CHƯƠNG 1: ÔN TẬP

1.1. Trung bình mẫu – Phương sai mẫu

1.1.1. Trung bình mẫu

Trong phân tích dữ liệu, cũng như trong cuộc sống hàng ngày, chúng ta thường nói đến chiều cao trung bình, thu nhập trung bình, vân vân. Đó chính là trung bình mẫu. Hãy xét ví dụ sau:

Ví dụ 1.1: Bảng quan sát nhiệt độ ở Đà Lạt

Thứ 2 (x_1)	Thứ 3 (x_2)	Thứ 4 (x_3)	Thứ 5 (x_4)
19^o	21^o	20^o	18^o

$$\Rightarrow \bar{x} = \tfrac{1}{4}(19+21+20+18) = 19.5^o$$

Một cách khái quát, *trung bình mẫu* được tính bằng công thức sau:

$$\bar{x} = \frac{1}{N}(x_1 + x_2 + x_3 + \ldots\ldots + x_N)$$

$$\textit{Hay:} \quad \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

1.1.2. Phương sai mẫu

Phương sai mẫu [ký hiệu s_X^2] bằng trung bình của tổng bình phương độ lệch giữa giá trị quan sát so với giá trị trung bình:

$$s_X^2 = \frac{1}{N}\left[(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots\ldots(x_N-\bar{x})^2\right]$$

$$\textit{Hay:} \quad s_X^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})^2$$

Chẳng hạn, về trung bình mà nói thì khí hậu ở sa mạc rất nóng. Hơn nữa nhiệt độ giao động rất lớn giữa ngày và đêm. Để thể hiện được sự khắc nghiệt của khí hậu sa mạc, chúng ta không những chỉ sử dụng trung bình (mẫu) về nhiệt độ, mà cả sự giao động của nhiệt độ theo từng thời điểm so với trung bình. Đó chính là khái niệm về phương sai mẫu nói trên.

1.2. Hàm mật độ xác suất, hàm phân bố xác suất

1.2.1. Tần suất và xác suất

Để có sự hình dung về tần suất, hãy xét ví dụ sau:

Ví dụ 1.2: Xếp hạng tốc độ gia tăng giá cổ phiếu trên thị trường chứng khoán Việt Nam.

Gọi X là tỉ lệ phần trăm mức tăng giá cổ phiếu trung bình trong 3 tháng đầu tiên sau khi "lên sàn"; gọi P là phần trăm các công ty có mức tăng giá cổ phiếu tương ứng với giá trị của X

	X	Y
(x_1)	50%	10%
(x_2)	40%	20%
(x_3)	30%	35%
(x_4)	20%	25%

Con số P= 10%, X= 50% có nghĩa là có 10% trong tổng số các công ty có mức tăng giá trong 3 tháng đầu sau khi phát hành cổ phiếu ra công chúng là 50%. Đó chính là ví dụ về tần suất

Ví dụ 1.3: Trò chơi tung đồng xu.

Giả sử bạn tham gia cuộc chơi tung đồng xu tại hội chợ. Nếu là mặt sấp, bạn sẽ được \$100. Ngược lại, nếu là mặt ngửa, bạn được \$0. Với thể lệ đó, bạn sẵn sàng trả bao nhiêu đôla để tham gia trò chơi?

Để cho tiện, hãy kí hiệu mặt sấp là 1, mặt ngửa là 0. Giả sử kết quả tung xu sau 10 lần là như sau:

X	P
1	3/10
0	7/10

Con số 3/10 chính là tần suất xuất hiện mặt sấp (X = 1). Nghĩa là, trong 10 lần tung xu, có 3 lần xuất hiện mặt sấp. Và do đó, có 7 lần xuất hiện mặt ngửa.

Số tiền bạn bỏ ra cho việc tham dự 10 lần tung xu là: \$50 x 10 = \$500.
Số tiền nhận được trong cuộc chơi: \$100 x 3 + \$0 x 7 = \$300.

→ Do vậy, cuộc chơi không hứng thú đối với bạn ($500 > $300).

Tuy nhiên, nếu giả sử rằng bạn tham dự cuộc chơi vô hạn lần. Khi đó, số lần xuất hiện mặt sấp và mặt ngửa là như nhau, và bằng ½. Khi đó, kỳ vọng đượccuộc sẽ là: $100x1/2 + $0x1/2 = $50; và bằng chính số tiền lớn nhất bạn sẵn sàng trả để tham dự cuộc chơi.

Điều chúng ta cần phân biệt là con số P = 3/10 trong ví dụ nêu trên là *tần suất* xuất hiện mặt sấp trong 10 lần thử. Và con số ½ là *xác suất* xuất hiện mặt sấp (hoặc ngửa). Khái niệm tần suất ứng với từng mẫu thử; còn xác suất tương ứng với tổng thể.

1.2.2. Biến ngẫu nhiên rời rạc và liên tục

2.2.1. Biến ngẫu nhiên rời rạc:

Một biến ngẫu nhiên là **rời rạc** nếu các giá trị có thể có của nó lập nên một tập hợp hữu hạn hoặc đếm được, nghĩa là có thể liệt kê được tất cả các giá trị có thể có của nó.

Cuộc chơi tung xu nêu trên là ví dụ về biến ngẫu nhiên rời rạc.

Một cách hình thức hóa, ta có thể nói như sau. Giả sử đối tượng quan sát X có thể xuất hiện trong K sự kiện khác nhau [trong ví dụ tung xu, K = 2]. Ta ký hiệu các sự kiện đó là $x_1, x_2, ..., x_K$.

Tần suất xuất hiện một biến cố x_k trong N phép thử, ký hiệu là p_k, là tỉ số giữa số lần xuất hiện biến cố cụ thể đó so với N phép thử được thực hiện.

Với mọi chỉ số, $k = 1,2,3,...,K$, ta có thể viết như sau:

X	x_1	x_2	x_3	…	x_K
P	p_1	p_2	p_3	…	p_K

$p_1, p_2, p_3, \ldots p_K > 0$, và

$p_1 + p_2 + p_3 + \ldots\ldots + p_K = 1$, hay cũng vậy,

$$\sum_{k=1}^{K} p_k = 1$$

Nếu số mẫu N là đủ lớn (tiến đến vô hạn), khái niệm tần suất xuất hiện một biến cố được thay bằng khái niệm **xác suất** xuất hiện biến cố, ký hiệu bởi: $f_k = f(x_k), k = 1,2,..,K$. Trong đó, $f(x_k)$ là hàm mật độ xác suất của $x_k, k = 1,2...K$.

Ta cũng có,

$$f_1, f_2, f_3, \ldots f_K > 0, \text{ và}$$

$$\sum_{k=1}^{K} f_k = 1$$

2.2.2. Biến ngẫu nhiên liên tục

Một biến ngẫu nhiên là liên tục nếu các giá trị có thể có của nó lấp đầy một khoảng trên trục số, nghĩa là không thể liệt kê và đếm được tất cả các giá trị có thể có của nó.

Tương tự với trường hợp phân bố xác suất rời rạc, nếu gọi X là một biến ngẫu nhiên liên tục; và f(x) là hàm mật độ xác suất của X. Khi đó:

$$f(x) \geq 0$$

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Ta định nghĩa hàm phân bố xác suất của X là:

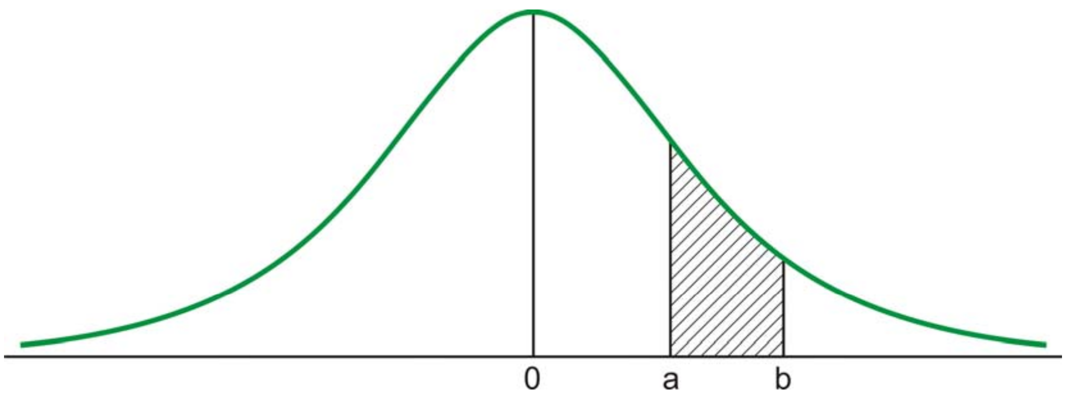
$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Điều đó có nghĩa là, xác suất của biến ngẫu nhiên X nhận giá trị trong khoảng $[a,b]$ sẽ là:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$$

Ví dụ, trong phân bố chuẩn, về đồ thị ta có thể biểu diễn công thức tính xác suất này như sau:

Đồ thị 1.1: Phân bố xác suất



Phần tô đậm chính là xác suất $P(a \leq X \leq b)$, được tính bởi tích phân:
$\int_a^b f(x)dx = F(b) - F(a)$.

1.3. Phân bố xác suất đồng thời

Nhiều khi chúng ta muốn đưa ra một đánh giá xác suất đồng thời cho một số biến lượng ngẫu nhiên. Ví dụ, bảng thống kê có ghi lại dữ kiện về thất nghiệp (u) và lạm phát (п). Cả hai biến lượng này đều là biến ngẫu nhiên, rất nhiều khả năng là chính phủ muốn hỏi những nhà kinh tế câu hỏi sau đây: *"Liệu khả năng lạm phát thấp hơn 8% và mức độ thất nghiệp nhỏ hơn 6% vào năm sau là bao nhiêu?"*. Điều đó có nghĩa là, ta cần phải xác định xác suất đồng thời:

$$P(\pi < 8, u < 6) = ?$$

Để trả lời được những câu hỏi như vậy, chúng ta cần phải xác định *hàm mật độ xác suất đồng thời* [joint probability density function].

1.3.1. Hàm mật độ xác suất đồng thời

Định nghĩa: Giả sử X và Y là 2 biến ngẫu nhiên. Hàm mật độ xác suất đồng thời của *x* và *y* là:

$$f(x, y) = P(X = x, Y = y)$$

Hàm số đó cần thỏa mãn điều kiện:

$f(x, y) \geq 0$, và

$\sum_x \sum_y f(x, y) = 1$ nếu X, Y rời rạc

$\int_x \int_y f(x, y)dy.dx$ nếu X, Y liên tục

Khi đó,

$P(a \leq x \leq b, c \leq y \leq d) = \sum_{a \leq x \leq b} \sum_{c \leq y \leq d} f(x, y)$, nếu X, Y là biến ngẫu nhiên rời rạc, và

$P(a \le x \le b, c \le y \le d) = \int_a^b \left[\int_c^d f(x,y)dy\right]dx$, nếu X,Y là biến ngẫu nhiên liên tục.

1.3.2. Hàm phân bố xác suất đồng thời F(x,y)

Tương tự như trường hợp biến ngẫu nhiên một biến, ta đưa ra định nghĩa sau về hàm phân bố xác suất đồng thời:

Định nghĩa: Gọi F(x,y) là hàm phân bố xác suất đồng thời của biến ngẫu nhiên x và y. Khi đó:

$$F(x,y) = \mathrm{Pr}ob(X \le x, Y \le y) = \sum_{X \le x}\sum_{Y \le y} f(x,y) \text{, nếu X, Y rời rạc}$$

$$F(x,y) = \mathrm{Pr}ob(X \le x, Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(t,s)ds.dt \text{ , nếu X, Y liên tục}$$

1.3.3. Phân phối xác suất cận biên

Hãy xét ví dụ sau:

Ví dụ 4: Xét một tổng thể, gồm có 1000 người. [Vì vậy ta nói về mật độ xác suất chứ không phải là tần suất]. Giả sử họ được phân loại theo 2 tiêu chuẩn:

Theo giới tính:

G = 1 nếu người đó là nam
G = 0 nếu người đó là nữ

Và theo trình độ học vấn:

D = 0 học xong trung học
D = 1 học xong đại học
D = 2 học xong cao học

Giả sử kết quả thống kê trên tổng thể 1000 người đó là như sau:

	Nam	Nữ	Học vị (tổng số)
Trung học	200	270	470
Đại học	300	100	400
Cao học	60	70	130
Giới tính(tổng số)	560	440	1000

Dựa trên bảng thống kê này, chúng ta có thể thấy xác suất 1 cá nhân là nữ, học xong đại học: f(0,1)= 100/1000 = 0.1. Một cách khái quát, chúng ta có thể viết hàm mật độ xác suất đồng thời $f(G,D)$ như sau:

		G		**Tổng**
		1	**2**	
D	**0**	0.2	0.27	0.47
	1	0.3	0.1	0.40
	2	0.06	0.07	0.13
Tổng		0.56	0.44	1

Bảng phân bố xác suất trên cho thấy, xác suất một cá nhân là nam trong tổng thể những người có học là: Prob(G=1) = 0.56. Tương tự, xác suất một cá nhân là nữ: Prob(G=0) = 440/1000 = 0.44.

Như vậy, ta có thể lập một biến ngẫu nhiên, thể hiện phân bố mật độ xác suất theo giới tính của tổng thể:

G	f(g)
1	0.56
0	0.44

Hàm f(G) được gọi là hàm mật độ xác suất cận biên. Hàm mật độ này được tính bằng cách cộng dồn theo cột qua tất cả mọi trình độ học vấn:

$f(g) = \sum_d f(g,d), \quad g = \overline{0,1,2}$. Tức là:

$$\begin{cases} f_G(1) = \sum_d f(1,d) = 0.56 \\ f_G(0) = \sum_d f(0,d) = 0.44 \end{cases}$$

Tương tự như vậy, ta cũng có thể tính được hàm mật độ xác suất cận biên theo học vấn:

$$f_D(d) = \sum_g f(g,d) \qquad\qquad d = \overline{0,1,2}$$

Hay cũng vậy,

$$\begin{cases} f_D(0) = \sum_g f(g,0) = 0.47 \\ f_D(1) = \sum_g f(g,1) = 0.4 \\ f_D(2) = \sum_g f(g,2) = 0.13 \end{cases}$$

Một cách tổng quát, gọi f(x,y) là hàm mật độ xác suất đồng thời của X và Y. Khi đó, hàm mật độ xác suất cận biên của X được xác định như sau:

$$f_X(x) = \sum_y f(x,y) \qquad \text{nếu X rời rạc}$$
$$f_X(x) = \int_y f(x,y)dy \qquad \text{nếu X liên tục}$$

Tương tự, ta xác định $f_Y(y)$

1.3.4. Các biến ngẫu nhiên độc lập

Định nghĩa: Hai biến ngẫu nhiên là độc lập khi và chỉ khi:

$$f(x,y) = f_X(x) \cdot f_Y(y)$$

$$\leftrightarrow F(x,y) = F_X(x) \cdot F_Y(y)$$

$$\leftrightarrow \Pr ob(X \le x, Y \le y) = \Pr ob(X \le x) \cdot \Pr ob(Y \le y)$$

1.4. Kỳ vọng – Phương sai

1.4.1. Khái niệm về Kỳ vọng của biến ngẫu nhiên:

Gọi X là một biến ngẫu nhiên rời rạc, nhận một trong các giá trị có thể có x_1, x_2, x_3,… x_K với xác suất tương ứng f_1, f_2, f_3,… f_K. Giá trị kỳ vọng của X được định nghĩa như sau:

$$E(X) = x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots\ldots + x_K f_K$$, hay cũng vậy:

$$E(X) = \sum_{k=1}^{K} x_k f_k$$

Tương tự, đối với biến ngẫu nhiên liên tục, giá trị kỳ vọng được định nghĩa như sau:

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx$$

Các tính chất của kỳ vọng:

1. $E(a) = a$, với a là hằng số
2. $E(a + bX) = a + bE(X)$
3. $E(XY) = E(X)E(Y)$

Định lý 1.1: Giả sử X là một biến ngẫu nhiên với hàm mật độ xác suất *f(x)* và *g(X)* là một hàm liên tục của X. Khi đó:

$$E[g(X)] = \sum_k g(x_k) f_k \qquad \text{nếu X rời rạc}$$

$$E[g(X)] = \int_{-\infty}^{+\infty} g(X) f(x)dx \qquad \text{nếu X liên tục}$$

1.4.2. Phương sai

Gọi X là một biến ngẫu nhiên với kỳ vọng EX. Để đo lường sự tán xạ của X so với giá trị trung bình (hay kỳ vọng) của nó, ta sử dụng phương sai, ký hiệu Var(X), được định nghĩa như sau:

$$Var(X) = \sigma_x^2 = E(X - E(X))^2$$

Với độ lệch chuẩn:

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

Sử dụng Định lý 1.1, phương sai của X được tính như sau:

$$Var(X) = \sum_k (x_k - EX)^2 f_k \qquad \text{nếu X rời rạc}$$

$$Var(X) = \int_{-\infty}^{+\infty} (X - E(X))^2 f(x)dx \qquad \text{nếu X liên tục}$$

Các tính chất của phương sai:

1. $VarX = E(X - E(X))^2 = E(X^2) - (E(X))^2$

2. $Var(a) = 0$, với a là hằng số
3. $Var(a + bX) = b^2 \cdot Var(X)$
4. $Var(X + Y) = Var(X) + Var(Y)$
   $Var(X - Y) = Var(X) + Var(Y)$
5. $Var(X - E(X)) = Var(X)$

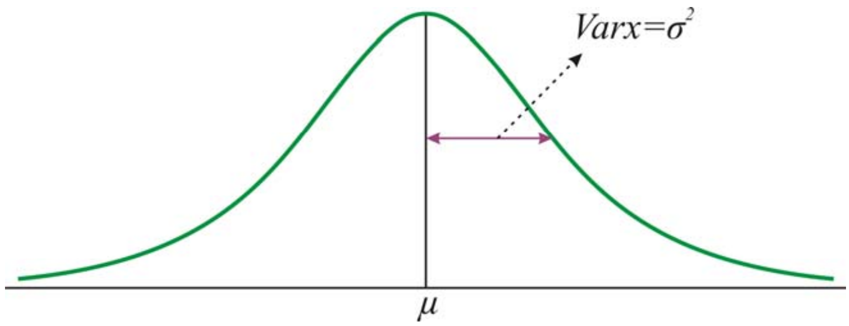
1.5. Hàm phân phối chuẩn

Biến ngẫu nhiên liên tục X nhận các giá trị trong khoảng $(-\infty, +\infty)$ có phân phối chuẩn với các tham số μ và σ^2, ký hiệu là: $X \sim N(\mu, \sigma^2)$, nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

với $\mu = E(X)$ và $\sigma^2 = Var(X)$

Đồ thị 1.2: Hàm phân phối chuẩn



Định lý 1.2: Giả sử X là biến ngẫu nhiên với phân bố chuẩn: $X \sim N(\mu, \sigma^2)$**.** Gọi $Z = (a + bx)$ là một biến đổi tuyến tính của X. Khi đó, Z cũng là hàm phân bố chuẩn: $Z \sim N(a + b\mu, b^2\sigma^2)$.

Hệ quả: Đặt $Z = \frac{x - \mu}{\sigma}$. Khi đó**,** $Z \sim N(0,1)$

Địnhlý 1.3: Cho trước chuỗi các biến ngẫu nhiên: $(x_1, x_2, x_3, ..., x_n) \sim N(\mu_n, \sigma_n^2)$

Khi đó, tổ hợp tuyến tính của chúng, cũng có phân bố chuẩn:

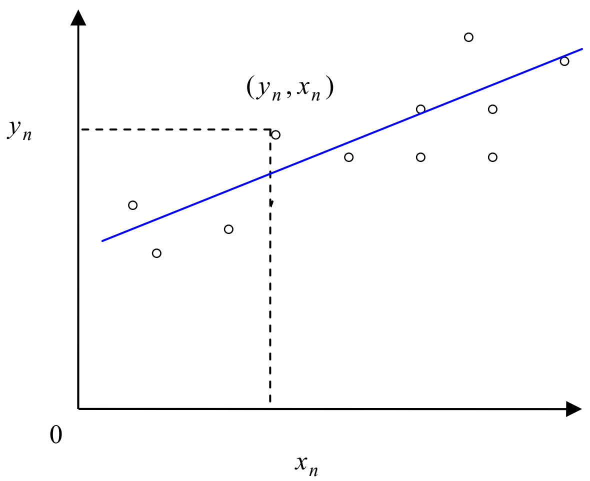
$$c_1x_1 + c_2x_2 + ... + c_nx_n \sim N\left(\sum \mu_n, \sum c_n^2\sigma_n^2\right)$$

1.6. Phân tích Covariance

Trong phần trên, chúng ta đã nói đến việc tồn tại hay không tính độc lập, hay quan hệ phụ thuộc giữa hai biến ngẫu nhiên *X* và *Y*. Nhưng nếu tồn tại quan hệ phụ thuộc lẫn nhau, thì quan hệ đó có thể mạnh hay yếu. Trong phần này, chúng ta sẽ đề cập tới hai thước đo mức độ liên quan giữa hai biến ngẫu nhiên**, tương quan** (hay **covariance**), và **hệ số tương quan** (hay **correlation**, ký hiệu là ρ_{XY}).

Để minh họa, giả sử *X* là trọng lượng của một mẫu nước lấy từ giếng lên, và *Y* là khối lượng của nó. Hiển nhiên là mối quan hệ rất chặt giữa *X* và *Y*. Nếu ta ký hiệu $\{x_n, y_n\}_{n=1}^{N}$ là các cặp đo lường với *N* mẫu thử; và vẽ chúng lên đồ thị, thì các quan sát dữ liệu này sẽ tạo thành một đường thẳng tuyến, thể hiện mối quan hệ vật lý của chúng. Nhưng chúng không rơi đúng vào các điểm dọc theo đường tuyến tính thể hiện quy luật liên hệ giữa khối lượng và trọng lượng nước. Chúng chỉ "bám" xung quanh cái trục tuyến tính đó, vì có sai số đo lường, hoặc các tạp chất trong nước làm các quan sát lệch khỏi quy luật vật lý, mô tả mối quan hệ ổn định giữa *X* và *Y*.

Đồ thị 1.3: Mối quan hệ giữa trọng lượng nước X và khối lượng nước Y



Câu hỏi đặt ra là làm sao chúng ta có thể đo lường mức độ tương quan mạnh hay yếu giữa hai biến X và Y này. Làm sao thể hiện mối quan hệ đó là đồng biến hay nghịch biến?

1.6.1. Covariance

Định nghĩa: *Covariance* giữa hai biến X và Y là hệ số đo:

$$Cov(X,Y) = E[(X - EX)(y - EY)]$$

Nếu như những giá trị lớn hơn trung bình của X được quan sát với những giá trị lớn hơn trung bình của Y; và những giá trị nhỏ của X cũng đi kèm với những giá trị nhỏ của Y, thì $Cov(X,Y) > 0$. Nói khác đi, nếu $(X - EX) > 0$ có xu hướng đi kèm với $(Y - EY) > 0$; hay ngược lại, khi $(X - EX) < 0$, thì $(Y - EY) < 0$, thì quan hệ đó có xu hướng tạo ra tích $(X - EX)(Y - EY) > 0$. Điều đó có nghĩa là $Cov(X,Y) > 0$, thể hiện rằng *X và Y* có mối quan hệ **đồng biến**. Ví dụ như quan hệ giữa khối lượng và trọng lượng các mẫu nước vừa nêu.

Nhiều khi, mối tương quan là **nghịch biến**, chứ không thuận. Chẳng hạn như chúng ta quan sát mối quan hệ giữa điều kiện bảo trợ quá dễ dàng cho một cá nhân, hay doanh nghiệp (ký hiệu là X); và nỗ lực tự vươn lên, tính khởi nghiệp của cá nhân, hay doanh nghiệp đó (ký hiệu là Y). Khi đó, mối quan hệ này thường là nghịch biến. Hỗ trợ nhiều làm chết tính tự chủ, tự vươn lên, tự chịu trách nhiệm của cá nhân. Nói khác đi, giá trị X rất lớn [được nâng đỡ, bảo trợ nhiều] thường đi với giá trị Y rất nhỏ [thiếu nỗ lực bản thân, hay ỉ lại]. Và giá trị X rất nhỏ [không được nâng đỡ] thường đi với giá trị Y rất lớn [tính tự lập, tự chủ cao]. Do vậy, $(X - EX) > 0$ thường đi kèm với $(Y - EY) < 0$, và $(X - EX) < 0$ thường xẩy ra với $(Y - EY) > 0$. Kết cục lại, chúng thường tạo ra tích $(X - EX)(Y - EY) < 0$. Hay cũng vậy, $Cov(X,Y) < 0$, thể hiện mối quan hệ nghịch biến giữa X và Y.

Chúng ta cũng nhận xét luôn rằng, mối quan hệ giữa việc được hỗ trợ, bảo trợ, với tính tự chủ, tự chịu trách nhiệm, ký hiệu là X và Y là nghịch biến. Nhưng về mức độ, nó có thể không mạnh như quan hệ vật lý giữa khối lượng và trọng lượng nước. Nếu chúng ta vẽ đồ thị các quan sát, mối quan hệ giữa việc được hỗ trợ với tính tự vươn lên sẽ **dốc** xuống, thể hiện mối quan hệ nghịch biến. Nhưng không nhất thiết nằm xung quanh một đường thẳng, trải dọc theo một đường cong phi tuyến, thể hiện mối quan hệ đó là yếu hơn so với quan hệ vật lý ở ví dụ đầu. Để đo lường sự khác biệt đó ta dùng hệ số tương quan.

1.6.2. Hệ số tương quan:

Định nghĩa: Hệ số tương quan giữa X và Y là hệ số đo $\rho(X,Y)$:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{VarX \cdot VarY}} \qquad \left(-1 \le \rho(X,Y) \le 1\right)$$

Ta có thể nói rằng, covariance cho phép xác định có mối quan hệ hay không giữa X và Y, và đó là quan hệ nghịch biến hay đồng biến. Hệ số tương quan lại cho phép đo lường mối quan hệ đó là mạnh tới mức nào. Nếu X và Y có quan hệ tuyến tính: $X = \alpha \pm \beta Y$, thì quan hệ đó là mạnh nhất. Và $|\rho(X,Y)| = 1$. Nếu đó là quan hệ phi tuyến, thì $|\rho(X,Y)| < 1$. Khi X và Y không có quan hệ tương quan: $Cov(X,Y) = 0$, khi đó, hệ số tương quan $\rho(X,Y) = 0$.

1.6.3. Hai đẳng thức với tương quan mẫu

Hai đẳng thức sau là hai đẳng thức thường sử dụng trong các chương tiếp theo.

1/ $\sum_n (x_n - \bar{x}) \cdot c = 0$, với c: const

2/ $\sum_n (x_n - \bar{x}) \cdot y_n = \sum_n [(x_n - \bar{x}) \cdot (y_n - \bar{y})]$

Chứng minh:

1/ $\sum_n (x_n - \bar{x}) \cdot c = c \cdot \sum_n (x_n - \bar{x}) = c \cdot (\sum_n x_n - \sum_n \bar{x}) = c \cdot (n \cdot \bar{x} - n \cdot \bar{x}) = 0$

2/ Vì $\bar{y}$ là hằng số nên theo chứng minh trên $\sum_n (x_n - \bar{x}) \cdot y_n = 0$, vì vậy:

$$\sum_n (x_n - \bar{x}) \cdot y_n = \sum_n (x_n - \bar{x}) \cdot y_n - \sum_n (x_n - \bar{x}) \cdot \bar{y}$$
$$= \sum_n [(x_n - \bar{x}) \cdot y_n - (x_n - \bar{x}) \cdot \bar{y}]$$
$$= \sum_n [(x_n - \bar{x}) \cdot (y_n - \bar{y})]$$

Chú ý rằng, dòng cuối cùng được gọi là tương quan mẫu giữa X và Y.