Bài giảng 2: Phân phối và các đặc tính của dữ liệu

TS. Tô Đức Khánh

Khoa Toán-Tin Học, Trường Đại Học Khoa Học Tự Nhiên
Đại Học Quốc Gia Tp. HCM

Học phần: Xử lý Số liệu Thống kê

Nội dung buổi học

Các đặc trưng của dữ liệu

Dữ liệu của một biến ngẫu nhiên được đặc trưng bởi ba yếu tố chính:

- giá trị trung tâm - central tendency,
- độ biến động - variability,
- phân phối dữ liệu.

Cụ thể

giá trị trung tâm cung cấp ước tính về nơi chứa hầu hết dữ liệu

giá trị biến động đo lường xem các giá trị dữ liệu được phân cụm chặt chẽ hay dàn trải.

phân phối dữ liệu cung cấp cái nhìn tổng quát về sự phân bố của dữ liệu (tập trung dày ở một vùng cụ thể, dàn trải trên vùng nào đó).

Giá trị trung tâm - Location

Có nhiều định nghĩa giá trị trung tâm khác nhau:

- trung bình cộng - mean (average)
- trung bình bị cắt bớt - trimmed mean (trimmed average)
- trung bình có trọng số - weighted mean (weighted average)
- trung vị - median (phân vị thứ 2 - second quantile)
- trung vị có trọng số - weighted median

Giá trị trung tâm - Location

Trung bình cộng - mean

Trung bình cộng - mean (average) là ước lượng cơ bản cho giá trị trung tâm của dữ liệu:

$$\text{mean} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

với n là số lượng dữ liệu được quan sát.

Trung bình bị cắt bớt - trimmed mean

Trung bình bị cắt bớt - trimmed mean là một biến thể của trung bình, nó tính dựa trên bộ dữ liệu đã được cắt bỏ đi $p\%$ dữ liệu nhỏ nhất và lớn nhất (hai đầu của dữ liệu được sắp xếp thứ tự):

$$\text{trimmed mean} = \overline{x}_p = \frac{1}{n - 2\lfloor np \rfloor}\sum_{i=\lfloor np \rfloor + 1}^{n-\lfloor np \rfloor} x_{(i)},$$

với $x_{(i)}$ là giá trị của dữ liệu sau khi được sắp xếp tăng dần.

Giá trị trung tâm - Location

Trung bình có trọng số - weighted mean

Trung bình có trọng số - weighted mean là một biến thể khác của trung bình, ở đó, các giá trị x_i được nhân thêm với một trọng số w_i:

$$\text{weighted mean} = \overline{x}_w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}.$$

Trung bình có trọng số thường được áp dụng khi:

- khi dữ liệu là kết quả từ các nhóm có kích thước khác nhau: tỷ lệ tội phạm của các thành phố;
- khi dữ liệu được thu thập từ các nguồn biến động không đồng nhất: dữ liệu thu được từ các sensor có chất lượng khác nhau.

Giá trị trung tâm - Location

Trung vị - median

Trung vị - median là một giá trị của dữ liệu quan sát, mà chia dữ liệu thành hai nửa đều nhau, khi dữ liệu được sắp xếp theo thứ tự tăng dần.

129 132 133 137 138 (138) | (140) 140 140 141 143 143

139

Trung vị của một bộ dữ liệu là nghiệm của

$$\text{median} = \arg\min_{x} \sum_{i=1}^{n} |x - x_i|$$

Giá trị trung tâm - Location

Trung vị có trọng số - weighted median

Trung vị có trọng số - weighted median là một biến thể của trung vị, được áp dụng cho trường hợp dữ liệu có trọng số.

Nó là điểm x_k trong dữ liệu sao cho

$$\sum_{x_i<x_k} w_i \leq \frac{1}{2}\sum_{i=1}^{n} w_i, \quad \sum_{x_i>x_k} w_i \leq \frac{1}{2}\sum_{i=1}^{n} w_i$$

Định nghĩa này tương ứng với trung vị có trọng số là nghiệm của

$$\text{weighted median} = \arg\min_x \sum_{i=1}^{n} w_i|x - x_i|$$

Ví dụ 1: Dữ liệu tỷ lệ giết người ở Mỹ

	State	Population	Murder.Rate	Abbreviation
1	Alabama	4779736	5.7	AL
2	Alaska	710231	5.6	AK
3	Arizona	6392017	4.7	AZ
4	Arkansas	2915918	5.6	AR
5	California	37253956	4.4	CA
6	Colorado	5029196	2.8	CO
7	Connecticut	3574097	2.4	CT
8	Delaware	897934	5.8	DE
9	Florida	18801310	5.8	FL
10	Georgia	9687653	5.7	GA
11	Hawaii	1360301	1.8	HI

Murder rate = (số vụ giết người/dân số) $\times$ 100,000 trong một năm của một bang.

Ví dụ 1: Dữ liệu tỷ lệ giết người ở Mỹ

Xét:

- Dân số - Population

	giá trị
mean	6162876
trimmed mean (10%)	4783697
median	4436370

- Murder rate

	giá trị
weighted mean	4.46
weighted median	4.4

Chú ý, murder rate được tính cho mỗi bang với kích thước dân số khác nhau, nên ta cần dùng Population như một trọng số.

Độ biến động - Variability

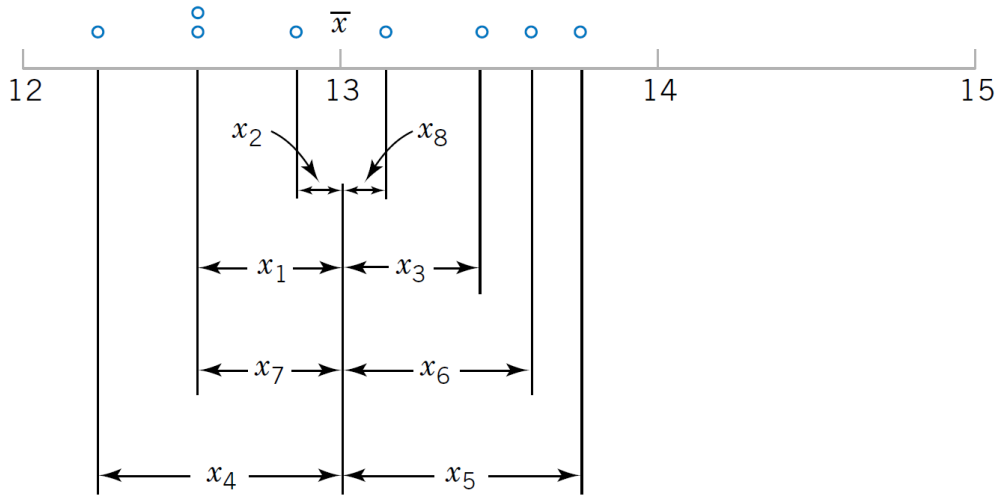
Độ biến động của dữ liệu được miêu tả thông qua các số đo:

- phương sai - variance
- độ lệch chuẩn - standard deviation
- trung bình độ lệch tuyệt đối - mean absolute deviation
- trung vị độ lệch tuyệt đối - median absolute deviation (MAD)
- khoảng biến động - range
- phân vị
- khoảng tứ phân vị - interquartile range (IQR)

Độ biến động - Variability

Phương sai, độ lệch chuẩn

Phương sai và độ lệch chuẩn (variance, standard deviation) là hai đại lượng đo độ biến động của dữ liệu xung quanh giá trị trung bình.



Phương sai được ký hiệu là s^2 và được xác định bởi:

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

Độ lệch chuẩn, được ký hiệu là s và được xác định bởi: $s = \sqrt{s^2}$.

Độ biến động - Variability

Trung bình độ lệch tuyệt đối

Trung bình độ lệch tuyệt đối (mean absolute deviation) là đại lượng biểu thị độ biến động của dữ liệu xung quanh giá trị trung bình thông qua trung bình khoảng cách Euclidean:

$$\text{mean absolute deviation} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|,$$

- Độ lệch chuẩn dễ diễn giải hơn nhiều so với phương sai vì nó có cùng tỷ lệ với dữ liệu gốc.
- Giá trị của phương sai, độ lệch chuẩn, trung bình độ lệch tuyệt đối càng nhỏ thì dữ liệu càng ít biến động so với trung bình.

Độ biến động - Variability

Trung vị độ lệch tuyệt đối - median absolute deviation (MAD)

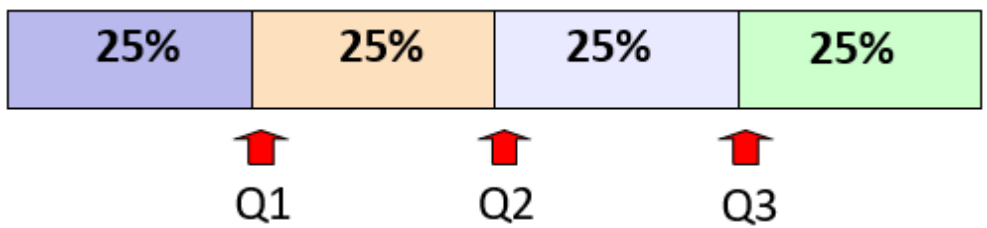
Trung vị độ lệch tuyệt đối - MAD là thước đo độ biến thiên dữ liệu xung quanh điểm trung vị:

$$\text{MAD} = \text{median}(|x_1 - m|, |x_2 - m|, \ldots, |x_n - m|),$$

với m là trung vị.

Khoảng tứ phân vị - interquartile range

Khoảng tứ phân vị - interquartile range là khoảng cách giữa phân vị thứ nhất (25% dữ liệu) và phân vị thứ ba (75% dữ liệu).



Độ biến động - Variability

Độ lệch chuẩn có trọng số - weighted standard deviation

Độ lệch chuẩn có trọng số - weighted standard deviation là một biến thể của độ lệch chuẩn, để tương thích với các trường hợp dữ liệu có trọng số.

$$s_w = \sqrt{\frac{\sum_{i=1}^{n} w_i (x_i - \overline{x}_w)^2}{\frac{n'}{n'-1} \sum_{i=1}^{n} w_i}},$$

trong đó, n' là số lượng trọng số khác 0.

Tương tự, ta có:

- phương sai có trọng số - weighted variance
- trung vị độ lệch tuyệt đối có trọng số - weighted median absolute deviation (weighted MAD).

Ví dụ 1: Dữ liệu tỷ lệ giết người ở Mỹ

Xét:

- Dân số - Population

	giá trị
standard deviation	6848235
MAD	3849870
IQR	4847308

- Murder rate

	giá trị
weighted standard deviation	1.504
weighted MAD	1.450

Chú ý, murder rate được tính cho mỗi bang với kích thước dân số khác nhau, nên ta cần dùng Population như một trọng số.

Giá trị ngoại lai - Outliers

Outliers - Extreme values

Giá trị ngoại lai (outliers/extreme values) là bất kỳ giá trị nào khác rất xa so với các giá trị còn lại trong tập dữ liệu.

129, (1.65), 132, 133, 137, 138, (1308), 140, 140, 141, 143, 143, (1405)

Giá trị ngoại lai - Outliers

Outliers - Extreme values

Giá trị ngoại lai (outliers/extreme values) là bất kỳ giá trị nào khác rất xa so với các giá trị còn lại trong tập dữ liệu.

129, (1.65), 132, 133, 137, 138, (1308), 140, 140, 141, 143, 143, (1405)

Giá trị ngoại lai có thể là:

- một quan sát hiếm khi xảy ra: lượng mưa lớn đột ngột, số lượng lớn khác hàng trong một giờ;
- kết quả của lỗi dữ liệu: đơn vị không đồng nhất giữa các giá trị, sai sót trong đo lường hoặc ghi chép.

Giá trị ngoại lai - Outliers

Outliers - Extreme values

Giá trị ngoại lai (outliers/extreme values) là bất kỳ giá trị nào khác rất xa so với các giá trị còn lại trong tập dữ liệu.

129, (1.65), 132, 133, 137, 138, (1308), 140, 140, 141, 143, 143, (1405)

Giá trị ngoại lai có thể là:

- một quan sát hiếm khi xảy ra: lượng mưa lớn đột ngột, số lượng lớn khác hàng trong một giờ;
- kết quả của lỗi dữ liệu: đơn vị không đồng nhất giữa các giá trị, sai sót trong đo lường hoặc ghi chép.

Nhận xét:

- các giá trị ngoại lai có ảnh hưởng xấu tới ước lượng của trung bình và phương sai cũng như độ lệch chuẩn;
- các ước lượng: median, MAD là các ước lượng "robust" với outliers (tức là ít bị ảnh hưởng bởi outliers).

Giá trị ngoại lai - Outliers

Để nhận dạng được sự hiện diện của outliers trong dữ liệu, ta có thể dùng:

- khoảng tứ phân vị - IQR, cụ thể, các điểm dữ liệu nằm ngoài khoảng

$$[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$$

  thì được coi là outliers trong dữ liệu;

- median và MAD, cụ thể, ta xét khoảng

$$[\text{median} - 3 \times MAD, \text{median} + 3 \times MAD]$$

  nếu các điểm dữ liệu không nằm trong khoảng này, thì được coi là outliers. Cách xét này được gọi là bộ lọc Hampel - Hampel filter.

Giá trị ngoại lai - Outliers

Ví dụ: Xét biến Population trong dữ liệu tỷ lệ tội phạm ở Mỹ.

	Khoảng	outliers
IQR	$[-5437958, 13951274]$	18801310 19378102 25145561 37253956
Hampel filter	$[-7113242, 15985981]$	18801310 19378102 25145561 37253956

Nhắc lại phân phối của biến ngẫu nhiên

Phân phối xác suất

Phân phối xác suất của một biến ngẫu nhiên là một hàm số toán học mô tả quy luật phân bổ giá trị (về mặt lý thuyết) của biến ngẫu nhiên.

Phân phối xác suất thường được đặc trưng bởi

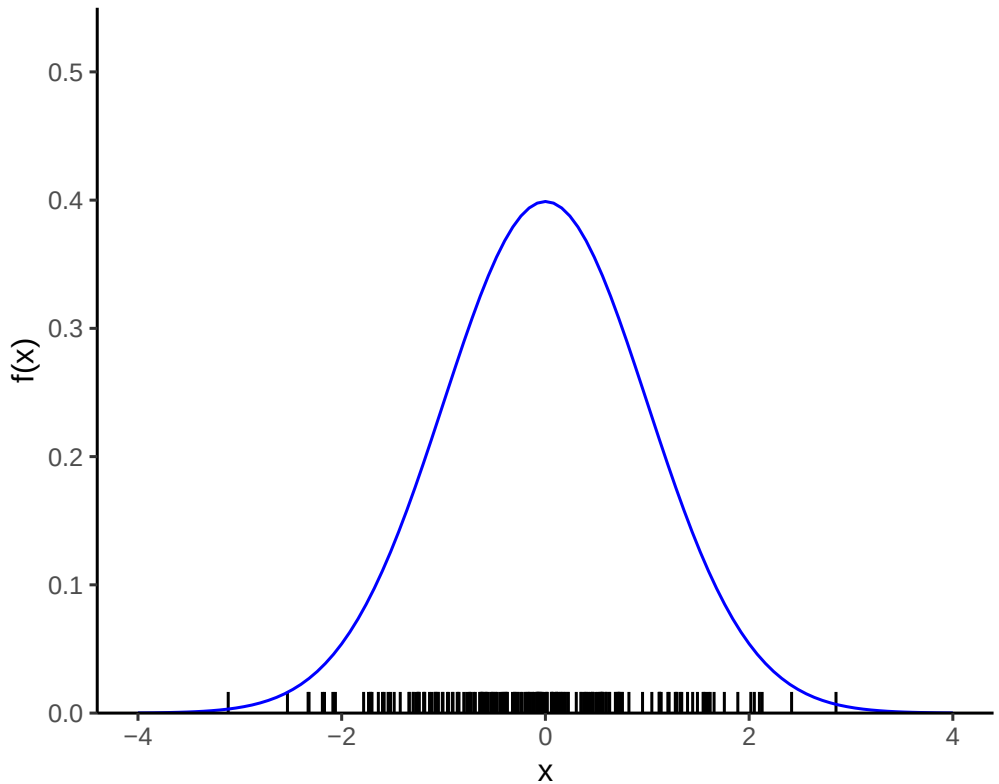
- hàm mật độ xác suất - probability density function đối với biến liên tục;
- hàm trọng lượng xác suất - probability mass function đối với biến rời rạc.

Các hàm số này là hàm dương.

Nhắc lại phân phối của biến ngẫu nhiên

Ví dụ: Xét biến ngẫu nhiên liên tục X có không gian giá trị $S \equiv \mathbb{R}$, với hàm mật độ xác suất:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$



Nhắc lại phân phối của biến ngẫu nhiên

Một số phân phối xác suất thường gặp:

- phân phối chuẩn - normal distribution;
- phân phối t-student - Student's t distribution;
- phân phối mũ - exponential distribution;
- phân phối đều - uniform distribution;
- phân phối gamma - gamma distribution;
- phân phối log chuẩn - log-normal distribution;
- phân phối Weibull - Weibull distribution;
- phân phối Bernoulli - Bernoulli distribution;
- phân phối nhị thức - binomial distribution;
- phân phối Poisson - Poisson distribution;
- phân phối nhị thức âm - negative binomial distribution.

Dự đoán phân phối xác suất

Trong thực tế, phân phối xác suất của dữ liệu là không được biết trước
↪ ta phải dự đoán.

Dự đoán phân phối xác suất

Trong thực tế, phân phối xác suất của dữ liệu là không được biết trước
↪ ta phải dự đoán.

Một số công cụ hay được sử dụng:

- bảng tần số - frequency table,
- biểu đồ tần số - histogram,
- biểu đồ hộp - boxplot,
- đồ thị hàm mật độ xác suất.

Bảng tần số

Bảng tần số - frequency table

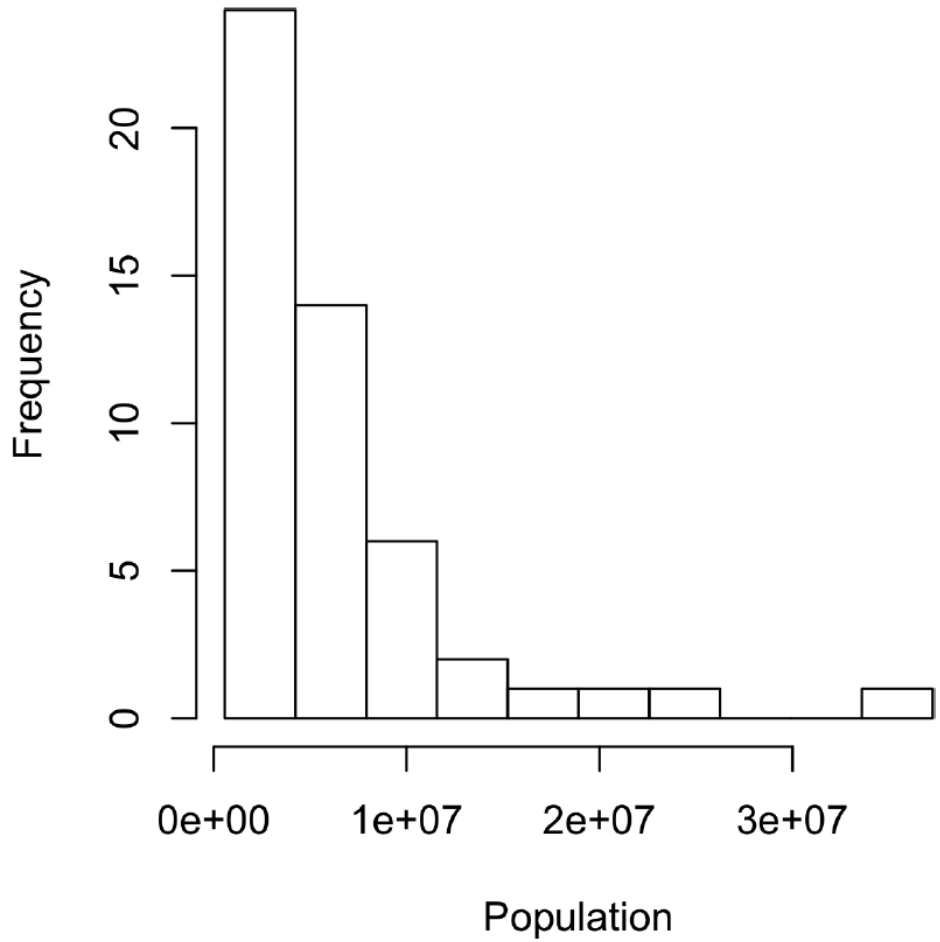
Bảng tần số của một biến chia phạm vi biến thành các phân đoạn cách đều nhau và cho chúng ta biết có bao nhiêu giá trị rơi vào mỗi phân đoạn.

BinNumber	BinRange	Count	States
1	563,626–4,232,658	24	WY,VT,ND,AK,SD,DE,MT,RI,NH,ME,HI,ID,NE,WV,NM,NV,UT,KS,AR,MS,IA,CT,OK,OR
2	4,232,659–7,901,691	14	KY,LA,SC,AL,CO,MN,WI,MD,MO,TN,AZ,IN,MA,WA
3	7,901,692–11,570,724	6	VA,NJ,NC,GA,MI,OH
4	11,570,725–15,239,757	2	PA,IL
5	15,239,758–18,908,790	1	FL
6	18,908,791–22,577,823	1	NY
7	22,577,824–26,246,856	1	TX
8	26,246,857–29,915,889	0	
9	29,915,890–33,584,922	0	
10	33,584,923–37,253,956	1	CA

Biểu đồ tần số

Biểu đồ tần số - histogram

Biểu đồ tần số là một cách để trực quan hóa bảng tần số, với các cột trên trục x và số lượng dữ liệu trên trục y.



Biểu đồ tần số

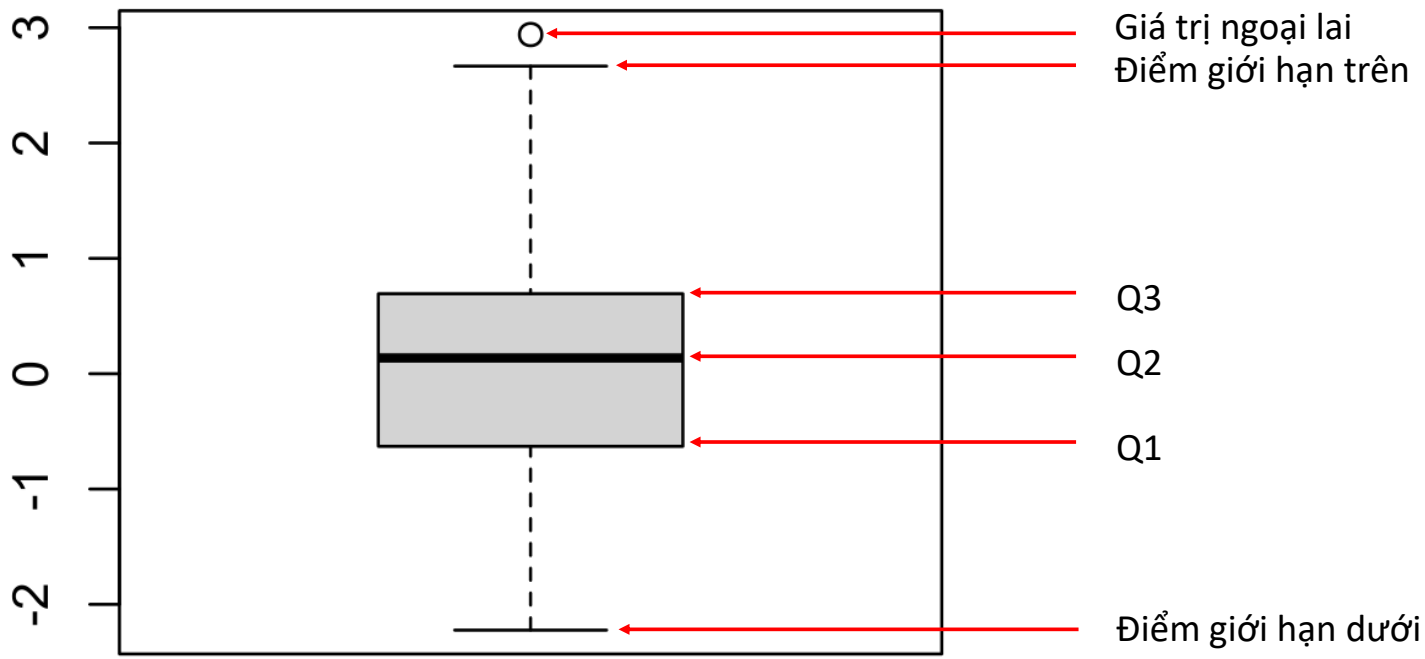
Về mặt kỹ thuật, một biểu đồ tần số được vẽ sao cho:

- Cột rỗng được bao gồm trong biểu đồ (khoảng không có dữ liệu).
- Các cột có chiều rộng bằng nhau.
- Số cột (hoặc tương đương, kích thước của cột) tùy thuộc vào người dùng.
- Các cột xếp kề nhau - không có khoảng trống nào hiển thị giữa các cột, trừ khi có cột trống.

Biểu đồ hộp

Biểu đồ hộp

Biểu đồ hộp hay **boxplot** là một biểu đồ dạng hộp dùng để miêu tả sự phân phối của dữ liệu, bằng việc biểu diễn các điểm tứ phân vị.



- Điểm giới hạn trên $= \min\{\max(x), Q_3 + 1.5 \times IQR\}$.
- Điểm giới hạn dưới $= \max\{\min(x), Q_1 - 1.5 \times IQR\}$.

Ước lượng hàm mật độ xác suất

Để ước lượng hàm mật độ xác suất, ta có một kỹ thuật được gọi tên là **kernel density estimation**:

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right),$$

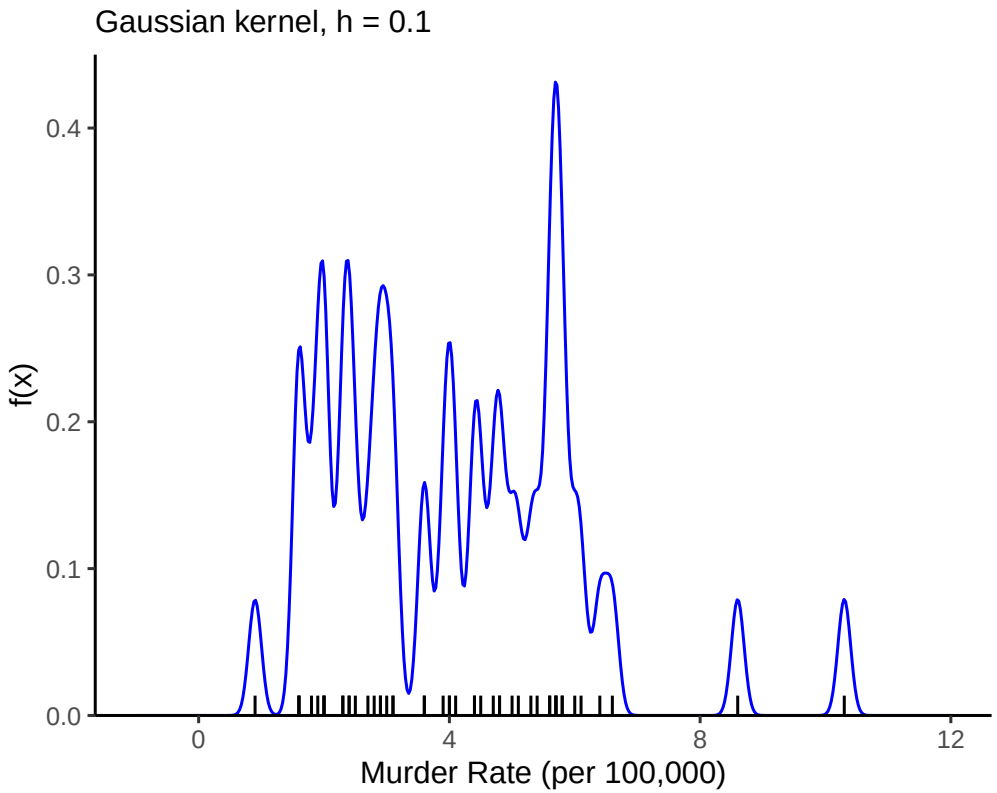
trong đó,

- $h > 0$ là bandwidth, được dùng để hiệu chỉnh độ trơn của hàm mật độ ước lượng;
- $K(\cdot)$ là hàm kernel, và là hàm dương.

Một số hàm kernel thường được sử dụng:

- Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-0.5u^2)$,
- Epanechnikov kernel: $K(u) = 0.75\left(1 - u^2\right)$ nếu $|u| \leq 1$.

Ước lượng hàm mật độ xác suất



Ước lượng hàm mật độ xác suất

Nhận xét:

- h càng nhỏ thì hàm mật độ ước lượng càng bị phân mảnh (quá chi tiết) - undersmooth;
- h càng lớn thì hàm mật độ ước lượng càng trơn - oversmooth.

Tìm h "tối ưu" là vấn đề then chốt trong ước lượng hàm mật độ.

Trong thống kê, có nhiều cách tìm h "tối ưu":

- rule-of-thumb
- unbiased cross validation
- biased cross validation
- Sheather & Jones method

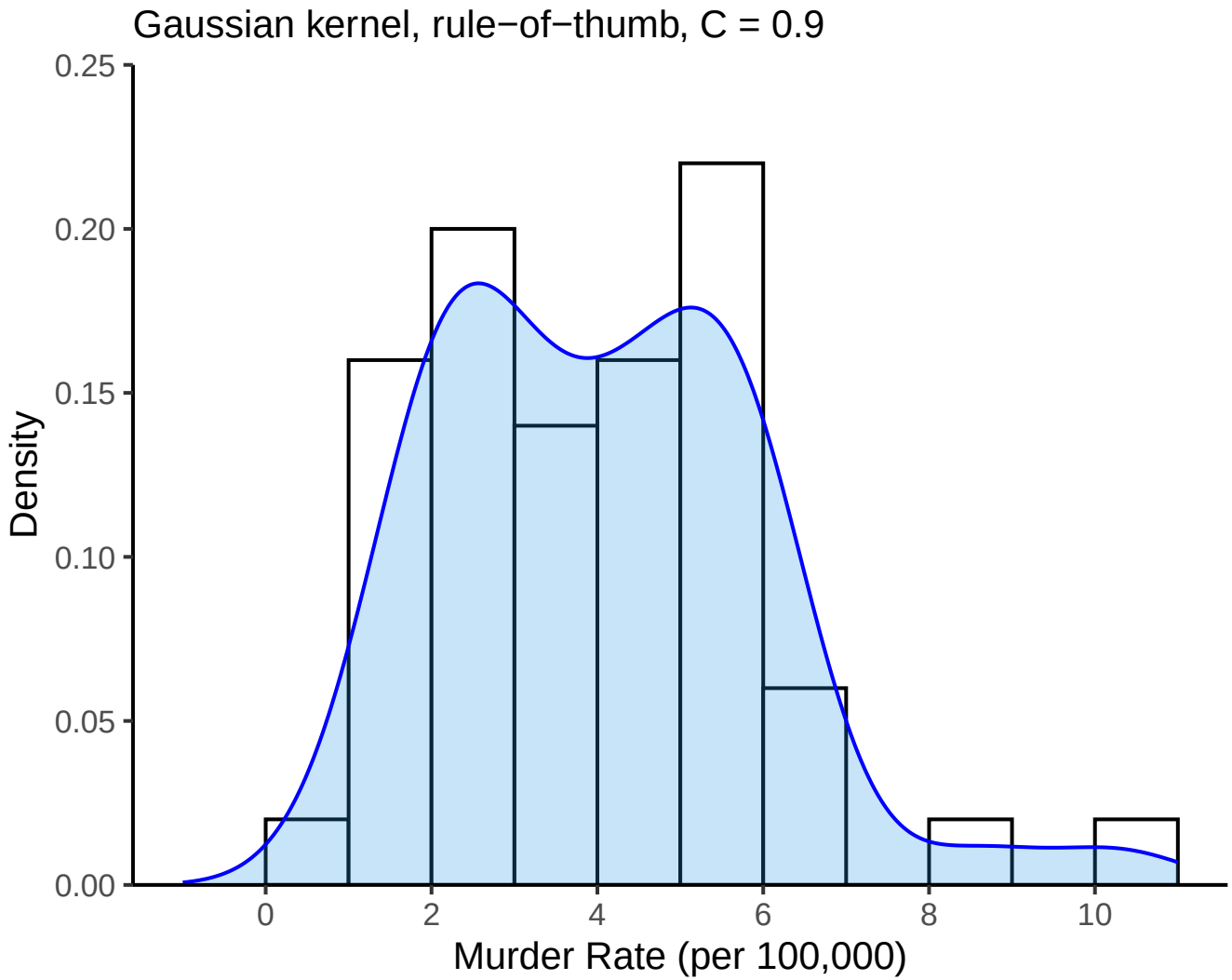
Trong đó, phương pháp rule-of-thumb được sử dụng rộng rãi nhất:

$$h_{opt} = C \times \min\{s, IQR/1.34\} \times n^{-0.2},$$

với $C = 0.9$ hoặc 1.06.

Thông thường, cách chọn h này sẽ đi kèm với Gaussian kernel.

Ước lượng hàm mật độ xác suất



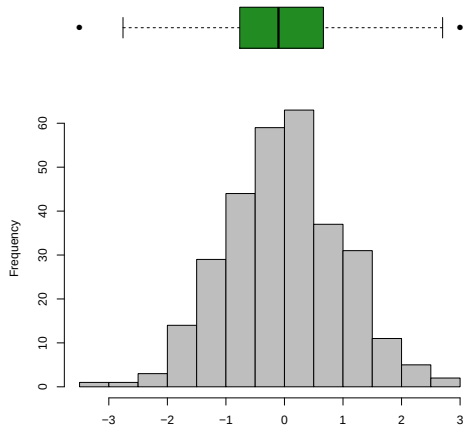
Nhận xét dạng điệu của phân phối

Các hình dạng phân bố thường gặp

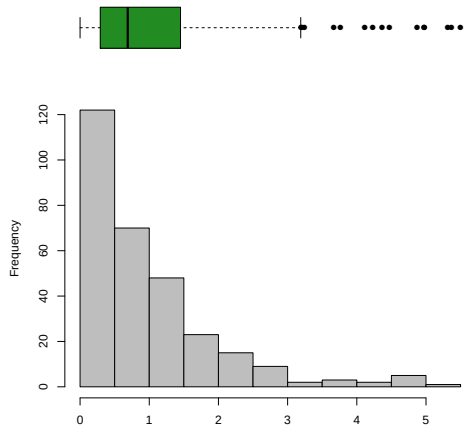
- Unimodal distribution: một đỉnh, hầu hết dữ liệu được phân bố xung quanh một giá trị;
- Bimodal (multimodal) distribution: hai (hoặc nhiều) đỉnh, dữ liệu được nhóm xung quanh hai (hoặc nhiều) đỉnh;
- Symmetric distribution: một đỉnh, phía bên trái của phân bố phản ánh phía bên phải.
- Positive or right-skewed distribution: đỉnh bên trái, đuôi dài bên phải, kéo dài (nghiêng) sang bên phải, một vài giá trị lớn, trung vị nhỏ hơn trung bình, nếu đơn hình thì mode nhỏ hơn trung vị nhỏ hơn trung bình;
- Negative of left-skewed distribution: đỉnh bên phải, đuôi dài bên trái, kéo dài (nghiêng) sang trái, một vài giá trị nhỏ, trung vị lớn hơn trung bình, nếu unimodal thì mode lớn hơn trung vị lớn hơn trung bình.

Nhận xét dạng điệu của phân phối

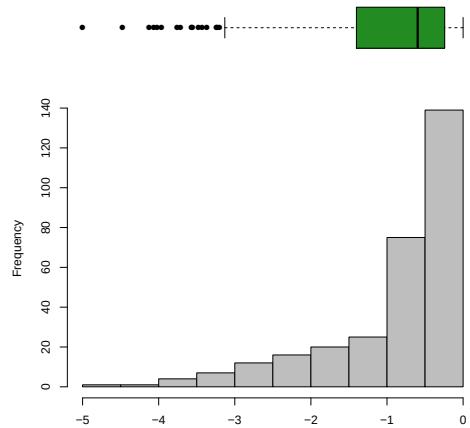
symmetric



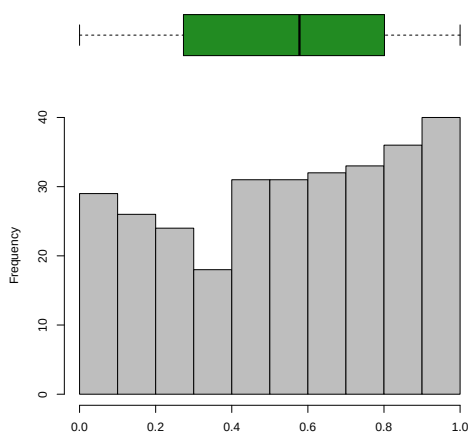
right-skewed



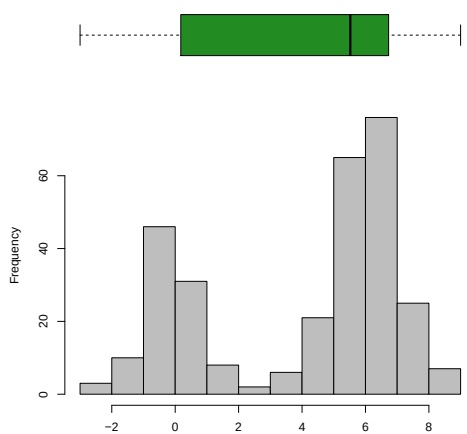
left-skewed



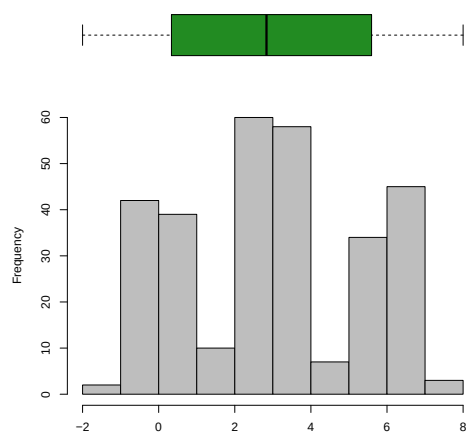
uniform



bimodal

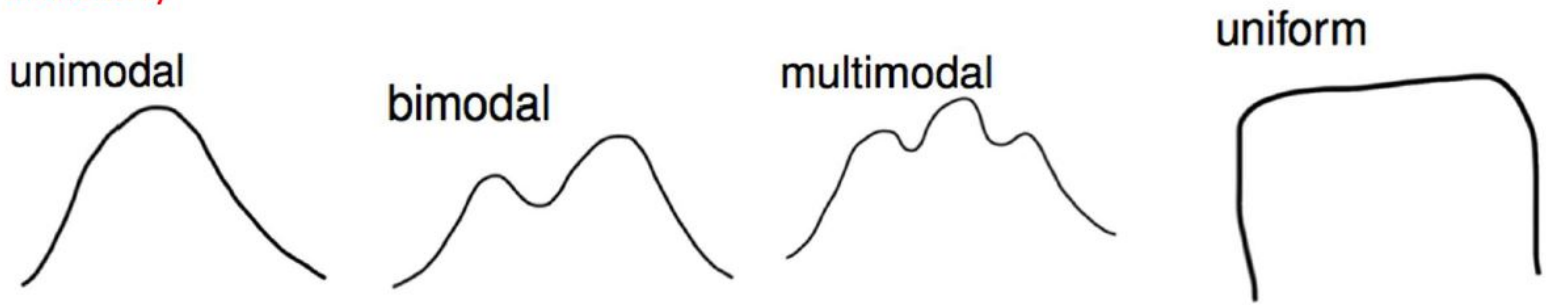


multimodal

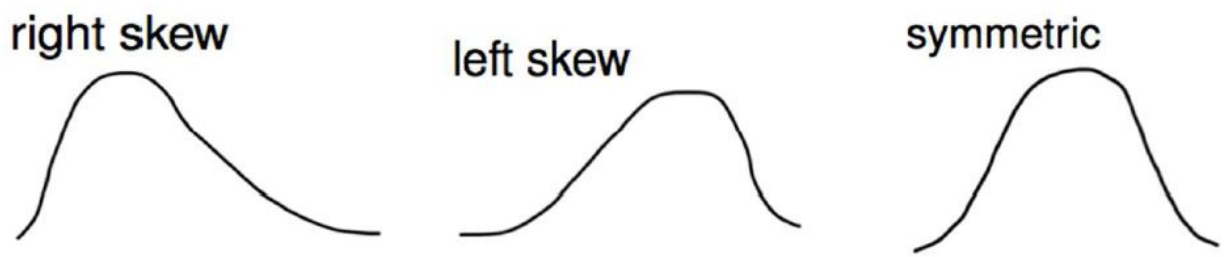


Nhận xét dạng điệu của phân phối

Modality



Skewness

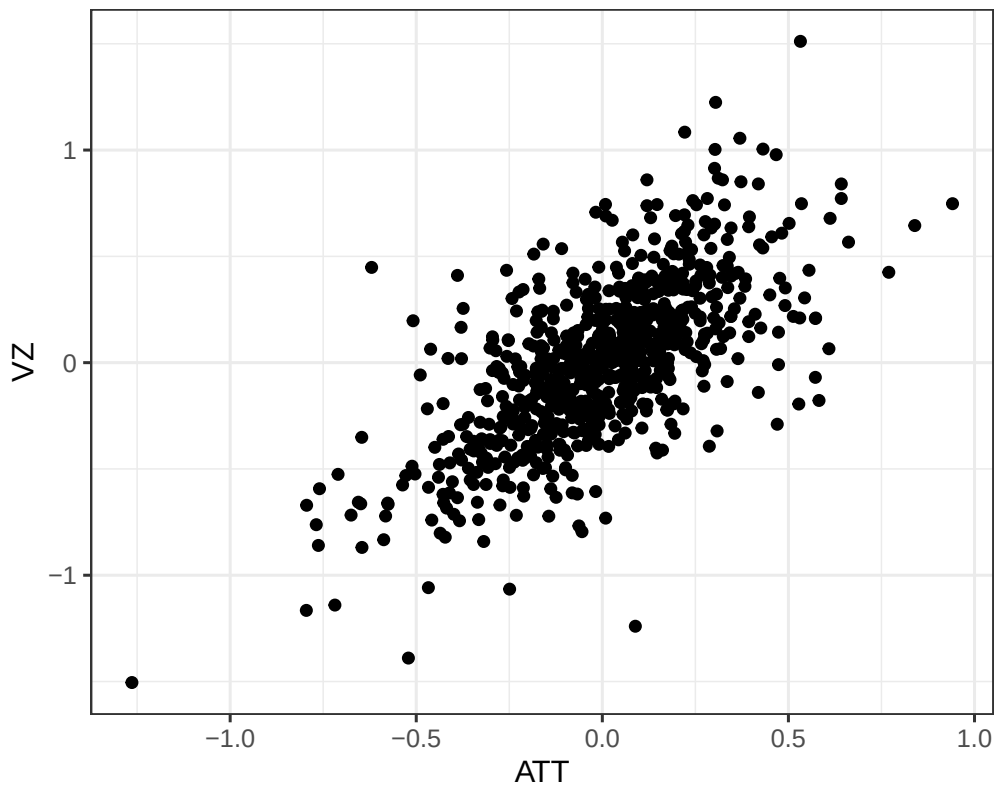


Sự tương quan - Correlation

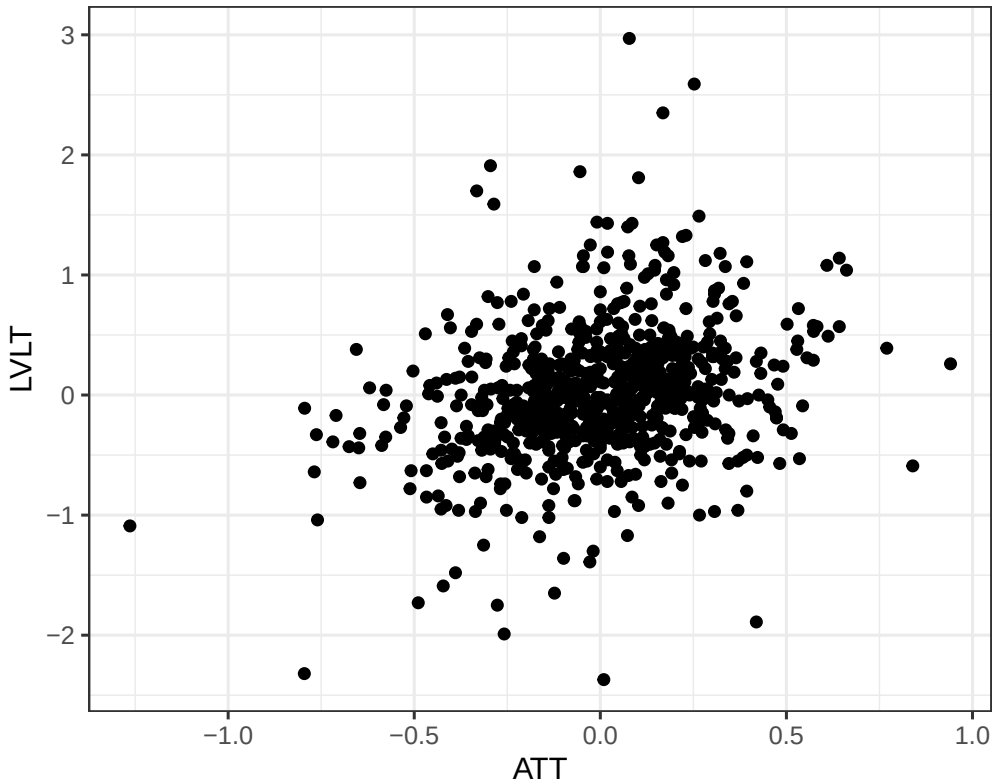
Sự tương quan - Correlation

Sự tương quan - Correlation là một khái niệm miêu tả mối liên hệ giữa hai biến dạng số (biến định lượng).

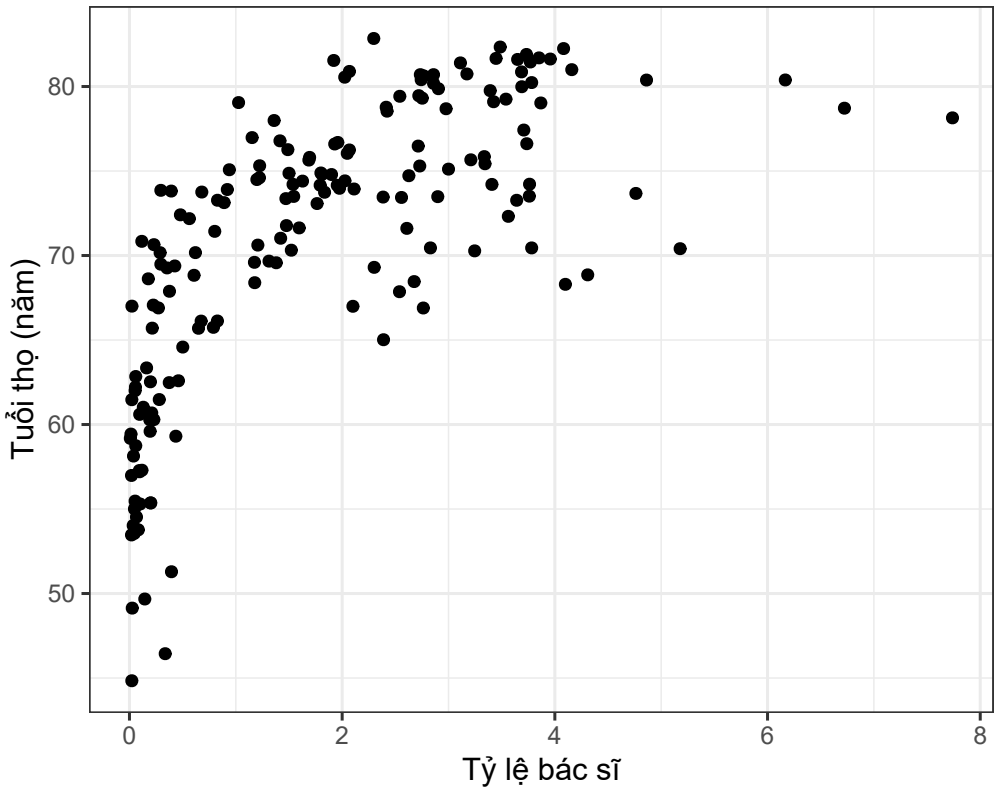
Để miêu tả mối liên hệ giữa hai biến dạng số, ta có thể dùng đồ thị phân tán (scatter plot).



Sự tương quan - Correlation



Sự tương quan - Correlation



Sự tương quan - Correlation

Ta có các dạng tương quan phổ biến sau:

- tương quan tuyến tính - linear correlation;
- tương quan phi tuyến - non-linear correlation.

Trong trường hợp phi tuyến

- tương quan đơn điệu - monotone correlation;
- tương quan phi đơn điệu - non-monotone correlation.

Sự tương quan - Correlation

Sự tương quan giữa hai biến định lượng, có thể được đo độ mạnh trong một số trường hợp cụ thể:

- tương quan tuyến tính,
- tương quan phi tuyến đơn điệu.

Thăng đo độ mạnh của sự tương quan, được gọi là **hệ số tương quan** - **correlation coefficient**.

Có nhiều hệ số tương quan khác nhau đã được phát triển, mỗi loại tương ứng cho một trường hợp cụ thể:

- hệ số tương quan tuyến tính Pearson - **Pearson correlation coefficient**, dành cho tương quan tuyến tính,
- hệ số tương quan hạng Spearman - **Spearman's correlation coefficient**, dành cho tương quan phi tuyến đơn điệu.

Trong khuôn khổ data science, hệ số tương quan tuyến tính Pearson được ưa chuộng sử dụng.

Hệ số tương quan tuyến tính

Hệ số tương quan tuyến tính Pearson

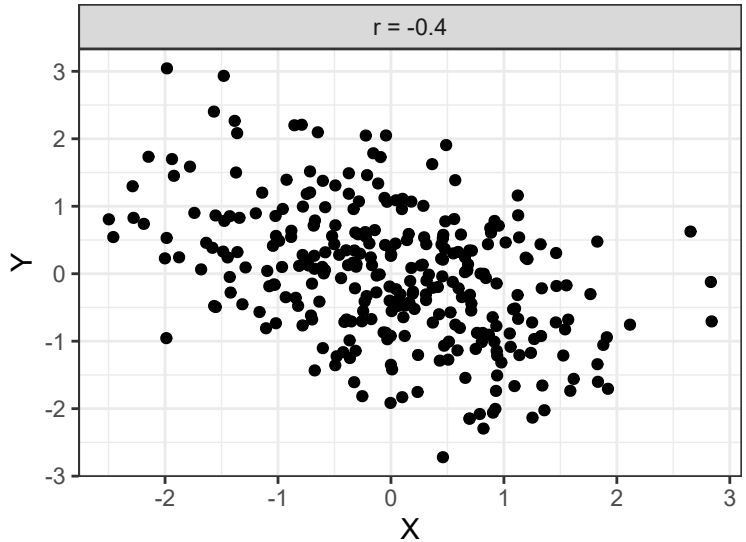
Hệ số tương quan tuyến tính Pearson - **Pearson correlation coefficient** là một hệ số dùng để đo độ mạnh của một tương quan tuyến tính giữa hai biến định lượng.

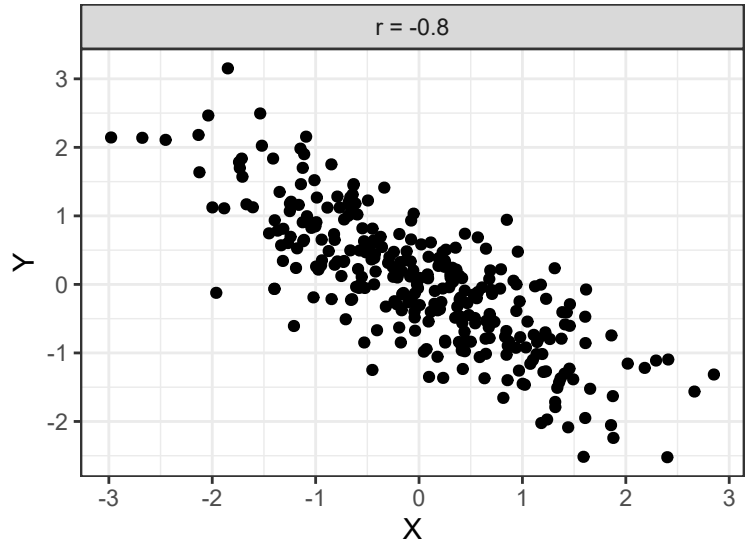
$$r = \frac{\mathbb{C}\mathrm{ov}(X, Y)}{\sqrt{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)}}.$$

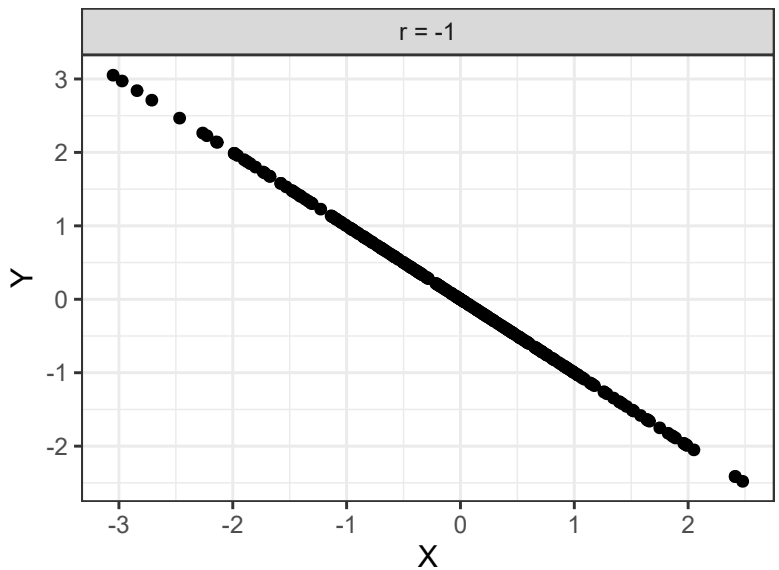
Hệ số r có giá trị nằm trong $[-1, 1]$:

- nếu $-1 < r < 0$: tương quan tuyến tính giữa X và Y là tương quan nghịch, tức là X tăng thì Y giảm;
- nếu $0 < r < 1$: tương quan tuyến tính giữa X và Y là tương quan thuận, tức là X tăng thì Y tăng;
- nếu $r = 0$: tương quan tuyến tính giữa X và Y là không tồn tại;
- nếu $r = -1$: tương quan tuyến tính giữa X và Y là tương quan nghịch ngặt;
- nếu $r = 1$: tương quan tuyến tính giữa X và Y là tương quan thuận ngặt;

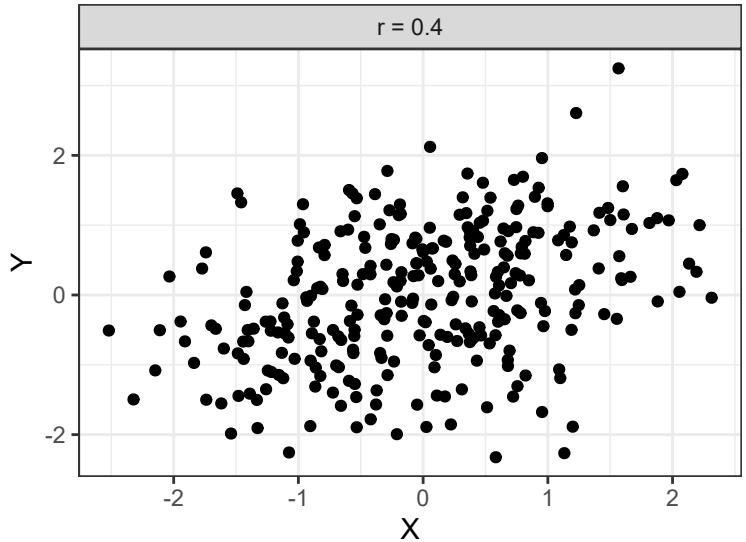
Hệ số tương quan tuyến tính

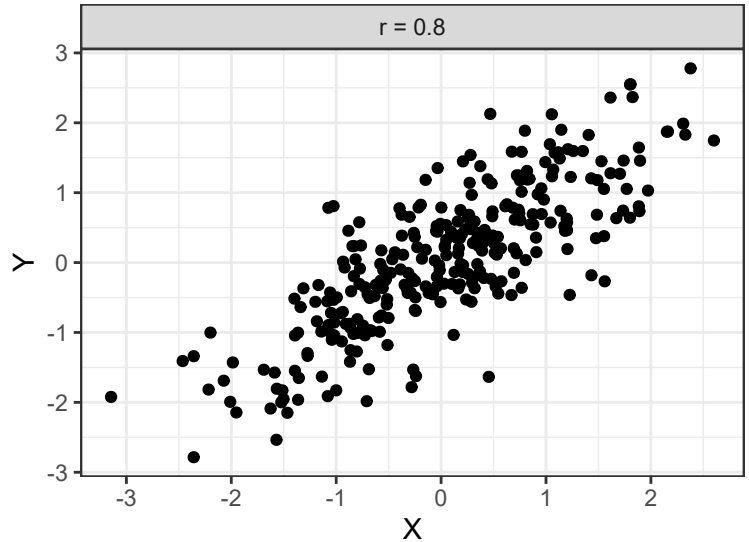


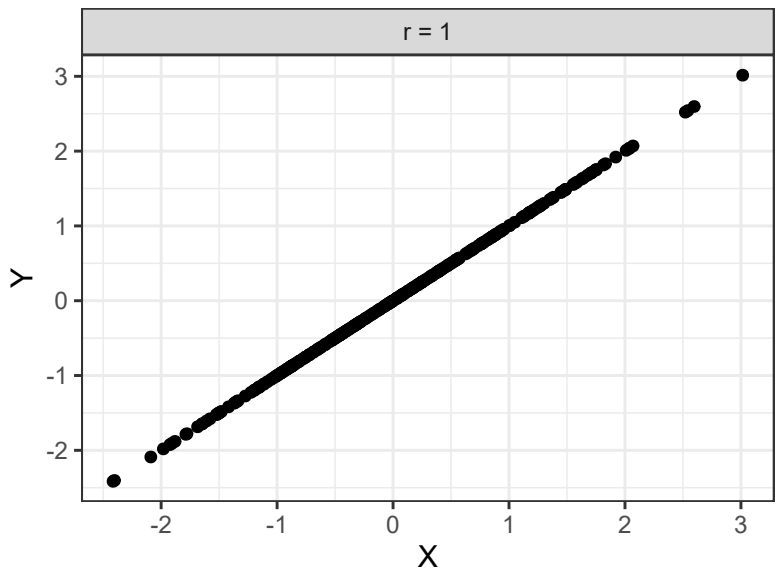




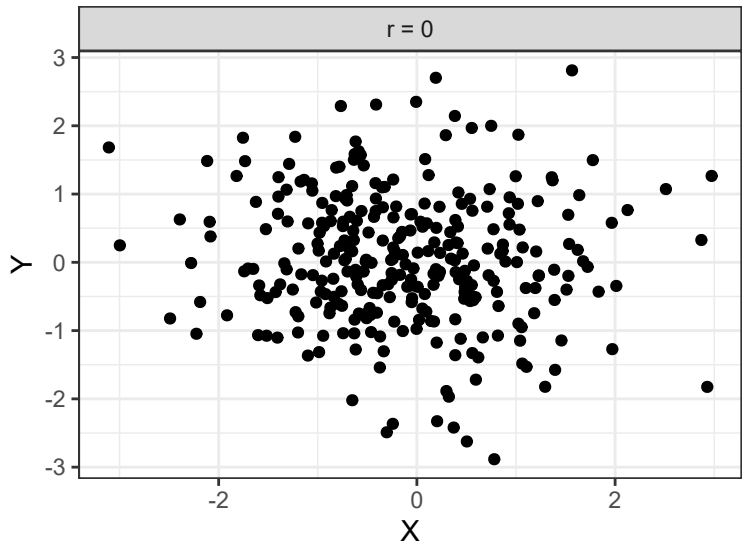
Hệ số tương quan tuyến tính

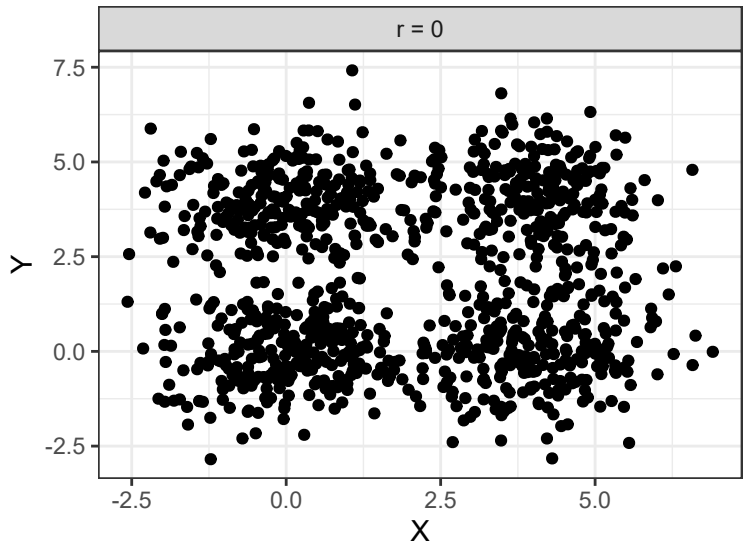


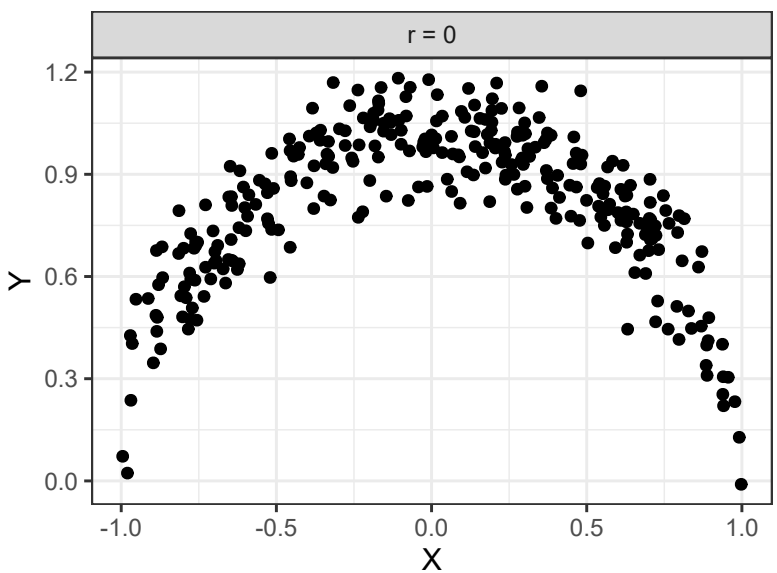




Hệ số tương quan tuyến tính







Ước lượng hệ số tương quan tuyến tính

Với dữ liệu quan sát $(x_1, y_1), \ldots, (x_n, y_n)$, ta có hệ số tương quan mẫu:

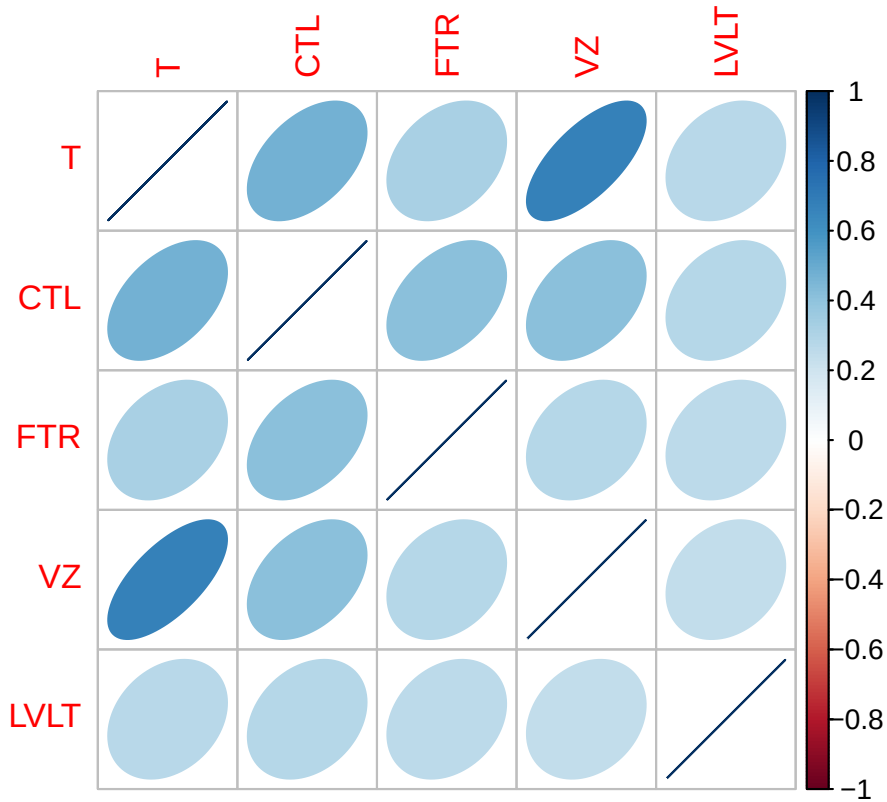
$$\widehat{r} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right).$$

Bảng dưới đây cung cấp **ma trận tương quan (correlation matrix)** giữa lợi nhuận hàng ngày của các cổ phiếu viễn thông từ tháng 7 năm 2012 đến tháng 6 năm 2015.

	ATT	CTL	FTR	VZ	LVLT
ATT	1.000	0.475	0.328	0.678	0.279
CTL	0.475	1.000	0.420	0.417	0.287
FTR	0.328	0.420	1.000	0.287	0.260
VZ	0.678	0.417	0.287	1.000	0.242
LVLT	0.279	0.287	0.260	0.242	1.000

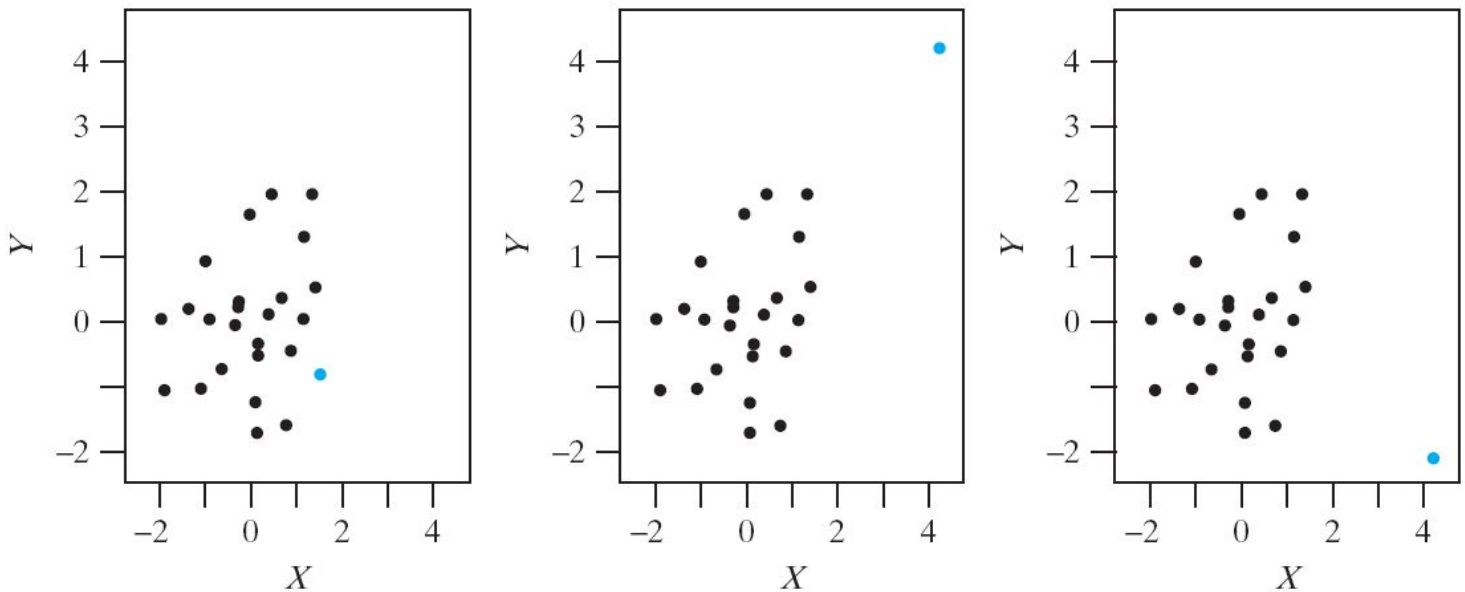
Ước lượng hệ số tương quan tuyến tính

Hoặc ta có thể biểu diễn sự tương quan tuyến tính giữa các biến bằng biểu đồ tương quan như sau:



Hệ số tương quan và giá trị ngoại lai

Hệ số tương quan rất nhạy cảm với các điểm cực trị.



(a) $r = 0.2$ (b) $r = 0.6$ (c) $r = -0.1$

Ba biểu đồ này minh họa cách một điểm có thể ảnh hưởng lớn đến độ lớn của hệ số tương quan tuyến tính.

↪ Điều quan trọng là luôn vẽ biểu đồ dữ liệu trước khi sử dụng r (hoặc bất kỳ số liệu thống kê nào khác) để tóm tắt dữ liệu.