Data Analysis Big Picture

Quách Đình Hoàng

2022/03/14

Nội dung

Data analysis

Data analysis and related terms

Thống kê (statistics), khai phá dữ liệu (data mining) và học máy (machine learning) đều liên quan đến việc thu thập và phân tích dữ liệu (data analysis).

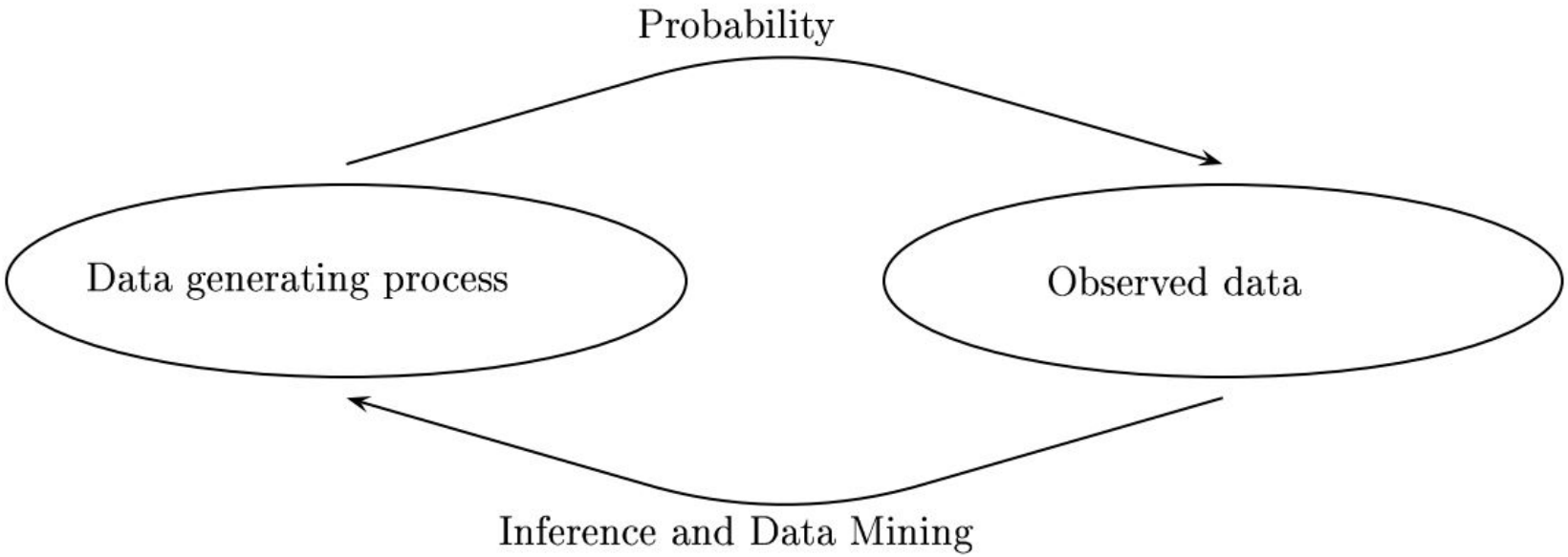
Phân tích dữ liệu, học máy và khai phá dữ liệu là những tên gọi khác nhau được đặt cho việc thực hành suy luận thống kê, tùy thuộc vào ngữ cảnh.

Prediction, classification, clustering, và estimation đều là những trường hợp đặc biệt của suy luận thống kê (statistical inference), tùy thuộc vào ngữ cảnh.

Source: Larry Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, 2004.

Probability and Inference

Xác suất và thống kê là nền tảng cho phân tích dữ liệu.



Source: Larry Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, 2004.

Data analysis big picture

Data analysis big picture

- Mục đích của phân tích dữ liệu là biến dữ liệu thành thông tin có ích.
- Quá trình này liên quan đến:
  - Thành lập câu hỏi nghiên cứu (identify a question),
  - Thu thập dữ liệu liên quan (collecting relevant data),
  - Phân tích dữ liệu (analyze the data), và
  - Diễn giải kết quả (interprete the results).

Thành lập câu hỏi nghiên cứu (identify a question)

- Quá trình phân tích dữ liệu bắt đầu với câu hỏi ta muốn biết gì?
- Ở bước này, ta cần xác định tất cả các đối tượng ta quan tâm, gọi là quần thể quần thể (population).
- Ví dụ: ta muốn biết ý kiến của người dân Hoa Kỳ (trưởng thành) về án tử hình
  - Quần thể (population) ở đây là ý kiến của tất cả người dân Hoa Kỳ (trưởng thành) về án tử hình

Thu thập dữ liệu liên quan (collecting relevant data)

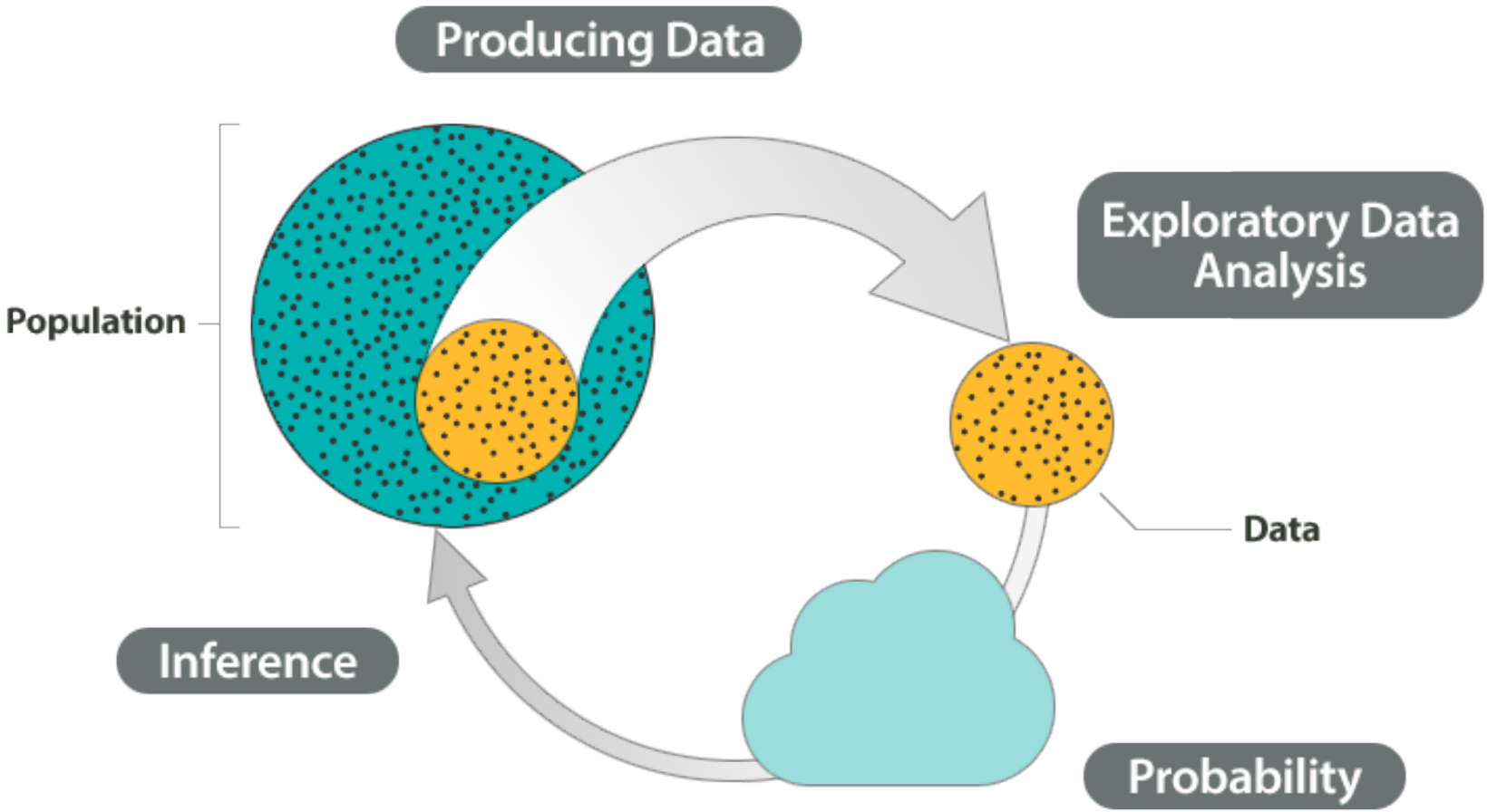
- Tuy nhiên, ta thường chỉ thu thập được một tập con của quần thể, gọi là mẫu (sample) để phân tích. Ta gọi quá trình này là chọn/lấy mẫu (sampling) hay sinh dữ liệu (producing data).
  - Để kết quả phân tích có ý nghĩa, mẫu được chọn nên là một đại diện tốt cho quần thể.
- Thay vì ý kiến của tất cả người dân Hoa Kỳ (trưởng thành) về án tử hình ta chỉ thu thập một mẫu ngẫu nhiên các ý kiến từ quần thể để phân tích.
  - Ngoài ra, chúng ta không muốn mẫu của chúng ta chỉ bao gồm ý kiến từ đảng Cộng hòa hoặc chỉ đảng Dân chủ.

Phân tích dữ liệu (analyze the data)

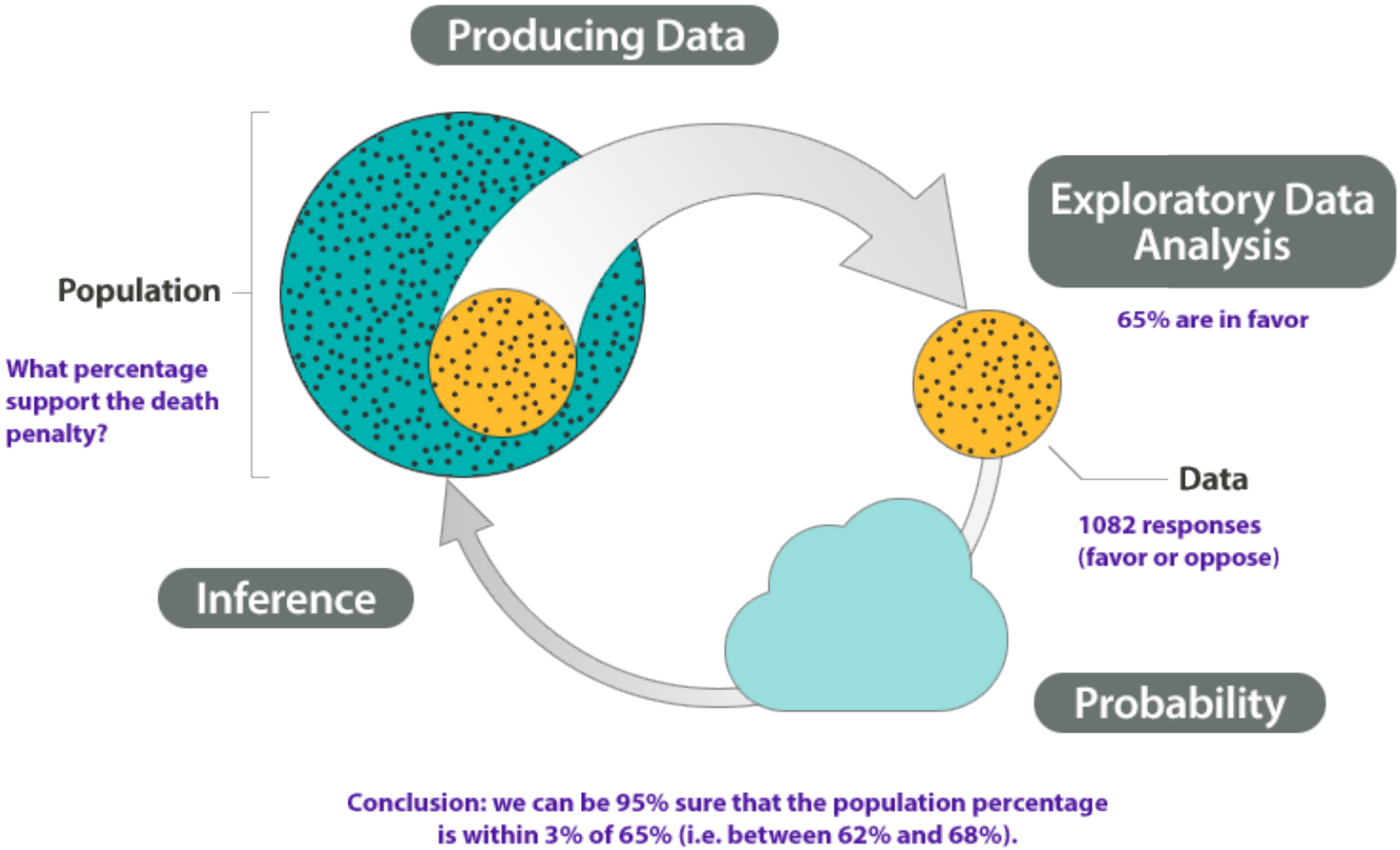
- Sau khi có dữ liệu, ta thường muốn tóm tắt và trực quan hóa chúng để có cái nhìn tổng quan.Việc này được gọi là phân tích dữ liệu thăm dò (exploratory data analysis).
- Tuy nhiên, mục đích của ta là hiểu về đặc tính của quần thể hơn là của mẫu ta đã thu thập.
- Để có thể làm như vậy, chúng ta cần xem xét mẫu mà chúng ta đang sử dụng có thể khác với quần thể như thế nào để có thể đưa điều đó vào phân tích của mình.
- Để hiểu được sự khác biệt giữa quần thể và mẫu ta cần sử dụng xác suất (probability).

Diễn giải kết quả (interpret the results)

- Ở bước này, ta muốn đưa ra kết luận về các đặc tính của quần thể dựa vào kết quả phân tích trên mẫu. Quá trình này được gọi là suy diễn (inference).
  - Xác suất là công cụ quan trọng để ta có thể thực hiện việc này.



Data analysis big picture example



Source: Probability & Statistics, http://oli.cmu.edu

Tham khảo

1. Larry Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, 2004.
2. Probability & Statistics. Provided by: Open Learning Initiative. Located at: http://oli.cmu.edu. License: CC BY: Attribution

Exploration Data Analysis (EDA)

Quách Đình Hoàng

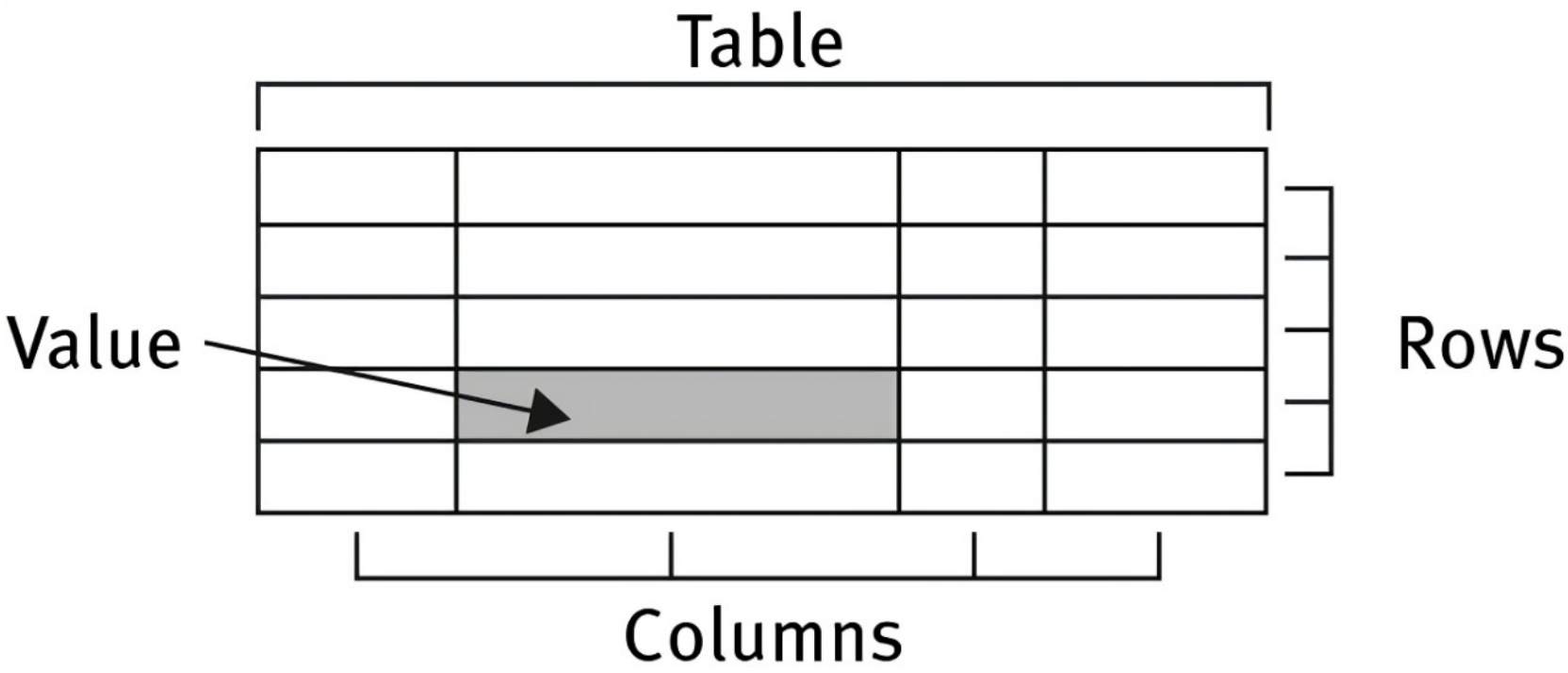
2022/03/14

Nội dung

Data

Dữ liệu và biến

- Dữ liệu (data) thường được biểu diễn ở dạng bảng (table), mỗi dòng (row) là một đối tượng (object), mỗi cột (column) là một biến (variable) của đối tượng tương ứng.
  - Biến (variable) còn được gọi là đặc trưng (feature) hay thuộc tính (atribute)



Các loại biến

Biến thường được chia làm hai loại: biến phân loại và biến số.

- Biến phân loại/định tính (category/qualitative variable), gồm:
  - Nominal: mô tả trạng thái hoặc tên gọi
    - Ví dụ: màu sắc, mã số, tình trạng hôn nhân
  - Ordinal: là nominal nhưng có thêm thứ tự
    - Ví dụ: xếp hạng, kích cỡ (lớn, trung, nhỏ)
- Biến số/định lượng (numeric/quantitative variable), gồm:
  - Interval: không có giá trị 0 thật sự (no true zero-point)
    - Ví dụ: ngày tháng, nhiệt độ C hoặc F, IQ.
  - Ratio: có giá trị 0 thật sự (inherent zero-point)
    - Ví dụ nhiệt độ K (Kelvin), chiều cao, cân nặng

Các loại biến

Loại biến sẽ qui định các phép toán mà ta có thể thực hiện

- Nominal: $=, \neq$
- Ordinal: $=, \neq, <, >$
- Intever: $=, \neq, <, >, +, -$
- Ratio: $=, \neq, <, >, +, -, \times, /$

Khi ta thực hiện những phép toán không phù hợp trên biến, kết quả sẽ không có ý nghĩa.

Phân tích thăm dò (EDA)

- Phân tích từng biến
  - Biến phân loại
  - Biến số
- Phân tích mối quan hệ giữa hai biến
  - Biến phân loại với biến số
  - Hai biến số
  - Hai biến phân loại

Phân tích một biến phân loại

Biến phân loại

- Bước đầu tiên trong EDA là tóm tắt dữ liệu và xác định phân bố (distribution) của dữ liệu.
- Phân bố của dữ liệu cho ta biết hai thông tin quan trọng:
  - Những giá trị mà một biến nhận
  - Những giá trị đó xuất hiện thường xuyên đến mức độ nào

Tần số

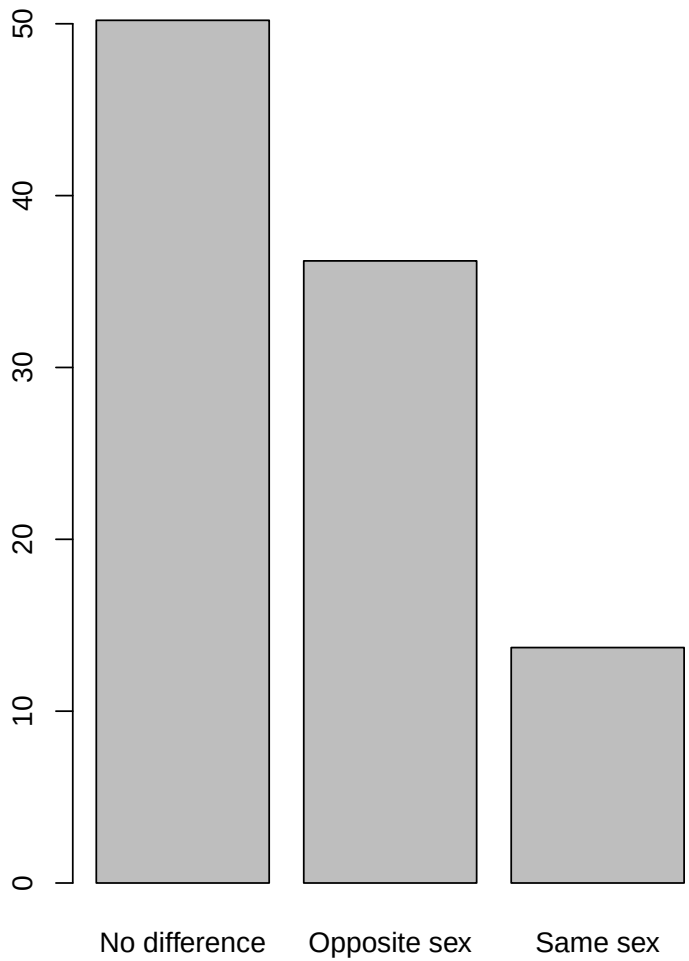
- Việc tóm tắt và nhìn vào phân bố của dữ liệu có thể giúp ta rút ra được các thông tin hữu ích
  - Chỉ nhìn vào tập các giá trị thường không giúp ta rút ra được các thông tin hữu ích
- Để tóm tắt một biến phân loại, ta thường dùng bảng phân bố tần số xuất hiện (frequency distribution)

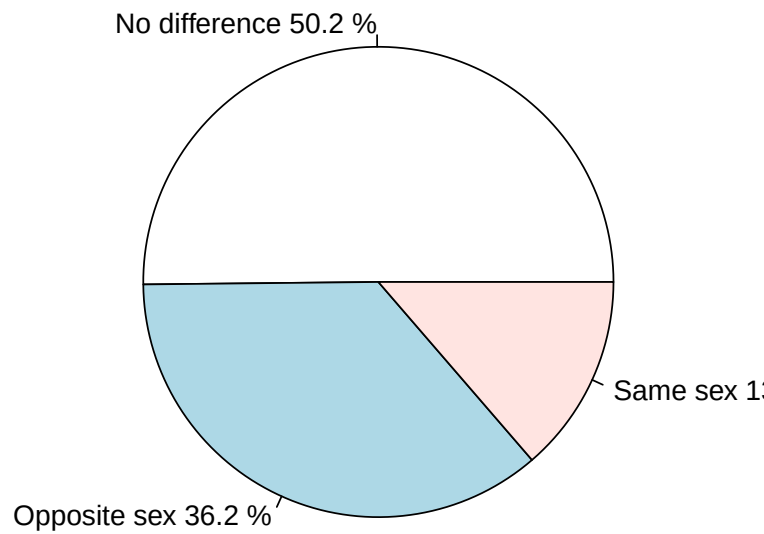
```
##
## No difference  Opposite sex       Same sex
##           602            434            164

##
## No difference  Opposite sex       Same sex
##            50.2           36.2           13.7
```

Pie chart và bar chart

- Pie chart và bar chart được dùng để trực quan hóa tóm tắt dạng số của biến phân loại





Phân tích một biến số

Biến số

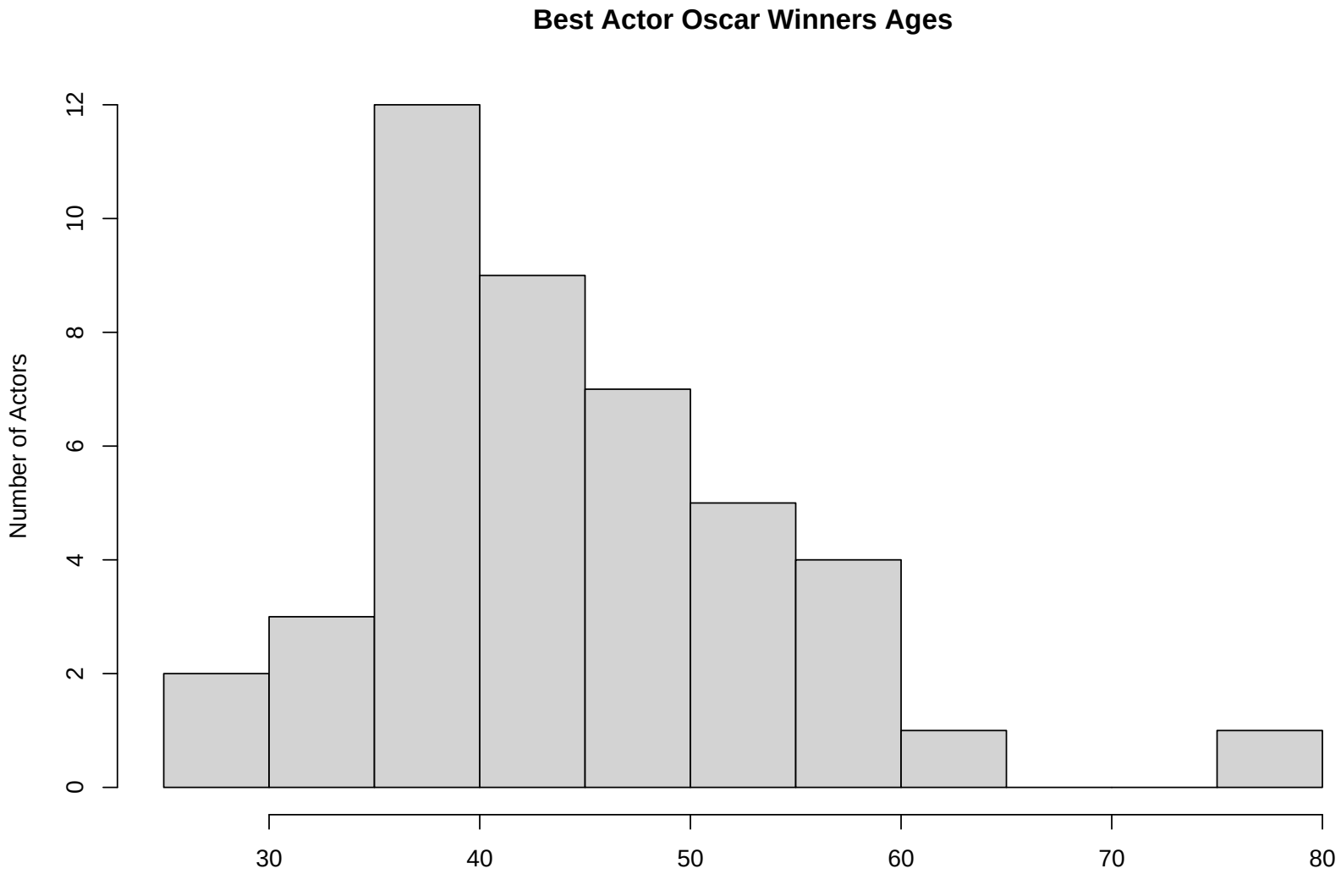
- Phân bố của biến số cung cấp cho ta các đặc trưng quan trọng:
  - Hình dạng (shape),
  - Khuynh hướng tập trung (center tendency), và
  - Sự phân tán (spead) của dữ liệu.
- Từ các thông tin trên có thể giúp ta suy ra các giá trị ngoại lệ (outlier) của dữ liệu.

Tóm tắt biến số

- Để tóm tắt biến số ta có thể dùng biểu đồ hoặc các giá trị số.
- Các biểu đồ phổ biến để trực quan hóa biến số là:
  - Biểu đồ tần số (histogram), biểu đồ thân cây (stemplot), và biểu đồ hộp (boxplot)
  - Hình dạng (shape) của biểu đồ giúp ta mô tả độ lệch (skewness) như và dạng thức (modality) của dữ liệu
    - Skewness: skewed right, skewed left, symmetric
    - Modality: unimodal, bimodal, multimodal, uniform
- Các giá trị số để tóm tắt biến số là các giá trị đo khuynh hướng tập trung (center tendency), mức độ phân tán (spread), và các ngoại lệ (outlier)
  - Center tendency: mean, median, mode
  - Spread: variance, range, inter-quartile range (IQR)
  - Outlier: các giá trị lớn hoặc nhỏ bất thường

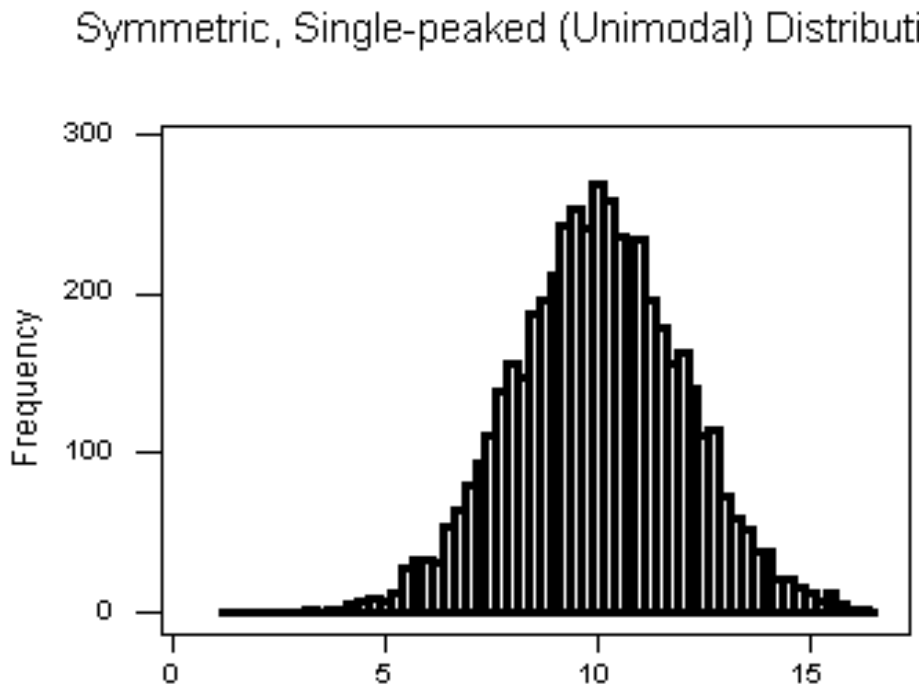
Biểu đồ tần số

```
##  [1] 43 42 48 49 56 38 60 30 40 42 37 76 39 53 45 36 62
## [26] 32 45 60 46 40 36 47 29 43 37 38 45 50 48 60 50 39
```

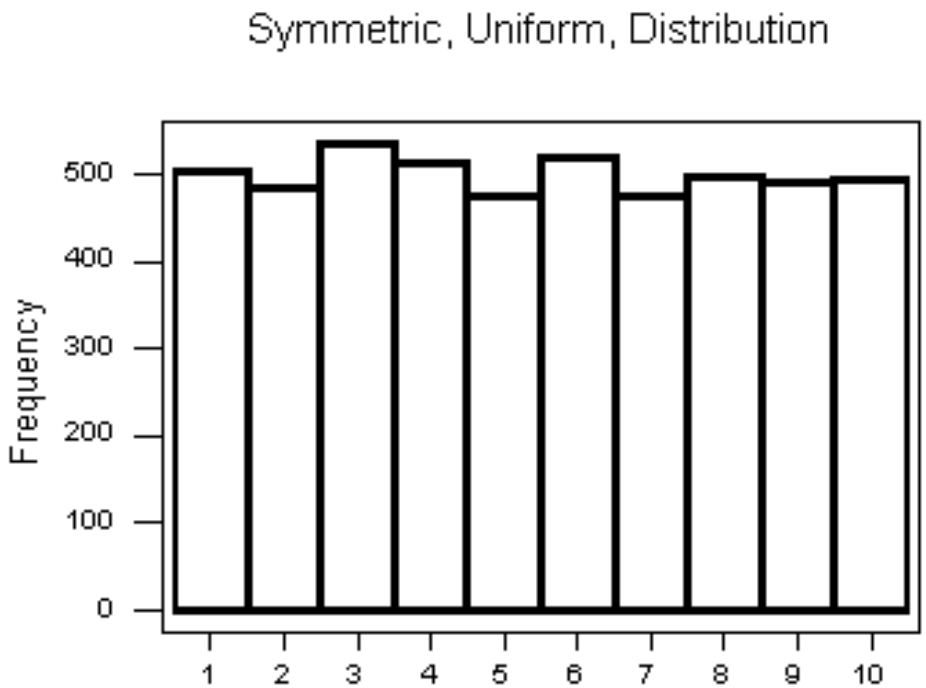


Phân bố đối xứng: symmetric, unimodal

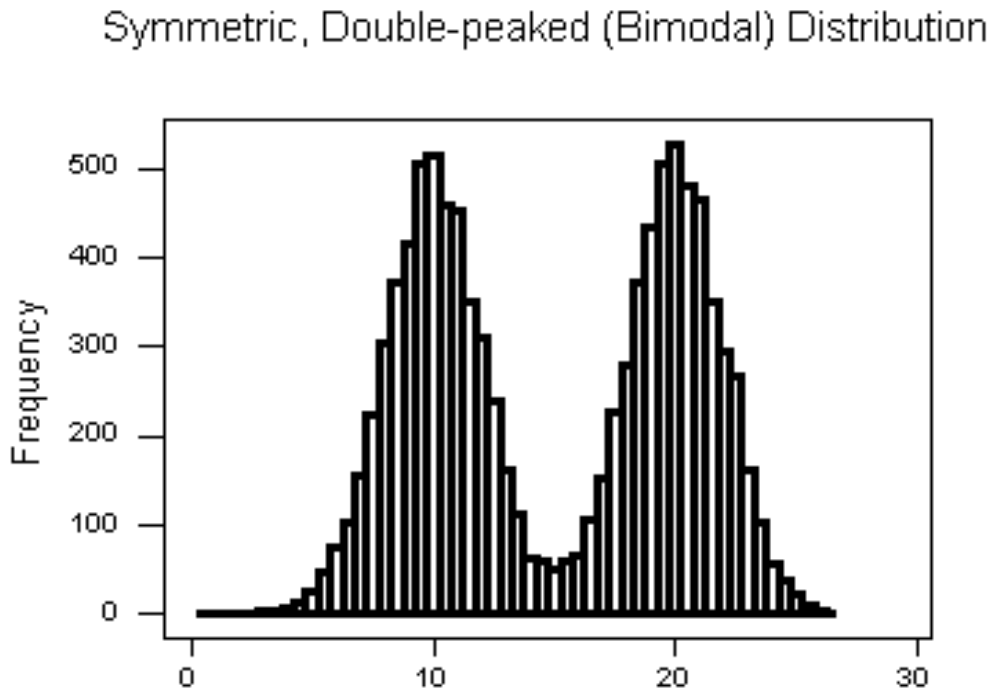
- Hình dạng (shape) của phân bố giúp ta mô tả độ lệch (skewness) như và dạng thức (modality) của dữ liệu



Phân bố đối xứng: symmetric, uniform

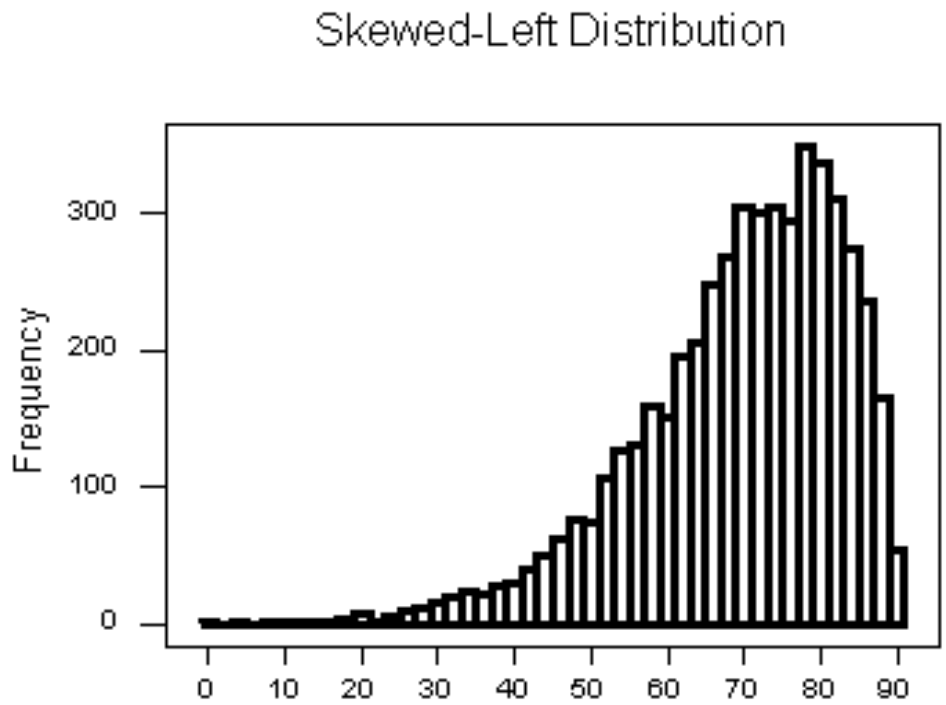


Phân bố đối xứng: symmetric, bimodal

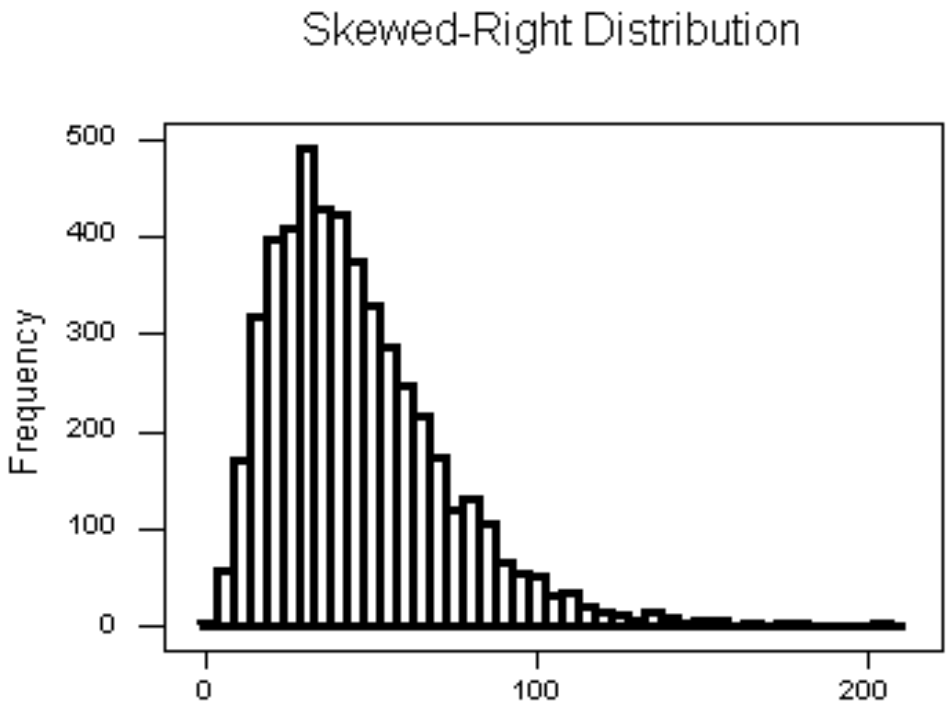


- Nếu dữ liệu có nhiều hơn hai mode, ta nói phân bố của nó là multimodal.

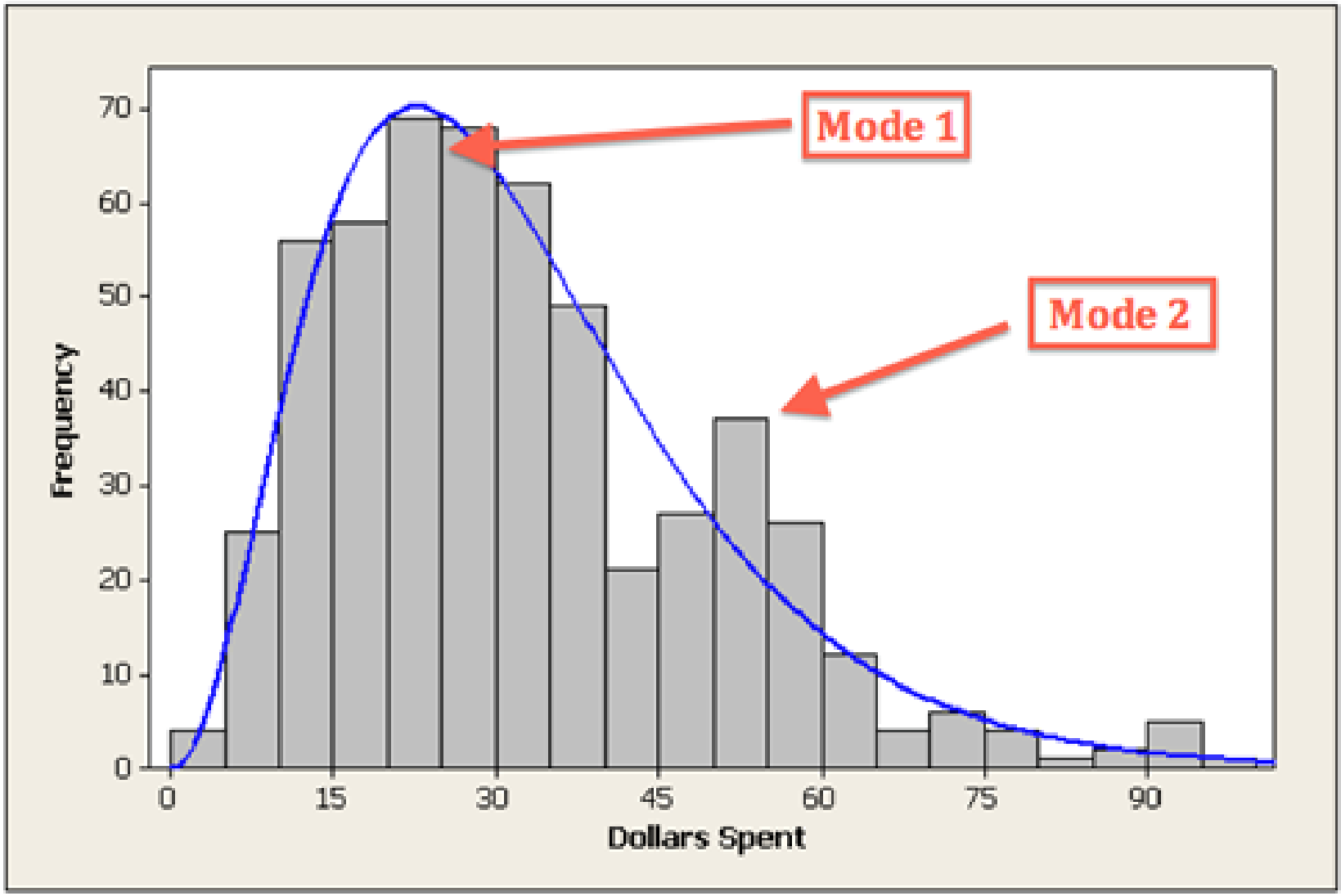
Phân bố lệch: skewed left



Phân bố lệch: skewed right



Phân bố lệch: skewed right, bimodel



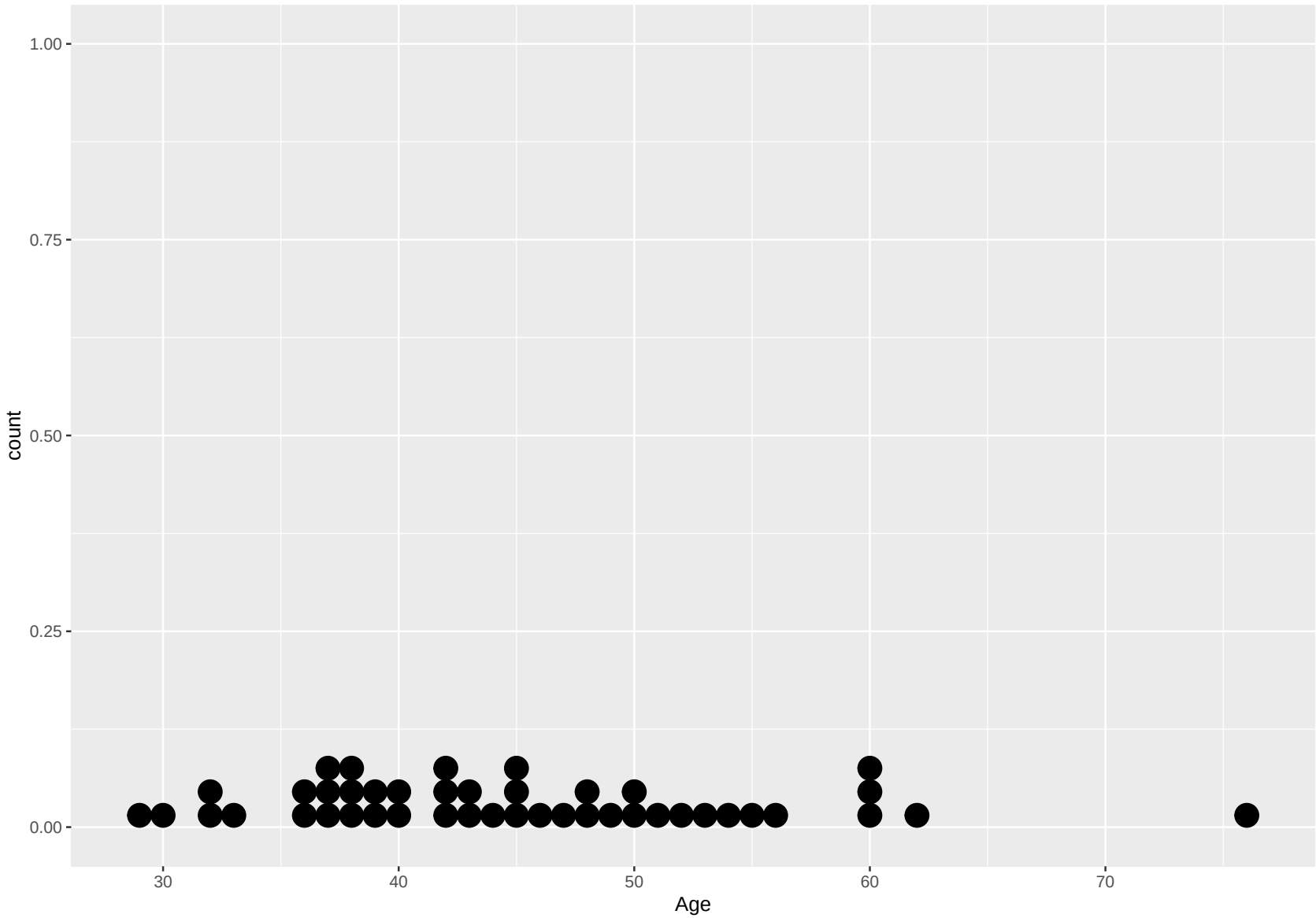
Biểu đồ stemplot

- Mỗi giá trị được phân thành stem và leaf như sau:
  - Leaf là chữ số bên phải nhất (right-most digit)
  - Stem là các số còn lại ngoại trừ chữ số bên phải nhất.

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   2 | 9
##   3 | 0223
##   3 | 6677788899
##   4 | 00222334
##   4 | 55567889
##   5 | 001234
##   5 | 56
##   6 | 0002
##   6 |
##   7 |
##   7 | 6
```

Biểu đồ dotplot

- Mỗi đối tượng là một dot.



Tóm tắt biến số bằng các giá trị số

- Phân bố của biến số giúp ta xác định được các thông tin quan trọng sau:
  - Hình dạng
  - Giá trị trung tâm
  - Mức độ phân tán
- Biểu đồ có thể cho ta thấy hình dạng của phân bố nhưng giá trị trung tâm và mức độ phân tán không thể hiện rõ lắm.

Các giá trị trung tâm

- Ba giá trị đo trung tâm của một phân bố là mean, median, và mode.
  - Mean: là giá trị được tính theo công thức sau:

$$mean(X) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - Mode: là giá trị xuất hiện nhiều lần nhất trong phân bố
  - Median: là giá trị nằm chính giữa phân bố.

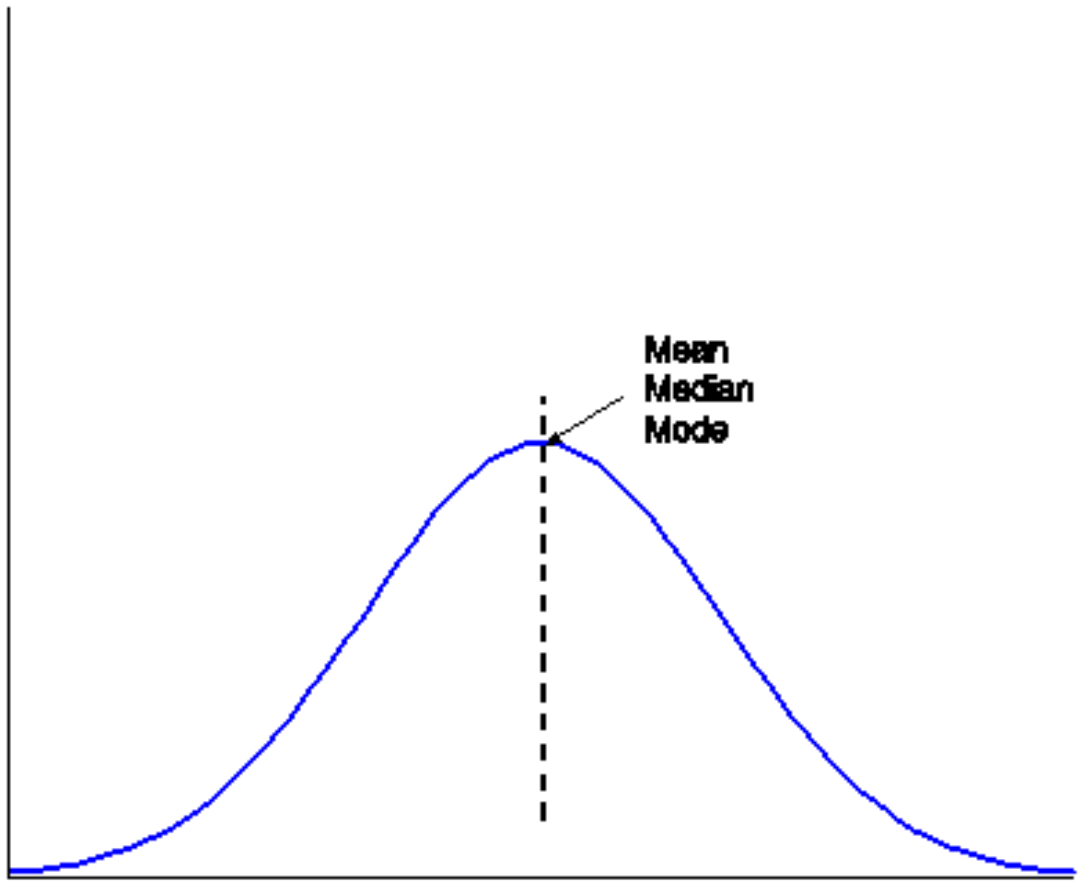
```
##  [1] 43 42 48 49 56 38 60 30 40 42 37 76 39 53 45 36 62
## [26] 32 45 60 46 40 36 47 29 43 37 38 45 50 48 60 50 39
```

```
## [1] "mean = 45"
```

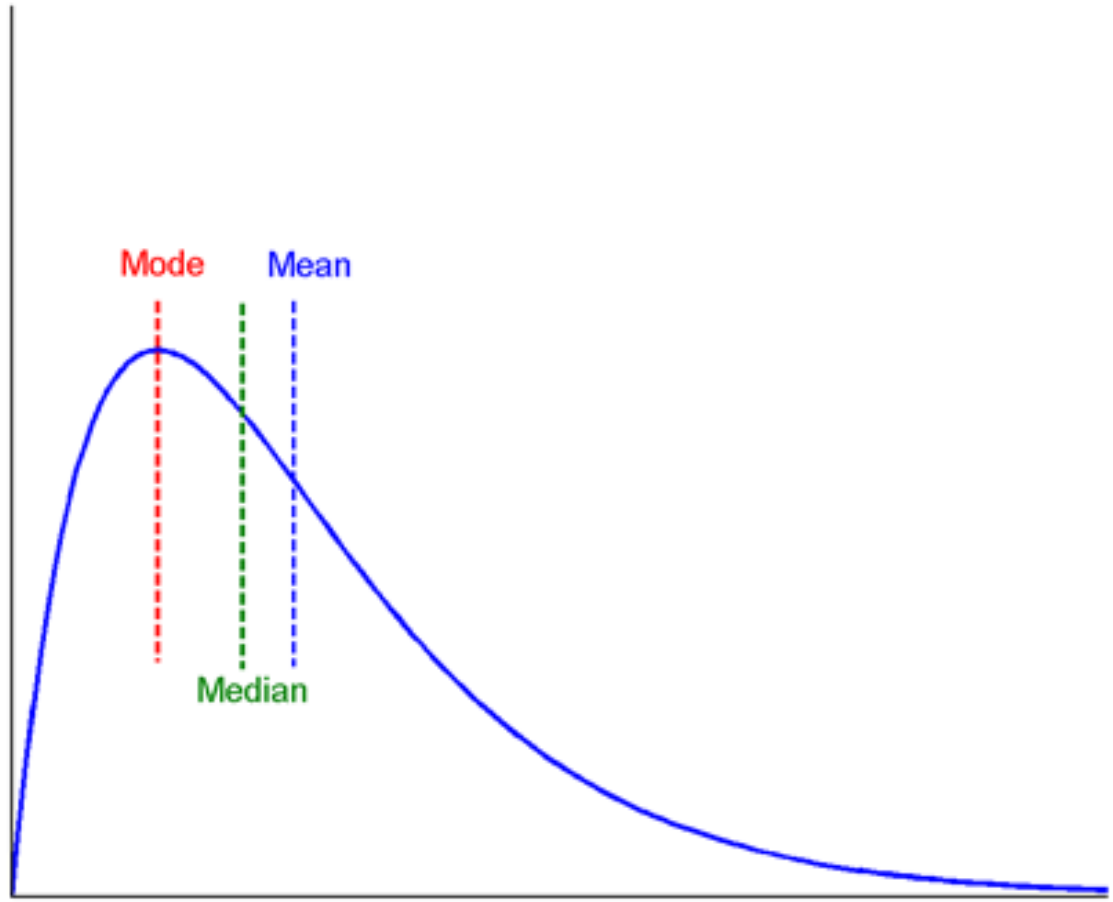
```
## [1] "modes:  42, 38, 60, 37, 45"
```

```
## [1] "median = 43.5"
```

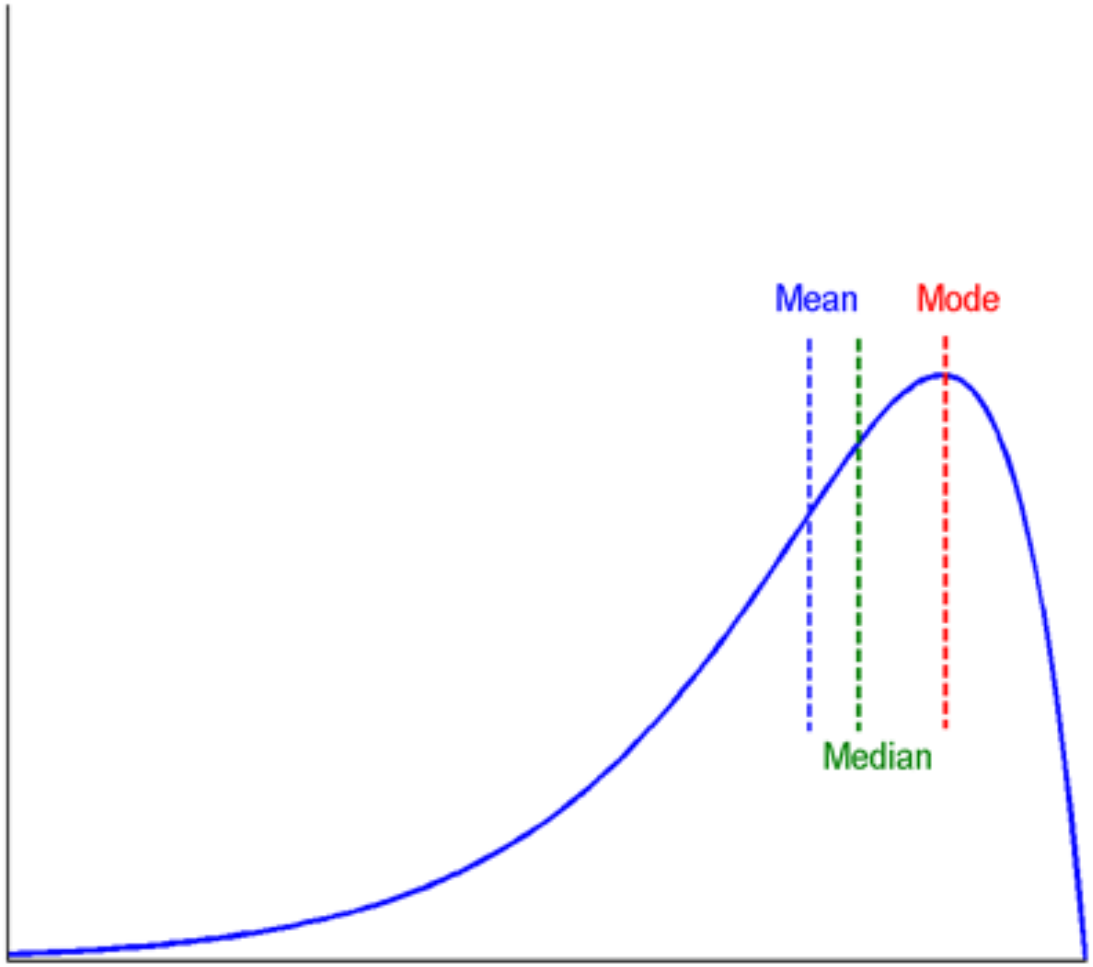
So sánh mean, mode, median: symmetric distribution



So sánh mean, mode, median: skewed right distribution

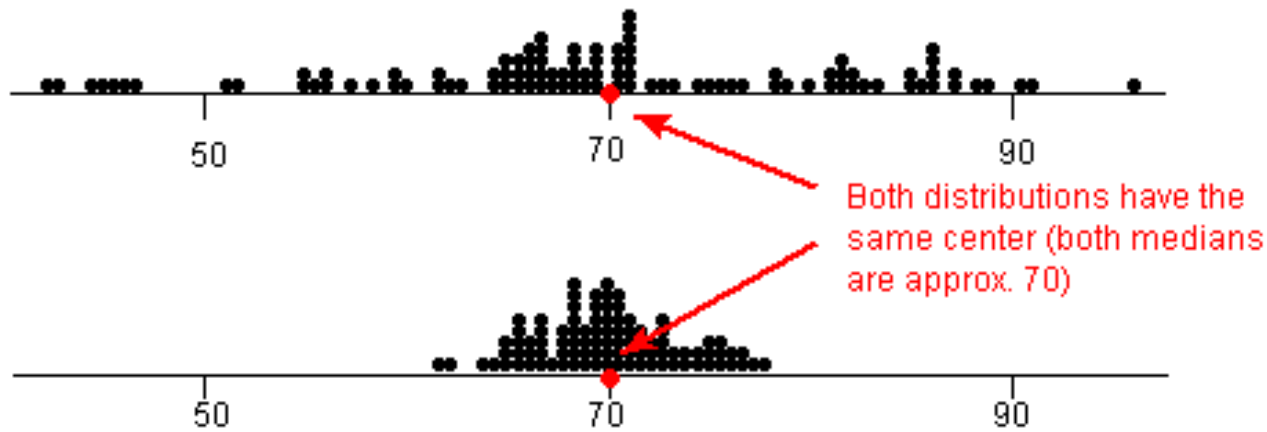


So sánh mean, mode, median: skewed left distribution



Mức độ phân tán

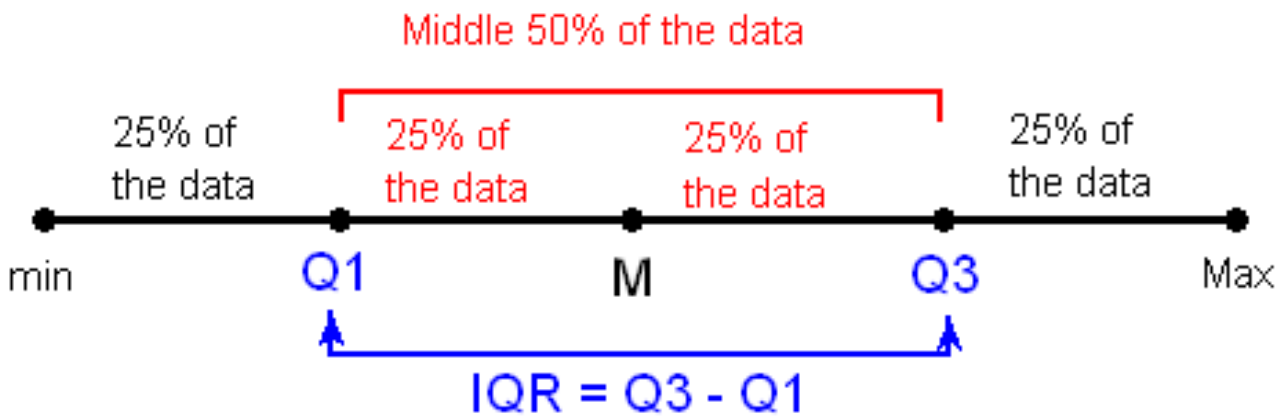
- Các giá trị trung tâm không đủ để đại diện cho một phân bố
  - Hai phân bố khác nhau có thể có các giá trị trung tâm giống nhau



- Các giá trị đo mức độ phân tán phổ biến là:
  - Variance, standard deviation
  - Range, inter-quartile range (IQR)

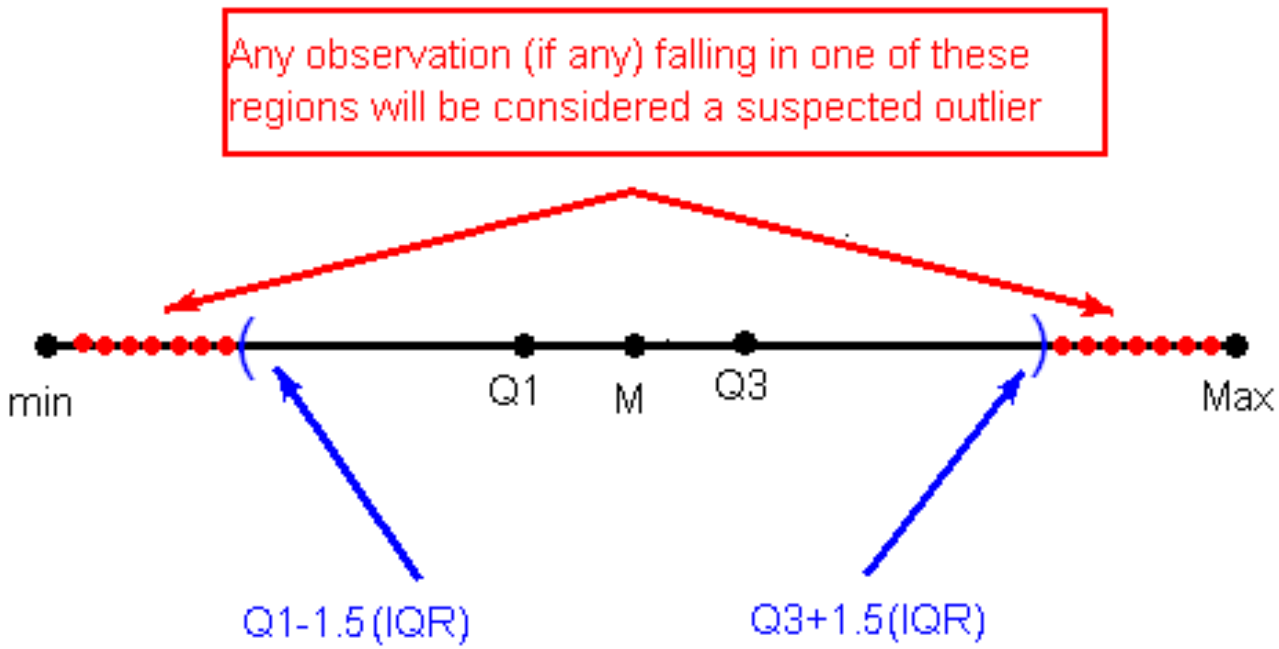
Range và inter-quartile range (IQR)

- Range = max - min
- Inter-Quartile Range (IQR)

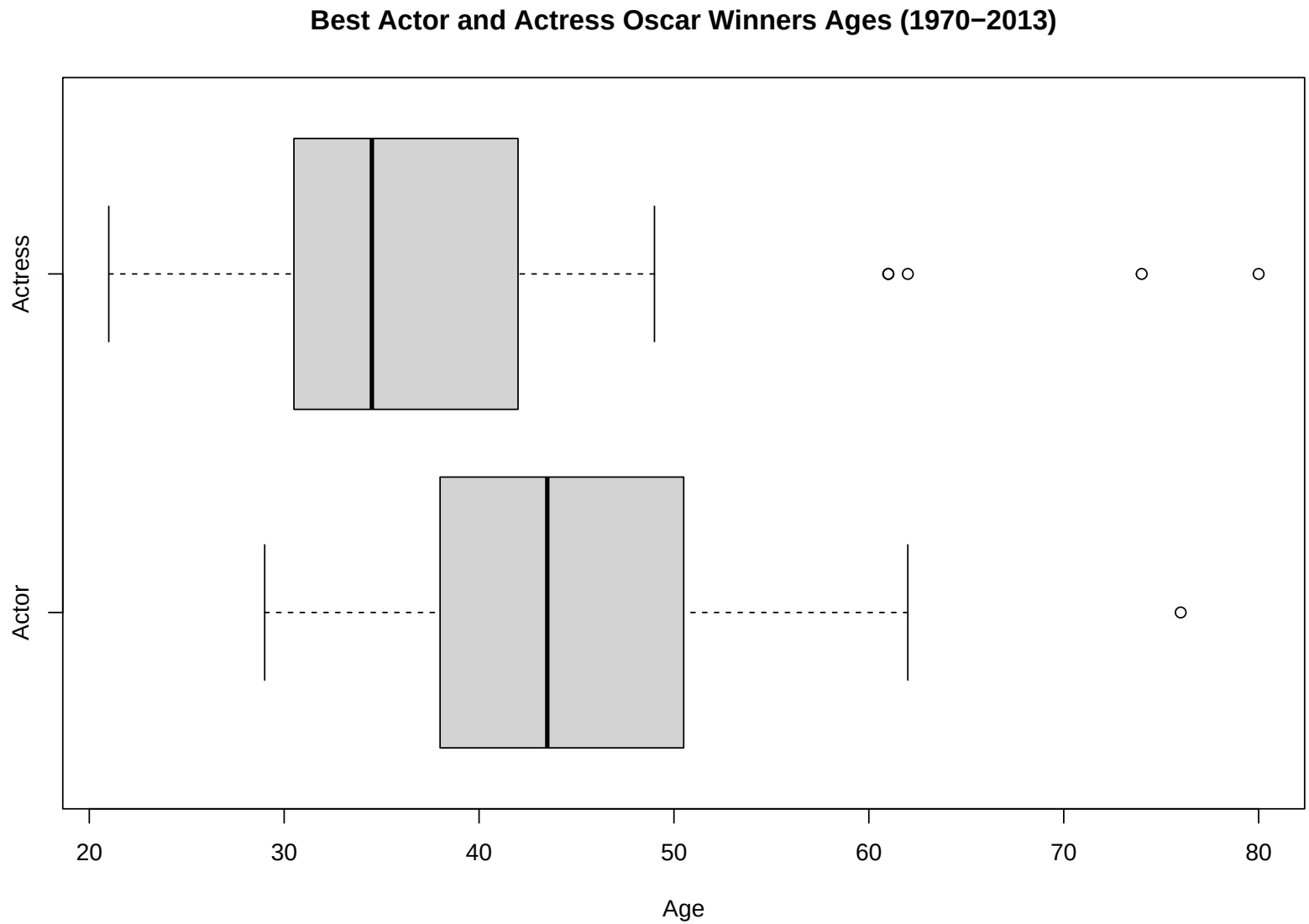


Phát hiện oulier dùng IQR

- Một giá trị là outlier nếu
  - Nhỏ hơn $Q_1 - 1.5 * IQR$, hoặc
  - Lớn hơn $Q_3 + 1.5 * IQR$



Biểu đồ boxplot



Variance and standard deviation

- Variance: đo mức độ phân tán của dữ liệu quanh giá trị trung tâm (trung bình).

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

- Standard deviation: cũng đo mức độ phân tán nhưng có cùng đơn vị với $mean(X)$.

$$sd(X) = \sqrt{var(X)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

Variance and standard deviation

```
##  [1] 43 42 48 49 56 38 60 30 40 42 37 76 39 53 45 36 62
## [26] 32 45 60 46 40 36 47 29 43 37 38 45 50 48 60 50 39
```

```
## [1] "mean = 45"
```

```
## [1] "modes:  42, 38, 60, 37, 45"
```

```
## [1] "median = 43.5"
```

```
## [1] "min = 29"
```

```
## [1] "max = 76"
```

```
## [1] "range = 47"
```

```
## [1] "IQR = 12.25"
```

```
## [1] "sd = 9.7"
```

```
## [1] "var = 95"
```

Phân tích mối quan hệ giữa hai biến

Biến giải thích và biến phản hồi

- Khi phân tích mối quan hệ giữa hai biến, mỗi biến có một vai trò, gồm:
  - Biến giải thích, còn được gọi là biến độc lập, (explanatory/independent variable) - nhằm giải thích hoặc dự đoán biến phản hồi, và
  - Biến phản hồi, còn được gọi là biến phụ thuộc (response/dependent variable) - là kết quả của nghiên cứu
- Biến giải thích thường được ký hiệu là X, biến phản hồi được ký hiệu là Y.

Phân tích mối quan hệ giữa hai biến

- Có 4 loại mối quan hệ giữa hai biến

		Response	
		Categorical	**Quantitative**
Explanatory	**Categorical**	C → C	C → Q
	Quantitative	Q → C	Q → Q

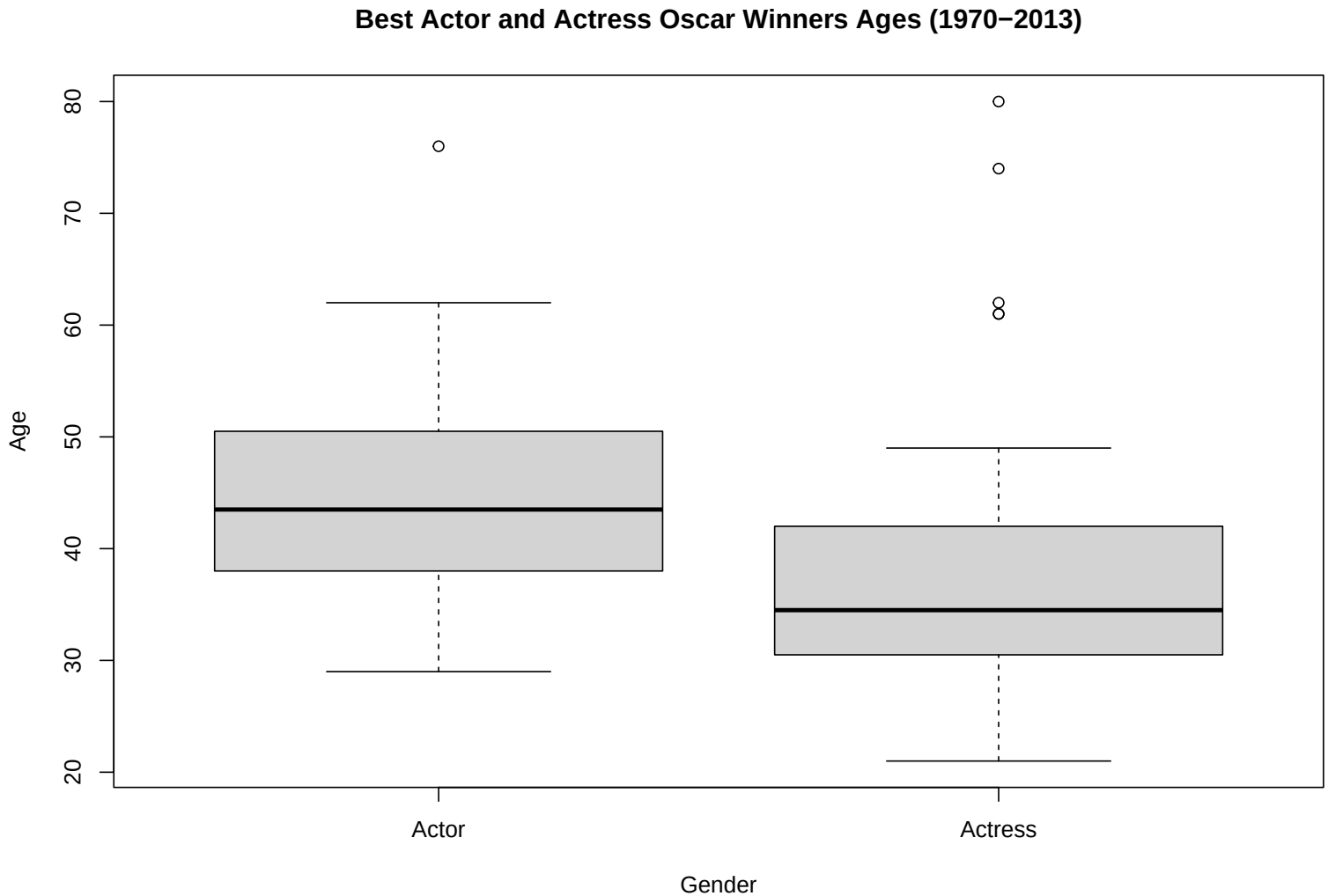
$$C \to Q$$

$C \to Q$

		Response	
		Categorical	**Quantitative**
Explanatory	**Categorical**	C → C	C → Q
	Quantitative	Q → C	Q → Q

$C \to Q$

- Ta có thể dùng side-by-side boxplot kết hợp với các thống kê mô tả để tóm tắt và trực quan hóa mối quan hệ.



$$C \to C$$

$C \to C$

		Response	
		Categorical	Quantitative
Explanatory	Categorical	C→C	√C→Q
	Quantitative	Q→C	Q→Q

Two-way table

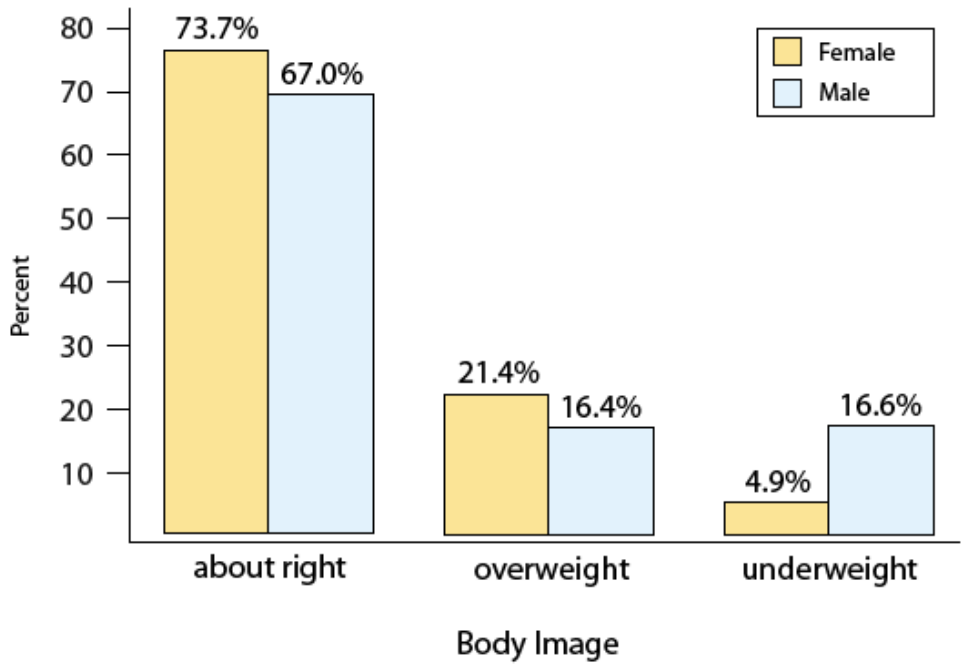
- Ta tóm tắt mối quan hệ bằng two-way table

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	560	163	37	760
	Male	295	72	73	440
	Total	855	235	110	1200

Two-way table kết hợp double bar chart

- Ta có thể dùng two-way table kết hợp double bar chart để tóm tắt và trực quan hóa mối quan hệ.

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	73.7%	21.4%	4.9%	100%
	Male	67.0%	16.4%	16.6%	100%

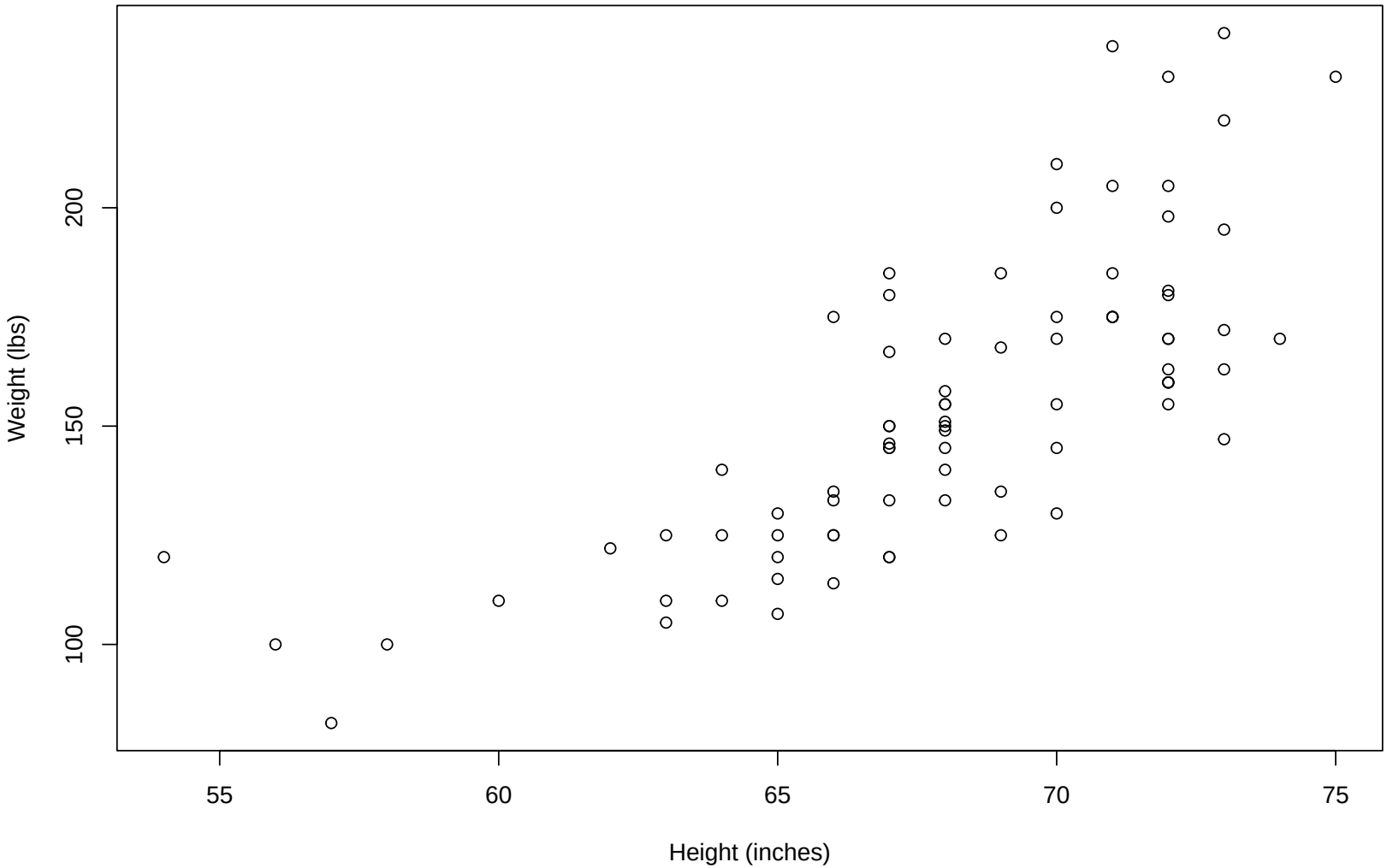


$$Q \to Q$$

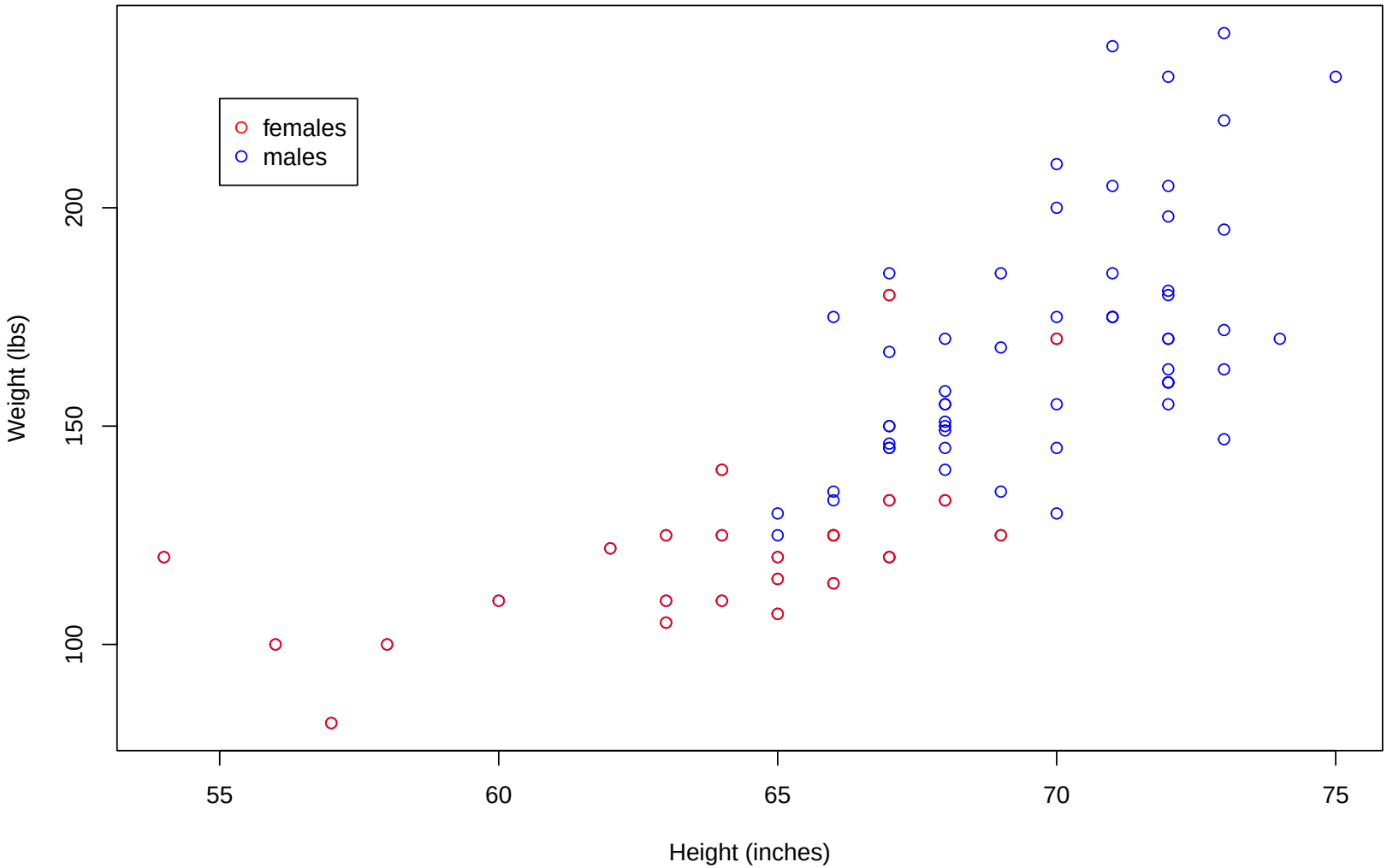
$Q \to Q$

		Response	
		Categorical	Quantitative
Explanatory	Categorical	√C→C	√C→Q
	Quantitative	Q→C	Q→Q

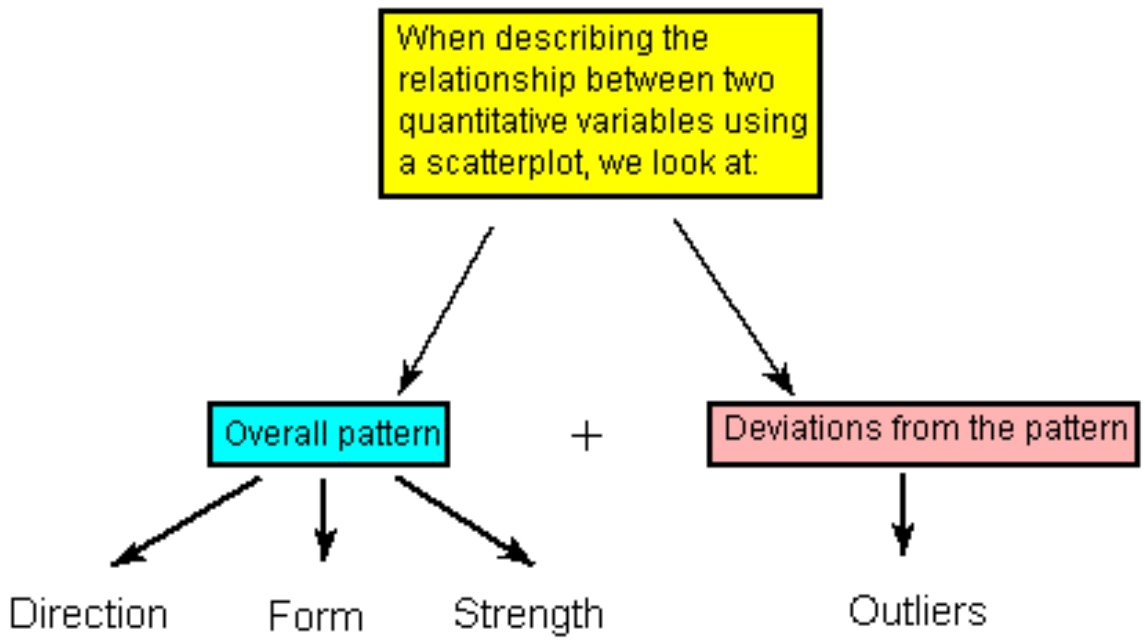
Biểu đồ scatterplot



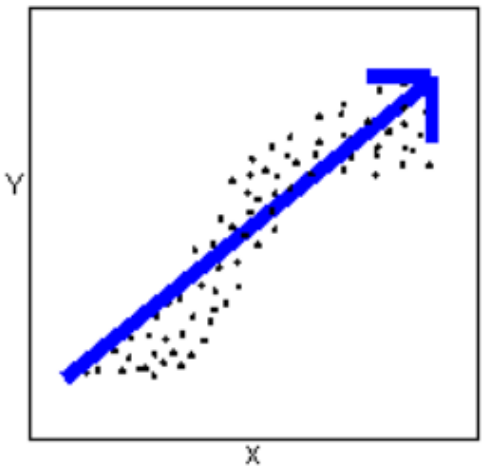
Biểu đồ scatterplot



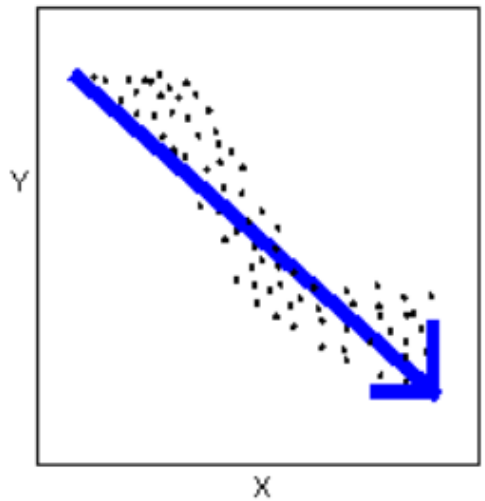
Hiểu biểu đồ scatterplot



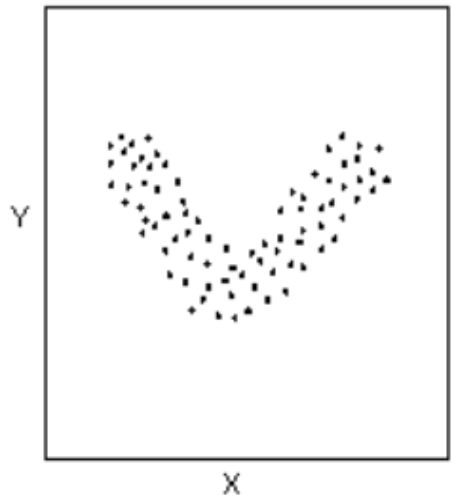
Direction của relationship



Positive relationship

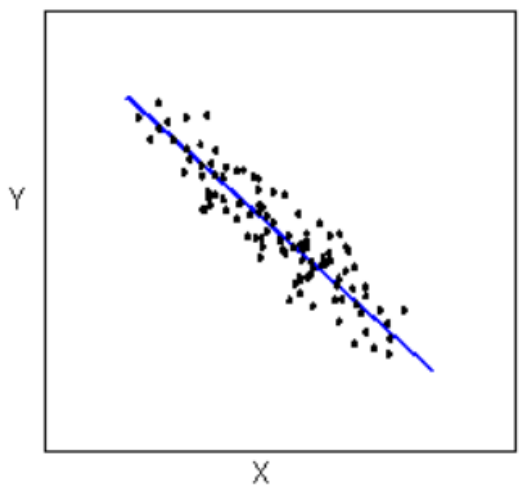


Negative relationship

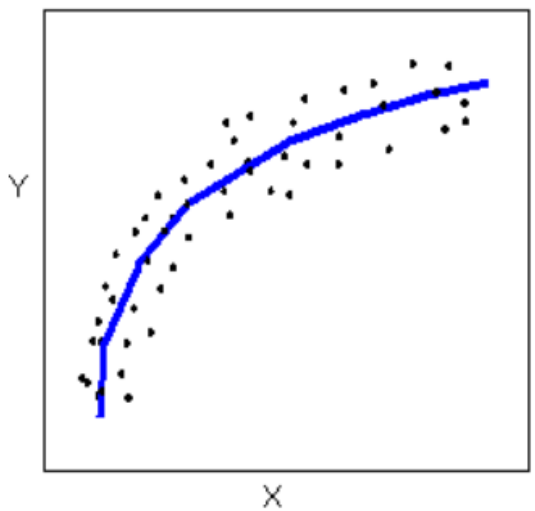


Neither positive nor negative

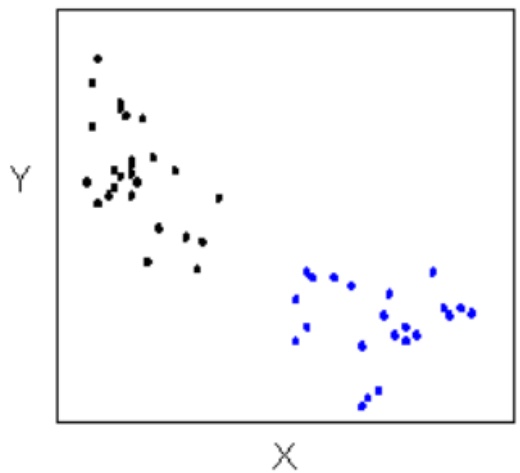
Form của relationship



linear

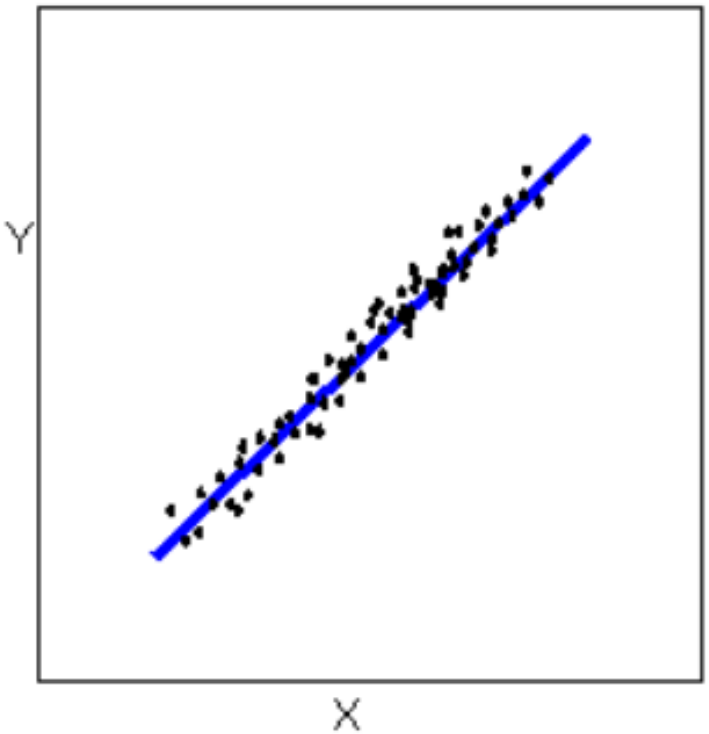


curvilinear

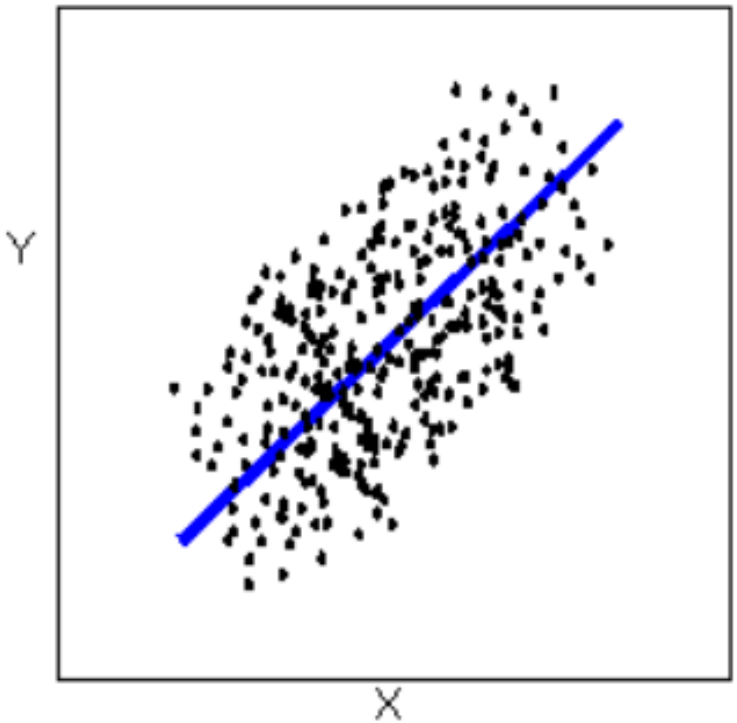


others

Strength của relationship



strong relationship



weaker relationship

Quan hệ tuyến tính

Linear relationship

- Biểu đồ scatterplot không thể hiện rõ strength của relationship
- Ta sẽ dùng giá trị số để mô tả rõ hơn strength của relationship
- Giá trị số này chỉ thích hợp để mô tả các linear relationship
- Không phải mọi quan hệ giữa hai biến định lượng đều có dạng linear

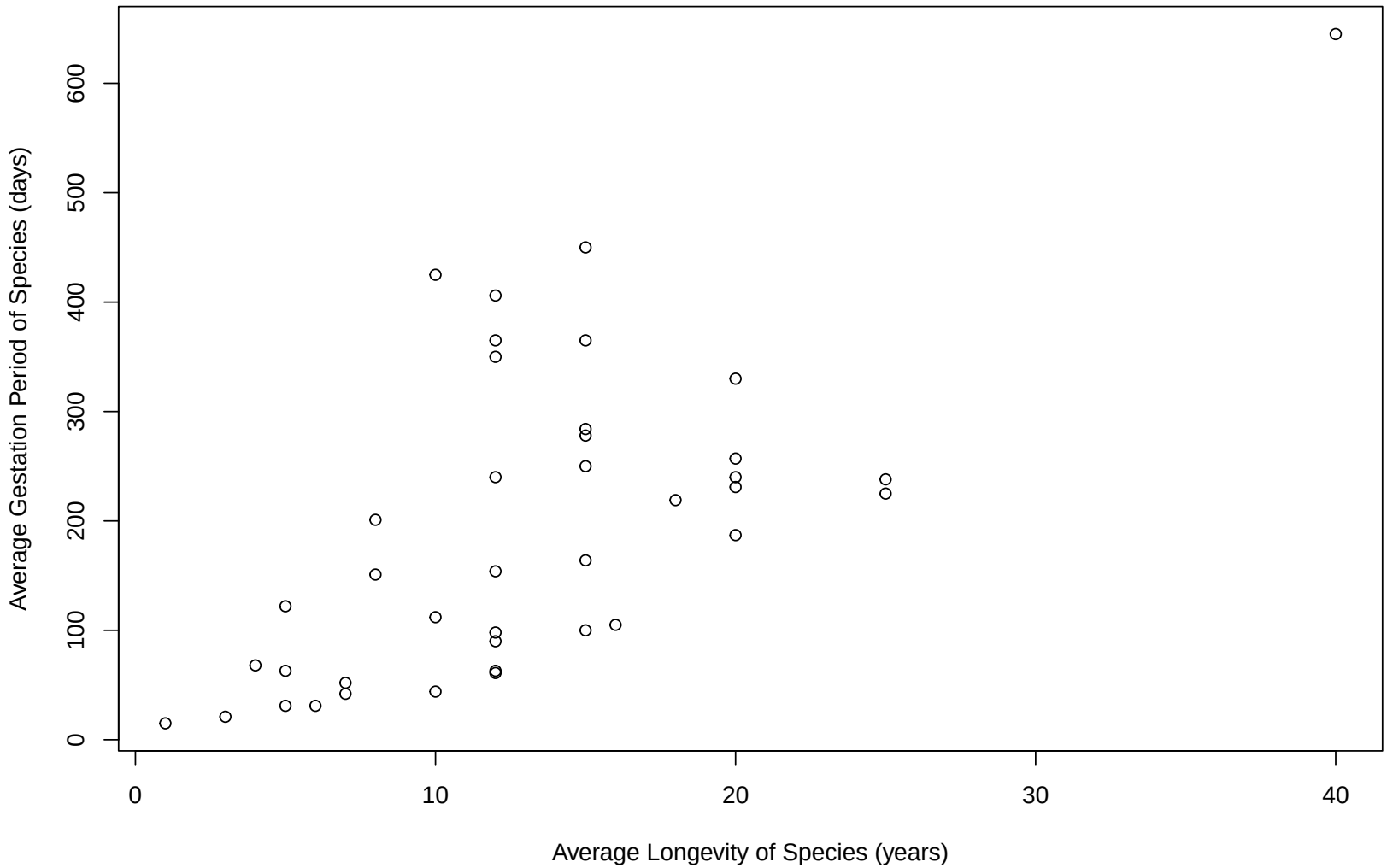
Sự tương quan (correlation)

- Giá trị số để đánh giá strength của một linear relationship được gọi là hệ số tương quan (correlation coefficient)
- Hệ số tương quan (r) giữa hai biến x, y được xác định bởi công thức

$$r_{X,Y} = r(X,Y) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{X}}{s_X} \right) \left(\frac{y_i - \overline{Y}}{s_Y} \right)$$

  - $\overline{x}, \overline{y}$ là giá trị trung bình của x và y
  - s_x, s_y là độ lệch chuẩn (standard deviation) của x và y

Hệ số tương quan



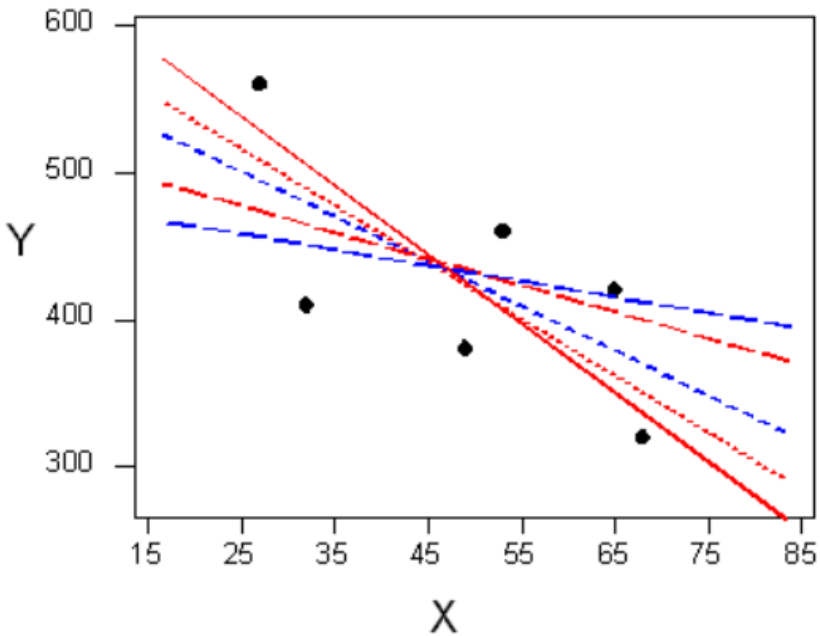

```
## [1] 0.6632397
```

Hồi quy (regression)

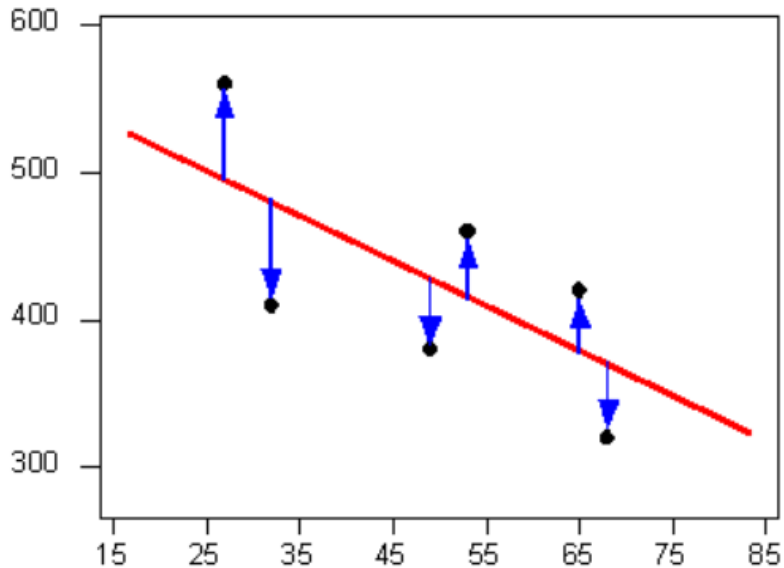
- Sự tương quan không mô tả hết mối quan hệ tuyến tính của hai biến định lượng
  - Nó chỉ mô tả strength và direction của quan hệ
- Ta thường muốn hiểu rõ hơn biến này ảnh hưởng đến biến kia như thế nào
  - Ta muốn dự đoán (predict) giá trị của response variable với giá trị của explanatory variable cho trước
- Để có thể làm được điều đó, ta cần tóm tắt mối quan hệ tuyến tính bằng một đường thẳng phù hợp nhất với dạng tuyến tính của dữ liệu

Hồi quy bình phương tối thiểu (Least squares regression)

- Kỹ thuật xác định sự phụ thuộc củaresponse variable vào explanatory variable gọi là hồi quy (regression)
- Khi sự phụ thuộc này là tuyến tính (linear), ta gọi nó là hồi quy tuyến tính (linear regression)



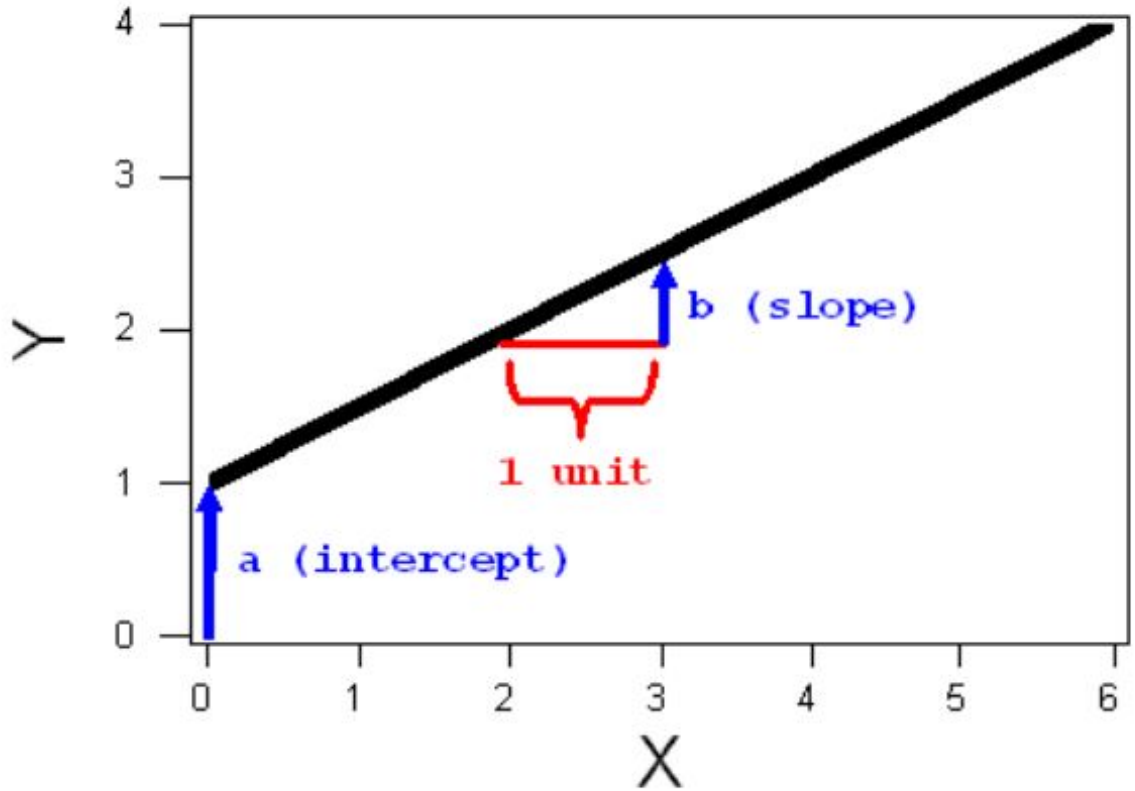
many candidates



best fit

Hồi quy bình phương tối thiểu (Least squares regression)

- Quan hệ tuyến tính được biểu diễn dưới dạng đường thẳng $Y = a + bX$
  - a là intercept (giá trị Y nhận khi $X = 0$)
  - b là slope (thay đổi của Y khi X tăng một đơn vị)



Intercept và slope

- Quan hệ tuyến tính được biểu diễn dưới dạng đường thẳng $Y = a + bX$
- Intercept và slope được tính như sau:

$$b = r\left(\frac{s_Y}{s_X}\right)$$

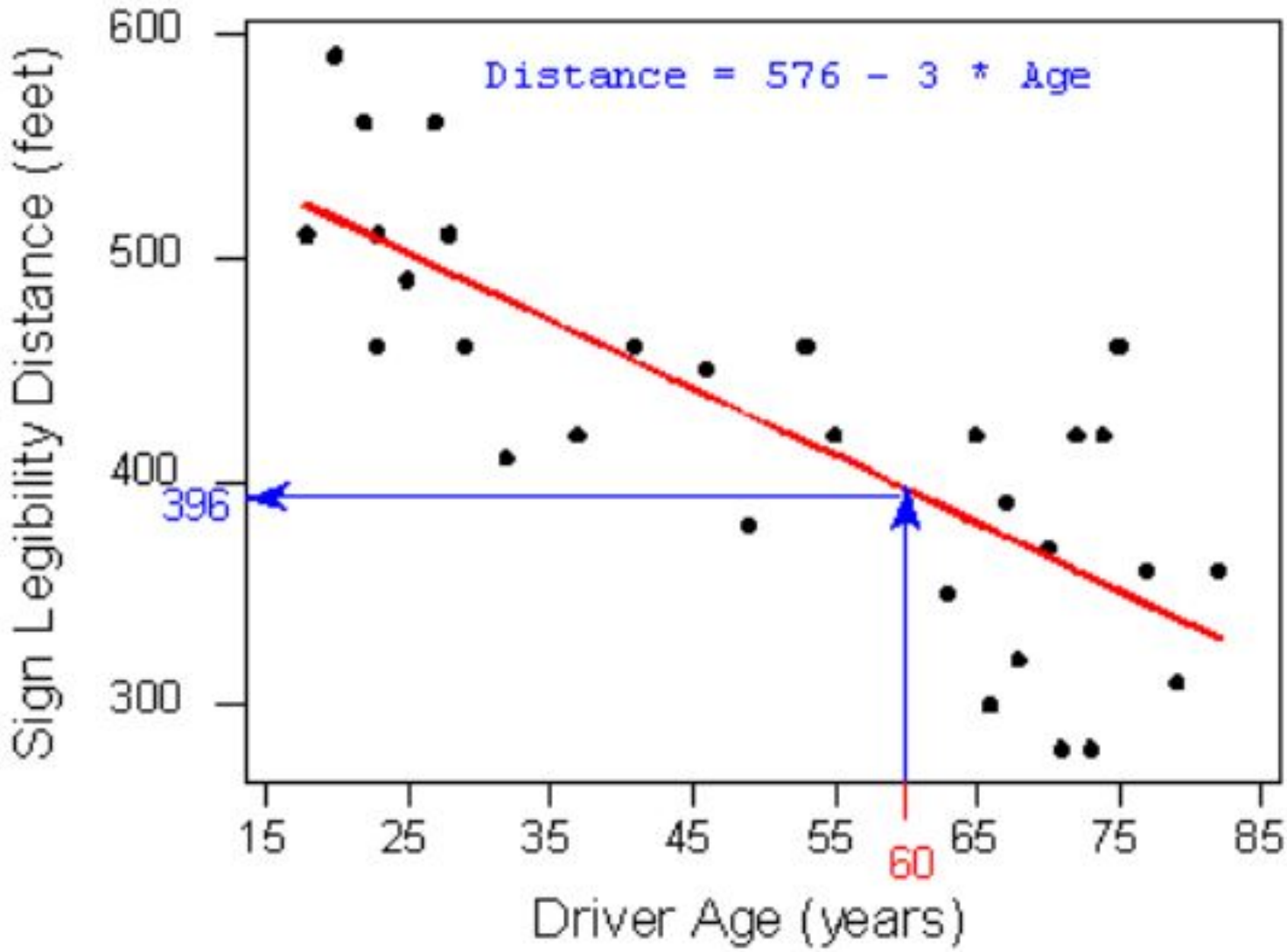
$$a = \overline{Y} - b\overline{X}$$

Trong đó:

- $\overline{X}, \overline{Y}$ là giá trị trung bình của X và Y
- s_X, s_Y là độ lệch chuẩn của X và Y
- r là hệ số tương quan của X và Y

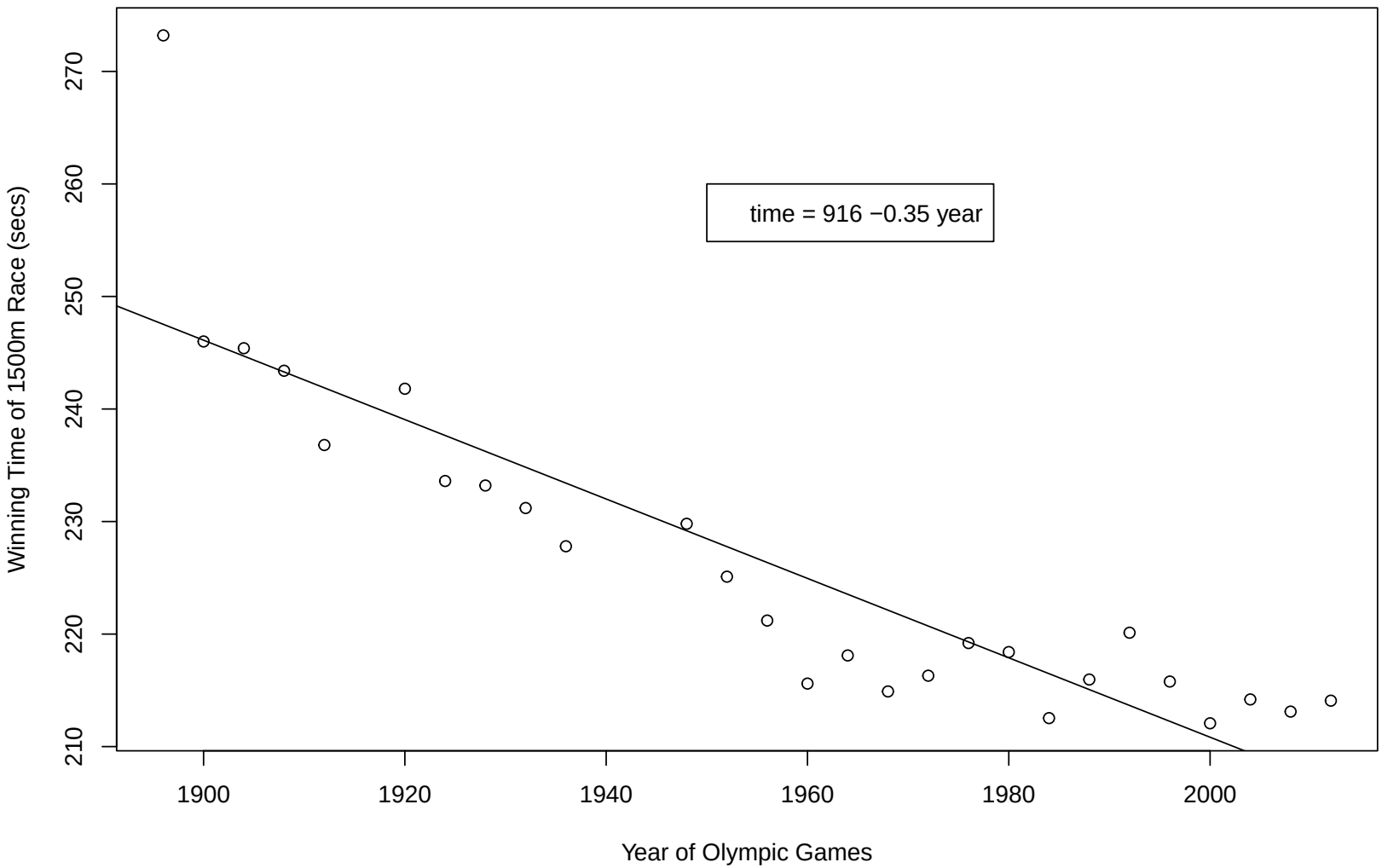
Dự đoán (prediction)

- Least squares regression line được dùng để đưa ra dự đoán.



Least squares regression line

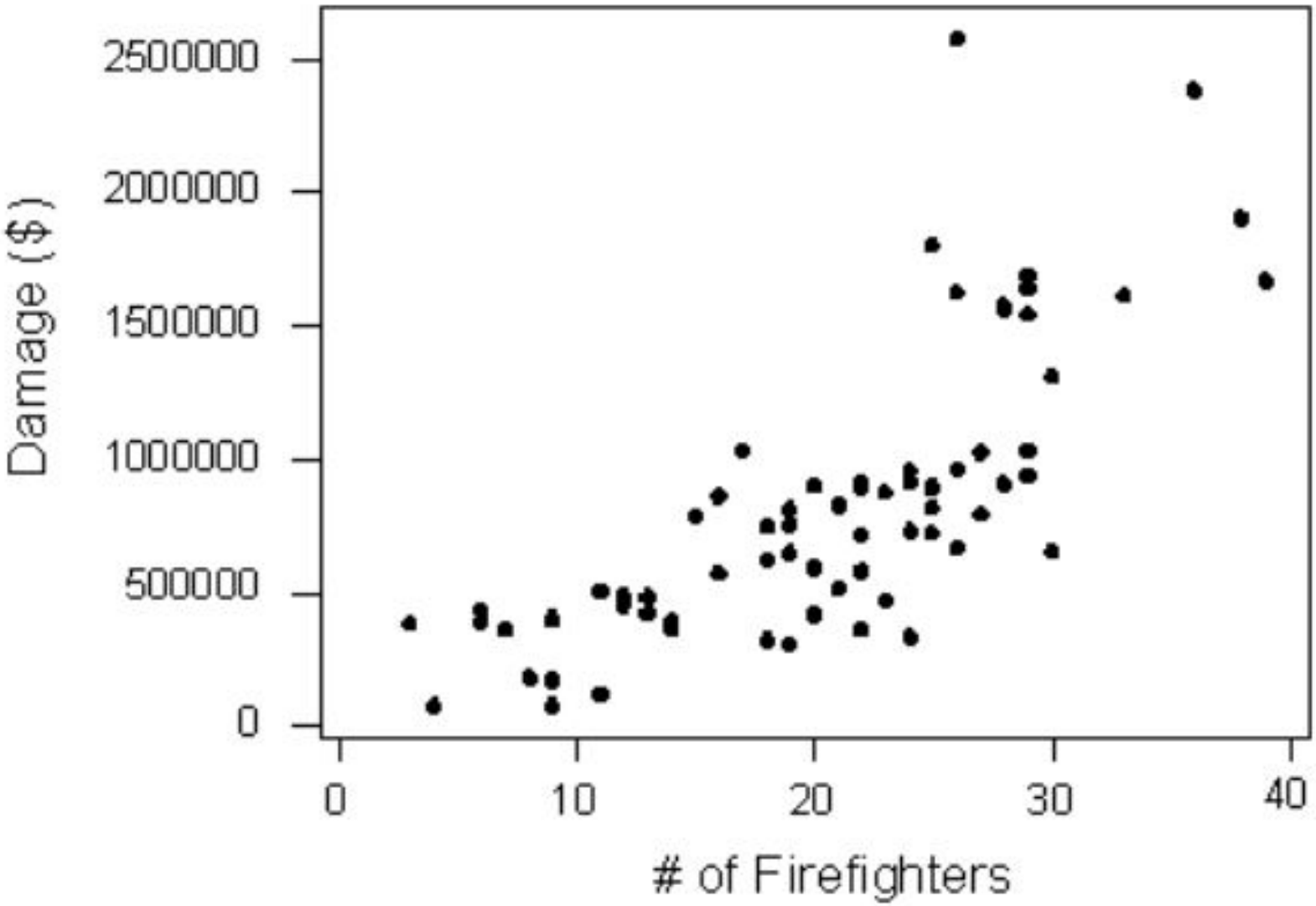
```
## (Intercept)   olym$Year
## 916.4323092  -0.3527988
```



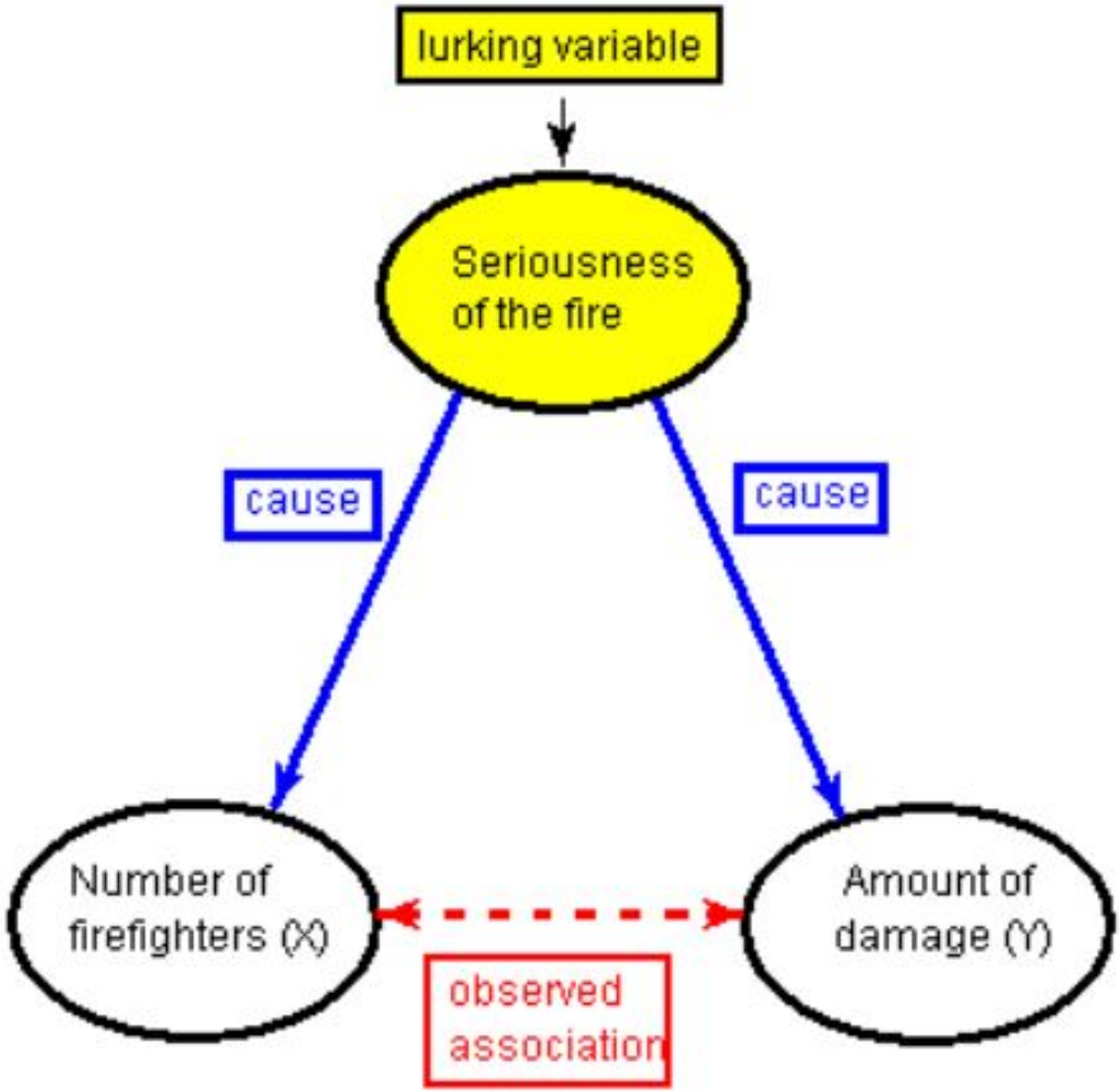
Quan hệ nhân quả (causation)

Sự tương quan không suy ra quan hệ nhân quả

- Sai lầm phổ biến là giải thích mối quan hệ là nhân quả khi thấy chúng có tương quan.



Causation and Lurking Variables



Simpson's Paradox

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	63	2037	2100
	Hospital B	16	784	800
	Total	79	2821	2900

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	**3%**	97%	100%
	Hospital B	**2%**	98%	100%

Simpson's Paradox

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	63	2037	2100
	Hospital B	16	784	800
	Total	79	2821	2900

Accounting for the lurking variable: **"severity of illness"**

Patients severly ill

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	57	1443	1500
	Hospital B	8	192	200
	Total	65	1635	1700

Patients *not* severly ill

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	6	594	600
	Hospital B	8	592	600
	Total	14	1186	1200

Tổng kết (1)

- Mục đích của EDA là biến dữ liệu thành thông tin có ý nghĩa
- Khi thực hiện EDA, chúng ta
  - Tóm tắt dữ liệu bằng biểu đồ và các giá trị số
  - Mô tả tổng thể dữ liệu và các ngoại lệ
- Biến phân loại
  - Biểu đồ: pie chart hoặc bar chart
  - Số: đếm, hoặc tỷ lệ
- Biến số
  - Biểu đồ: histogram, stemplot, dot plot, boxplot
    - Shape, center, spread, outlier
  - Số: center, spread
    - Center: mean, mode, median
    - Spread: standard deviation, range, IQR

Tổng kết (2)

- Phân tích mối quan hệ của hai biến

		Response	
		Categorical	Quantitative
Explanatory	Categorical	✓ C→C	✓ C→Q
	Quantitative	✗ Q→C	✓ Q→Q

Tổng kết (3)

- Mối quan hệ tuyến tính
  - Tương quan (correlation)
  - Hồi quy (regression)
  - Binary ralationship
    - Least squares regression
- Mối quan hệ nhân quả
  - Lurking Variables
  - Simpson's paradox

Tham khảo

Probabilit & Statistics. Provided by: Open Learning Initiative. Located at: http://oli.cmu.edu. License: CC BY: Attribution

Xác suất (probability)

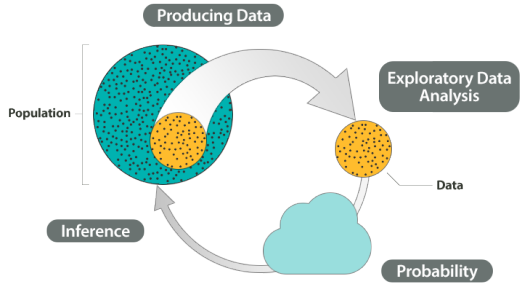
Quách Đình Hoàng

2022/03/21

Nội dung

Định nghĩa Xác suất

Xác suất (probability)



- Mục tiêu cuối cùng của phân tích dữ liệu là rút ra kết luận đáng tin cậy về quần thể dựa trên những gì ta đã khám phá trên mẫu.
- Xác suất là nền tảng cơ bản cho các phương pháp suy luận thống kê.

Xác suất (probability)

- Chúng ta sử dụng một mẫu để tìm hiểu về quần thể mà từ đó nó được rút ra.
- Lý tưởng nhất là mẫu nên ngẫu nhiên để nó đại diện tốt cho quần thể.
- Tuy nhiên, điều này không có nghĩa là tất cả các mẫu ngẫu nhiên nhất thiết phải hoàn hảo.
- Khi nhìn vào một mẫu cụ thể, chúng ta sẽ không bao giờ biết nó khác với quần thể như thế nào. Sự không chắc chắn này là nơi xác suất xuất hiện trong bức tranh.
- Chúng ta sử dụng xác suất để định lượng mức độ mà chúng ta mong đợi các mẫu ngẫu nhiên sẽ thay đổi. Điều này cho chúng ta một cách để đưa ra kết luận về quần thể khi đối mặt với sự không chắc chắn được tạo ra bởi việc sử dụng một mẫu ngẫu nhiên.

Xác suất (probability)

- Xét ví dụ sau: Chọn ngẫu nhiên 1000 nam sinh và đo chiều cao ta thu được chiều cao trung bình là 170 cm. Thống kê này có đúng cho toàn bộ nam sinh Việt Nam?
- Ta không chắc vì một mẫu ngẫu nhiên khác có thể cho ra kết quả khác.
- Nhưng ta có thể sử dụng xác suất để mô tả khả năng thống kê trên mẫu của ta nằm trong mức độ chính xác mong muốn.
  - Ví dụ: Khả năng thống kê mẫu của chúng ta không quá 5 cm so với chiều cao trung bình của tất cả nam sinh Việt Nam?
  - Câu trả lời cho câu hỏi này (dùng xác suất) sẽ có ảnh hưởng quan trọng đến độ tin cậy ở bước suy luận. Đặt biệt, nếu chúng ta thấy trung bình mẫu không khác nhiều so với trung bình quần thể, thì chúng ta rất tự tin rằng chúng ta có thể đưa ra kết luận về quần thể dựa trên mẫu.

Xác suất (probability)

- Xác suất là một cách để đo lường sự không chắc chắn.
- Gọi A là điều ta muốn tìm xác xuất, A được gọi là sự kiện (event).
- $P(A)$ thể hiện xác suất của sự kiện A, nó đo khả năng sự kiện A xảy ra.
- $0 \leq P(A) \leq 1, \forall A$, xác suất của sự kiện A bất kỳ luôn nằm ở giữa 0 và 1.

Xác định xác suất

- Có hai cách để xác định xác suất: lý thuyết (theoretical / classical) và thực nghiệm (empirical / observational).
- Phương pháp lý thuyết sử dụng bản chất của tình huống để xác định xác suất.
  - Tung một đồng xu cân bằng, $P(H) = P(T) = 0.5$.
- Phương pháp thực nghiệm dùng các thử nghiệm để tạo ra các kết quả không thể dự đoán trước
  - Tung đồng xu cân bằng 10000 lần và ghi lại kết quả để tính $P(H)$ và $P(T)$.
  - Ta khó thể đạt chính xác $P(H) = P(T) = 0.5$ như phương pháp lý thuyết nhưng thường cũng khá gần với nó.
  - Trong nhiều trường hợp, ta khó thể tìm được xác suất lý thuyết và phải dùng xác suất thực nghiệm để thay thế.

Không gian mẫu (sample space)

- Thí nghiệm ngẫu nhiên (random experiment): là một thí nghiệm tạo ra kết quả không thể dự đoán trước.
  - Ví dụ: tung đồng xu và ghi lại kết quả, chọn ngẫu nhiên một SV và ghi lại ngày sinh, ...
- Mỗi thí nghiệm ngẫu nhiên có một tập hợp các kết quả có thể xảy ra được gọi là không gian mẫu (sample space).
  - Ta không chắc chắn về kết quả nào sẽ nhận được từ thí nghiệm nhưng chắc chắn về các kết quả có thể xảy ra.
  - Tình huống thường gặp là tất cả các kết quả trong không gian mẫu đều có khả năng xảy ra như nhau
- Một sự kiện (event) là một tập con các kết quả của không gian mẫu.
- Khi một sự kiện được xác định, chúng ta có thể tính xác suất (probability) của nó.

Ví dụ

- Thí nghiệm (Experiment)
  - Tung một đồng xu cân bằng 3 lần
- Không gian mẫu (Sample space)
  - S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
- Sự kiện (Event)
  - E = "Có 2 mặt H" = {HHT, HHT, HTH, THH}
- Hàm xác suất (Probability function)
  - Mỗi kết quả (outcome) có xác suất là 1/8

Outcome	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Probability	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Hàm xác suất

- Không gian mẫu rời rạc (discrete sample space)
  - Là một không gian mẫu mà các kết quả có thể liệt kê được, nó có thể hữu hạn hoặc vô hạn.
  - Ví dụ: {H, T}, {1, 2, 3, ⋯}
- Hàm xác suất (probability function)
  - Hàm xác suất P trên một không gian mẫu rời rạc Ω là hàm gán cho mỗi outcome ω một giá trị $P(\omega)$ gọi là xác suất của ω. Hàm P thỏa:

$$\begin{cases} 0 \leq P(\omega) \leq 1 \\ \sum_{i=1}^{n} P(\omega_i) = 1, \Omega = \{\omega_1, \omega_2, ... \omega_n\} \end{cases}$$

- Xác suất của sự kiện
  - Xác suất của sự kiện E là tổng xác suất của tất cả các outcome trong E.

$$P(E) = \sum_{\omega \in E} P(\omega)$$

Các qui tắc xác suất cơ bản

Các qui tắc xác suất cơ bản

1. $0 \leq P(A) \leq 1$, với mọi sự kiện A
2. $P(S) = 1$, với S là không gian mẫu của thí nghiệm
3. $P(\overline{A}) = 1 - P(A)$
4. $P(A \cup B) = P(A) + P(B)$, nếu $A \cap B = \emptyset$
5. $P(A \cap B) = P(A) * P(B)$, nếu A và B là độc lập

- Hai sự kiện A và B là độc lập (independent) nếu A xảy ra không ảnh hưởng đến xác suất B xảy ra và ngược lại.

6. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, với mọi sự kiện A, B

Ví dụ 1

Một cặp vợ chồng lên kế hoạch đến có 3 con. Ký hiệu B cho con trai và G cho con gái. Cho biết:

1. Không gian mẫu S của tất cả các kết quả có thể có
2. Sự kiện A - con giữa là con gái
3. Sự kiện B - 3 con có cùng giới tính
4. Sự kiện C - chính xác một trong ba con là con gái

Ví dụ 2

Một cặp vợ chồng quyết định có con cho đến khi họ có một trai và một gái, nhưng họ sẽ không có nhiều hơn ba người con. Cho biết:

1. Không gian mẫu S của tất cả các kết quả có thể có
2. Sự kiện A - cặp vợ chồng này có một bé trai
3. Sự kiện B – cặp vợ chồng này có ba người con
4. Sự kiện C – tất cả con đều có cùng giới tính

Ví dụ 3

Cho bảng phân bố xác xuất về nhóm máu trong dân số như sau:

Type	O	A	B	AB
Prob	0.44	0.42	0.10	0.04

Ngoài ra, tính chất của các nhóm máu như sau:

- Người nhóm A có thể hiến máu cho người nhóm A hoặc AB.
- Người nhóm B có thể hiến máu cho người nhóm B hoặc AB.
- Người nhóm AB chỉ có thể hiến máu cho người nhóm AB.
- Người nhóm O có thể hiến cho bất kỳ ai.

Giả sử rằng có 2 bệnh nhân cần được hiến máu. Bệnh nhân 1 có nhóm máu A và bệnh nhân 2 có nhóm máu B. Tìm xác suất để một người được chọn ngẫu nhiên có thể hiến máu cho bệnh nhân 1 hoặc bệnh nhân 2.

Ví dụ 4 (1)

Cho bảng phân bố xác xuất về nhóm máu trong dân số như sau:

Type	O	A	B	AB
Prob	0.44	0.42	0.10	0.04

Chọn đồng thời và ngẫu nhiên 2 người, xác suất để cả hai đều có nhóm máu O là bao nhiêu?

Xác suất có điều kiện

Xác suất có điều kiện (conditional probability)

- Xác suất có điều kiện của sự kiện B cho trước A, ký hiệu $P(B|A)$, được định nghĩa:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- Nếu hai sự kiện A và B là độc lập thì

$$P(B|A) = P(B)$$

.

  - Điều kiện tương đương: $P(B|A) = P(B|\overline{A})$

Luật xác suất toàn phần

- Luật nhân (multiplication rule)

$$P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$$

- Luật xác suất toàn phần (Law of total probability)
  - Cho $A_1, A_2, \cdots, A_n$ là một phân hoạch (partition) của không gian mẫu Ω, tức $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$ và $A_i \cap A_j = \emptyset, \forall i, j$. Khi đó, với mọi sự kiện B:

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i)$$

  - Áp dụng luật nhân ta có thể viết lại

$$P(B) = \sum_{i=1}^{n} P(A_i) \times P(B|A_i)$$

Sự độc lập (independence)

- Hai sự kiện được gọi là độc lập (independent) nếu việc biết sự kiện này xảy ra không ảnh hưởng đến xác suất của sự kiện còn lại.
- Định nghĩa: A độc lập với B nếu

$$P(A \cap B) = P(A) \times P(B)$$

- Hệ quả: A độc lập với B thì:

$$P(A|B) = P(A)$$

Định lý Bayes (Bayes' theorem)

- Định lý Bayes: Cho hai sự kiện A và B, khi đó:

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}$$

- Ví dụ: Xét thí nghiệm tung 3 đồng xu cân bằng
  - Gọi A là sự kiện đồng xu thứ nhất có mặt sấp
  - B là sự kiện cả 3 đồng xu đều mặt sấp
  - Khi đó: $P(A) = 1/2, P(B) = 1/8, P(A|B) = 1$
  - Dùng định lý Bayes ta có thể tính $P(B|A)$ như sau:

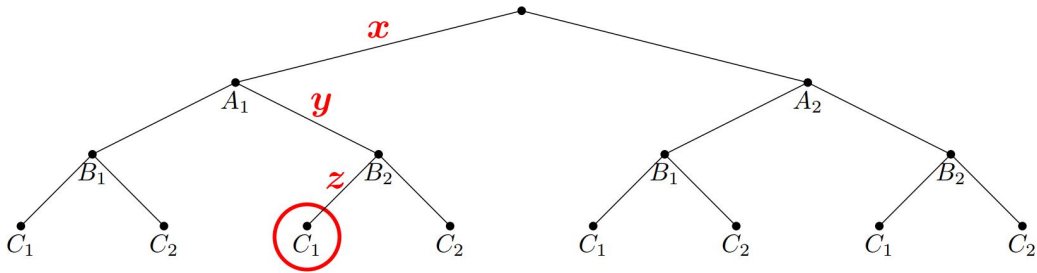
$$P(B|A) = \frac{P(B) \times P(A|B)}{P(A)} = \frac{1/8 \times 1}{1/2} = \frac{1}{4}$$

Ví dụ 1

- Có 5 bi đỏ và 2 bi xanh trong hộp. Chọn ra ngẫu nhiên một viên và thay thế bằng viên có màu kia rồi sau đó chọn ra ngẫu nhiên một viên thứ hai.
  1. Xác suất viên thứ hai có màu đỏ
  2. Xác suất viên thứ nhất có màu đỏ cho biết viên thứ hai có màu đỏ.

Ví dụ 2

Cho cây xác suất



1. Xác suất mà x thể hiện là gì?
2. Xác suất mà y thể hiện là gì?
3. Xác suất mà z thể hiện là gì?
4. Nút được bao quanh bởi đường tròn thể hiện sự kiện gì?

Ví dụ 3

Monty Hall problem

- Một cửa có quà (C), 2 cửa không có (G).
- Người chơi chọn một cửa
- Monty mở cửa không có quà trong hai cửa còn lại
- Người chơi được cho phép đổi sự lựa chọn nếu muốn.

Đâu là chiến lược tốt nhất của người chơi?

1. Giữ nguyên
2. Đổi
3. Đổi hay không cũng vậy

Ví dụ 4

- $P(HIV) = 0.0001 \rightarrow P(\overline{HIV}) = 0.9999$
- $P(DT|\overline{HIV}) = 0.01 \rightarrow P(DT|HIV) = 0.99$

Cho biết $P(HIV|DT) = ?$

Biến ngẫu nhiên và phân bố xác suất

Quách Đình Hoàng

2022/03/28

Nội dung

Biến ngẫu nhiên

Biến ngẫu nhiên

- Một biến ngẫu nhiên (random variable) gán một giá trị số duy nhất cho kết quả (outcome) của một thí nghiệm ngẫu nhiên (random experiment).
  - Ta thường dùng chữ cái in hoa, như $X, Y, ...$, để ký hiệu biến ngẫu nhiên và dùng chữ cái thường, như $x, y, ...$, để ký hiệu giá trị của biến ngẫu nhiên.
  - $P(X = x)$: xác xuất biến ngẫu nhiên X nhận giá trị x.
- Có 2 loại biến ngẫu nhiên:
  - Biến ngẫu nhiên liên tục: thường chỉ nhận chỉ những giá trị nguyên
    - Số loại thẻ tín dụng mà một người sở hữu, số con trong một gia đình
  - Biến ngẫu nhiên liên tục: nhận giá trị thực
    - Chiều cao của một người, thời gian chạy của 100 m của một vận động viên

Ví dụ về biến ngẫu nhiên

- Ví dụ 1: Xét thí nghiệm ngẫu nghiên tung đồng xu cân bằng 2 lần.
  - Không gian mẫu: $S = \{HH, HT, TH, TT\}$
  - Ta định nghĩa biến ngẫu nhiên X là số lượng mặt T mà mỗi thí nghiệm tạo ra.
    - Nếu kết quả (outcome) là HH thì $X = 0$.
    - Nếu kết quả (outcome) là HT thì $X = 1$.
    - Nếu kết quả (outcome) là TH thì $X = 1$.
    - Nếu kết quả (outcome) là TT thì $X = 2$.
- Như vậy, X là một biến định lượng có thể nhận các giá trị là 0, 1 hoặc 2. X là ngẫu nhiên vì chúng ta không biết nó sẽ nhận giá trị nào trong ba giá trị.

Ví dụ về biến ngẫu nhiên

- Ví dụ 2: Giả sử chúng ta chọn ngẫu nhiên một võ sĩ nam hạng nhẹ và ghi lại trọng lượng chính xác của anh ta.
  - Không gian mẫu: $S = \{$Tất cả võ sĩ quyền anh nặng khoảng 130-135 pound$\}$
  - Ta định nghĩa biến ngẫu nhiên X là là trọng lượng của võ sĩ quyền Anh.
    - X có thể nhận bất kỳ giá trị nào trong khoảng từ 130 đến 135.
- Biến ngẫu nhiên trong ví dụ 1 có thể nhận một danh sách các giá trị riêng biệt, được gọi là biến ngẫu nhiên rời rạc.
- Biến ngẫu nhiên trong ví dụ 2 có thể nhận bất kỳ giá trị nào trong một khoảng, được gọi là biến ngẫu nhiên liên tục.

Phân biệt loại biến ngẫu nhiên

- Đôi khi, các biến ngẫu nhiên liên tục được làm tròn và do đó trông giống rời rạc. Mặc dù vậy, đây vẫn là các biến ngẫu nhiên liên tục.
  - Thời gian xem TV trong một tuần, được làm tròn đến giờ (hoặc phút) gần nhất
  - Nhiệt độ môi trường, chính xác đến mức độ C.
- Mặt khác, có một số biến có bản chất là rời rạc, nhưng có thể nhận rất nhiều giá trị khác nhau nên sẽ dễ dàng hơn nhiều nếu coi chúng là liên tục thay vì rời rạc.
  - Điểm TOEIC của một học sinh
  - Lương hàng năm của một CEO (làm tròn đến đô hay xu gần nhất)
- Một nguyên tắc nhanh để xác định loại biến là: biến rời rạc là những gì ta đếm (count), biến liên tục là những gì ta đo (measure).

Biến ngẫu nhiên rời rạc

Biến ngẫu nhiên rời rạc

- Cho Ω là không gian mẫu. Một biến ngẫu nhiên rời rạc (discrete random variable) là hàm

$$X : \Omega \to \mathbb{R}$$

nhận các giá trị rời rạc.

- Ví dụ: Xét thí nghiệm tung 2 con xúc xắc
  - $\Omega = \{(\omega_1, \omega_2) | \omega_1, \omega_2 \in \{1, 2, 3, 4, 5, 6\}\}$

$$X : \Omega \to R$$
$$\omega = (\omega_1, \omega_2) \mapsto X(\omega) = X(\omega_1, \omega_2) = \omega_1 + \omega_2$$

Probability mass function (pmf)

- Probability mass function (pmf) của một biến ngẫu nhiên rời rạc (discrete random variable) X là hàm

$$p(a) = P(X = a)$$

  - Khi cần nhấn mạnh biến X ta viết $p_X(a)$.
- Tính chất
  - $0 \leq p(a) \leq 1, \forall a$
  - $p(a) = 0$ khi X không bao giờ nhận giá trị a.
- $X \leq a = \{\omega | X(\omega) \leq a\}$: tập tất cả các outcome ω sao cho $X(\omega) \leq a$.

$$X \leq 4 = \{(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)\}$$

Cumulative distribution function (cdf)

- Hàm phân bố tích lũy (cumulative distribution function - cdf) của một biến ngẫu nhiên rời rạc (discrete random variable) X là hàm

$$F(a) = P(X \leq a)$$

- Tính chất
  - $F(a) \leq F(b), \forall a \leq b$
  - $0 \leq F(a) \leq 1, \forall a$
  - $\lim_{x \to \infty} F(x) = 1, \lim_{x \to -\infty} F(x) = 0$
- Ví dụ: Xét thí nghiệm tung hai xúc xắc và X là biến ngẫu nhiên mô tả giá trị của mặt lớn hơn trong hai xúc xắc.

Value	a	1	2	3	4	5	6
pmf	p(a)	1/36	3/36	5/36	7/36	9/36	11/36
cdf	F(a)	1/36	4/36	9/36	16/36	25/36	36/36

Kỳ vọng của biến ngẫu nhiên rời rạc

- Định nghĩa: Giả sử X là biến ngẫu nhiên rời rạc nhận các giá trị x_1, x_2, $\cdots$, x_n với các xác suất tương ứng là $p(x_1)$, $p(x_2)$, $\cdots$, $p(x_n)$. Khi đó, giá trị kỳ vọng (expected value) của X, ký hiệu $E(X)$ được định nghĩa như sau:

$$E(X) = \sum_{i=1}^{n} p(x_i)x_i = p(x_1)x_1 + p(x_2)x_2 + \cdots + p(x_n)x_n$$

- Giá trị kỳ vọng (expected value) của X còn được gọi là giá trị trung bình có trọng số (weighted mean/average) của X và thường được ký hiệu là μ.
  - Nếu các xác suất $p(x_i)$ bằng nhau, nó trở thành giá trị trung bình thông thường.
- Giá trị kỳ vọng (expected value) cho to biết vị trí (location) hoặc xu hướng tập trung (center tendency) của một biến ngẫu nhiên.

Ví dụ

- Cho biến ngẫu nhiên X có bảng xác suất sau:

X	2	4	6
pmf	1/6	1/6	2/3

$$E(X) = \frac{1}{6} \times 2 + \frac{1}{6} \times 4 + \frac{2}{3} \times 6 = 5$$

- Cho X là biến ngẫu nhiên có phân bố Bernoulli, $X \sim Bernoulli(p)$

$$E(X) = p \times 1 + (1 - p) \times 0 = p$$

Tính chất của kỳ vọng

- Cho X, Y là hai biến ngẫu nhiên trên không gian mẫu Ω

$$E(X+Y) = E(X) + E(Y)$$

- Cho biến ngẫu nhiên X và các hằng số a, b

$$E(aX+b) = aE(X) + b$$

- Ví dụ: X là tổng kết quả của hai lần tung xúc xắc.
  - Gọi X_1, X_2 là lần lượt là kết quả của lần tung xúc xắc thứ nhất và thứ hai. Khi đó, $X = X_1 + X_2$. Do đó,

$$E(X) = E(X_1 + X_2) = E(X_1) + E(X_2) = 3.5 + 3.5 = 7$$

Kỳ vọng của một hàm trên biến ngẫu nhiên

- Cho X là biến ngẫu nhiên rời rạc nhận các giá trị x_1, x_2, $\cdots$, x_n với các xác suất tương ứng $p(x_1)$, $p(x_2)$, $\cdots$, $p(x_n)$ và $h(X)$ là một biến ngẫu nhiên khác. Khi đó,

$$E(h(X)) = \sum_{i=1}^{n} h(x_i)p(x_i)$$

- Ví dụ: Cho X là kết quả của một lần tung xúc xắc và $Y = X^2$.

X	1	2	3	4	5	6
Y	1	4	9	16	25	36
pmf	1/6	1/6	1/6	1/6	1/6	1/6

$$E(Y) = E(X^2) = 1 \times \frac{1}{6} + 4 \times \frac{1}{6} + 9 \times \frac{1}{6} + 16 \times \frac{1}{6} + 25 \times \frac{1}{6} + 36 \times \frac{1}{6} = 15.17$$

Phương sai của biến ngẫu nhiên rời rạc

- Cho biến ngẫu nhiên X có kỳ vọng $E(X) = \mu$, phương sai (variance) của X được định nghĩa bởi:

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

- Độ lệch chuẩn (standard deviation) của X được được định nghĩa bởi:

$$SD(X) = \sigma = \sqrt{Var(X)}$$

- Nếu X nhận các giá trị x_1, x_2, $\cdots$, x_n với xác suất tương ứng $p(x_1)$, $p(x_2)$, $\cdots$, $p(x_n)$ thì

$$Var(X) = E((X - \mu)^2) = \sum_{i=1}^{n} p(x_i)(x_i - \mu)^2$$

Ví dụ

- Cho biến ngẫu nhiên X có bảng xác suất sau:

X	2	4	6
pmf	1/6	1/6	2/3

$$E(X) = \frac{1}{6} \times 2 + \frac{1}{6} \times 4 + \frac{2}{3} \times 6 = 5$$

$$Var(X) = \frac{1}{6} \times (2-5)^2 + \frac{1}{6} \times (4-5)^2 + \frac{2}{3} \times (6-5)^2 = \frac{7}{3}$$

Tính chất của phương sai

1. Nếu X và Y là độc lập, tức $P(X = a, Y = b) = P(X = a)P(Y = b)$, thì

$$Var(X + Y) = Var(X) + Var(Y)$$

2. Với mọi hằng số a, b thì

$$Var(aX + b) = a^2 Var(X)$$

3.

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

Kỳ vọng và phương sai của biến ngẫu nhiên rời rạc

Expected Value:		**Variance:**
Synonyms:	mean, average	
Notation:	$E(X)$, μ	$\mathrm{Var}(X)$, σ^2
Definition:	$E(X) = \sum_j p(x_j)x_j$	$E((X-\mu)^2) = \sum_j p(x_j)(x_j-\mu)^2$
Scale and shift:	$E(aX+b) = aE(X)+b$	$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$
Linearity:	(for any X, Y) $E(X+Y) = E(X)+E(Y)$	(for X, Y independent) $\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$
Functions of X:	$E(h(X)) = \sum p(x_j)\,h(x_j)$	
Alternative formula:		$\mathrm{Var}(X) = E(X^2) - E(X)^2 = E(X^2) - \mu^2$

Các phân bố rời rạc phổ biến

Các phân bố rời rạc phổ biến

1. Phân bố Bernoulli (Bernoulli distribution)
2. Phân bố nhị thức (Binomial distribution)
3. Phân bố hình học (geometric distribution)
4. Phân bố đều (uniform distribution)

Phân bố Bernoulli (Bernoulli distribution)

- Phân bố Bernoulli mô hình hóa một phép thử (trial) trong một thí nghiệm (experiment) có thể dẫn đến thành công (success) hoặc thất bại (failure).
- Biến ngẫu nhiên X có phân bố Bernoulli với tham số p, $X \sim Bernoulli(p)$, nếu
  1. X chỉ nhận hai giá trị là 1 (success) hoặc 0 (failure)
  2. $P(X = 1) = p$ và $P(X = 0) = 1 - p$.
- Nếu X có (tuân theo/được sinh từ) phân bố Bernoulli với tham số p, ta viết $X \sim Bernoulli(p)$.
- Ví dụ: Xét thí nghiệm tung đồng xu với xác suất mặt sấp là p, X là biến ngẫu nhiên nhận giá trị 1 nếu đồng xu có mặt sấp, và 0 nếu ngược lại. Khi đó, $X \sim Bernoulli(p)$.

value	a	0	1
pmf	p(a)	1 - p	p
cdf	F(a)	1 - p	1

Phân bố nhị thức (Binomial distribution)

- Phân bố nhị thức $Binomial(n, p)$ mô hình hóa số lần thành công (success) trong n phép thử Bernoulli(p) độc lập.
- Ví dụ: Xét thí nghiệm tung đồng xu với xác suất mặt sấp là p, X là biến ngẫu nhiên mô tả số lần đồng xu có mặt sấp trong n lần tung. Khi đó, $X \sim Binomial(n, p)$

value	a	k (k = 0, 1, 2, ..., n)
pmf	p(a)	$\binom{n}{k} p^k (1-p)^{n-k}$

Phân bố nhị thức (Binomial distribution)

- Một cặp vơ chồng dự định có 2 con. Giả định xác suất sinh con trai là $p = 0.5$.

1. Tìm xác suất cặp vơ chồng không có con trai?
2. Tìm xác suất cặp vơ chồng có 1 con trai?
3. Tìm xác suất cặp vơ chồng có 2 con trai?

Phân bố hình học (geometric distribution)

- Phân bố geometric mô hình hóa số lần thất bại trước khi gặp thành công (hoặc ngược lại) trong một chuỗi các phép thử Bernoulli (Bernoulli trial).
- Biến ngẫu nhiên X có phân bố hình học với tham số p, $X \sim geometric(p)$, nếu
  1. X nhận các giá trị 0, 1, 2, ...
  2. Hàm pmf của X được xác định bởi $p(k) = P(X = k) = (1-p)^k p$.

Phân bố hình học (geometric distribution)

- Tung đồng xu cân bằng cho đến khi xuất hiện mặt H.

1. Tìm xác suất xuất hiện mặt H ở lần tung đầu tiên?
2. Tìm xác suất xuất hiện mặt H ở lần tung thứ 4?
3. Số lần tung kỳ vọng cho đến khi xuất hiện mặt H là bao nhiêu?
4. Tìm xác suất xuất phải thực hiên hơn 5 lần tung để hiện mặt H?

Phân bố đều (uniform distribution)

- Phân bố đều mô hình thí nghiệm mà tất cả outcome đều có xác xuất như nhau.
- Biến ngẫu nhiên X có phân bố đều với tham số N, $X \sim uniform(N)$, nếu X nhận các giá trị $1, 2, \cdots, N$ với cùng xác suất $1/N$.

Kỳ vọng và phương sai của các phân bố phổ biến

Distribution	range X	pmf $p(x)$	mean $E(X)$	variance $\mathrm{Var}(X)$
Bernoulli(p)	0, 1	$p(0) = 1-p, \quad p(1) = p$	p	$p(1-p)$
Binomial(n, p)	0, 1,..., n	$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Uniform(n)	1, 2, ..., n	$p(k) = \frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Geometric(p)	0, 1, 2,...	$p(k) = p(1-p)^k$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

Biến ngẫu nhiên liên tục

Biến ngẫu nhiên liên tục

- Một biến ngẫu nhiên liên tục (continuous random variable) nhận một miền các giá trị liên tục (ví dụ: $[a, b], [0, \infty]$).
- Một biến ngẫu nhiên X là liên tục nếu có hàm $f(x)$ sao cho với mọi giá trị a, b, ta có

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

  - Hàm $f(x)$ được gọi là hàm mật độ xác xuất (probability density function - pdf).
- Hàm mật độ xác suất luôn thỏa mãn các tính chất sau:
  1. $f(x) \geq 0$
  2. $\int_{-\infty}^{\infty} f(x)dx = 1$
- Chú ý: Hàm mật độ xác suất không phải là hàm xác suất.

Cumulative distribution function (cdf)

- Hàm phân bố tích lũy (cumulative distribution function - cdf) của một biến ngẫu nhiên liên tục (continuous random variable) X là hàm

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

  Trong đó, $f(x)$ là hàm mật độ xác suất của X.

- Ta thường gọi X có phân bố $F(x)$ thay vì X có hàm phân bố tích lũy $F(x)$

Tính chất của hàm phân bố

1. $F(x) = P(X \leq x)$
2. $0 \leq F(x) \leq 1$
3. Nếu $a \leq b$ thì $F(a) \leq F(b)$
4. $\lim\limits_{x \to \infty} F(x) = 1, \lim\limits_{x \to -\infty} F(x) = 0$
5. $P(a \leq X \leq b) = F(b) - F(a)$
6. $F'(x) = f(x)$

Kỳ vọng của biến ngẫu nhiên liên tục

- Định nghĩa: Cho X là biến ngẫu nhiên liên tục với miền giá trị $[a, b]$ và hàm mật độ xác suất $f(x)$. Giá trị kỳ vọng của X được định nghĩa bởi:

$$E(X) = \int_a^b x f(x) dx$$

  - $f(x)dx$ biểu diễn xác suất X thuộc một lân cận của x với độ rộng vô cùng nhỏ dx
- Ví dụ 1: Cho $X \sim Uniform(0, 1) \rightarrow f(x) = 1$

$$E(X) = \int_0^1 x f(x) dx = \int_0^1 x dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$$

Kỳ vọng của biến ngẫu nhiên liên tục

- Ví dụ 2: Cho $X \sim Exp(\lambda) \rightarrow f(x) = \lambda e^{-\lambda x}, x \geq 0$

$$E(X) = \int_0^{\infty} x f(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx$$

  - Dùng tích phân từng phần (integration by parts) với $u = x, v' = \lambda e^{-\lambda x}$ $\Rightarrow u' = 1, v = -e^{-\lambda x}$

$$E(X) = \int_0^{\infty} x \lambda e^{-\lambda x} dx = -x e^{-\lambda x} \Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = 0 - \frac{e^{-\lambda x}}{\lambda} \Big|_0^{\infty} = \frac{1}{\lambda}$$

Tính chất của kỳ vọng

- Cho X, Y là hai biến ngẫu nhiên trên không gian mẫu Ω

$$E(X+Y) = E(X) + E(Y)$$

- Cho biến ngẫu nhiên X và các hằng số a, b

$$E(aX+b) = aE(X) + b$$

- Cho X là biến ngẫu nhiên liên tục với hàm mật độ $f(x)$ và $h(X)$ là một biến ngẫu nhiên khác. Khi đó,

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

Phương sai của biến ngẫu nhiên liên tục

- **Định nghĩa**: Cho X là biến ngẫu nhiên liên tục với kỳ vọng μ. **Phương sai** của X được định nghĩa bởi:

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

- **Tính chất**

1. Nếu X và Y là **độc lập** thì $Var(X + Y) = Var(X) + Var(Y)$
2. Với mọi hằng số $a, b, Var(aX + b) = a^2 Var(X)$
3. $Var(X) = E(X^2) - E(X)^2 = E(X^2) - \mu^2$

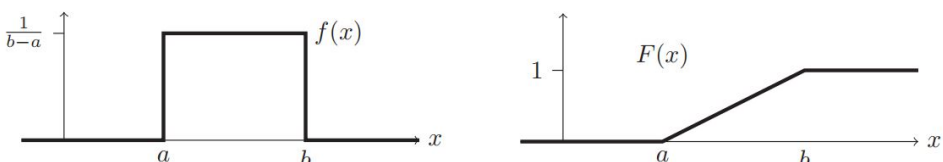
Các phân bố liên tục phổ biến

Các phân bố liên tục phổ biến

1. Phân bố đều (Uniform distribution)
2. Phân bố mũ (Exponential distribution)
3. Phân bố chuẩn (Normal distribution)

Phân bố đều (Uniform distribution)

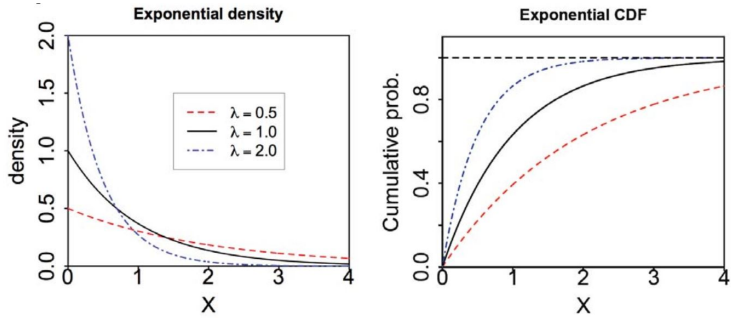
1. Tham số (parameter): a, b
2. Miền giá trị (range): $[a, b]$
3. Ký hiệu (notation): $uniform(a, b)$ hay $U(a, b)$
4. Hàm mật độ (pdf): $f(x) = \frac{1}{b-a}, a \leq x \leq b$
5. Hàm phân bố (cdf): $F(x) = \frac{x-a}{b-a}, a \leq x \leq b$
6. Mô hình (model): Tất cả outcome trong range $[a, b]$ có xác suất bằng nhau.



pdf and cdf for uniform(a,b) distribution.

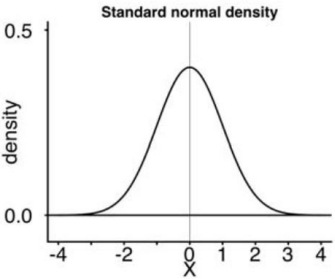
Phân bố mũ (Exponential distribution)

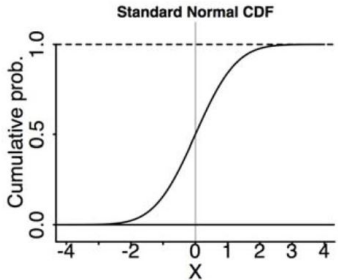
1. Tham số (parameter): λ
2. Miền giá trị (range): $[0, \infty)$
3. Ký hiệu (notation): $exponential(a, b)$ hay $exp(a, b)$
4. Hàm mật độ (pdf): $f(x) = \lambda e^{-\lambda x}, x \geq 0$
5. Hàm phân bố (cdf): $F(x) = 1 - e^{-\lambda x}, x \geq 0$
6. Mô hình (model): Thời gian chờ cho đến khi một quá trình liên tục thay đổi trạng thái.

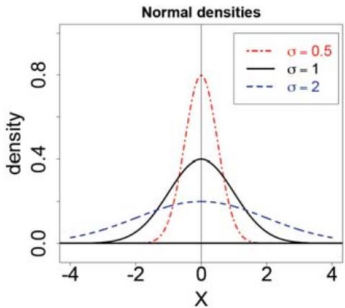


Phân bố chuẩn (Normal distribution)

1. Tham số (parameter): μ, σ
2. Miền giá trị (range): $(-\infty, \infty)$
3. Ký hiệu (notation): $normal(\mu, \sigma^2)$ hay $N(\mu, \sigma^2)$
4. Hàm mật độ (pdf): $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
5. Hàm phân bố (cdf): Không có công thức (sử dụng bảng để tính F(x))
6. Mô hình (model): Sai số đo đạc, IQ, chiều cao, ...







Phân bố chuẩn tắc (standard normal distribution)

- Phân bố chuẩn với $\mu = 0, \sigma = 1$ được gọi là standard normal distribution, ký hiệu: $N(0, 1)$.
- Các ký hiệu sau được dành riêng cho standard normal distribution
  - Z để chỉ biến ngẫu nhiên có standard normal distribution,
  - $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ để chỉ standard normal density function,
  - $\Phi(z)$ để chỉ standard normal cumulative distribution function.

Chỉ số z (z score)

- Điểm SAT (Scholastic Assessment Test) và điểm ACT (American College Testing) thường được dùng để xét tuyển vào đại học Mỹ.
- Cho biết điểm SAT tuân theo phân bố chuẩn với trung bình là 1500 và độ lệch chuẩn là 300.
- Cho biết điểm ACT tuân theo phân bố chuẩn với trung bình là 21 và độ lệch chuẩn là 5.
- Cho biết A được 1800 điểm SAT, B được 24 điểm ACT.
- Câu hỏi: Thí sinh nào (A hay B) có năng lực đầu vào tốt hơn dựa trên hai bài thi của họ?

Chỉ số z (z score)

- Ta không thể so sánh điểm thô của hai thí sinh vì chúng khác thang đo. Do đó, ta cần chuẩn hóa chúng về cùng thang đo rồi mới có thể so sánh.
- z score hay standardized score cho phép ta đo điểm số của họ lệch với điểm trung bình bao nhiêu lần độ lệch chuẩn.
  - z score độc lập với đơn vị đo.

$$z = \frac{x - mean}{sd}$$

$$z_A = \frac{1800 - 1500}{300} = 1 \qquad z_B = \frac{24 - 21}{5} = 0.6$$

Percentile

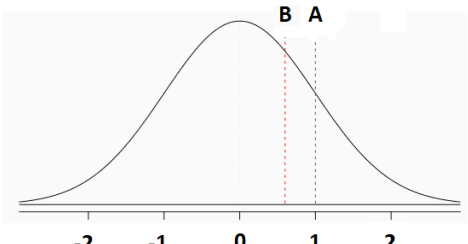
- Percentile là phần trăm (tỷ lệ) các giá trị của một biến nhỏ hơn hoặc bằng một giá trị cho trước.

```
from scipy.stats import norm
norm.cdf(1800, loc=1500, scale=300)
```

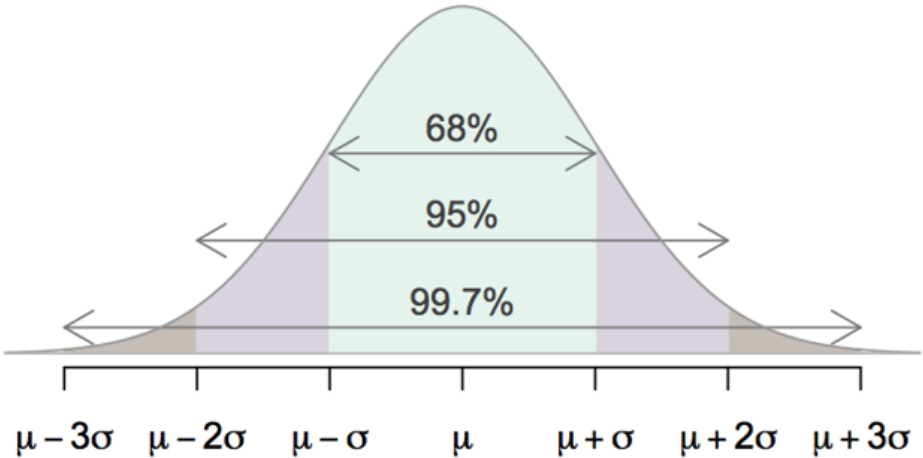
```
## 0.8413447460685429
```

```
norm.cdf(24, loc=21, scale=5)
```

```
## 0.7257468822499265
```

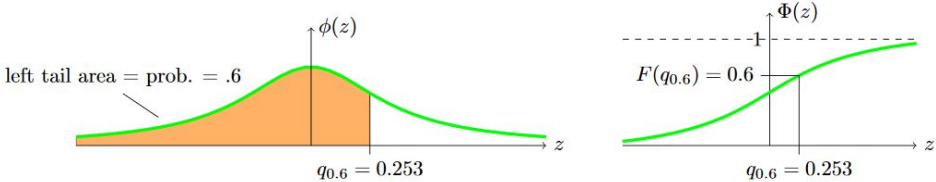


Qui tắc 68-95-99.7



Quantile

- Định nghĩa: p quantile của X là giá trị q_p sao cho $F(q_p) = P(X \leq q_p) = p$.
  - Median là $q_{0.5}$ (0.5 quantile)
  - Quantile có sự tương ứng với percentile, ví dụ, 60^{th} percentile giống với 0.6 quantile.
- Ví dụ: Cho $X \sim N(0,1), q_{0.6} = stats.norm.ppf(0.6, 0, 1) = 0.25335$



$q_{0.6}$: left tail area $= 0.6 \Leftrightarrow F(q_6) = 0.6$

Định lý giới hạn trung tâm

Quách Đình Hoàng

2022/04/04

Nội dung

Luật số lớn (The law of large number - LoLN)

Định lý giới hạn trung tâm (central limit theorem - CLT)

Phân bố mẫu và sai số chuẩn (sampling distribution and standard error)

Luật số lớn (The law of large number - LoLN)

Independent and identically distributed (IID)

- Giả sử các biến ngẫu nhiên $X_1, X_2, \cdots$ là độc lập và có phân bố giống nhau. Các X_i có cùng mean μ và standard deviation σ. Khi đó ta gọi các X_i là iid (independent and identically distributed).
  - Ta thường dùng X để thể hiện biến ngẫu nhiên có cùng phân bố với các biến ngẫu nhiên $X_1, X_2, \cdots$
- Ví dụ: Tung đồng xu cân bằng 100 lần. Gọi X_i là biến ngẫu nhiên có giá trị 1 nếu đồng xu thứ i mặt sấp và 0 nếu mặt ngửa. Khi đó,
  - X_i là độc lập và có phân bố giống nhau (iid), ta viết
  - X_i có phân bố Bernoulli với xác suất $p = 0.5$,

$$X_i \sim^{iid} Bernoulli(0.5)$$

Mẫu ngẫu nhiên (random sample)

- Định nghĩa: Ta nói $X_1, X_2, \cdots, X_n$ là một mẫu ngẫu nhiên (random sample) có kích cỡ n từ một quần thể nếu các X_i là iid và phân bố chung của chúng giống với phân bố của quần thể.
- Một mẫu ngẫu nhiên (random sample) được tạo bằng cách:
  1. Thực hiện lấy mẫu có hoàn lại (sampling with replacement)
  2. Mỗi X_i được sinh từ quần thể có xác suất như nhau
  - Bước 1 làm các X_i là độc lập, bước 2 làm các X_i có cùng phân bố.
- Nếu ở bước 1 ta thực hiện lấy mẫu không hoàn lại (sampling without replacement) thì ta gọi mẫu đó là mẫu ngẫu nhiên đơn giản (simple random sample).
- Trong nhiều trường hợp, quần thể thường rất lớn, thậm chí vô hạn, rất khó có thể lấy mẫu cùng một đối tượng hai lần. Do đó, mẫu ngẫu nhiên đơn giản cũng là mẫu ngẫu nhiên.

Luật số lớn (the law of large number - LoLN)

- Cho các biến ngẫu nhiên $X_1, X_2, \cdots, X_n$ là iid. X_i có kỳ vọng μ và độ lệch chuẩn σ. Khi đó, trung bình mẫu $\overline{X_n}$ là biến ngẫu nhiên được định nghĩa bởi:

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

- Luật số lớn (law of large number - LoLN)

$$\lim_{n\to\infty} P(|\overline{X_n} - \mu| < \epsilon) = 1, \forall \epsilon > 0$$

  - Khi n lớn, xác suất $\overline{X_n}$ gần μ tiến tới 1.

Ví dụ

- Xét thí nghiệm tung đồng xu cân bằng, gọi X_i là biến ngẫu nhiên có giá trị 1 nếu đồng xu thứ i mặt sấp và 0 nếu mặt ngửa. Khi đó, $X_i \sim^{iid} Bernoulli(0.5), \mu = 0.5$. Gọi $\overline{X_n}$ là tỷ lệ mặt sấp sau n lần tung.
- Luật số lớn nói rằng, khi n đủ lớn, $\overline{X_n}$ sẽ gần với $\mu = 0.5$.
- Ta sẽ dùng Python để tính xác suất để $\overline{X_n}$ gần với $\mu = 0.5$ với bán kính 0.01.

$$P(|\overline{X_n} - 0.5| \leq 0.01) \Leftrightarrow P(0.49 \leq \overline{X_n} \leq 0.51)$$

```
from scipy.stats import binom
binom.cdf(510,1000,0.5) - binom.cdf(489,1000,0.5) # n=1000
```

```
## 0.49333995737505015
```

```
binom.cdf(5100,10000,0.5) - binom.cdf(4899,10000,0.5) # n=10000
```

```
## 0.9555742009542736
```

Định lý giới hạn trung tâm (central limit theorem - CLT)

Sự chuẩn hóa (standardization)

- Cho biến ngẫu nhiên X với mean μ và standard deviation σ, ta định nghĩa sự chuẩn hóa (standardization) của X là biến ngẫu nhiên mới

$$Z = \frac{X - \mu}{\sigma}$$

- Z có mean $\mu = 0$ và standard deviation $\sigma = 1$.
- Nếu $X \sim N(\mu, \sigma)$ thì $Z \sim N(0, 1)$.

Định lý giới hạn trung tâm (central limit theorem - CLT)

- Cho các biến ngẫu nhiên $X_1, X_2, \cdots, X_n$ là iid. Các X_i có cùng mean μ và standard deviation σ. Đặt S_n và $\overline{X_n}$ là tổng và trung bình của các X_i.

$$S_n = \sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$$

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

Khi đó:

$$E(S_n) = n\mu, \quad Var(S_n) = n\sigma^2, \quad \sigma_{S_n} = \sqrt{n}\sigma$$

$$E(\overline{X_n}) = \mu, \quad Var(\overline{X_n}) = \frac{\sigma^2}{n}, \quad \sigma_{\overline{X_n}} = \frac{\sigma}{\sqrt{n}}$$

- Chuẩn hóa của S_n và $\overline{X_n}$ là giống nhau

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}$$

Định lý giới hạn trung tâm (central limit theorem - CLT)

- Khi n lớn,

$$\overline{X_n} \approx N(\mu, \sigma/\sqrt{n}), \quad S_n \approx N(n\mu, \sigma\sqrt{n}), \quad Z_n \approx N(0,1)$$

- Định lý giới hạn trung tâm (central limit theorem - CLT) dạng chính thức

$$\lim_{n\to\infty} P\left(\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} < z\right) = \Phi(z)$$

  - $\Phi(z)$ là giá trị của standard normal cdf tại z.
  - Khi n lớn, phân bố của $\frac{\overline{X_n}-\mu}{\sigma/\sqrt{n}}$ hội tụ đến $N(0,1)$ hay $\overline{X_n}$ hội tụ đến $N(\mu, \sigma^2/n)$.

Ví dụ

- Tung đồng xu cân bằng 100 lần, gọi X_i là biến ngẫu nhiên có giá trị 1 nếu đồng xu thứ i mặt sấp và 0 nếu mặt ngửa. Khi đó, $X_i \sim^{iid} Bernoulli(0.5), \mu = 0.5$. Gọi $\overline{S}$ là số mặt sấp sau 100 lần tung, $S = X_1 + X_2 + \cdots + X_{100}$.
- Ta có: $E(X_i) = 0.5, Var(X_i) = 0.25$
- $E(S) = 100 * 0.5 = 50, Var(S) = 100 * 0.25 = 25, \sigma_S = \sqrt{Var(S)} = 5$
- Tính $P(S > 60)$?
  - Theo CLT, $S \sim N(50, 5) \rightarrow \frac{S-50}{5} \sim N(0, 1)$

$$P(S > 60) = P\left(\frac{S-50}{5} > \frac{60-50}{5}\right) \approx P(Z > 2) = 1 - \Phi(2)$$

```
from scipy.stats import norm
1 - norm.cdf(2,0,1)
```

```
## 0.02275013194817921
```

Phân bố mẫu và sai số chuẩn (sampling distribution and standard error)

Phân bố mẫu (sampling distribution) và sai số chuẩn (standard error)

- Phân bố mẫu (sampling distribution)
  - Lặp lại việc lấy mẫu ngẫu nhiên của biến X nhiều lần (1000 lần chẳng hạn), mỗi lần như vậy ta sẽ tạo ra một mẫu ngẫu nhiên.
  - Với mỗi mẫu, ta tính giá trị thống kê mình quan tâm (trung bình chẳng hạn).
  - Ta sẽ thu được một phân bố mẫu của giá trị thống kê đã tính (phân bố mẫu của trung bình chẳng hạn).
- Sai số chuẩn (standard error)
  - Ta gọi độ lệch chuẩn của phân bố mẫu là sai số chuẩn.

Phân bố mẫu của tỷ lệ

- Định lý: Giả sử X là biến phân loại (categorical variable) có tỷ lệ quần thể (population propotion) là p. Nếu ta lặp lại việc lấy mẫu ngẫu nhiên biến X với cỡ mẫu n vô hạn lần thì phân bố mẫu của tỷ lệ $\hat{p_n}$ tuân theo phân bố chuẩn với giá trị trung bình là p và độ lệch chuẩn $\sqrt{\frac{p(1-p)}{n}}$ nếu $np \geq 10$ và $n(1-p) \geq 10$.
- Ví dụ: cho biến phân loại X có $p = 0.6$, $n = 25$, khi đó: $np \geq 10$ và $n(1-p) \geq 10$. Do đó, phân bố mẫu của $\hat{p_n}$ tuân theo phân bố chuẩn với trung bình là $E(\hat{p_n}) = 0.6$ và độ lệch chuẩn là $SD(\hat{p_n}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{25}} = 0.097$.

Phân bố mẫu của trung bình

- Định lý: Giả sử X là biến số (numerical variable) có trung bình quần thể (population mean) là μ và độ lệch chuẩn quần thể (population standard deviation) là σ. Nếu ta lặp lại việc lấy mẫu ngẫu nhiên biến X với cỡ mẫu n vô hạn lần thì phân bố mẫu của $\overline{X_n}$ tuân theo phân bố chuẩn với giá trị trung bình là μ và độ lệch chuẩn $\frac{\sigma}{\sqrt{n}}$ nếu $\overline{X_n}$ có phân bố chuẩn hoặc cỡ mẫu n đủ lớn (thường $n \geq 30$ là đủ lớn).
- Ví dụ: cho biến số X có $\mu = 3000$, $\sigma = 500$, và cỡ mẫu $n = 100$. Phân bố mẫu của $\overline{X_n}$ tuân theo phân bố chuẩn với trung bình là $E(\overline{X_n}) = 3000$ và độ lệch chuẩn là $SD(\overline{X_n}) = \frac{\sigma}{\sqrt{n}} = \frac{500}{\sqrt{100}} = 50$.

Statistics Notes

(Dựa trên chương 5 sách OpenIntro Statistics, 4th Edition)

Quách Đình Hoàng

2021/10/18

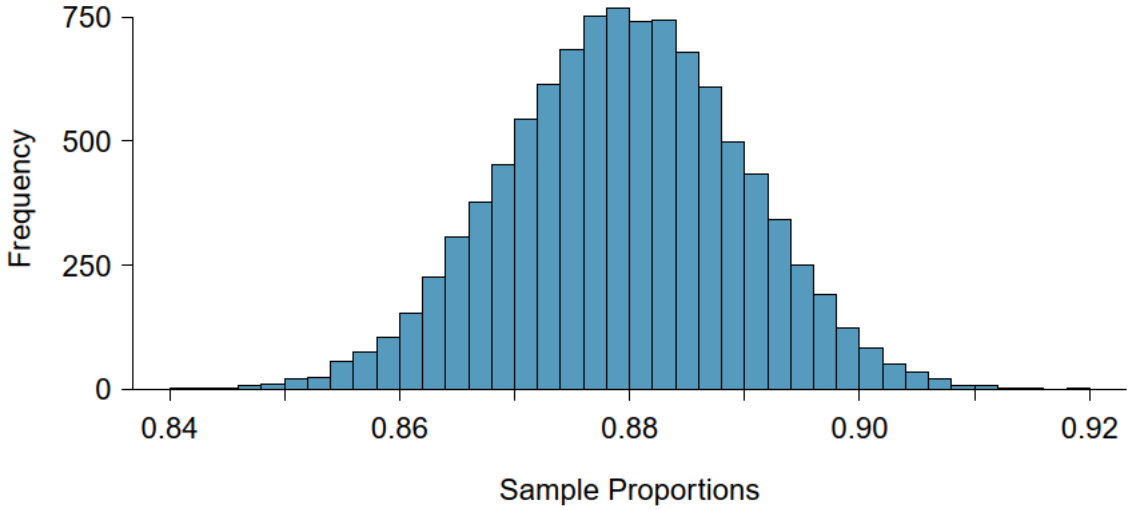
1. Các cơ sở cho suy luận

1.1. Phân bố mẫu (sampling distribution)

Chúng ta thường quan tâm đến các tham số quần thể (population parameters). Vì rất khó (hoặc không thể) để thu thập dữ liệu của cả quần thể, chúng ta sử dụng các thống kê mẫu (sample statistics) làm các ước lượng điểm (point estimates) cho các tham số quần thể chưa biết mà ta quan tâm. Các số liệu thống kê mẫu hầu như luôn khác nhau giữa các mẫu. Việc xác định các thống kê mẫu khác nhau như thế nào cung cấp một cách để ước tính biên độ sai số (margin of error) tương ứng với ước lượng điểm của chúng ta. Nhưng trước khi chúng ta xác định sự biến thiên giữa các mẫu, chúng ta hãy cố gắng hiểu cách thức và lý do tại sao các ước lượng điểm khác nhau giữa các mẫu.

- Giả sử chúng ta lấy mẫu ngẫu nhiên 1.000 người lớn từ mỗi tiểu bang của Hoa Kỳ. Bạn mong đợi trung bình chiều cao của các mẫu giống nhau, hơi khác hoặc rất khác nhau? Thực sự, chúng ta không kỳ vọng chúng giống nhau, nhưng mong đợi chúng chỉ khác nhau đôi chút.
- Giả sử bạn không thể tiếp cận được tất cả những người trưởng thành ở Mỹ. Để ước tính tỷ lệ người Mỹ trưởng thành ủng hộ việc mở rộng việc khai thác năng lượng mặt trời, bạn có thể lấy mẫu từ quần thể tất cả những người trưởng thành ở Mỹ và sử dụng tỷ lệ mẫu của bạn làm dự đoán cho tỷ lệ quần thể chưa biết.
  - Lấy mẫu có hoàn lại, 1000 người Mỹ trưởng thành và ghi lại xem họ có đồng ý việc khai thác năng lượng mặt trời hay không.
  - Tìm tỷ lệ mẫu.
  - Giả sử bạn lặp lại quá trình này nhiều lần và vẽ biểu đồ kết quả. Những gì bạn vừa xây dựng được gọi là phân bố mẫu (sampling distribution).

Hình dưới đây là một biểu đồ tần số (histogram) của $10,000$ tỷ lệ mẫu (sample proportion), với mỗi mẫu được lấy ngẫu nhiên từ một quần thể (population) có tỷ lệ quần thể (population proportion) là 0.88 và cỡ mẫu (sample size) là $n = 1000$.



Biểu đồ (phân bố mẫu) trông đối xứng và có dạng hình chuông. Giá trị trung tâm của phân bố khoảng 0.88.

Trong thực tế, chúng ta không bao giờ thực sự quan sát được phân bố mẫu, nhưng sẽ rất hữu ích khi luôn nghĩ về một ước lượng điểm đến từ một phân bố giả định như vậy. Hiểu được phân bố mẫu sẽ giúp chúng ta mô tả đặc điểm và ý nghĩa về các ước lượng điểm mà chúng ta quan sát được.

1.1.1 Định lý giới hạn trung tâm (Central Limit Theorem) cho tỷ lệ

Định lý Các tỷ lệ mẫu sẽ có phân bố gần như phân bố chuẩn với giá trị trung bình bằng với tỷ lệ quần thể, p và sai số chuẩn bằng $\sqrt{\frac{p(1-p)}{n}}$.

$$\hat{p} \sim N(mean = p, SE = \sqrt{\frac{p(1-p)}{n}})$$

Không phải ngẫu nhiên mà phân bố mẫu mà chúng ta đã thấy ở trên là đối xứng và tập trung quanh tham số quần thể thực sự.

Khi n tăng thì SE sẽ giảm, tức sự biến thiên của các $\hat{p}$ sẽ nhỏ hơn.

Điều kiện Một số điều kiện sau cần thỏa để có thể áp dụng định lý giới hạn trung tâm:

1. Sự độc lập (Independence)

- Các đối tượng được lấy mẫu phải độc lập. Điều này rất khó xác minh, nhưng nhiều khả năng sẽ đạt được nếu:
  - Lấy mẫu / chỉ định ngẫu nhiên được sử dụng, và
  - Lấy mẫu không hoàn lại với cỡ mẫu $n < 0.1N$ (N số đối tượng của quần thể).

2. Cỡ mẫu (Sample size)

- Cần có ít nhất 10 thành công (success) và 10 thất bại (failure) trong mẫu.
  - Điều này khó xác minh nếu ta không biết tỷ lệ quần thể (hoặc không thể giả định một giá trị cho nó). Trong những trường hợp đó, chúng ta cần số thành công và thất bại quan sát được ít nhất là 10.

Trường hợp p chưa biết Định lý giới hạn trung tâm (CLT) nói rằng

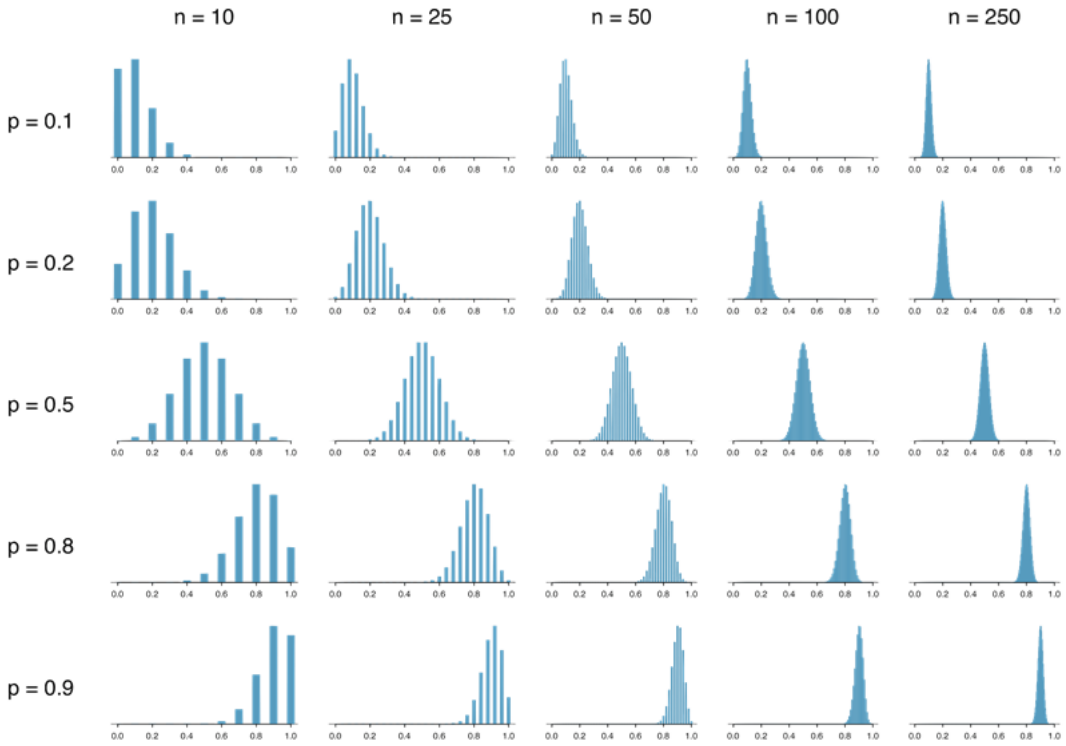
$$SE = \sqrt{\frac{p(1-p)}{n}}$$

với điều kiện $np \geq 10$ và $n(1-p) \geq 10$.

Tuy nhiên, nếu ta không biết tỷ lệ quần thể p, ta có thể dùng tỷ lệ mẫu $\hat{p}$ để thay thế.

Trường hợp np hoặc $n(1-p)$ nhỏ Giả sử chúng ta có một quần thể mà tỷ lệ quần thể thực là $p = 0.05$ và chúng ta lấy mẫu ngẫu nhiên có cỡ mẫu $n = 50$ từ quần thể này. Điều kiện thành công-thất bại không được đáp ứng ($50 \times 0.05 = 2.5 < 10$), vì vậy chúng ta không mong đợi phân bố mẫu gần gần với phân bố chuẩn.

Biểu đồ dưới đây mô tả phân bố mẫu cho các giá trị n và p khác nhau.



Ta quan sát thấy một số điều sau:

- Khi np hoặc $n(1-p)$ nhỏ, phân bố mẫu rời rạc hơn.
- Khi $np < 10$ hoặc $n(1-p) < 10$, phân bố mẫu bị lệch hơn.
- Khi cả np và $n(1-p)$ lớn hơn, phân bố mẫu gần với phân bố chuẩn hơn.
- Khi cả np và $n(1-p)$ rất lớn, phân bố mẫu càng trơn (không còn rời rạc) và trông giống như một phân phối chuẩn.

Chiến lược sử dụng thống kê mẫu để ước tính tham số quần thể khá phổ biến và đó là chiến lược mà chúng ta có thể áp dụng cho các thống kê khác ngoài tỷ lệ.

1.1.2. Định lý giới hạn trung tâm (Central Limit Theorem) cho trung bình

Định lý Phân bố của trung bình mẫu (sample mean) xấp xỷ phân bố chuẩn

$$\overline{x} \sim N(mean = \mu, SE = \frac{\sigma}{\sqrt{n}})$$

với SE là sai số chuẩn (sandard eror), được định nghĩa là độ lệch chuẩn (standard deviation) của phân bố mẫu. Nếu độ lệch chuẩn quần thể σ chưa biết, ta sử dụng độ lệch chuẩn mẫu s để thay thế.

Khi n tăng thì SE sẽ giảm, tức sự biến thiên của các $\overline{x}$ sẽ nhỏ hơn.

Điều kiện Một số điều kiện sau cần thỏa để có thể áp dụng định lý giới hạn trung tâm:

1. Sự độc lập (Independence)

- Các đối tượng được lấy mẫu phải độc lập. Điều này rất khó xác minh, nhưng nhiều khả năng sẽ đạt được nếu:
  - Lấy mẫu / chỉ định ngẫu nhiên được sử dụng, và
  - Lấy mẫu không hoàn lại với cỡ mẫu $n < 0.1N$ (N số đối tượng của quần thể).

2. Cỡ mẫu/độ lệch (Sample size/skew)

- Quần thể có phân bố chuẩn hoặc nếu phân bố quần thể bị lệch thì cỡ mẫu phải lớn.
  - Phân bố quần thể càng lệch thì cỡ mẫu chúng ta cần để có thể áp dụng định lý CLT phải càng lớn
  - Đối với phân phối lệch vừa phải, $n > 30$ là quy tắc đơn giản thường được sử dụng

Điều này cũng khó xác minh đối với quần thể, nhưng chúng ta có thể kiểm tra nó bằng cách sử dụng dữ liệu mẫu và giả định rằng mẫu phản ánh tổng thể.

1.2. Khoảng tin cậy cho tỷ lệ (Confidence Intervals for a Proportion)

1.2.1. Khoảng tin cậy (Confidence intervals)

Một khoảng giá trị hợp lý cho tham số quần thể được gọi là khoảng tin cậy (confidence interval).

Chỉ sử dụng một thống kê mẫu để ước tính một tham số giống như câu cá trong hồ nước âm u bằng một ngọn giáo và sử dụng khoảng tin cậy giống như câu cá bằng lưới. Chúng ta có thể phóng một ngọn giáo vào nơi chúng ta nhìn thấy một con cá nhưng chúng ta có thể sẽ trượt. Nếu chúng ta quăng lưới ở khu vực đó, chúng ta có cơ hội bắt được cá.

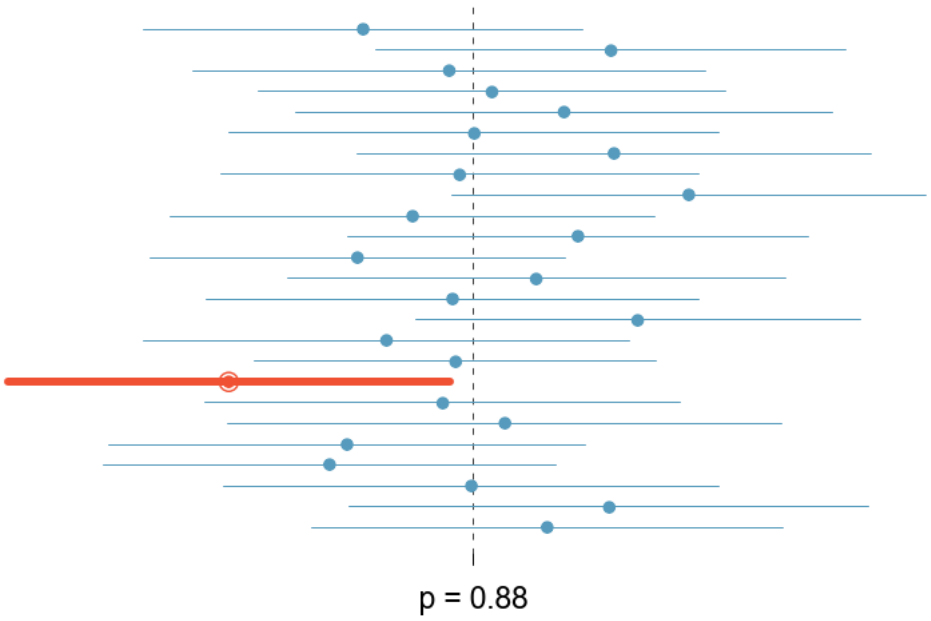
Nếu chúng ta chỉ báo cáo một ước lượng điểm, chúng ta có thể sẽ không dự đoán đúng tham số quần thể. Nếu chúng ta báo cáo một khoảng các giá trị hợp lý, chúng ta có một cơ hội tốt hơn để nắm bắt tham số.

1.2.2. Ý nghĩa của khoảng tin cậy 95%

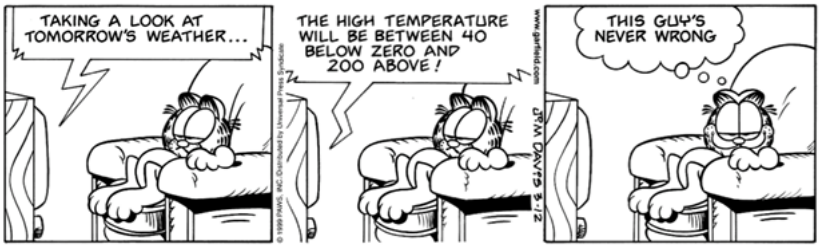
Giả sử ta lấy nhiều mẫu ngẫu nhiên từ quần thể và xây dựng một khoảng tin cậy 95% (95% CI) cho mỗi mẫu dùng công thức:

$$\text{point estimate} \pm 1.96 \times SE$$

thì khoảng 95% khoảng tin cậy đó sẽ chứa tỷ lệ quần thể thật sự (true population proportion).



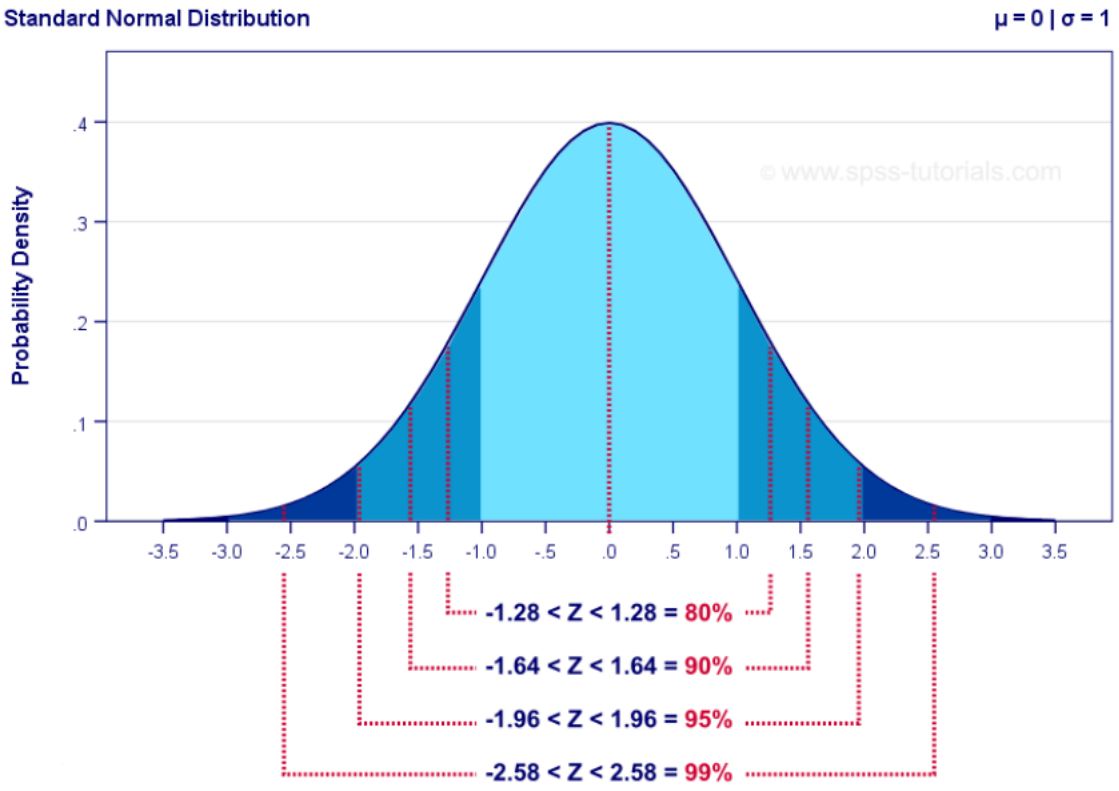
Nếu chúng ta muốn chắc chắn hơn rằng khoảng tin cậy của chúng ta sẽ bắt được (chứa) tham số quần thể, tức là tăng mức độ tin cậy của chúng ta, chúng ta có thể sử dụng một khoảng rộng hơn. Tuy nhiên, nếu khoảng tin cậy quá rộng nó có thể không cung cấp cho ta thông tin thật sự hữu ích.



Trong khoảng tin cậy 95%, point estimate $\pm 1.96 \times SE$ hay tổng quát hơn, point estimate $\pm z^* \times SE$, $z^* \times SE$ là biên lỗi (margin of error). Ứng với một mẫu cụ thể, biên lỗi thay đổi khi mức độ tin cậy (confidence level) thay đổi.

Khi thay đổi mức độ tin cậy, ta cần thay đổi z^* trong công thức trên. Các mức độ tin cậy phổ biến là 90%, 95%, 98%, 99%. Với mức độ tin cậy 95%, $z^* = 1.96$.

Dùng phân bố chuẩn, ta có thể tìm giá trị z^* phù hợp cho bất cứ mức độ tin cậy nào.



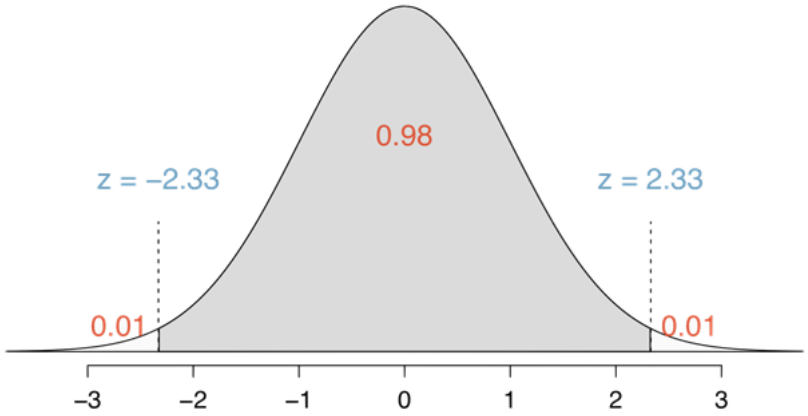
Trong R, ta có thể dùng hàm `qnorm()` để tìm giá trị z^*.

Cho biết giá trị z^* cho khoảng tin cậy 98%:

$$z^* = qnorm(0.01, lower.tail = F) \sim 2.33$$

hoặc

$$z^* = qnorm(0.99) \sim 2.33$$



1.2.3. Diễn giải khoảng tin cậy

Các khoảng tin cậy luôn luôn là các phát biểu về quần thể. Một **cách diễn giải sai phổ biến** cho khoảng tin cậy 95% là xem khoảng tin cậy đó chứa tham số quần thể với xác xuất 95%.

Giả sử khoảng tin cậy 90% cho khảo sát về năng lượng mặt trời là $(87.1\%, 90.4\%)$. Điều này có nghĩa là, "**chúng ta** 90% **tin rằng** 87.1% **đến** 90.4% **người lớn ở Mỹ đồng ý mở rộng việc sử dụng năng

lượng mặt trời". Phát biểu này dễ bị nhầm lẫn với phát biểu sai "**xác xuất** 90% **là từ** 87.1% **đến** 90.4% **người lớn ở Mỹ đồng ý mở rộng việc sử dụng năng lượng mặt trời**".

1.3. Kiểm định giả thuyết

1.3.1. Kiểm định giả thuyết cho một tỷ lệ

- Giả sử chúng ta muốn biết liệu có sự phân biệt giới tính khi đề bạt chức vụ. Bảng sau mô tả dữ liệu ta thu thập được.

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

Từ bảng dữ liệu trên ta tính được:

$$\hat{p}_{male} = 21/24 = 0.88$$

$$\hat{p}_{female} = 14/24 = 0.58$$

Từ kết quả trên có thể các các giải thích:

- Sự đề bạt (thăng tiến trong công việc) và giới tính là độc lập, không phân biệt giới tính, sự khác biệt về tỷ lệ quan sát được chỉ đơn giản là do ngẫu nhiên. → null hypothesis (không có gì xảy ra)
- Sự thăng tiến và giới tính phụ thuộc vào nhau, có sự phân biệt giới tính, sự khác biệt về tỷ lệ quan sát được không phải do ngẫu nhiên. → alternative hypothesis (điều gì đó đang diễn ra)

Hypothesis testing framework

- Chúng ta bắt đầu với giả thuyết vô hiệu (null hypothesis) H_0 mô tả rằng không có sự khác biệt.
- Chúng ta cũng có một giả thuyết thay thế (alternative hypothesis) H_A đại diện cho câu hỏi nghiên cứu của chúng ta, tức là những gì chúng ta đang thử nghiệm.
- Chúng ta tiến hành kiểm định giả thuyết với giả định rằng giả thuyết vô hiệu là đúng, thông qua mô phỏng hoặc các phương pháp truyền thống dựa trên định lý giới hạn trung tâm.
- Nếu kết quả kiểm định cho thấy dữ liệu không cung cấp bằng chứng thuyết phục cho giả thuyết thay thế, chúng ta giữ lại giả thuyết vô hiệu. Nếu ngược lại thì chúng ta bác bỏ giả thuyết vô hiệu để ủng hộ giả thuyết thay thế.

Kiểm định giả thuyết dùng khoảng tin cậy Giả sử ta muốn biết số phần trăm người lớn trả lời đúng câu hỏi có bao nhiêu phần trăm trẻ em trên thế giới được tiêm phòng một bệnh nào đó (20% hay 50%, hay 80%) có khác với 33% hay không? Khảo sát ngẫu nhiên 50 người lớn ta thu được 24% câu trả lời đúng (80% có trẻ em trên thế giới được tiêm phòng một bệnh nào đó). Dữ liệu quan sát được này có cung cấp bằng chứng mạnh rằng tỷ lệ người trưởng thành trả lời đúng câu hỏi chênh lệch so với $33,3\%$ hay không?

Chúng ta biết rằng có sự thay đổi từ mẫu này sang mẫu khác, và không chắc rằng tỷ lệ mẫu, $\hat{p}$, sẽ chính xác bằng p, nhưng chúng ta muốn đưa ra kết luận về p. Chúng ta muốn biết liệu sự khác biệt giữa 24% so với $33,3\%$ này có đơn giản là do ngẫu nhiên không, hay dữ liệu cung cấp bằng chứng chắc chắn rằng tỷ lệ dân số đã khác xa so với $33,3\%$?

Chúng ta đã biết cách định lượng độ không chắc chắn trong ước tính của mình bằng cách sử dụng khoảng tin cậy. Phương pháp tương tự để đo độ biến thiên có thể hữu ích cho việc kiểm định giả thuyết.

Các điều kiện để p có phân bố xấp xỉ phân bố chuẩn đều thỏa: dữ liệu đến từ một mẫu ngẫu nhiên đơn giản (thỏa mãn tính độc lập) và $n\hat{p} = 12 \geq 10$ và $n(1-\hat{p}) = 38 \geq 10$ (thỏa mãn cỡ mẫu).

Để xây dựng khoảng tin cậy, ta cần ước lượng điểm ($\hat{p} = 0.24$), critical value cho khoảng tin cậy 95% ($z^* = 1.96$), và sai số chuẩn của $\hat{p}$ ($SE = \sqrt{\hat{p}(1-\hat{p})/n} = 0.06$). Khoảng tin cậy 95% cho p là:

$$\hat{p} \pm z^* \times SE_{\hat{p}} = 0.24 \pm 1.96 \times 0.06 = (0.122, 0.358)$$

Chúng ta 95% tin tưởng rằng tỷ lệ tất cả người trưởng thành trả lời đúng câu hỏi về tiêm chủng cho trẻ sơ sinh là từ $12,2\%$ đến $35,8\%$.

Bởi vì giá trị null (null value) trong kiểm định giả thuyết là $p_0 = 0,333$, thuộc khoảng tin cậy, chúng ta không thể nói rằng giá trị này là không đáng tin cậy. Điều đó nghĩa là, dữ liệu không cung cấp đủ bằng chứng xác thực để bác bỏ giả thuyết vô hiệu, H_0.

Đây là một cách tiếp cận nhanh chóng và dễ hiểu để kiểm định giả thuyết, nhưng nó không cho chúng ta biết khả năng xảy ra các kết quả nhất định theo giả thuyết vô hiệu (p-value).

Các lỗi khi ra quyết định (decision errors) Các kiểm định giả thuyết không hoàn hảo.

Trong hệ thống tòa án, người vô tội đôi khi bị kết án sai, và người có tội đôi khi được tự do. Tương tự, chúng ta cũng có thể đưa ra một quyết định sai lầm trong các kiểm đỉnh giả thuyết thống kê.

Sự khác biệt là chúng ta có các công cụ cần thiết để định lượng tần suất chúng ta mắc lỗi trong thống kê.

Ta có hai giả thuyết đối ngược: giá trị vô hiệu và giả thuyết thay thế. Trong một kiểm định giả thuyết, chúng ta đưa ra quyết định có thể đúng, nhưng lựa chọn của chúng ta cũng có thể không chính xác. Có 4 trường hợp có thể xảy ra được mô tả ở bảng sau:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	**Type 1 Error**
	H_A true	**Type 2 Error**	okay

Một lỗi loại 1 (Type 1 Error) là lỗi xảy ra khi ta bác bỏ giả thuyết vô hiệu H_0 khi nó thật sự đúng. Một lỗi loại 2 (Type 2 Error) là lỗi xảy ra khi ta không thể bác bỏ giả thuyết vô hiệu H_0 trong khi giả thuyết thay thế H_A là đúng.

Nếu chúng ta nghĩ kiểm định giả thuyết như một phiên tòa hình sự thì sẽ hợp lý khi định khung phán quyết theo các giả thuyết vô hiệu và giả thuyết thay thế:

- H_0: Bị cáo vô tội
- H_A: Bị cáo có tội

Khi đó:

- Tuyên bố bị cáo vô tội khi họ thực sự có tội là lỗi loại 2.
- Tuyên bố bị cáo có tội khi họ thực sự vô tội là lỗi loại 1.

Trong nhiều trường hợp, lỗi loại 1 là nghiêm trọng hơn và ta muốn tránh.

> "Thà mười người có tội trốn thoát còn hơn một người vô tội phải chịu đựng" - William Blackstone

Kiểm định giả thuyết được xây dựng xung quanh việc bác bỏ hoặc không thể bác bỏ giả thuyết vô hiệu. Ta không bác bỏ giả thuyết vô hiệu trừ khi dữ liệu quan sát được cung cấp một bằng chứng mạnh mẽ hỗ trợ cho điều đó. Làm thế nào để biết được khi nào bằng chứng đủ mạnh?

Nguyên tắc chung được sử dụng phổ biến là, đối với những trường hợp giả thuyết vô hiệu thực sự đúng, chúng ta không muốn bác bỏ H_0 sai quá 5% lần. Điều này tương ứng với một mức ý nghĩa (significance level) $0,05$. Nghĩa là, nếu giả thuyết vô hiệu là đúng, thì mức ý nghĩa cho biết tần suất dữ liệu đưa chúng ta đi đến kết luận bác bỏ H_0 một cách không chính xác. Chúng ta thường viết mức ý nghĩa dùng ký hiệu α. Đây là lý do tại sao chúng ta thích các giá trị α nhỏ, tăng α sẽ làm tăng tỷ lệ lỗi loại 1.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

p-values Giá trị p (p-value) là một cách để định lượng độ mạnh của bằng chứng chống lại giả thuyết vô hiệu và ủng hộ giả thuyết thay thế. Kiểm định giả thuyết thống kê thường sử dụng giá trị p thay vì đưa ra quyết định dựa trên khoảng tin cậy.

Giá trị p là xác suất thu được kết quả ít nhất là như kết quả quan sát được hiện tại, nếu giả thuyết vô hiệu là đúng. Một giá trị p rất nhỏ có nghĩa là một kết quả quan sát được như vậy sẽ rất khó xảy ra nếu giả thuyết vô hiệu là đúng.

- Nếu giá trị p thấp (thấp hơn mức ý nghĩa, α, thường là 5%), chúng ta nói rằng sẽ rất khó quan sát được dữ liệu như đã thấy nếu giả thuyết vô hiệu là đúng và do đó bác bỏ H_0.
- Nếu giá trị p cao (cao hơn α), chúng ta nói rằng có khả năng quan sát được dữ liệu như đã thấy ngay cả khi giả thuyết vô hiệu là đúng, và do đó không bác bỏ H_0.

Chọn mức ý nghĩa α Việc chọn một mức ý nghĩa cho một kiểm định (test) là quan trọng trong nhiều ngữ cảnh và mức ý nghĩa truyền thống là $0,05$. Tuy nhiên, việc điều chỉnh mức ý nghĩa (nhỏ hơn hoặc lớn hơn $0,05$) dựa trên từng trường hợp cụ thể thường hữu ích.

- Nếu việc tạo ra lỗi loại 1 là nguy hiểm hoặc đặc biệt tốn kém, chúng ta nên chọn mức ý nghĩa nhỏ (ví dụ: $0,01$). Theo kịch bản này, chúng ta muốn rất thận trọng trong việc bác bỏ giả thuyết vô hiệu H_0, vì vậy chúng ta yêu cầu bằng chứng mạnh mẽ ủng hộ H_A trước khi bác bỏ H_0.
- Nếu lỗi loại 2 tương đối nguy hiểm hơn hoặc tốn kém hơn nhiều so với lỗi loại 1, thì chúng ta nên chọn mức ý nghĩa cao hơn (ví dụ: $0,10$). Ở đây chúng ta muốn thận trọng về việc không bác bỏ H_0 khi nó thực sự là sai.

Hãy quay lại câu hỏi lỗi nào, loại 1 hay loại 2, là nặng hơn. Rất khó có câu trả lời dứt khoát vì nó còn tùy vào tình huống cụ thể.

Xét ví dụ một bị cáo bị cáo buộc phạm tội và có cần mức án cực kỳ nghiêm khắc. Giả thuyết vô hiệu là bị cáo vô tội. Tất nhiên ta không muốn bỏ qua một người có tội (lỗi loại 2), nhưng hầu hết mọi người cho rằng việc kết án một người vô tội với hình phạt như vậy (lỗi loại 1) là một hậu quả tồi tệ hơn. Do đó, trong tường hợp này lỗi loại 1 (dương tính giả) nặng hơn lỗi loại 2 (âm tính giả).

Xét một ví dụ khác, giả sử bạn đang thiết kế một cuộc kiểm tra y tế cho một căn bệnh. Giả thuyết vô hiệu là người xét nghiệm không có bệnh. Việc mắc lỗi loại 1 (dương tính giả) có thể khiến bệnh nhân lo lắng, nhưng điều này sẽ dẫn đến các quy trình xét nghiệm khác mà cuối cùng sẽ cho thấy xét nghiệm ban đầu không chính xác. Ngược lại, lỗi loại 2 (âm tính giả) sẽ cung cấp cho bệnh nhân sự đảm bảo (không chính xác) rằng họ không mắc bệnh trong khi thực tế ngược lại. Do thông tin không chính xác này, bệnh sẽ không được điều trị và có thể dẫn đến những hậu quả nghiêm trọng tiếp theo. Do đó, trong tường hợp này lỗi loại 2 (âm tính giả) nặng hơn lỗi loại 1 (dương tính giả).

Kiểm định 2 phía (two-sided) hay một phía (one-sided) Trong kiểm định giả thuyết hai phía (two-sided), chúng ta quan tâm đến việc liệu p có trên hay dưới null value p_0 nào đó: $H_A : p \neq p_0$.

Trong kiểm định giả thuyết một phía (one-sided), chúng ta quan tâm đến việc p khác với null value p_0 theo một hướng (chứ không phải hướng khác): $H_A : p < p_0$ hay $H_A : p > p_0$.

Kiểm tra hai phía thường thích hợp hơn vì chúng ta thường muốn phát hiện xem liệu dữ liệu có đi theo hướng ngược lại với hướng giả thuyết vô hiệu một cách rõ ràng hay không.

Ý nghĩa thống kê và ý nghĩa thực tế Khi cỡ mẫu trở nên lớn hơn, các ước lượng điểm trở nên chính xác hơn và bất kỳ khác biệt thực nào giữa giá trị trung bình mẫu và null value trở nên dễ phát hiện hơn. Ngay cả một sự chênh lệch rất nhỏ cũng có thể được phát hiện nếu chúng ta lấy một mẫu đủ lớn. Đôi khi các nhà nghiên cứu sẽ lấy những mẫu lớn đến mức phát hiện được ngay cả những sai lệch nhỏ nhất, thậm chí những sai lệch không có giá trị thực tế. Trong những trường hợp như vậy, chúng ta vẫn nói rằng sự khác biệt là có ý nghĩa thống kê (statistical significance), nhưng nó không có ý nghĩa thực tế (practical significance). Ví dụ: một thử nghiệm trực tuyến có thể xác định rằng việc đặt thêm quảng cáo trên trang web đánh giá phim về mặt thống kê có thể làm tăng lượng người xem một chương trình truyền hình lên $0,001\%$, nhưng sự gia tăng này có thể không có bất kỳ giá trị thực tế nào.

Một vai trò của nhà khoa học dữ liệu trong việc thực hiện một nghiên cứu thường bao gồm việc lập kế hoạch về quy mô của nghiên cứu. Trước tiên, nhà khoa học dữ liệu có thể tham khảo ý kiến của các chuyên gia hoặc tài liệu khoa học để tìm hiểu sự khác biệt nhỏ nhất có ý nghĩa so với null value. Cô/anh ấy cũng sẽ thu được các thông tin khác, chẳng hạn như ước tính rất sơ bộ về tỷ lệ quần thể thực sự p, để có thể ước lượng một cách gần đúng sai số chuẩn. Từ đây, cô/anh ấy có thể đề xuất một cỡ mẫu phù hợp để có thể phát hiện ra nó, nếu có một sự khác biệt có ý nghĩa thực sự. Mặc dù cỡ mẫu lớn hơn vẫn có thể được sử dụng, nhưng những tính toán này đặc biệt hữu ích khi xem xét chi phí hoặc rủi ro tiềm ẩn, chẳng hạn như tác động sức khỏe có thể có đối với tình nguyện viên trong một nghiên cứu y tế.

Ước lượng và kiểm định giả thuyết

Quách Đình Hoàng

2022/04/25

Nội dung

Ước lượng điểm

Ước lượng điểm (point estimation)

- Ước lượng tham số chưa biết sử dụng một giá trị duy nhất dựa trên mẫu
- Ví dụ
  - Ta quan tâm chiều cao trung bình của SV nam ĐH SPKT (μ). Ta chọn một mẫu ngẫu nhiên gồm 100 SV nam ĐH SPKT và tính chiều cao trung bình ($\bar{x}$), giả sử ta được $\bar{x} = 1.7$ m.
  - Nếu ta muốn ước lượng μ bằng một giá trị duy nhất dựa trên mẫu, sẽ trực quan khi dùng $\bar{x} = 1.7$ m. Ta gọi $\bar{x}$ là ước lượng điểm cho μ.

Ước lượng điểm (point estimation)

- Ta sẽ dùng thống kê mẫu để làm ước lượng điểm cho tham số quần thể.
  - Trung bình mẫu $\bar{X}$ được dùng để ước lượng cho trung bình quần thể μ.
  - Phương sai mẫu s^2 được dùng để ước lượng cho phương sai quần thể σ^2
  - Tỷ lệ mẫu $\hat{p}$ được dùng để ước lượng cho tỷ lệ quần thể p

Ví dụ 1

Một nghiên cứu về thói quen tập thể dục đã sử dụng một mẫu ngẫu nhiên gồm 2540 sinh viên đại học (1220 nữ và 1320 nam giới). Nghiên cứu cho thấy những điều sau đây:

- 818 nữ trong mẫu tập thể dục một cách thường xuyên.
- 924 nam trong mẫu tập thể dục một cách thường xuyên.

Cho biết:

- Ước lượng điểm cho tỷ lệ sinh viên nữ tập thể dục thường xuyên là bao nhiêu?
- Ước lượng điểm cho tỷ lệ sinh viên tập thể dục thường xuyên là bao nhiêu?

Ước lượng không chệch (unbiased estimator)

- Nếu một thống kê (statistic) được sử dụng để ước lượng (estimate) một tham số (parameter), thống kê đó được gọi là một estimator của tham số.
  - Trung bình mẫu $\bar{X}$ là một estimator cho trung bình quần thể μ.
  - Tỷ lệ mẫu $\hat{p}$ là một estimator cho tỷ lệ quần thể p.
- Nếu một estimator S được dùng để ước lượng (estimate) một tham số θ, thì S là một ước lượng không chệnh (unbiased estimator) của θ nếu

$$E(S) = \theta$$

Ước lượng không chệch (unbiased estimator)

- Tỷ lệ mẫu $\hat{p}$ là một unbiased estimator cho tỷ lệ quần thể p.
- Trung bình mẫu $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ là một unbiased estimator cho trung bình quần thể μ.
- Phương sai mẫu $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n-1}$ là một unbiased estimator phương sai quần thể σ^2

Ví dụ 2

Một nhà nghiên cứu muốn ước tính μ, số giờ trung bình mà sinh viên tại một trường đại học nhà nước lớn dành để tập thể dục mỗi tuần. Nhà nghiên cứu thu thập dữ liệu từ một mẫu 150 sinh viên rời khỏi phòng tập thể dục của trường đại học sau khi tập luyện.

Điều nào trong số những điều sau đây là đúng về $\bar{x}$, số giờ trung bình mà 150 sinh viên được lấy mẫu tập thể dục mỗi tuần?

A. Đó là một ước tính không thiên vị cho μ.

B. Đó không phải là một ước tính không thiên vị đối với μ và có lẽ đánh giá thấp μ.

C. Nó không phải là một ước tính không thiên vị cho μ và có lẽ đánh giá quá cao μ.

Ví dụ 3

Dựa trên kết quả khảo sát, tỷ lệ người trưởng thành ở Hoa Kỳ sử dụng Internet hàng ngày là 0.37. Ước lượng điểm này sẽ không thiên vị và chính xác nhất nếu cuộc khảo sát dựa trên:

A. Một mẫu ngẫu nhiên của 1000 người Trưởng thành ở Hoa Kỳ.

B. Một mẫu ngẫu nhiên của 2500 người trưởng thành ở Hoa Kỳ.

C. Một mẫu ngẫu nhiên của 1000 sinh viên đại học.

D. Một mẫu ngẫu nhiên của 2500 sinh viên đại học.

Các nhận xét

- Các nguyên tắc lấy mẫu và thiết kế nghiên cứu là quan trọng đến kết quả ước lượng.
  - Ước lượng điểm chỉ thực sự là ước lượng không chệch (không thiên vị) cho tham số quần thể khi mẫu là ngẫu nhiên và thiết kế nghiên cứu không thiếu sót.
- Độ chính xác của ước lượng điểm được cải thiện khi cỡ mẫu tăng lên.
  - Cỡ mẫu lớn hơn cung cấp cho ta nhiều thông tin hơn để xác định bản chất thực sự của quần thể.

Ước lượng khoảng

Ước lượng khoảng

- Ước lượng điểm tuy trực quan và đơn giản nhưng có hạn chế.
  - Trung bình mẫu $\bar{X}$ gần như không bằng chính xác μ.
- Do đó, ước lượng khoảng sẽ không thực sự hữu ích nếu ta không xác định mức độ sai số của ước lượng.
- Ước lượng khoảng bổ sung cho ước lượng điểm bằng cách cung cấp thông tin về sai số của ước lượng.
  - Ví dụ: ta tin tưởng 95% (95% confident) chiều cao trung bình của nam sinh ĐH SPKT là $\bar{X} = 1.7 \pm 0.2$ m, tức trong khoảng $(1.68, 1.72)$.
    - 95% là mức tin cậy mà giá trị của tham số nằm trong khoảng tin cậy

Khoảng tin cậy cho trung bình

- Định lý giới hạn trung tâm cho trung bình: phân bố mẫu (sampling distribution) của trung bình mẫu (sample mean) $\bar{X}$ xấp xỉ phân bố chuẩn với trung bình μ là độ lệch chuẩn $\sigma/\sqrt{n}$. Trong đó:
  - μ là trung bình quần thể
  - σ là độ lệch chuẩn quần thể
  - n là cỡ mẫu
- Khoảng tin cậy 95%
  - $P(\mu - 1.96\sigma \leq \bar{X} \leq \mu + 1.96\sigma) \approx 0.95$: Xác xuất trung bình mẫu $\bar{X}$ nằm trong khoảng 1.96σ so với trung bình quần thể μ là 95%, hay
  - Ta 95% tin tưởng (95% confident) trung bình quần thể μ nằm trong khoảng 1.96σ so với trung bình mẫu X.
  - Phát biểu trên là về $\bar{X}$ (một biến ngẫu nhiên), phát hiểu dưới là về μ (một tham số, một giá trị chưa biết nhưng cố định).

Khoảng tin cậy cho trung bình

- Giả sử dữ liệu $x_1, x_2, \cdots, x_n$ được sinh từ phân bố chuẩn $N(\mu, \sigma)$, μ chưa biết và σ đã biết. Khoảng tin cậy $(1-\alpha)$ cho μ là

$$\left[\overline{x} - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}, \overline{x} + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right]$$

  Trong đó, $z_{\alpha/2}$ là giá trị thỏa $P(Z > z_{\alpha/2}) = \alpha/2$, được gọi là standard deviation multiplier.

- Standard deviation multiplier phụ thuộc vào confidence level $(1-\alpha)$.
  - Nếu $\alpha = 0.1$ thì $z_{\alpha/2} \approx 1.645$.
  - Nếu $\alpha = 0.05$ thì $z_{\alpha/2} \approx 1.96$.
  - Nếu $\alpha = 0.01$ thì $z_{\alpha/2} \approx 2.576$.

Khoảng tin cậy t cho trung bình

- Nếu σ cũng chưa biết, ta sẽ
  1. Dùng $s_{\overline{x}} = \frac{s}{\sqrt{n}}$ thay cho $\frac{\sigma}{\sqrt{n}}$, s là độ lệch chuẩn trên mẫu.
  2. Sử dụng giá trị t (t critical value) thay cho giá trị z (z critical value).
- Giả sử dữ liệu $x_1, x_2, \cdots, x_n$ được sinh từ phân bố chuẩn $N(\mu, \sigma)$, μ và σ đều chưa biết. Khoảng tin cậy $(1-\alpha)$ cho μ là

$$\left[\overline{x} - \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}, \overline{x} + \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}\right]$$

Trong đó, $t_{\alpha/2}$ là giá trị thỏa $P(T > t_{\alpha/2}) = \alpha/2$, với $T \sim t(n-1)$ và s là độ lệch chuẩn trên mẫu.

$T \sim t(n-1)$: T có phân bố t với bậc tự do $n-1$

Khoảng tin cậy cho trung bình với mẫu lớn

- Giả sử dữ liệu $x_1, x_2, \cdots, x_n$ được sinh từ một phân bố bất kỳ với trung bình và phương sai hữu hạn. Nếu n đủ lớn, khoảng tin cậy $(1-\alpha)$ cho μ là xấp xỉ

$$\left[\overline{x} - \frac{z_{\alpha/2} \cdot s}{\sqrt{n}}, \overline{x} + \frac{z_{\alpha/2} \cdot s}{\sqrt{n}}\right]$$

Trong đó, $z_{\alpha/2}$ là giá trị thỏa $P(Z > z_{\alpha/2}) = \alpha/2$.

Khoảng tin cậy cho tỷ lệ

- Giả sử dữ liệu $x_1, x_2, \cdots, x_n$ được sinh từ phân bố bất kỳ của biến phân loại (categorical variable) có tỷ lệ quần thể (population propotion) là p (p chưa biết). Khoảng tin cậy $(1-\alpha)$ cho p khi $np \geq 10$ và $n(1-p) \geq 10$ là

$$\left[\hat{p}_n - z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$$

  Trong đó, $z_{\alpha/2}$ là giá trị thỏa $P(Z > z_{\alpha/2}) = \alpha/2$

- Ví dụ: Cho biết $n = 1082$, $\hat{p}_n = 0.65$, (tức $703/1082$). Ước lượng p và 95% CI cho p?
  - Vì $np = 1082 * 0.65 \geq 10$ và $n(1-p) = 1082 * 0.35 \geq 10$. Áp dụng định lý trên, 95% CI cho p là

$$0.65 \pm 1.96\sqrt{\frac{0.65(1-0.65)}{1082}} = 0.65 \pm 0.03$$

Áp dụng: khoảng tin cậy cho tỷ lệ quần thể

- Cho $n = 670, \hat{p} = 0.85$
- Sai số chuẩn (standard eror) của tỷ lệ mẫu (sample proportion) là:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- Tính khoảng tin cậy (confident interval) 95% cho tỷ lệ quần thể (population proportion)?

Áp dụng: chọn cỡ mẫu

- Với $\hat{p} = 0.85$, cần cỡ mẫu bao nhiêu để biên lỗi (margin of error) của khoảng tin cậy 95% là 0.1?

$$ME = z\text{-value} \times SE$$

Kiểm định giả thuyết

Kiểm định giả thuyết

- Ví dụ dẫn nhập
  - Quần thể: các thanh niên Việt Nam từ 18-35 tuổi
  - Mẫu: 100 người được chọn ngẫu nhiên từ quần thể trên
  - Biến: IQ
  - Thống kê: trung bình mẫu là 110, độ lệch chuẩn là 5
  - Tham số: trung bình quần thể là 100 (giả sử ta biết được điều này do nhiều nghiên cứu trước đây)
  - Câu hỏi:
    1. Xác suất nhận được trung bình mẫu là 110 một cách tình cờ, nếu trung bình quần thể thực sự là 100 là bao nhiêu? Xác xuất này chính là p-value.
    2. Sự khác biệt giữa trung bình mẫu và trung bình quần thể có quá lớn để có thể xảy ra tình cờ hay không? Hay nói cách khác, sự khác biệt đó có ý nghĩa thống kê (statistically significant) hay không?

Kiểm định giả thuyết

- H_0: Trung bình mẫu và trung bình quần thể không khác nhau (no effect). Giả thuyết này được đặt ra trước khi ta lựa chọn mẫu và tính giá trị thống kê.

$$H_0 : \mu = \overline{X}$$

- H_a: Trung bình mẫu và trung bình quần thể là khác nhau. Đây là giả thuyết hai phía - (two-sided)

$$H_a : \mu \neq \overline{X}$$

  - μ là trung bình quần thể
  - $\overline{X}$ là trung bình mẫu
- Ta cũng có thể đặt giả thuyết một phía (one-sided)

$$H_a : \mu < \overline{X}, H_a : \mu > \overline{X}$$

Mức ý nghĩa alpha

- Mức ý nghĩa alpha, (α), là một tiêu chí mà chúng ta sẽ sử dụng để quyết định có nên giữ lại hay loại bỏ giả thuyết đặt ra
  - Thông thường α được chọn là 0.05.
- Khi ta đã chọn α, nếu sự khác biệt giữa thống kê trên mẫu và tham số của quần thể nhỏ hơn α, chúng ta có thể bác bỏ giả thuyết H_0 và kết luận rằng sự khác biệt này có lẽ không phải do tình cờ.
- Khi ta bác bỏ giả thuyết H_0, ta có thể sai (bác bỏ H_0 mặc dù nó đúng). Lỗi như vậy được gọi là lỗi loại 1.
- Mức độ alpha, (α), đại diện cho tỷ lệ lỗi loại 1 mà chúng ta sẵn sàng chấp nhận trước khi tiến hành phân tích thống kê.

Phân tích thống kê

- Khi ta làm suy luận thống kê, ta muốn biết một hiện tượng mà ta quan sát được trên mẫu có đại diện cho một hiện tượng thực tế trên quần thể hay không.
- Ta lập giả thuyết vô hiệu H_0 là không có sự khác biệt
- Ta chọn một mức ý nghĩa α làm tiêu chuẩn để chấp nhận hay bác bỏ giả thuyết
- Tính giá trị p (p-value)
  - Nếu $p < \alpha$, ta bác bỏ giả thuyết H_0 và kết luận sự khác biệt nhiều khả năng không phải do tình cờ.
    - Khi bác bỏ H_0 ta có khả năng mắc sai lầm, đây là sai lầm loại 1.
  - Nếu $p > \alpha$, ta không bác bỏ được H_0 và kết luận sự khác biệt nhiều khả năng là do tình cờ hoặc dữ liệu quan sát được là không đủ để chứng tỏ rằng có sự khác biệt.

z-test cho giá trị trung bình

- Giả sử dữ liệu $x_1, x_2, \cdots, x_n \sim N(\mu, \sigma^2)$, μ chưa biết, σ đã biết.
- Null hypothesis: $H_0 : \mu = \mu_0$, với μ_0 cho trước.
- Alternative hypothesis:
  - Two-sided: $H_a : \mu \neq \mu_0$
  - Right-sided: $H_a : \mu > \mu_0$
  - Left-sided: $H_a : \mu < \mu_0$
- Test statistic:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

- p-value
  - Two-sided: $p = P(Z > |z||H_0)$
  - Right-sided: $p = P(Z > z|H_0)$
  - Left-sided: $p = P(Z < z|H_0)$

Ví dụ z-test giá trị trung bình

Giả sử dữ liệu có phân bố $N(\mu, 4)$. $H_0 : \mu = 0$, $H_a : \mu > 0$. Dữ liệu mẫu $1, 2, 3, 6, -1$, $\alpha = 0.05$. Tính p-value?

- $\overline{x} = 2.2$
- $$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.2 - 0}{2/\sqrt{5}} = 2.46$$
- p-value: $p = P(Z > z) = P(Z > 2.46)$
  - Ta tính $P(Z > z)$ dùng hàm `stats.norm.cdf()` của thư viện `scipy`, $cdf(z)$ là $P(Z < z)$.
  - $P(Z > 2.46) = 1 - stats.norm.cdf(2.46) = 0.007$

t-test cho giá trị trung bình

- Giả sử dữ liệu $x_1, x_2, \cdots, x_n \sim N(\mu, \sigma^2)$, μ và σ chưa biết.
- Null hypothesis: $H_0 : \mu = \mu_0$, với μ_0 cho trước.
- Alternative hypothesis:
  - Two-sided: $H_a : \mu \neq \mu_0$
  - Right-sided: $H_a : \mu > \mu_0$
  - Left-sided: $H_a : \mu < \mu_0$
- Test statistic:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  - Giá trị t ứng với phân bố t với bậc tự do $(n-1)$.
- p-value
  - Two-sided: $p = P(T > |t||H_0)$
  - Right-sided: $p = P(T > t|H_0)$
  - Left-sided: $p = P(T < t|H_0)$

Ví dụ t-test cho giá trị trung bình

Giả sử dữ liệu có phân bố $N(\mu, \sigma^2)$. $H_0 : \mu = 0$, $H_a : \mu > 0$. Dữ liệu mẫu $1, 2, 3, 6, -1$, $\alpha = 0.05$. Tính p-value?

- $\overline{x} = 2.2, s^2 = 6.7$
- $$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{2.2 - 0}{2.59/\sqrt{5}} = 1.9$$
- p-value: $p = P(T > t) = P(T > 1.9)$
  - Ta tính $P(T > t)$ dùng hàm `stats.t.cdf()` của thư viện `scipy`, $stats.t.cdf(t, n)$ là $P(T < t)$ với bậc tự do n.
  - $p = P(T > 1.9) = 1 - stats.t.cdf(1.9, 4) = 0.065$

t-test trung bình với mẫu lớn

- Giả sử dữ liệu $x_1, x_2, \cdots, x_n$ được sinh từ một phân bố bất kỳ với trung bình và phương sai hữu hạn, n đủ lớn.
- Null hypothesis: $H_0 : \mu = \mu_0$, với μ_0 cho trước.
- Alternative hypothesis:
  - Two-sided: $H_a : \mu \neq \mu_0$
  - Right-sided: $H_a : \mu > \mu_0$
  - Left-sided: $H_a : \mu < \mu_0$
- Test statistic:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  - Giá trị t ứng với phân bố t với bậc tự do $(n-1)$.
- p-value
  - Two-sided: $p = P(T > |t||H_0)$
  - Right-sided: $p = P(T > t|H_0)$

z-test cho tỷ lệ

- Giả sử dữ liệu $x_1, x_2, \cdots, x_n$ được sinh từ phân bố bất kỳ của biến phân loại (categorical variable) có tỷ lệ quần thể (population propotion) là p (p chưa biết), $np \geq 10$ và $n(1-p) \geq 10$.
- Null hypothesis: $H_0 : p = p_0$, với p_0 cho trước.
- Alternative hypothesis:
  - Two-sided: $H_a : p \neq p_0$
  - Right-sided: $H_a : p > p_0$
  - Left-sided: $H_a : p < p_0$
- Test statistic:

$$z = \frac{\hat{p} - p_0}{SE}, \qquad SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  - Giá trị t ứng với phân bố t với bậc tự do $(n-1)$.
- p-value
  - Two-sided: $p = P(T > |t||H_0)$

Tham khảo

Probability and Statistics. Provided by: Open Learning Initiative. Located at: http://oli.cmu.edu. License: CC BY: Attribution

Statistical Inference for Relationships

Quách Đình Hoàng

2022/05/07

Nội dung

Suy diễn về mối quan hệ giữa hai biến

$C \to Q$

$C \to Q$: 2 nhóm độc lập

$C \to Q$: 2 nhóm phụ thuộc

$C \to Q$: nhiều nhóm độc lập

$C \to C$

$Q \to Q$

Suy diễn về mối quan hệ giữa hai biến

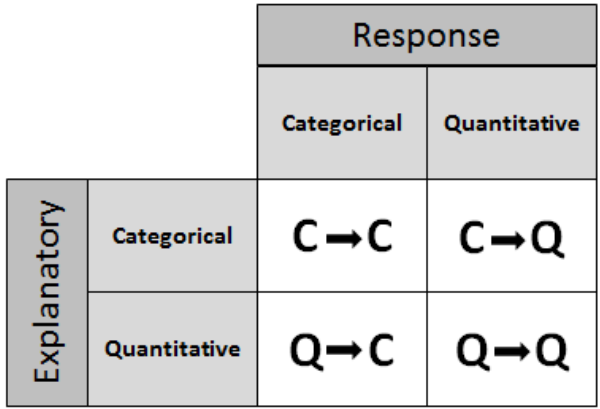
Suy diễn cho một biến

- Ta đã đề cập đến suy luận cho population proportion p (categorical variable) và suy luận cho population mean μ (quantitative variable).
- Ta tập trung vào 3 dạng suy luận
  - Point estimation: Ước lượng tham số chưa biết sử dụng một giá trị duy nhất dựa trên mẫu.
  - Interval estimation: Ước lượng tham số chưa biết sử dụng một khoảng các giá trị có thể với một level of confidence.
  - Hypothesis testing: Đánh giá bằng chứng được cung cấp bởi dữ liệu có chống lại một phát biểu về population parameter.

Suy diễn về mối quan hệ giữa hai biến

- Mục tiêu: Suy luận về mối quan hệ giữa hai biến X và Y trong population, dựa trên mối quan hệ quan sát được giữa chúng trong sample.
  - Ta muốn biết dữ liệu có cung cấp bằng chứng đủ mạnh để có thể tổng quát hóa mối quan hệ quan sát được trong mẫu và kết luận (với một số mức độ không chắc chắn xác định trước) rằng mối quan hệ giữa X và Y tồn tại trong quần thể hay không.
- Công cụ chính để thực hiện việc này là hypothesis testing.
  - H_0: Không có mối quan hệ giữa X và Y
  - H_a: Có mối quan hệ đáng kể giữa X và Y

Suy diễn về mối quan hệ giữa hai biến



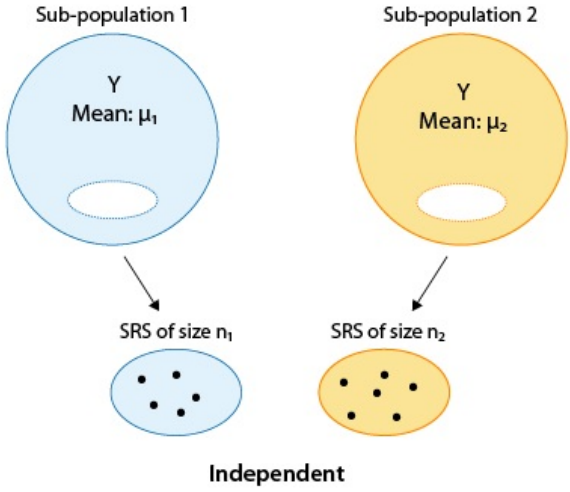
- Ta sẽ đề cập đến các suy diễn $C \to Q$, $C \to C$, và $Q \to Q$

$$C \to Q$$

$C \to Q$

- Biến giải thích X kiểu phân loại và biến kết quả Y kiểu định lượng.
- Trong EDA, ta so sánh phân bố của Y ứng với mỗi giá trị của X và thể hiện bằng các biểu đồ hộp (boxplot).
  - Ví dụ: IQ theo giới tính, IQ theo theo các mức tuổi.
- Trong suy diễn, ta muốn biết X có mối quan hệ với Y hay không dựa trên phân tích một mẫu từ quần thể.
  - Ta phân các giá trị Y theo các nhóm ứng với các giá trị X và so sánh giá trị trung bình của các nhóm.
- Ta sẽ xét mối quan hệ $C \to Q$ trong 3 trường hợp
  - Có 2 nhóm độc lập
  - Có 2 nhóm phụ thuộc
  - Có hơn hai nhóm độc lập

$C \to Q$: 2 nhóm độc lập



$C \to Q$: 2 nhóm độc lập

Điều kiện (giả định) cần thỏa trước khi áp dụng

1. Hai mẫu là độc lập và được chọn ngẫu nhiên
2. Một trong hai trường hợp sau:

2.1. Cả hai quần thể tuân theo phân bố chuẩn, hoặc

2.2. Cỡ mẫu đủ lớn (n $>$ 30)

$C \to Q$: 2 nhóm độc lập

$C \to Q$: 2 nhóm độc lập

- $H_0 : \overline{y}_1 = \overline{y}_2$
- $H_a : \overline{y}_1 \neq \overline{y}_2$ (kiểm định hai phía)
- $H_a : \overline{y}_1 > \overline{y}_2$ (kiểm định một phía)
- $H_a : \overline{y}_1 < \overline{y}_2$ (kiểm định một phía)
- Two-sample t-test statistic:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(1-\alpha/2, v), \qquad v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

- Khoảng tin cậy $(1-\alpha)$ ($(1-\alpha)$ CI) cho $\overline{x}_1 - \overline{x}_2$:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$C \to Q$: 2 nhóm độc lập

- Cho mẫu ngẫu nhiên 239 SV làm bài thi có phổ điểm từ 1 đến 25. Ta muốn biết liệu điểm của nam và nữ có sự khác biệt?
  - SV nam: $n_1 = 150, \bar{x_1} = 13.33, s_1 = 4.25$
  - SV nữ: $n_2 = 85, \bar{x_2} = 10.73, s_2 = 4.02$

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{13.33 - 10.73}{\sqrt{\frac{4.25^2}{150} + \frac{4.02^2}{85}}} = 4.66$$

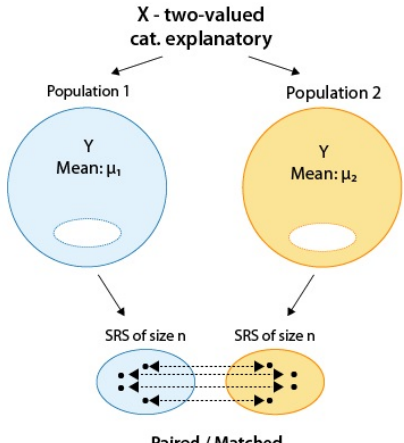
$$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{n_2-1}} = \frac{(\frac{4.25^2}{150} + \frac{4.02^2}{85})^2}{\frac{(4.25^2/150)^2}{(150-1)} + \frac{(4.02^2/85)^2}{85-1}} \sim 183$$

- $p = 1 - stats.t.cdf(4.66, 183) \sim 0 \to$ bác bỏ H_0.
- Khoảng tin cậy 95% CI) cho $\overline{y}_1 - \overline{y}_2$:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (13.33 - 10.73) \pm 1.97 \sqrt{\frac{4.25^2}{150} + \frac{4.02^2}{85}} = 2.6 \pm 1.1$$

$C \to Q$: 2 nhóm phụ thuộc

$C \to Q$: 2 nhóm phụ thuộc



- Trong trường hợp 2 nhóm phụ thuộc, một đối tượng của của mẫu này mathched/paired với một đối tượng của mẫu kia.

$C \to Q$: 2 nhóm phụ thuộc

- Đây là trường hợp đặt biệt của one-sample t-test
- Điều kiện (giả định) cần thỏa trước khi áp dụng

1. Mẫu được chọn ngẫu nhiên
2. Một trong hai trường hợp sau:

2.1. Sự khác biệt (the sample of differences) tuân theo phân bố chuẩn, hoặc

2.2. Cỡ mẫu đủ lớn (n $>$ 30)

$C \to Q$: 2 nhóm phụ thuộc

- $H_0 : \overline{x}_d = 0$
- $H_a : \overline{x}_d \neq 0$ (kiểm định hai phía)
- $H_a : \overline{x}_d > 0$ (kiểm định một phía)
- $H_a : \overline{x}_d < 0$ (kiểm định một phía)
- Paired t-test statistic

$$t = \frac{\overline{x}_d - 0}{\frac{s_d}{\sqrt{n}}} \sim t(n-1)$$

- Khoảng tin cậy $(1-\alpha)$ ($(1-\alpha)$ CI) cho $\overline{x}_d$:

$$\overline{x}_d \pm t_{1-\alpha/2} \frac{s_d}{\sqrt{n}}$$

$C \to Q$: 2 nhóm phụ thuộc

- Ta muốn thời gian phản ứng trước và sau khi uống 2 chai bia.
- $n = 20, \overline{x}_d = -0.5015, s_d = 0.8686$

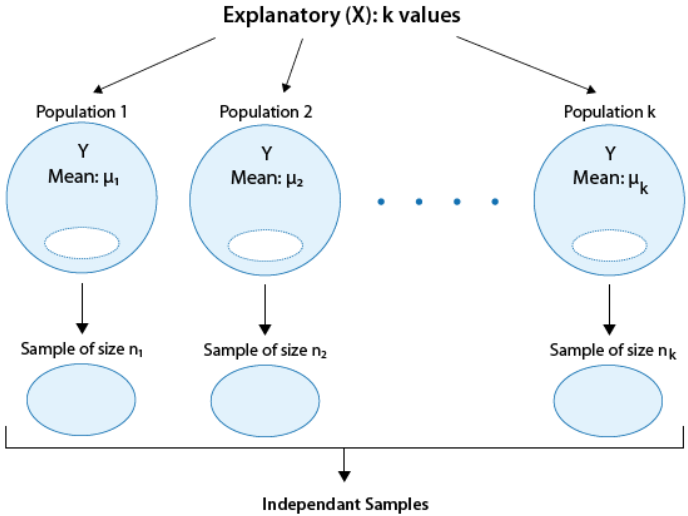
$$t = \frac{\overline{x}_d - 0}{\frac{s_d}{\sqrt{n}}} = \frac{-0.5015 - 0}{\frac{0.8686}{\sqrt{20}}} = 2.58$$

- $p = 1 - stats.t.cdf(2.58, 19) \sim 0.009 \to$ bác bỏ H_0
- Khoảng tin cậy 95% cho $\overline{x}_d$:

$$\overline{x}_d \pm t_{1-\alpha/2}\frac{s_d}{\sqrt{n}} = 0.5015 \pm 2.09\frac{0.8686}{\sqrt{20}} = 0.5015 \pm 0.4059$$

$C \to Q$: nhiều nhóm độc lập

$C \to Q$: nhiều nhóm độc lập



$C \rightarrow Q$: nhiều nhóm độc lập

- Điều kiện (giả định) cần thỏa trước khi áp dụng

1. Các mẫu là độc lập và được chọn ngẫu nhiên
2. Sự khác biệt trong mỗi nhóm tuân theo phân bố chuẩn (nếu cỡ mẫu đủ lớn thì điều này không quan trọng lắm)
3. Các quần thể của mỗi nhóm có cùng độ lệch chuẩn

$C \to Q$: nhiều nhóm độc lập

- $H_0 : \overline{x}_1 = \overline{x}_2 = \cdots = \overline{x}_k$
- $H_a : \exists i, j \in \{1, 2, \cdots, k\} : \overline{x}_i \neq \overline{x}_j$
- ANOVA F-test statistic:

$$F = \frac{\sum_{i=1}^{k} n_i \frac{(\overline{x}_i - \overline{x})^2}{k-1}}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(x_{ij} - \overline{x}_i)^2}{n-k}} \sim F(k-1, n-k)$$

- n là cỡ mẫu, k là số nhóm
- n_i là số phần tử của nhóm i
- $\overline{x}_i$ là trung bình của nhóm i
- $\overline{x}$ là trung bình của n phần tử
- x_{ij} là đối tượng thứ j của nhóm thứ i

$C \to Q$: nhiều nhóm độc lập

- Ví dụ: Mức độ thất vọng của SV ở các ngành học

$$F = \frac{\text{variation among sample means}}{\text{variation within groups}}$$

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Business	35	256	7.31428571	8.398319
English	35	412	11.7714286	4.357983
Mathematics	35	462	13.2	4.635294
Psychology	35	491	14.0285714	9.49916

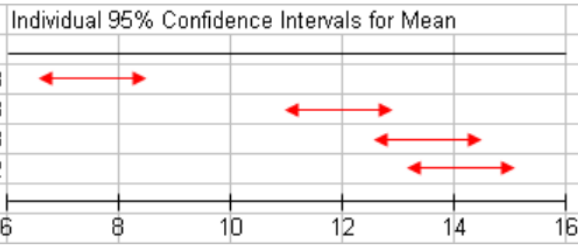
ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	939.85	3	313.283333	46.6009	8.8416E-21	2.6711788
Within Groups	914.2857143	136	6.72268908			
Total	1854.135714	139				

$C \to Q$: nhiều nhóm độc lập

- Ví dụ: Mức độ thất vọng của SV ở các ngành học

	Mean	StDev
Business	7.314	2.898
English	11.771	2.088
Mathematics	13.2	2.153
Psychology	14.029	3.082



$$C \to C$$

$C \to C$

- Cả biến giải thích X và biến kết quả Y đều có kiểu phân loại.
- Trong EDA, ta biểu diễn mối quan hệ giữa hai biến phân loại bằng bảng 2 chiều.

X/Y	y_1	y_2	...	y_k
x_1	n_{11}	n_{12}	...	n_{1k}
x_2	n_{21}	n_{22}	...	n_{2k}
...	...	...	...	...
x_p	n_{p1}	n_{p2}	...	n_{pk}

- Trong suy diễn, ta muốn biết X có mối quan hệ với Y hay không dựa trên phân tích một mẫu từ quần thể.

$C \to C$: Chi-Square test

- H_0: hai biến không có quan hệ
- H_a: hai biến có quan hệ

X/Y	Y_1	Y_2	...	Y_k
X_1	n_{11}	n_{12}	...	n_{1k}
X_2	n_{21}	n_{22}	...	n_{2k}
...	...	...	...	...
X_p	n_{p1}	n_{p2}	...	n_{pk}

- Chi-square test statistic:

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{k} \frac{(n_{ij} - Expected_{ij})^2}{Expected_{ij}} \sim \chi^2((p-1)\times(k-1))$$

$$Expected_{ij} = \frac{(\sum_{j=1}^{k} n_{ij}) \times (\sum_{i=1}^{p} n_{ij})}{\sum_{i=1}^{p}\sum_{j=1}^{k} n_{ij}}$$

$C \to C$: Chi-Square test

Gender ↓	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Expected count

Gender ↓	Drank Alcohol in Last 2 Hours? Yes	No	Total	
Male	(93*481)/619 →		481	row total
Female	↓		138	
Total	93	526	619	
	column total		table total	

$C \rightarrow C$: Chi-Square test

Gender ↓	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Gender	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	(93*481)/619=72.3	(526*481)/619=408.7	481
Female	(93*138)/619=20.7	(526*138)/619=117.3	138
Total	93	526	619

$C \to C$: Chi-Square test

Observed counts / Expected counts

Gender	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77 / 72.3	404 / 408.7	481
Female	16 / 20.7	122 / 117.3	138
Total	93	526	619

$$\chi^2 = \frac{(77-72.3)^2}{72.3} + \frac{(404-408.7)^2}{408.7} + \frac{(16-20.7)^2}{20.7} + \frac{(122-117.3)^2}{117.3} = 1.62$$

- $(p-1) \times (k-1) = (2-1) \times (2-1) = 1$
- $p = 1 - stats.chi2.cdf(1.62, 1) = 0.203$

$$Q \to Q$$

$Q \to Q$

- Cả biến giải thích X và biến kết quả Y đều có kiểu định lượng.
- Trong EDA, ta biểu diễn mối quan hệ giữa hai biến định lượng bằng biểu đồ tán xạ (scatterplot) hoặc hệ số tương quan.

$$Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- Trong suy diễn, ta muốn biết X có mối quan hệ với Y hay không dựa trên phân tích một mẫu từ quần thể.

$Q \to Q$

- $H_0 : \beta = 0$: Không có quan hệ tuyến tính giữa X và Y

$$Y = \alpha + \beta X$$

- $H_a : \beta \neq 0$: Có quan hệ tuyến tính giữa X và Y
- Least squares estimates:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

$Q \to Q$

- t-test statistic:

$$t = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} \sim t(n-2)$$

$$SE(\hat{\beta}) = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-2}}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

- Khoảng tin cậy $(1-\alpha)$ ($(1-\alpha)$ CI) cho β:

$$\hat{\beta} \pm t_{1-\alpha/2} \times SE(\hat{\beta})$$

Tham khảo

Probability and Statistics. Provided by: Open Learning Initiative. Located at: http://oli.cmu.edu. License: CC BY: Attribution