frac{p(1-p)}{n}}$$

Với mẫu cụ thể, số lần biến cố A xuất hiện là k_A, tần suất mẫu cụ thể:

$$f = \frac{k_A}{n}$$

Ví dụ 5.14. Điều tra về thu nhập (đơn vị: triệu đồng) của một số hộ gia đình, coi thu nhập là biến liên tục, ta được số liệu gộp theo nhóm như sau:

Thu nhập	10 – 14	14 – 18	18 – 22	22 – 26	26 – 30
Số hộ	2	5	8	7	3

(a) Hãy tính các thống kê đặc trưng mẫu gồm trung bình, phương sai, độ lệch chuẩn.

(b) Tính tỷ lệ mẫu hộ gia đình có thu nhập ít hơn 18 triệu.

Giải:

(a) Số liệu đã cho là một mẫu cụ thể, đại lượng nghiên cứu là thu nhập. Đặt X là thu nhập, đơn vị triệu đồng. Giá trị của X dưới dạng khoảng, có tổng cộng 5 khoảng. Ta lấy giá trị ở giữa các khoảng làm đại diện, số hộ chính là tần số trong mẫu, được bảng tần số như sau:

x_i	12	16	20	24	28
n_i	2	5	8	7	3

Để tính các thống kê, ta viết lại các công thức dưới dạng mẫu đã phân nhóm. Vì đây là mẫu cụ thể nên các công thức dùng dưới dạng chữ viết thường.

Kích thước mẫu: $n = \sum_{i=1}^{k} n_i$

Trung bình mẫu: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$

Phương sai mẫu có hai công thức:

Kích thước mẫu: $s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2$

$$s^2 = \frac{n}{n-1}\left[\overline{x^2} - (\bar{x})^2\right] \text{ với } \overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2$$

Để đơn giản, trong phần sau ta sẽ bỏ bớt chỉ số i chạy từ 1 đến k trong các dấu tổng Σ và chỉ giữ chỉ số i trong n_i, x_i.

Có nhiều cách tính phương sai mẫu, ở đây ta sẽ theo công thức thứ hai. Sinh viên có thể sử dụng máy tính tính trực tiếp các đại lượng. Dưới đây trình bày cách tính bằng lập bảng. Cách này có thể mất thời gian hơn nhưng đảm bảo tính chính xác hơn, cũng như dễ kiểm tra khi có sai sót.

Lập một bảng gồm bốn cột, hai cột [1] và [2] chính là xoay dọc bảng số đã cho, cột [1] là x_i và cột [2] là n_i. Thêm dòng Tổng ở dưới.

Cột [3] bằng cột [1] nhân với cột [2], tức là $n_i x_i$

Cột [4] bằng cột [1] bình phương nhân với cột [2], tức là $n_i x_i^2$

Cột [1]	Cột [2]	Cột [3] = [1]×[2]	Cột [4] = [1]^2×[2]
x_i	n_i	$n_i x_i$	$n_i x_i^2$
12	**2**	24	288
16	**5**	80	1280
20	**8**	160	3200
24	**7**	168	4032
28	**3**	84	2352
Tổng (Σ)	25	516	11152

Theo bảng tính, dòng cuối cho các giá trị là:

$$\sum n_i = 25\,,\ \sum n_i x_i = 516\,,\ \sum n_i x_i^2 = 11152$$

Thay vào công thức, ta được:

Kích thước mẫu: $n = 25$

Trung bình mẫu: $\bar{x} = \dfrac{\sum n_i x_i}{n} = \dfrac{516}{25} = 20,64$

Phương sai mẫu: $s^2 = \dfrac{n}{n-1}\left[\overline{x^2} - (\bar{x})^2\right] = \dfrac{25}{25-1}\left[\dfrac{11152}{25} - (20,64)^2\right] \approx 20,907$

Độ lệch chuẩn mẫu: $s = \sqrt{s^2} = \sqrt{20,907} \approx 4,572$

Vậy trung bình mẫu là 20,64 (triệu đồng), phương sai mẫu là 20,907 (triệu đồng)^2, độ lệch chuẩn mẫu là 4,572 (triệu đồng).

(b) Với câu hỏi tính tỷ lệ hộ có thu nhập dưới 18 triệu, vì coi thu nhập là biến ngẫu nhiên liên tục, việc nhỏ hơn hay nhỏ hơn hoặc bằng 18 triệu được coi là không khác nhau. Khi đó tần số hay số hộ thu nhập nhỏ hơn 18 triệu chính là tổng tần số của hai nhóm đầu tiên (nhóm 10 – 14 và 14 – 18), và bằng 2 + 5 = 7.

Vậy tỷ lệ mẫu là:

$$f = \frac{k_{(X<18)}}{n} = \frac{2+5}{25} = 0,28$$

Vậy tỷ lệ mẫu của hộ thu nhập dưới 18 triệu là 0,28 hay 28%.

Ví dụ 5.15 (tình huống dẫn nhập). Tính các thống kê đặc trưng mẫu với ví dụ tình huống trong trường hợp số liệu được liệt kê chi tiết và trường hợp đã gộp thành nhóm.

Với số liệu gốc:

169	210	160	196	203	221	208	174	260	164
119	177	248	208	321	214	283	234	197	221
234	118	191	141	60	197	182	195	287	311
299	219	174	179	165	237	156	225	68	138
181	294	116	173	234	211	223	256	338	175
204	220	159	171	258	174	184	189	182	301
130	152	210	157	297	195	65	101	216	227
221	304	174	106	198	160	252	198	153	139
176	234	281	232	196	165	301	211	222	170
175	184	127	215	227	258	195	160	219	231

Việc tính trên số liệu gốc khá dài, ở đây chỉ viết kết quả:

Kích thước mẫu: $n = 100$

Trung bình mẫu: $\bar{x} = \frac{\sum x_i}{n} = \frac{20040}{100} = 200,4$

Phương sai mẫu: $s^2 = \frac{n}{n-1}\left[\overline{x^2} - (\bar{x})^2\right] = \frac{100}{100-1}\left[\frac{4320366}{100} - (200,4)^2\right] \approx 3074,24$

Độ lệch chuẩn mẫu: $s = \sqrt{s^2} = \sqrt{3074,24} \approx 55,446$

Khi gộp thành các nhóm, thông tin đã bị mất đi do quá trình gộp, do đó kết quả tính toán có thể không còn chính xác như với số liệu chi tiết.

Gộp thành nhóm:

Giá trị	60–100	100–140	140–180	180–220	220–260	260–300	300–340
x_i	80	120	160	200	240	280	320
n_i	3	9	25	29	21	7	6

Lập bảng tính:

x_i	n_i	$n_i x_i$	$n_i x_i^2$
80	3	240	19200
120	9	1080	129600
160	25	4000	640000
200	29	5800	1160000
240	21	5040	1209600
280	7	1960	548800
320	6	1920	614400
Tổng	**100**	**20040**	**4321600**

Kích thước mẫu: n = 100

Trung bình mẫu: $\bar{x} = \frac{\sum x_i}{n} = \frac{20040}{100} = 200,4$

Phương sai mẫu: $s^2 = \frac{n}{n-1}\left[\overline{x^2} - (\bar{x})^2\right] = \frac{100}{100-1}\left[\frac{4321600}{100} - (200,4)^2\right] \approx 3086,71$

Độ lệch chuẩn mẫu: $s = \sqrt{s^2} = \sqrt{3086,71} \approx 55,558$

Vậy so sánh với kết quả khi tính chi tiết thì khi gộp thành nhóm, phương sai và độ lệch chuẩn có sai lệch một chút. Trong tính toán, nếu có số liệu chi tiết thì kết quả chính xác hơn. Số liệu gộp chỉ nên dùng khi không thể có số chi tiết.

5.5. Quy luật phân phối xác suất liên hệ

Thống kê ngẫu nhiên trong mẫu là biến ngẫu nhiên, và có quy luật phân phối xác suất. Phân phối mẫu của một thống kê phụ thuộc vào phân phối của biến ngẫu nhiên gốc, kích thước của mẫu và phương pháp lựa chọn mẫu. Phần này giới thiệu phân phối mẫu của một số thống kê quan trong có nhiều ứng dụng trong các bài tập tiếp theo.

Với dấu hiệu nghiên cứu định lượng

Để xác định quy luật phân phối xác suất, cần biết phân phối của tổng thể. Với các biến định lượng, thường giả định phân phối Chuẩn được chấp nhận.

Trường hợp biến ngẫu nhiên gốc phân phối chuẩn, $X \sim N(\mu, \sigma^2)$, trung bình tổng thể m chính là μ, với mẫu ngẫu nhiên kích thước n: $W = (X_1, X_2, \ldots, X_n)$, các biến ngẫu nhiên thành phần cũng có cùng quy luật phân phối chuẩn, cùng kỳ vọng và phương sai, $i = 1 \div n$. Do tính chất tổ hợp bậc nhất của các biến ngẫu nhiên phân phối chuẩn cũng sẽ phân phối chuẩn với kỳ vọng và phương sai tương ứng, có thể xét quy luật phân phối xác suất của một số thống kê.

Xét trung bình mẫu $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ là tổ hợp tuyến tính của các X_i, và, $E(\overline{X}) = m = \mu$, $V(\overline{X}) = \frac{\sigma^2}{n}$ nên $\overline{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$

Thống kê liên quan với trung bình mẫu và trung bình tổng thể

Chứng minh được khi tổng thể phân phối Chuẩn thì:

Đại lượng $U = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma}$ sẽ phân phối Chuẩn hóa: $U \sim N(0\ ;\ 1)$.

Điều này có nghĩa là khi muốn phản ánh về trung bình tổng thể μ cần phải có các giá trị là $\overline{X}$ (lấy từ mẫu), độ lệch chuẩn tổng thể σ, kích thước mẫu, và các giá trị tới hạn Chuẩn u_α. Tuy nhiên thông thường độ lệch chuẩn tổng thể là không biết, do đó chứng minh được:

Thống kê $T = \frac{(\overline{X} - \mu)\sqrt{n}}{S}$ sẽ phân phối Student với $(n - 1)$ bậc tự do: $T \sim T(n - 1)$.

Điều này có nghĩa là khi muốn phản ánh về trung bình tổng thể μ cần phải có các giá trị là trung bình mẫu $\overline{X}$, độ lệch chuẩn mẫu S (lấy từ mẫu) kích thước mẫu n, và các giá trị tới hạn $t_\alpha^{(n-1)}$. Đây là cơ sở cho các bài toán thống kê về trung bình tổng thể.

Thống kê liên quan với phương sai mẫu và phương sai tổng thể

Chứng minh được khi tổng thể phân phối Chuẩn thì:

Thống kê $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ sẽ phân phối Khi – bình phương $(n - 1)$ bậc tự do: $\chi^2 \sim \chi^2(n-1)$

Điều này có nghĩa là khi muốn phản ánh về phương sai tổng thể σ^2 thì cần có các giá trị là phương sai mẫu S^2, kích thước mẫu n, và các giá trị tới hạn $\chi^2_\alpha(n-1)$.

Với dấu hiệu định tính

Dấu hiệu định tính chỉ có hai trạng thái Không và Có, biến ngẫu nhiên gốc có dạng Không – một, tỷ lệ tổng thể hay xác suất bằng p.

Tỷ lệ mẫu hay tần suất mẫu của mẫu ngẫu nhiên kích thước n là f, thì với $n \geq 100$, chứng minh được:

$U = \frac{(f - p)\sqrt{n}}{\sqrt{p(1-p)}}$ sẽ phân phối xấp xỉ chuẩn hóa: $U \sim N(0\ ;\ 1)$

Có nghĩa là muốn phản ánh về tỷ lệ tổng thể p cần có tỷ lệ mẫu f, kích thước mẫu n, và các giá trị tới hạn Chuẩn u_α.

Tổng kết:

	Tổng thể	Mẫu ngẫu nhiên: *W*	Mẫu cụ thể: *w*	Phân phối xác suất
Kích thước	N	n	n	
Giá trị biến	x_i	X_i	x_i	$X \sim N(\mu;\ \sigma^2)$
Trung bình	$m = \mu$	$\overline{X}$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$	N(0;1) T(n – 1)
Phương sai	σ^2	S^2	$s^2 = \frac{n}{n-1}\left[\overline{x^2} - (\bar{x})^2\right]$	$\chi^2(n-1)$
Độ lệch chuẩn	σ	S	s	
Tần số	M	X_A	k_A	
Tần suất	$p = \frac{M}{N}$	$f = \frac{X_A}{n}$	$f = \frac{k_A}{n}$	$N(0;1)$
Tính chất	Là giá trị xác định, chưa biết	Là biến ngẫu nhiên	Là những con số, tính toán được	Tra bảng giá trị tới hạn

Tóm lược cuối bài

- Để nghiên cứu về một hiện tượng, cần xét qua đối tượng nghiên cứu, dấu hiệu nghiên cứu và đại lượng nghiên cứu. Đại lượng nghiên cứu còn gọi là Biến ngẫu nhiên gốc.
- Tổng thể là tập hợp tất cả các phần tử chứa dấu hiệu nghiên cứu, kích thước tổng thể có thể rất lớn, thậm chí là vô hạn. Trên tổng thể có quy luật phân phối xác suất của biến ngẫu nhiên gốc. Với tổng thể có các tham số: trung bình tổng thể, phương sai tổng thể, độ lệch chuẩn tổng thể đều là tham số của biến ngẫu nhiên gốc. Tỷ lệ tổng thể cũng là trung bình tổng thể khi biến ngẫu nhiên gốc dạng Không – một.
- Mẫu là bộ phận thông tin rút ra từ tổng thể, gồm mẫu ngẫu nhiên và mẫu cụ thể, mẫu cụ thể là các bộ số liệu. Các thống kê trên mẫu ngẫu nhiên là biến ngẫu nhiên, trên mẫu cụ thể là các con số. Các thống kê phải tương ứng với tham số trong tổng thể, do đó có trung bình mẫu, phương sai mẫu, độ lệch chuẩn mẫu, tỷ lệ mẫu.
- Mối quan hệ giữa các thống kê trong mẫu và tham số trong tổng thể được thể hiện qua các quy luật phân phối Chuẩn hóa, Khi – bình phương, Student.

Câu hỏi ôn tập

1. Dấu hiệu nghiên cứu và đại lượng nghiên cứu là gì, có những loại nào?

2. Thế nào là nghiên cứu tổng thể? Ưu nhược điểm của nghiên cứu tổng thể là gì?

3. Các tham số đặc trưng của tổng thể gồm những gì? Ký hiệu là gì?

4. Mẫu là gì? Mẫu ngẫu nhiên khác mẫu cụ thể như thế nào?

5. Những thống kê như thế nào mới được xét đến trong mẫu?

6. Có những thống kê cơ bản nào? Những thống kê đó tương ứng với tham số nào của tổng thể?

7. Trong cách tính phương sai tổng thể và phương sai mẫu có gì khác nhau?

8. Mối quan hệ giữa trung bình tổng thể và trung bình mẫu thể hiện qua quy luật gì?

9. Mối quan hệ giữa phương sai tổng thể và phương sai mẫu thể hiện qua quy luật gì?

10. Mối quan hệ giữa tỷ lệ tổng thể và tỷ lệ mẫu thể hiện qua quy luật gì?