tiên phong trong công nghệ nhận biết danh tính bằng màng mắt. Thuật toán của ông ta giúp tạo ra từ ảnh màng mắt một mã với 266 đơn vị thông tin có thể coi là ngẫu nhiên và độc lập với nhau (mỗi đơn vị là một BNN nhận giá trị 0 và 1, với xác suất 50%-50%). Để tìm ra 266 đơn vị thông tin độc lập đó (xuất phát từ $2^{11} = 2048$ đơn vị thông tin không độc lập với nhau) và kiểm định sự độc lập của chúng, Daugman đã làm thống kê so sánh hơn 222 000 lần cặp ảnh màng mắt khác chủ (2 mắt trong 1 cặp là của hai người khác nhau), và hơn 500 cặp ảnh màng mắt cùng chủ. Một trong các kết quả là, tỷ lệ đơn vị thông tin chệch nhau giữa mã của 2 mắt khác chủ tuân theo phân phối chuẩn với kì vọng là 45.6% (tức là trung bình hai mắt khác chủ thì có 45.6% đơn vị thông tin chệch nhau) với độ lệch chuẩn là 0.18%, và không có cặp mắt khác chủ nào có dưới 37% đơn vị thông tin lệch nhau. Mặt khác, hai ảnh màng mắt khác nhau của cùng một chủ thì trung bình chỉ có 9% các đơn vị thông tin bị lệch nhau trong số 266 đơn vị, và không có cặp ảnh mắt cùng chủ nào bị lệch nhau quá 31% đơn vị thông tin. Từ đó dẫn đến thuật toán phân biệt: coi rằng nếu hai mã bị lệch nhau không quá 34% số đơn vị thông tin, thì vẫn là của cùng một người, còn nếu trên 34% thì coi là của hai người khác nhau.

Một điều cần chú ý là, thống kê thường bị các tổ chức hay cá nhân lạm dụng để bóp méo sự thật theo hướng có lợi cho mình, hoặc có khi tự dối mình, nếu như làm không đúng cách. Có rất nhiều cách nói dối khác nhau bằng thống kê, chẳng hạn như: bịa đặt các con số không có thật, lựa chọn các con số có lợi, giấu đi các con số bất lợi, thiên vị (bias) trong việc chọn mẫu thí nghiệm... Chẳng hạn như: Bộ quốc phòng Mỹ đã tuyên bố rằng, trong cuộc chiến với Irac năm 1991, các tên lửa Patriot của Mỹ đã bắn rơi 41 tên lửa Scud của Irac, nhưng khi Quốc hội Mỹ điều tra lại thấy chỉ có 4 tên lửa Scud bị bắn rơi; hay ví dụ về *bias* làm hỏng kết quả thống kê: Báo Literacy Digest thăm dò ý kiến cử tri về bầu cử tổng thống ở Mỹ năm 1936, qua điện thoại và qua các độc giả đặt báo. Kết quả thăm dò trên phạm vi rất rộng cho dự đoán là Landon sẽ được 370 phiếu (đại cử tri) còn Roosevelt sẽ chỉ được 161 phiếu. Thế nhưng lúc bầu thật thì Roosevelt thắng. Hoá ra, đối tượng mà Literacy Digest thăm dò năm đó, những người có tiền đặt điện thoại hay đặt báo, là những người thuộc tầng lớp khá giả, có bias theo phía Landon (Đảng Cộng hòa), không đặc trưng cho toàn dân chúng Mỹ.

Nói chung, để thống kê toán học cho ra được các kết quả đáng tin cậy, ngoài các công thức toán học đúng đắn, còn cần đảm bảo sự trung thực của các số liệu, có mẫu thực nghiệm (lượng số liệu) đủ lớn, và loại đi được ảnh hưởng của các bias để đảm bảo tính ngẫu nhiên của số liệu. Nhiều khi việc loại đi các kết quả có bias cao từ mẫu thực nghiệm là công việc hiệu quả, cho ra kết luận thống kê chính xác và đỡ tốn kém hơn là tăng cỡ của mẫu thực nghiệm lên thêm nhiều.

3.2. CƠ SỞ LÝ THUYẾT MẪU

3.2.1. Mẫu ngẫu nhiên

a) Mẫu ngẫu nhiên và tập tổng thể

Bài toán: Một nhà sản xuất dưa chuột muối đóng hộp muốn biết phân phối chiều dài các quả dưa chuột (chiều dài trung bình, độ lệch chuẩn...), để làm vỏ hộp

với kích thước thích hợp. Nhà sản xuất này không thể đo hết chiều dài của hàng triệu quả dưa chuột sẽ được đóng hộp. Họ chỉ đo chiều dài của n quả dưa chuột được chọn một cách ngẫu nhiên, rồi từ đó ước lượng ra phân phối chiều dài. Số n ở đây có thể là một số khá lớn, ví dụ 100 quả hay 1000 quả, nhưng nó là một phần rất nhỏ của tổng số các quả dưa chuột.

Để mô hình hóa bài toán ước lượng trên, ta gọi X là BNN "chiều dài của quả dưa chuột". Chúng ta muốn ước lượng phân phối xác suất của X, hoặc là ước lượng những đại lượng đặc trưng của X, ví dụ như kì vọng và phương sai. Để ước lượng, chúng ta sẽ lấy ra n giá trị của X một cách ngẫu nhiên và gọi các giá trị được lấy ra là $x_1, \cdots, x_n$. Bộ $(x_1, \cdots, x_n)$ được gọi là một mẫu ngẫu nhiên cỡ n của BNN X.

Tổng quát, một mẫu ngẫu nhiên cỡ n của BNN X là giá trị $x = (x_1, \cdots, x_n)$ của véctơ ngẫu nhiên $X = (X_1, \cdots, X_n)$, trong đó các BNN $X_1, \cdots, X_n$ độc lập và có cùng phân phối xác suất với X. Như vậy, mẫu ngẫu nhiên có nguồn gốc từ một tập lớn hơn mà ta sẽ gọi là *tập tổng thể* và mang thông tin nào đó về tập tổng thể, mặc dù các thông tin đó có thể khác nhau ở những mẫu khác nhau (Trong ví dụ, X_i là biến ngẫu nhiên "chiều dài của quả dưa chuột thứ i được chọn", còn x_i là giá trị nhận được của X_i; ta cũng có thể lấy mẫu về khối lượng, đường kính mặt cắt...).

Để ý rằng giả thiết độc lập cho phép làm đơn giản rất nhiều các tính toán sau này. Chẳng hạn nếu biến gốc X rời rạc, có hàm xác suất $p(x)$, thì hàm xác suất đồng thời của $(X_1, \cdots, X_n)$ sẽ là

$$p_n(x_1, \cdots, x_n) = P(X_1 = x_1, \cdots, X_n = x_n) = \prod_{i=1}^{n} p(x_i).$$

Tương tự nếu BNN X liên tục có mật độ $f(x)$ thì

$$f_n(x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i).$$

Như vậy, các phân phối đồng thời đã được biểu diễn đơn giản qua các phân phối biến thành phần. Muốn có đầy đủ thông tin về đối tượng nào đó, ta phải làm việc với tập tổng thể. Tuy nhiên việc nghiên cứu tập tổng thể sẽ vô cùng khó khăn vì:

- tập tổng thể quá lớn dẫn đến đòi hỏi quá nhiều chi phí vật chất và thời gian;

- trình độ tổ chức và nghiên cứu hạn chế của đội ngũ khi làm việc với quy mô lớn, không nắm bắt và kiểm soát được quá trình nghiên cứu;

- nhiều trường hợp không khả thi khi tập tổng thể biến động nhanh, các phần tử thay đổi thường xuyên. Chẳng hạn việc xét tuổi thọ của tất cả công dân Việt Nam là một việc làm không khả thi;

- có thể trong quá trình điều tra sẽ phá hủy đối tượng nghiên cứu. Chẳng hạn, để đánh giá chất lượng bia của nhà máy bia Hà Nội sản xuất trong một tháng mà đem mở tất cả các chai bia này để kiểm tra thì sau khi kiểm tra sẽ không còn bia để bán.

Như vậy việc nghiên cứu trên tập tổng thể, trừ các tập đủ bé, thường không thể thực hiện được. Từ đó đặt ra vấn đề chọn mẫu và nghiên cứu trên tập mẫu.

Nếu mẫu được chọn ngẫu nhiên và với số lượng đủ, chúng ta hy vọng rằng việc xử lí chúng sẽ cho ta kết quả vừa nhanh vừa đỡ tốn kém mà vẫn đạt được độ chính xác và tin cậy cần thiết.

b) Vấn đề chọn mẫu

Ta mong muốn mẫu có tính đại diện tốt cho tập tổng thể bởi vì việc nghiên cứu với mẫu như vậy cho ta độ tin cậy cao. Hiện nay có nhiều phương pháp khác nhau để chọn mẫu, nhưng khó có thể nói rằng phương pháp nào là tốt nhất. Việc chọn phương pháp lấy mẫu phù hợp phụ thuộc vào chính tập đối tượng cụ thể và vào sở trường của nhà nghiên cứu.

- **Chọn mẫu ngẫu nhiên**: Trong phương pháp chọn mẫu ngẫu nhiên, mỗi phần tử của tập tổng thể đã có xác suất chọn xác định từ trước cả khi chọn mẫu. Mẫu ngẫu nhiên cho phép đánh giá khách quan hơn các đặc trưng của tập tổng thể. Có 3 cách chọn như sau:

(i) Chọn mẫu ngẫu nhiên đơn giản: là phương pháp chọn mẫu có tính chất mọi mẫu có cùng kích cỡ (cùng số phần tử) có cùng xác suất được chọn và mọi phần tử của tập tổng thể có đồng khả năng lọt vào mẫu. Để việc chọn hoàn toàn ngẫu nhiên, ta có thể tiến hành theo kiểu bốc thăm hoặc dùng bảng số ngẫu nhiên, ở đây có hai phương thức chọn là không hoàn lại (mỗi phần tử chỉ được chọn một lần) và có hoàn lại (chọn xong trả lại tổng thể để chọn tiếp). Nếu số lượng phần tử của mẫu khá bé so với tập tổng thể thì kết quả lấy mẫu theo hai phương thức sai lệch không đáng kể. Do tính ngẫu nhiên nên mẫu có tính đại diện cao và tin cậy. Tuy nhiên phương pháp này đòi hỏi phải biết toàn bộ tập tổng thể, vì thế chi phí chọn mẫu khá lớn.

(ii) Chọn mẫu phân nhóm: Đầu tiên ta chia tập tổng thể thành các nhóm tương đối thuần nhất, sau đó từ mỗi nhóm trích ra một mẫu ngẫu nhiên; tập hợp tất cả các mẫu đó cho ta một mẫu (ngẫu nhiên) phân nhóm. Người ta dùng phương pháp này khi trong nội bộ tập tổng thể có những sai khác lớn. Nhà nghiên cứu phải có hiểu biết nhất định về cấu trúc tập tổng thể để phân chia nhóm hợp lý. Sau này mỗi nhóm sẽ có vai trò khác nhau phụ thuộc vào độ quan trọng của chúng trong tập tổng thể. Hạn chế của phương pháp là tính chủ quan khi phân chia nhóm. Nhưng nó vẫn hay được dùng do cách thức đơn giản khi làm việc với các nhóm khá bé và thuần nhất.

(iii) Chọn mẫu chùm: là chọn một mẫu ngẫu nhiên của các tập con của tập tổng thể, được gọi là các chùm. Ta cũng giả sử rằng các phần tử của mỗi chùm mang tính đại diện cho tập tổng thể. Ngoài ra ta cố gắng sao cho mỗi chùm vẫn có độ phân tán cao như tập tổng thể và đồng đều nhau về quy mô. Chẳng hạn ta muốn nghiên cứu nhu cầu tiêu thụ một mặt hàng nào đó bằng phương pháp chọn mẫu chùm: đầu tiên ta chia thành phố thành các khu dân cư, sau đó chọn ra một số khu làm phần tử của mẫu, cuối cùng ta nghiên cứu tất cả các gia đình sống trong các khu được chọn. Phương pháp này cho ta tiết kiệm kinh phí và thời gian (vì không phải di chuyển trên toàn thành phố), nhưng sai số có thể lớn hơn hai phương pháp trên.

- **Chọn mẫu có suy luận**: Phương pháp chọn mẫu này dựa trên ý kiến các chuyên gia về đối tượng nghiên cứu. Như vậy việc chọn mẫu ở đây dựa trên hiểu

biết và kinh nghiệm của một vài nhà chuyên môn. Do đó phương pháp này cũng có hạn chế cơ bản là: khi không có sự tham gia của các công cụ thống kê vào việc chọn mẫu nên tính khách quan rất khó được bảo đảm, từ đó kéo theo các kết luận mang nặng tính chủ quan. Tất nhiên điều đó không có nghĩa là không nên dùng các phương pháp chuyên gia. Rõ ràng là chất lượng mẫu phụ thuộc nhiều vào trình độ của nhà nghiên cứu và kinh nghiệm của họ sẽ trở thành một công cụ hữu hiệu.

- **Sai số trong lấy mẫu**: Khi lấy mẫu, do nhiều nguyên nhân khác nhau, sẽ không tránh khỏi những sai số trong các số liệu mẫu. Do đó, trước khi dùng các phương pháp thống kê để phân tích, xử lí ta cần loại bỏ các sai số không đáng có ở trong mẫu đã cho, có như vậy các thông tin thu được sau xử lí mới đảm bảo tính chính xác với độ tin cậy cao. Để thuận lợi cho việc xử lí, ta phân loại các sai số như sau:

(i) Sai số thô : sinh ra do phạm vi các điều kiện cơ bản của việc lấy mẫu hoặc do sơ suất của người thực hiện, chẳng hạn, người kiểm tra cố ý chọn ra các sản phẩm tốt để kiểm tra khi đánh giá chất lượng của lô sản phẩm, hoặc người lấy mẫu ghi nhầm kết quả thu được...

(ii) Sai số hệ thống: là sai số do không điều chỉnh chính xác dụng cụ hoặc không thống nhất giữa những người lấy mẫu về cách xác định một đại lượng nào đó... dẫn đến các kết quả quan sát được bị sai lệch.

(iii) Sai số ngẫu nhiên: sinh ra do một số lớn các nguyên nhân mà tác động của chúng nhỏ đến mức không thể tách riêng và tính riêng biệt cho từng nguyên nhân được. Chẳng hạn, trong các cuộc thi thể thao, khi từng thành viên trong Ban giám khảo đánh giá bằng cho điểm (các môn võ, thể dục dụng cụ...), sẽ có giám khảo cho hơi cao, lại có người cho thấp hơn một chút, đó chính là sai số ngẫu nhiên.

Trong ba loại sai số trên, sai số thô, sai số hệ thống cần phát hiện sớm và loại bỏ, còn sai số ngẫu nhiên không thể loại bỏ được trong mỗi lần lấy mẫu.

- **Phương pháp loại bỏ sai số thô**: Khi tiến hành loại bỏ sai số thô (số liệu lạ) ta cần chú ý:

(i) Trước tiên cần kiểm tra xem có sơ suất hoặc có vi phạm các nguyên tắc cơ bản khi thu thập số liệu không?

(ii) Thử loại bỏ x_0 là số liệu bị nghi ngờ rồi tiến hành xử lí số liệu xem kết luận có khác so với khi giữ lại x_0 hay không? Nếu không có sai khác đáng kể thì nên giữ lại số liệu x_0.

(iii) Nên tham khảo các tài liệu chuyên môn liên quan có thể giải thích cho việc xuất hiện số liệu lạ này sau đó mới quyết định nên giữ hay nên bỏ.

Giả sử ta có dãy số liệu: $x_0, x_1, \cdots, x_n$ ở đó x_0 bị nghi ngờ là số dị thường (giá trị nhỏ nhất hoặc lớn nhất) trong dãy số trên. Khi đó ta xét đại lượng: $T = \dfrac{x_0 - \overline{x}}{s}$ ($\overline{x}$: trung bình mẫu, s: độ lệch chuẩn mẫu hiệu chỉnh, xem phần **3.2.3**).

(i) Nếu $T > t_{\alpha/2}(n-1)$ thì loại bỏ giá trị x_0 ra khỏi dãy các số liệu trên.

(ii) Nếu $T \leq t_{\alpha/2}(n-1)$ ta kết luận dãy số liệu trên không có số dị thường, trong đó $t_\alpha(n)$ là giá trị tới hạn mức α của phân phối Student n bậc tự do (xem phần **3.4.2**). Trong thực tế tùy yêu cầu chính xác của việc xử lí số liệu người ta thường lấy α ở các mức từ 0,01 đến 0,05. Việc đưa ra tiêu chuẩn loại bỏ sai

số thô nói trên dựa trên giả thiết các số liệu mẫu lấy từ tổng thể có phân phối chuẩn $\mathcal{N}(\mu;\sigma^2)$.

3.2.2. Phân loại và mô tả số liệu mẫu

a) Phân loại

Giả sử từ một tập tổng thể có N phần tử, chọn ra một mẫu có kích thước n, các phần tử của mẫu gồm n giá trị $x_1, \cdots, x_n$ tạo ra một *mẫu đơn*. Nếu trong mẫu có nhiều giá trị giống nhau: chẳng hạn giá trị x_1 xuất hiện n_1 lần, x_2 xuất hiện n_2 lần,..., x_k xuất hiện n_k lần; khi đó $n_1 + n_2 + \cdots + n_k = n$.

Ví dụ 3.5. Kiểm tra ngẫu nhiên 50 học viên. Ta sắp xếp điểm số môn Giải tích theo thứ tự tăng dần và số học viên có điểm tương ứng vào bảng như sau:

Điểm số	4	5	6	7	8	9	10
Số học viên	6	20	12	7	2	2	1

Trong thực hành, nhiều khi mẫu điều tra có kích thước lớn, hoặc khi các giá trị cụ thể của BNN X lấy giá trị khác nhau song lại khá gần nhau, người ta thường xác định một số các khoảng $I_1, I_2, \cdots, I_k$ sao cho mỗi giá trị của dấu hiệu điều tra thuộc vào một khoảng nào đó. Trong trường hợp này ta có *mẫu lớp* (mẫu cho dưới dạng nhiều lớp là các khoảng không cắt nhau). Việc chọn số khoảng và độ rộng khoảng là tuỳ thuộc vào kinh nghiệm của người nghiên cứu, nhưng nói chung không nên chia quá ít khoảng. Ngoài ra độ rộng các khoảng cũng không nhất thiết phải bằng nhau.

Ví dụ 3.6. Một mẫu cụ thể về cân nặng (đơn vị kg) của 40 nam sinh viên đại học được cho trong bảng với trọng lượng lớn nhất là 73 và nhỏ nhất 47, vậy khoảng cách tối đa là $73 - 47 = 26$.

54	67	60	51	57	47	59,5	63,5
55	58	62,4	59	61	57	69	48,2
66,8	49,7	62	67,1	58	71,5	56	58
55	52,3	65	58,2	53,1	56	60	63
58	64	54,6	73	52,6	61,2	57,9	49

Nếu chia độ rộng khoảng 3 thì có xấp xỉ $26/3 \approx 9$ khoảng. Nếu chia độ rộng khoảng 8 thì có xấp xỉ $26/8 \approx 4$ khoảng. Ta có hai bảng phân phối ghép lớp với độ rộng khoảng bằng 3 có 9 khoảng và độ rộng khoảng bằng 4 có 7 khoảng (xem Bảng 3.1).

Bảng 3.1: Bảng số liệu thu gọn

(a) 9 khoảng

Trọng lượng (kg)	Số lượng
47-50	4
50-53	3
53-56	5
56-59	10
59-62	6
62-65	5
65-68	4
68-71	1
71-74	2

(b) 7 khoảng

Trọng lượng (kg)	Số lượng
47-51	4
51-55	6
55-59	12
59-63	8
63-67	5
67-71	3
71-75	2

b) Tần số, bảng phân phối tần số, tần suất và phân phối thực nghiệm

Số lần xuất hiện x_i hoặc một khoảng thứ i nào đó, ký hiệu là n_i, được gọi là tần số. Sau khi sắp xếp số liệu theo thứ tự tăng dần của giá trị mẫu, ta có thể xây dựng bảng tần số. Bảng số liệu trong Ví dụ 3.5 được gọi là bảng phân phối tần số dạng điểm, hai bảng số trong Ví dụ 3.6 chính là bảng phân phối tần số dạng khoảng.

Bảng 3.2: Bảng tần số

x_i	x_1	x_2	$\cdots$	x_i	$\cdots$	x_k	
n_i	n_1	n_2	$\cdots$	n_i	$\cdots$	n_k	$\sum_{i=1}^{k} n_i = n$

Từ Bảng 3.2, nếu ta đặt $f_i = \dfrac{n_i}{n}, i = \overline{1,k}$ là tần suất xuất hiện giá trị x_i ở trong mẫu thì ta có thể mô tả bảng tần suất tương ứng. Như vậy

$$\sum_{i=1}^{k} f_i = \frac{1}{n}\sum_{i=1}^{k} n_i = 1$$

và bảng tần suất là

x_i	x_1	x_2	$\cdots$	x_i	$\cdots$	x_k	
f_i	f_1	f_2	$\cdots$	f_i	$\cdots$	f_k	$\sum_{i=1}^{k} f_i = 1$

rất giống với bảng phân phối xác suất của một BNN rời rạc.

Nếu đặt w là tần số tích lũy của $x \in \mathbb{R}$ và $F_n(x)$ là tần suất tích lũy của x, tức là

$$w = \sum_{x_j<x} n_j \quad \text{và} \quad F_n(x) = \frac{w}{n} = \sum_{x_j<x} f_j,$$

thì $F_n(x)$ là một hàm của x và được gọi là *hàm phân phối thực nghiệm* của mẫu hay là *hàm phân phối mẫu*. Người ta chứng minh được rằng hàm phân phối thực nghiệm $F_n(x)$ xấp xỉ với phân phối lí thuyết $F(x) = P(X \leq x)$ khi n đủ lớn, trong đó X là BNN gốc của tập tổng thể.

Ví dụ 3.7. Bảng tần suất được xây dựng từ Ví dụ 3.5 là

Điểm số	4	5	6	7	8	9	10
Tần số	6	20	12	7	2	2	1
Tần suất	$\frac{3}{25}$	$\frac{2}{5}$	$\frac{6}{25}$	$\frac{7}{50}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{50}$

Chú ý: (áp dụng cho bảng tần số dạng khoảng):

(i) Quy ước đầu mút bên phải của mỗi khoảng không thuộc vào khoảng đó khi tính tần số của mỗi khoảng, trong Ví dụ 3.6 ta có các khoảng $[47; 51), [51; 55), [55; 59)$...

(ii) Một trong những gợi ý để chọn số khoảng k tối ưu là hãy chọn $k \in \mathbb{Z}_+$ nhỏ nhất sao cho $2^k \geq n$ (n: kích thước mẫu).

(iii) Độ rộng các khoảng không đòi hỏi bằng nhau.

(iv) Khi tính toán ta đưa về bảng tần số dạng điểm bằng cách lấy giá trị chính giữa của mỗi khoảng, chẳng hạn khoảng $[a_1; a_2)$ ta sẽ lấy điểm $x_1 = \dfrac{a_1 + a_2}{2}$.

3.2.3. Các đặc trưng của mẫu ngẫu nhiên

Định nghĩa 3.1. Hàm $g(X_1, \cdots, X_n)$ với $(X_1, \cdots, X_n)$ là một mẫu ngẫu nhiên được gọi là một hàm mẫu hay một thống kê.

Vì mẫu $(X_1, \cdots, X_n)$ là một véctơ ngẫu nhiên nên $g(X_1, \cdots, X_n)$ là một BNN. Với mẫu cụ thể, BNN $X_i = x_i$ $(i = \overline{1, n})$, thì $g(x_1, \cdots, x_n)$ là giá trị cụ thể mà thống kê $g(X_1, \cdots, X_n)$ nhận tương ứng với mẫu đã cho. Phân phối xác suất của thống kê $g(X_1, \cdots, X_n)$ phụ thuộc vào phân phối xác suất của BNN X ở tổng thể. Các thống kê mẫu cùng với quy luật phân phối xác suất của chúng là cơ sở để khảo sát dấu hiệu nghiên cứu của tổng thể từ các thông tin của mẫu.

Có hai nhóm thống kê mẫu quan trọng đặc trưng cho BNN của tổng thể:

(i) Các số đặc trưng cho ta hình ảnh về vị trí trung tâm của mẫu, tức là xu thế các số liệu trong mẫu tụ tập xung quanh những con số nào đó. Chẳng hạn trung bình mẫu, trung vị mẫu, Mode mẫu...

(ii) Các số đặc trưng cho sự phân tán của các số liệu: độ lệch trung bình, độ lệch tiêu chuẩn và phương sai mẫu.

Ta sẽ xem xét một số thống kê đặc trưng mẫu quan trọng sau:

a) Trung bình mẫu (kì vọng mẫu)

Xét mẫu ngẫu nhiên $(X_1, \cdots, X_n)$ của BNN X, thống kê

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

gọi là trung bình mẫu. Với mẫu cụ thể $(x_1, \cdots, x_n)$ thì $\overline{x} = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i$ là giá trị mà trung bình mẫu nhận được ứng với mẫu đã cho.

Nếu số liệu cho dưới dạng Bảng 3.2 thì $\overline{X} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$.

Do $X_1, \cdots, X_n$ là các BNN độc lập cùng phân phối như X nên $\overline{X}$ là một BNN. Theo tính chất của kì vọng và phương sai ta có:

$$\begin{aligned} E(\overline{X}) &= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n.E(X)}{n} = E(X), \\ D(\overline{X}) &= \frac{1}{n^2} D\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{n.D(X)}{n^2} = \frac{D(X)}{n}. \end{aligned} \tag{3.1}$$

Từ công thức trên, do phương sai $D(\overline{X})$ bé hơn n lần $D(X)$ nên các giá trị có thể có của $\overline{X}$ sẽ ổn định quanh kì vọng hơn các giá trị của X.

b) Phương sai mẫu, độ lệch chuẩn mẫu

Một cách tương tự trung bình mẫu, phương sai mẫu được định nghĩa là kì vọng của độ lệch bình phương các thành phần của mẫu với trung bình mẫu và kí hiệu

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

Nếu mẫu cho dưới dạng Bảng 3.2 thì

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - (\overline{X})^2.$$

Do $\hat{S}^2$ là BNN, sử dụng các tính chất kì vọng, ta có:

$$\begin{aligned} E(\hat{S}^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2\right) = E\left(\frac{n-1}{n^2}\sum_{i=1}^{n} X_i^2 - \frac{1}{n^2}\sum_{i\neq j} X_i X_j\right) \\ &= \frac{n-1}{n^2}.n.E(X^2) - \frac{n(n-1)}{n^2}[E(X)]^2 = \frac{n-1}{n}D(X), \end{aligned} \tag{3.2}$$

do $X_i, i = \overline{1,n}$ là độc lập, cùng phân phối với X nên

$$E(X_i X_j) = E(X_i).E(X_j) = [E(X)]^2.$$

Để kì vọng của phương sai mẫu trùng với phương sai $D(X)$ của BNN gốc ta cần phương sai mẫu có hiệu chỉnh là

$$S^2 = \frac{n}{n-1}\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}(\overline{X})^2. \tag{3.3}$$

Để phân biệt trong phần còn lại của giáo trình ta sử dụng các chữ viết hoa chỉ các thống kê của mẫu ngẫu nhiên, các chữ viết thường chỉ các giá trị tương ứng:

(i) Thống kê $\hat{S}$ gọi là độ lệch chuẩn mẫu chưa hiệu chỉnh và $\hat{s}$ là giá trị của $\hat{S}$ với mẫu đã cho.

(ii) Thống kê S gọi là độ lệch chuẩn mẫu đã hiệu chỉnh và s là giá trị của S với mẫu đã cho.

Ví dụ 3.8. Tuổi thọ (đơn vị: 100 giờ) một loại linh kiện do công ty A sản xuất ra được kiểm tra ngẫu nhiên, kết quả ghi thành bảng.

Tuổi	≤ 7	7-7,5	7,5-8	8-8,5	8,5-9	≥ 9
n_i	3	15	18	14	20	4

Tính $\overline{x}, \hat{s}^2, s^2$.

Lời giải. Dùng giá trị chính giữa của mỗi khoảng làm đại diện cho khoảng đó, riêng với khoảng đầu và cuối ta chọn giá trị hợp lí nào đó, ví dụ 6,5 và 9,5 tương ứng, ta có bảng số liệu

Tuổi	≤ 7	7-7,5	7,5-8	8-8,5	8,5-9	≥ 9	
x_i	6,5	7,25	7,75	8,25	8,75	9,5	
n_i	3	15	18	14	20	4	$\sum = 74$
$x_i n_i$	19,5	108,75	139,5	111,5	175	38	$\sum = 596,25$
$x_i^2 n_i$	58,5	1631,25	2511	1617	3500	152	$\sum = 4841,438$

Từ bảng trên ta tính được:

$$\overline{x} = \frac{596,25}{74} \approx 8,06; \hat{s}^2 = \frac{4841,438}{74} - (8,06)^2 \approx 0,503 \text{ và } s^2 = \frac{74}{73}.0,503 \approx 0,51.$$

c) Tần suất mẫu

Trường hợp cần nghiên cứu một dấu hiệu định tính A nào đó mà mỗi cá thể của tổng thể có thể có hoặc không, giả sử p là tần suất có dấu hiệu A của tổng thể. Nếu cá thể có dấu hiệu A ta cho nhận giá trị 1, trường hợp ngược lại ta cho nhận giá trị 0. Lúc đó dấu hiệu nghiên cứu có thể xem là BNN X có phân phối Bernoulli tham số p có kì vọng $E(X) = p$ và phương sai $D(X) = p(1-p)$ (xem lại phân phối Bernoulli).

Lấy mẫu ngẫu nhiên kích thước n, trong đó $X_1, X_2, \cdots, X_n$ là dãy các BNN độc lập có cùng phân phối $\mathcal{B}(1; p)$. Tần số xuất hiện dấu hiệu A của mẫu là

$$r = X_1 + X_2 + \cdots + X_n.$$

Tần suất mẫu kí hiệu là f và được xác định bởi

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

Như vậy tần suất mẫu là trung bình mẫu của BNN X có phân phối Bernoulli tham số p. Sử dụng các công thức của kì vọng và phương sai, ta có:

$$E(f) = E(\overline{X}) = E(X) = p \quad \text{và} \quad D(f) = D(\overline{X}) = \frac{D(X)}{n} = \frac{p(1-p)}{n}. \qquad (3.4)$$