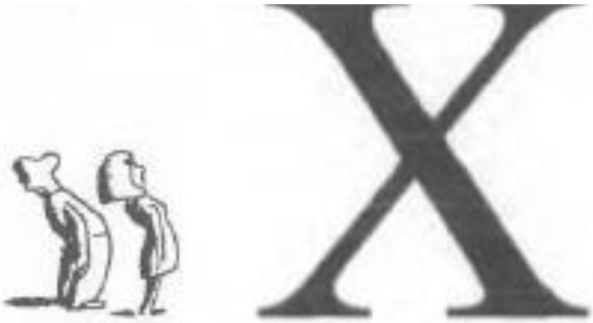
# CHƯƠNG 4:
# BIẾN NGẪU NHIÊN

Trong Chương 2, chúng ta đã tìm hiểu về các quan sát của các dữ liệu dạng số như là cân nặng của sinh viên. Dữ liệu có thể được mô tả bằng biểu đồ và phân tích thông qua điểm chính giữa, khoảng biến thiên, độ phân tán, giá trị ngoại lai của dữ liệu… Trong Chương 3, chúng ta đã biết cách gán xác suất cho các kết quả của một thử nghiệm ngẫu nhiên.


Thử hình dung 1 thử nghiệm ngẫu nhiên được lặp lại rất nhiều lần, chúng ta mong các kết quả thực tế sẽ luôn bị chi phối bởi xác suất của chúng, đó chính là mô hình xác suất của các thí nghiệm thực tế trong cuộc sống… Vậy tại sao chúng ta không thử đi tìm mô hình để mô tả các dữ liệu đó?

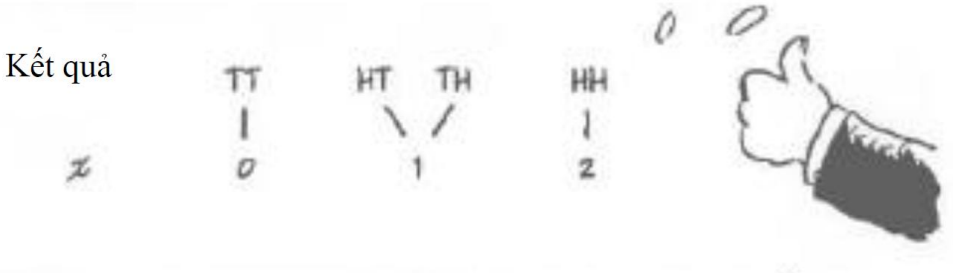
Mục đích chính là biến ngẫu nhiên, ta kí hiệu biến ngẫu nhiên bằng một kí tự lớn.



Một biến ngẫu nhiên được định nghĩa là kết quả dạng số của một hiện tượng ngẫu nhiên.

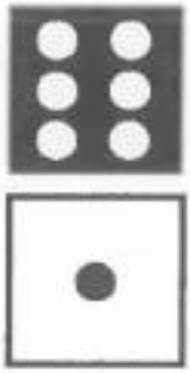
Ví dụ, hãy thử tượng tượng chúng ta đang phác họa một học sinh bất kì. Đó chính là một thử nghiệm ngẫu nhiên. Chiều cao, cân nặng, thu nhập gia đình, điểm S.A.T, điểm trung bình ở lớp của học sinh đều là các biến dạng số, mô tả các đặc điểm của học sinh được chọn lựa ngẫu nhiên đó. Chúng đều là các biến ngẫu nhiên.


Một ví dụ khác là: tung 2 đồng xu (thí nghiệm ngẫu nhiên) và ghi chép lại số lần xuất hiện của mặt ngửa H: 0, 1 hoặc 2.




Lưu ý: Biến được kí hiệu bởi chữ X in hoa, x thường biểu hiện các giá trị cá biệt của biến X. Ví dụ x = 2 nếu mặt ngửa H xuất hiện 2 lần (HH).

Một ví dụ nữa dựa trên việc tung hai con xúc xắc quen thuộc. Gọi Y là tổng số các chấm trên 2 con xúc xắc. Y chính là biến ngẫu nhiên nhận các giá trị từ 2 đến 12.



Y=7

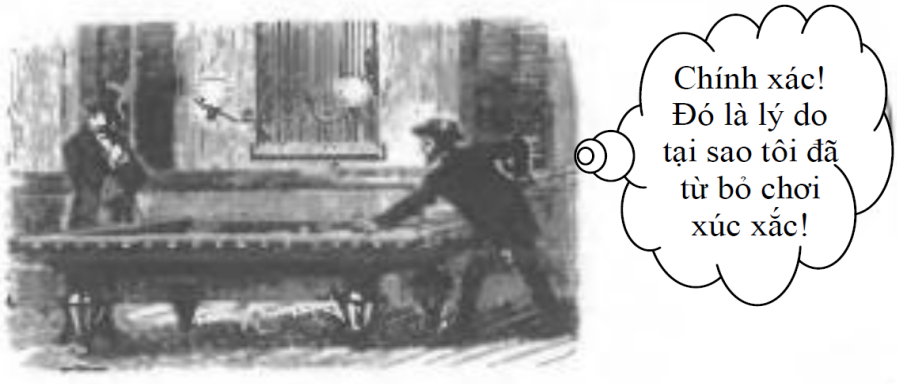
Bây giờ chúng ta cùng nhìn vào xác suất của các kết quả. Xác suất để biến ngẫu nhiên X nhận giá trị là x được viết như sau: Pr(X = x) hoặc P(x). Với biến ngẫu nhiên của việc tung đồng xu, chúng ta có thể lập được bảng như sau:

| $x$ | 0 | 1 | 2 |
|---|---|---|---|
| $Pr(X=x)$ | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |

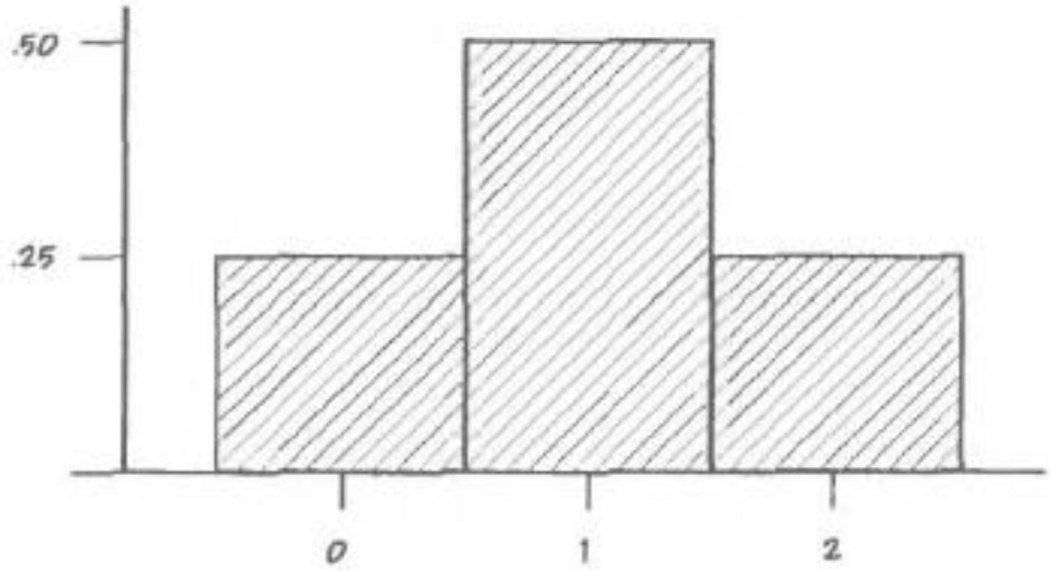
Đây được gọi là bảng phân bố xác suất của biến ngẫu nhiên X.

Với biến ngẫu nhiên Y (Y là tổng số chấm xuất hiện sau mỗi lần tung của hai con xúc xắc), sự phân bố xác suất được mô tả như sau:

| $y$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $Pr(Y=y)$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

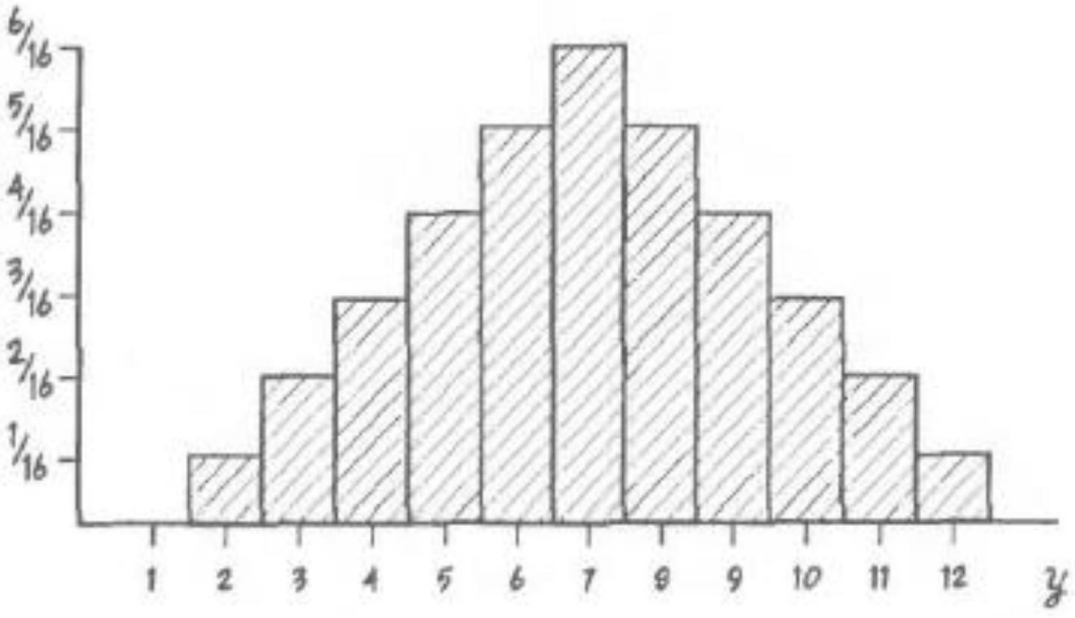


Giờ thì hãy chúng ta hãy cùng vẽ các đồ thị hoặc biểu đồ biểu diễn sự phân bố của xác suất. Với mỗi giá trị của X, chúng ta vẽ 1 cột có độ cao bằng p(x).

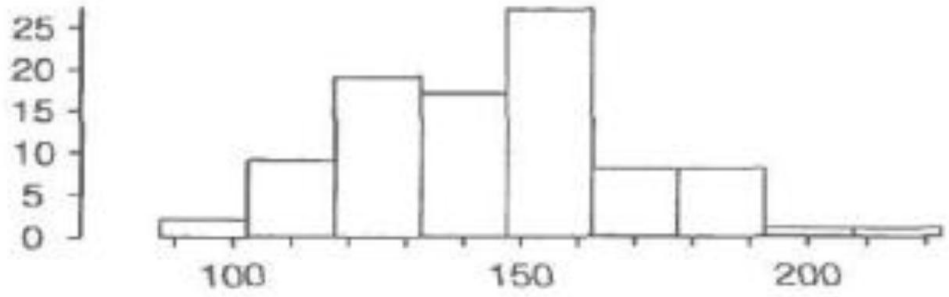



Dễ dàng nhận thấy tổng chiều cao của các cột này bằng 1: Mỗi cột có đáy rộng bằng 1 và chiều cao là p(x), bởi vậy tổng chiều cao của các cột này chính là tổng xác suất của tất cả các kết quả, ví dụ 1.

Còn đây chính là biểu đồ xác suất của biến ngẫu nhiên Y, biểu thị sự phân bố xác suất của tổng hai con xúc xắc:

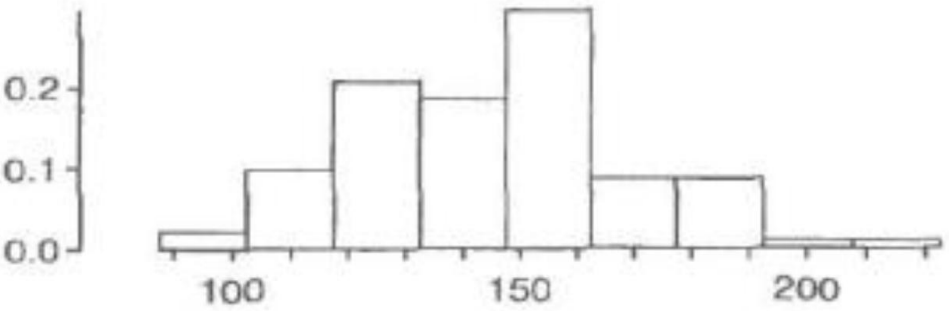


Tại sao chúng ta lại gọi những đồ thị này là biểu đồ tần suất? Nhắc lại, trong Chương 2, biểu đồ tần số chính là một biểu đồ cho biết có bao nhiêu số điểm dữ liệu nằm trong chuỗi của các khoảng dữ liệu:



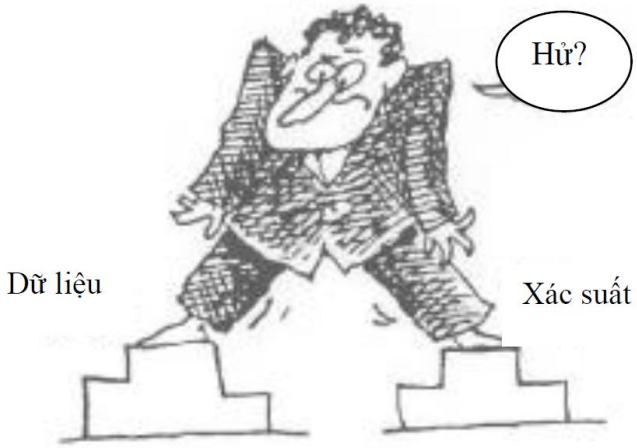

Cân nặng (pound)

Từ biểu đồ tần số, chúng ta đưa ra biểu đồ tần suất, biểu thị tỷ lệ của dữ liệu trong mỗi khoảng:




Cân nặng (pound)

Nhưng bạn cũng hãy nhớ rằng định nghĩa xác suất chính là tần suất của một sự kiện "trong dài hạn". Nếu chúng ta lặp lại thí nghiệm ngẫu nhiên này nhiều lần thì biểu đồ tần suất của các kết quả sẽ giống như biểu đồ xác suất của các biến ngẫu nhiên!



Chúng tôi minh họa việc sử dụng biến ngẫu nhiên X với việc một người ngốc nghếch đang ngồi tung đồng xu.


Cô ta bắt đầu bằng việc tung hai đồng xu, và cứ thế lặp đi lặp lại hành động đó, kết hợp với việc ghi chép kết quả.


Chúng ta đã biết tới sự phân bố xác suất của X và chúng ta cũng đã biết việc tung đồng xu thực tế có liên quan tới việc ước lượng xác suất. Sau 1000 lần tung, cô gái đếm được các dữ liệu như sau:

| Mô hình xác suất | | Dữ liệu được quan sát | |
|---|---|---|---|
| p(x) | x | $n_x$ = số khả năng xảy ra | $\frac{n_x}{n}$ = tần suất |
| .25 | 0 | 260 | .260 |
| .5 | 1 | 517 | .517 |
| .25 | 2 | 223 | .223 |

Và chúng ta thấy rằng biểu đồ xác suất của X giống như một "dạng thuần túy" hoặc giống như biểu đồ tần suất của dữ liệu.

Để mở rộng việc phân tích giữa tần suất và dữ liệu chúng ta nên nói về trung bình và phương sai (hoặc độ lệch chuẩn) của phân phối xác suất.


Và nhắc lại một chút là chúng ta hiện đang ở trong một vương quốc của sự trừu tượng, chúng ta đã giải mã được một số kí tự Hi lạp…

## TRUNG BÌNH VÀ PHƯƠNG SAI CỦA CÁC BIẾN NGẪU NHIÊN

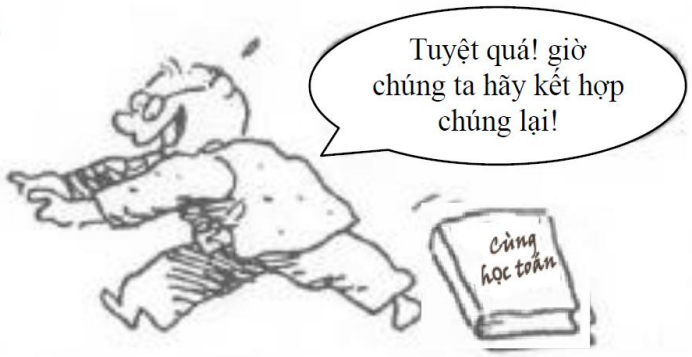
Chúng tôi sử dụng các thuật ngữ và các biểu tượng đặc biệt để phân biệt các đặc trưng của dữ liệu với sự phân phối của xác suất:

Các đặc trưng của dữ liệu được gọi là các đặc trưng mẫu, trong khi các đặc trưng sự phân phối của xác suất được gọi là mô hình hoặc các tổng thể đặc trưng.

Chúng ta sử dụng kí tự Hi lạp μ (mu) để kí hiệu cho trung bình của tổng thể, và σ (sigma thường) cho độ lệch chuẩn của tổng thể (với dữ liệu ta sử dụng kí tự của Roma là $\bar{x}$ và s).

Trung bình mẫu được xác định bởi công thức sau:

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




Giờ đây một vài điểm dữ liệu $x_i$ có thể có giá trị bằng nhau. Hãy nhớ lại người tung đồng xu: Các giá trị có khả năng xảy ra chỉ là 0, 1, 2 và cô ấy đã tung tới 1000 lần. Giá trị 0 đầu xuất hiện 260 lần, 1 đầu xuất hiện nhiều nhất là 517 lần, và 2 đầu là 223 lần.

Chúng ta cho x nhận tất cả các giá trị của X, gọi $n_x$ là số lượng điểm dữ liệu mang giá trị x, công thức được viết lại như sau:

$$\bar{x} = \frac{1}{n}\sum_{\text{mọi x}} n_x\, x$$

Hoặc

$$\bar{x} = \sum_{\text{mọi x}} x\,\frac{n_x}{n}$$


Ah! Nhưng giờ thì $\frac{n_x}{n}$ n chính là tần suất… "xác suất ước lượng…" con số đạt được p(x)…, tương tự chúng ta có công thức:

$$\sum_{\text{mọi x}} xp(x)$$


Và được định nghĩa là trung bình của sự phân bố xác suất.

**Định nghĩa: Trung bình**
của biến ngẫu nhiên X được xác định như sau:

$$\mu = \sum_{\text{mọi x}} x p(x)$$


μ cũng được gọi là giá trị kì vọng của X hoặc E[X]. Được hiểu là tổng của các giá trị tồn tại, quyền số của mỗi giá trị chính là xác suất của nó.

Thí nghiệm của người tung đồng xu cho phép so sánh trung bình mẫu $\bar{x}$ với trung bình của mô hình μ

Mẫu

| x | $\frac{n_x}{n}$ | $x\frac{n_x}{n}$ |
|---|---|---|
| 0 | 0.26 | 0 |
| 1 | 0.517 | 0.517 |
| 2 | 0.223 | 0.446 |
| | | 0.963 = $\bar{x}$ |

Mô hình

| x | p(x) | xp(x) |
|---|---|---|
| 0 | 0.25 | 0 |
| 1 | 0.5 | 0.5 |
| 2 | 0.25 | 0.5 |
| | | 1 = μ |

Bây giờ chúng ta hãy thực hiện điều tương tự với phương sai. Chắc hẳn bạn vẫn còn nhớ công thức này:

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

Nó đo lường trung bình của bình phương khoảng cách các điểm dữ liệu tới giá trị trung bình. Giống như trên $s^2$ được viết như sau:

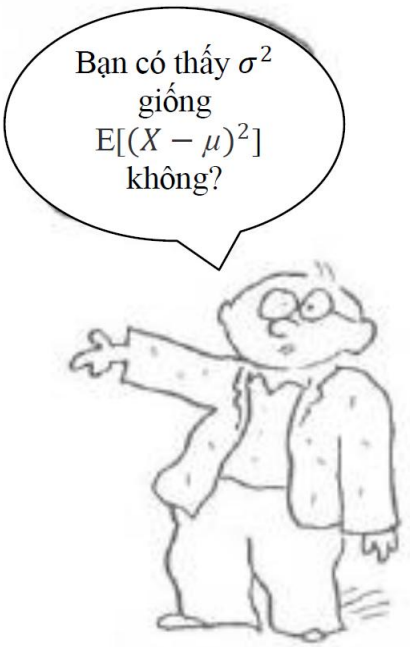
$$s^2 = \sum_{\text{mọi x}}(x-\bar{x})^2\frac{n_x}{n-1}$$

Mẫu số là (n-1) chứ không phải là n, điều này cũng giống như quyền số của tổng các bình phương các khoảng cách… Chúng ta xây dựng được một định nghĩa khác:

**Phương sai** của biến ngẫu nhiên X là trung bình của bình phương độ lệch chuẩn biến ngẫu nhiên X so với trung bình tổng thể:

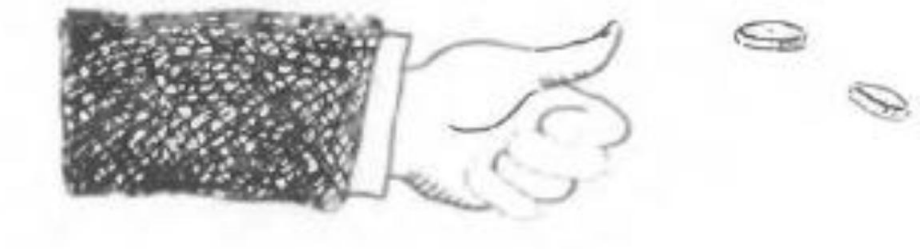
$$\sigma^2 = \sum_{\text{mọi x}} (x-\mu)^2 p(x)$$

**Độ lệch chuẩn σ** chính là căn bậc hai của phương sai.




Chúng ta sử dụng bảng sau để tìm phương sai của thí nghiệm tung hai đồng xu (cho μ = 1).

| x | p(x) | $(x-\mu)^2 p(x)$ |
|---|---|---|
| 0 | 0.25 | $(0-1)^2\ 0.25 = 0.25$ |
| 1 | 0.5 | $(1-1)^2\ 0.5 = 0$ |
| 2 | 0.25 | $(2-1)^2\ 0.25 = 0.25$ |
| Tổng | | $0.5 = \sigma^2$ |



Tóm lại: μ và σ (trung bình tổng thể và độ lệch chuẩn) là các đặc trưng chúng ta có thể ước tính được từ sự phân phối xác suất. Chúng hoàn toàn tương tự với trung bình mẫu $\bar{x}$ và độ lệch chuẩn s được ước tính từ dữ liệu mẫu.

***(Còn nữa)***

***Biên dịch: Minh Ánh và các nghiên cứu viên, Viện Khoa học Thống kê***