Exploration Data Analysis (EDA)

Quách Đình Hoàng

23/9/2019

Data analysis big picture

- Mục đích của phân tích dữ liệu là biến dữ liệu thành thông tin có ích.
- Quá trình này thường liên quan đến:
  - Thu thập dữ liệu (collecting data),
  - Tóm tắt dữ liệu (summarizing data), và
  - Diễn giải dữ liệu (interpreting data).

Thu thập dữ liệu: quần thể và mẫu

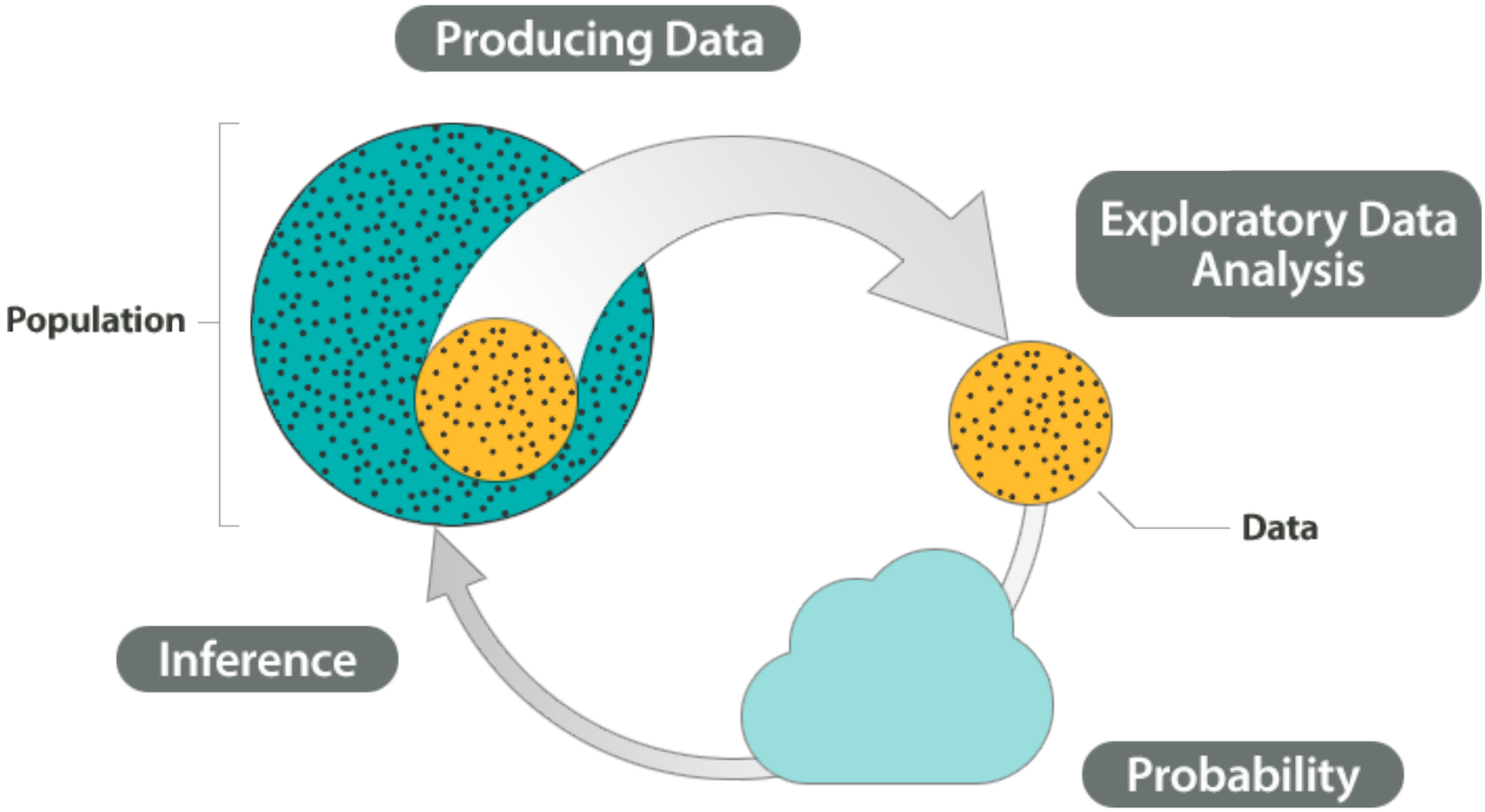
- Quá trình phân tích dữ liệu bắt đầu với việc xác định tất cả những đối tượng mà ta quan tâm, gọi là quần thể (population).
  - Ví dụ: tập tất cả các người nam từ 18 đến 25 tuổi ở Việt Nam.
- Tuy nhiên, ta thường chỉ thu thập được một tập con của quần thể, gọi là mẫu (sample) để phân tích. Ta gọi quá 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.
  - Ví dụ: trong tập tất cả các người nam từ 18 đến 25 tuổi ở Việt Nam ta chỉ chọn ra ngẫu nhiên 1000 người để phân tích.

Tóm tắt dữ liệu

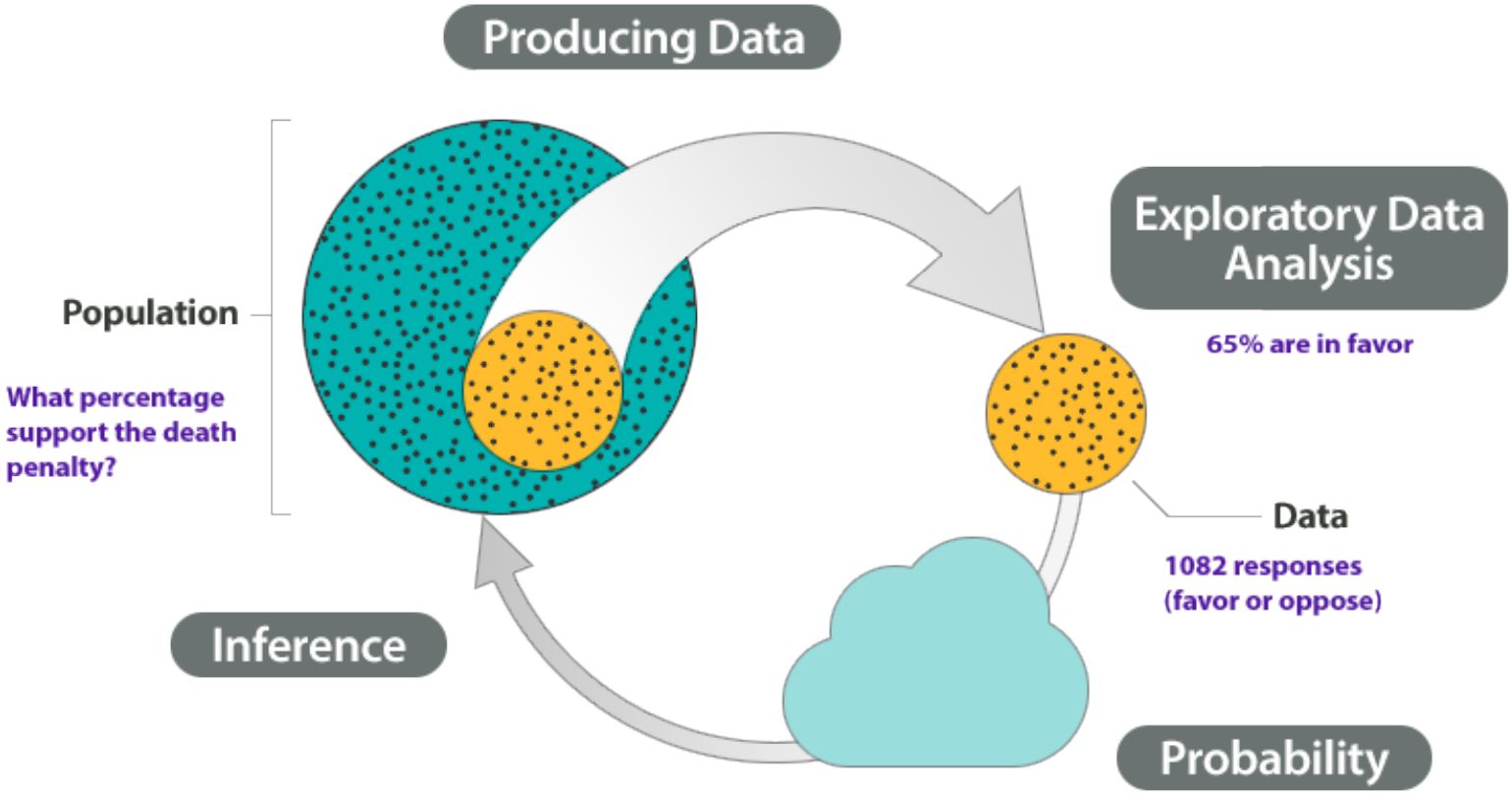
- Sau khi có dữ liệu, ta thường muốn tóm tắt chúng để có cái nhìn tổng quan, gọi là phân tích thăm dò (exploratory data analysis).
- Tuy nhiên, mục đích của ta là hiểu về đặc tính của quần thể hơn là của mẫu ta đã thu thập.
  - Ta muốn đưa ra kết luận về các đặc tính của quần thể dựa vào kết quả phân tích trên mẫu.
- Để hiểu được sự khác biệt giữa quần thể và mẫu ta cần sử dụng xác suất (probability).

Diễn giải dữ liệu

- Quá trình đưa ra kết luận về quần thể dựa trên kết quả trên mẫu 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



Conclusion: we can be 95% sure that the population percentage is within 3% of 65% (i.e. between 62% and 68%).

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)

Mã SV	Họ và tên lót	Tên	Ngày sinh	Tên lớp
17133001	Võ Đình	An	12/03/1999	171330B
17133002	Trần Gia	Bảo	03/10/1999	171330A
17133003	Phạm Hoàng Quang	Cảnh	08/09/1999	171330A
17133004	Vỏ Phú	Cường	08/10/1999	171330C
17133005	Ngô Thành	Danh	19/09/1999	171330B
17133006	Võ Trọng	Diện	04/03/1999	171330A

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

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

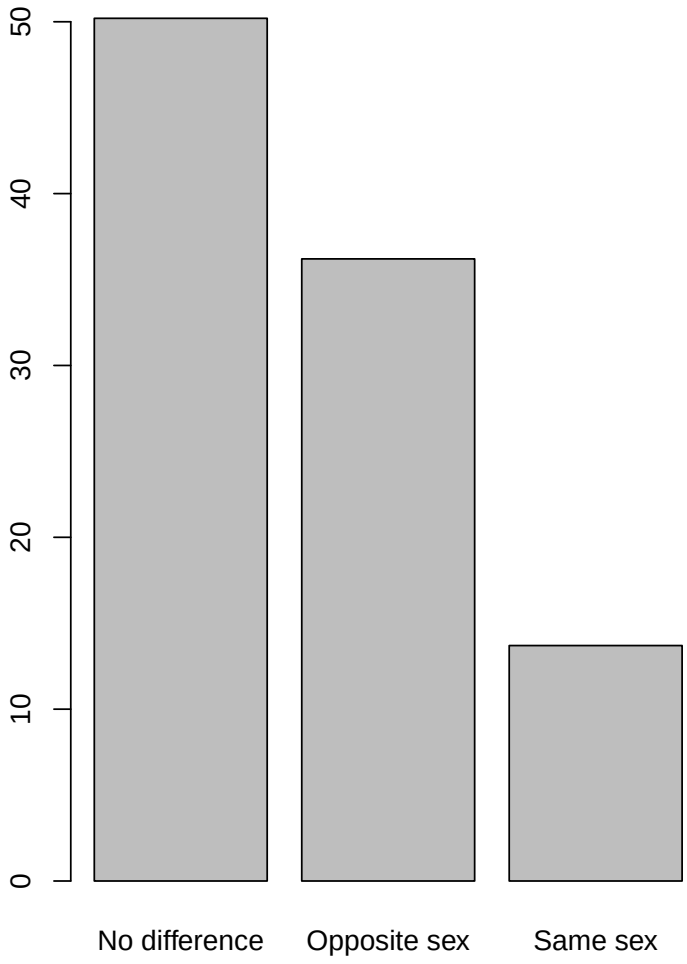
- Việc tóm tắt và nhìn vào phân bố của dữ liệu có thể giúp ta rút ra được các thông tin hữu ích
  - Chỉ nhìn vào tập các giá trị thường không giúp ta rút ra được các thông tin hữu ích
- Để tóm tắt một biến phân loại, ta thường dùng bảng phân bố tần số xuất hiện (frequency distribution)

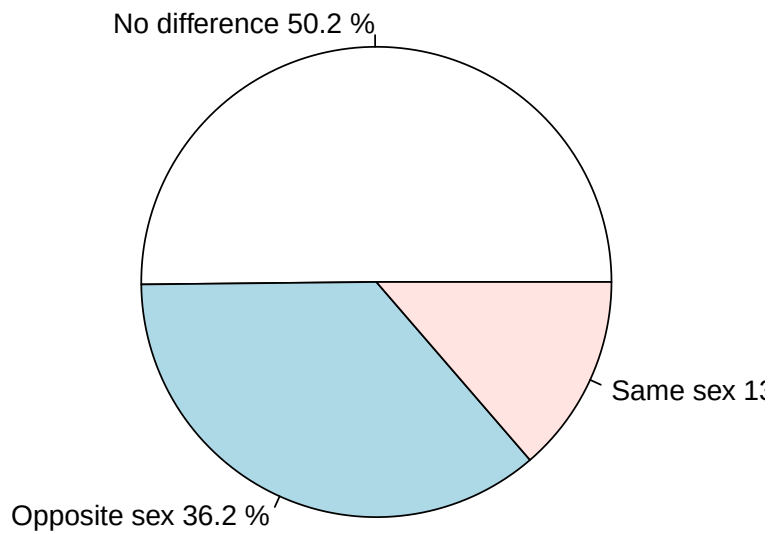
```
##
## No difference  Opposite sex       Same sex
##           602            434            164

##
## 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





Biến số

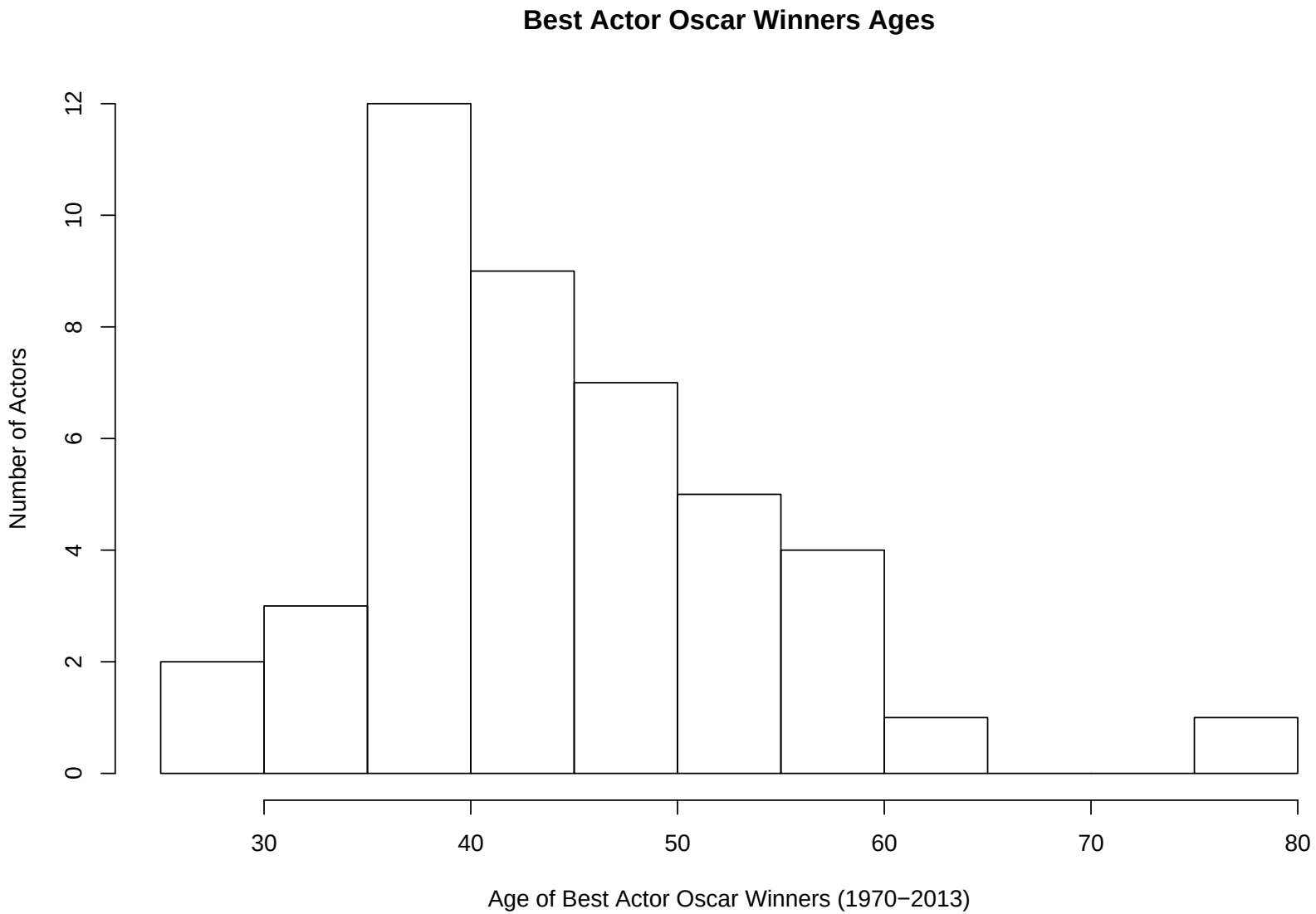
- Phân bố của biến số cung cấp cho ta các đặc trưng quan trọng:
  - Hình dạng (shape),
  - Khuynh hướng tập trung (center tendency), và
  - Sự phân tán (spead) của dữ liệu.
- Từ các thông tin trên có thể giúp ta suy ra các giá trị ngoại lệ (outlier) của dữ liệu.

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

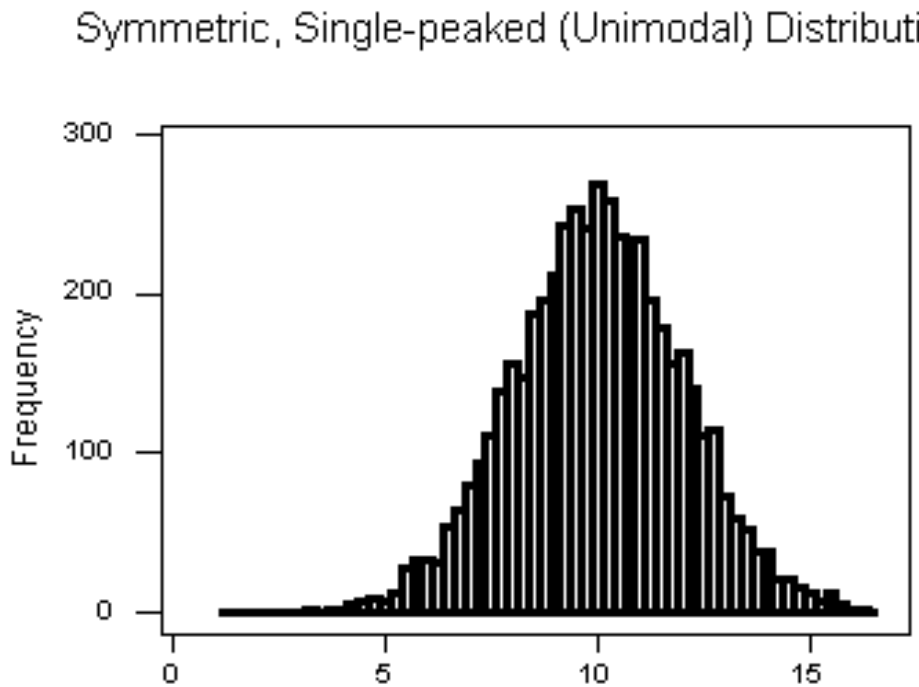
Biểu đồ tần số

```
##  [1] 43 42 48 49 56 38 60 30 40 42 37 76 39 53 45 36 62
## [24] 37 38 32 45 60 46 40 36 47 29 43 37 38 45 50 48 60
```

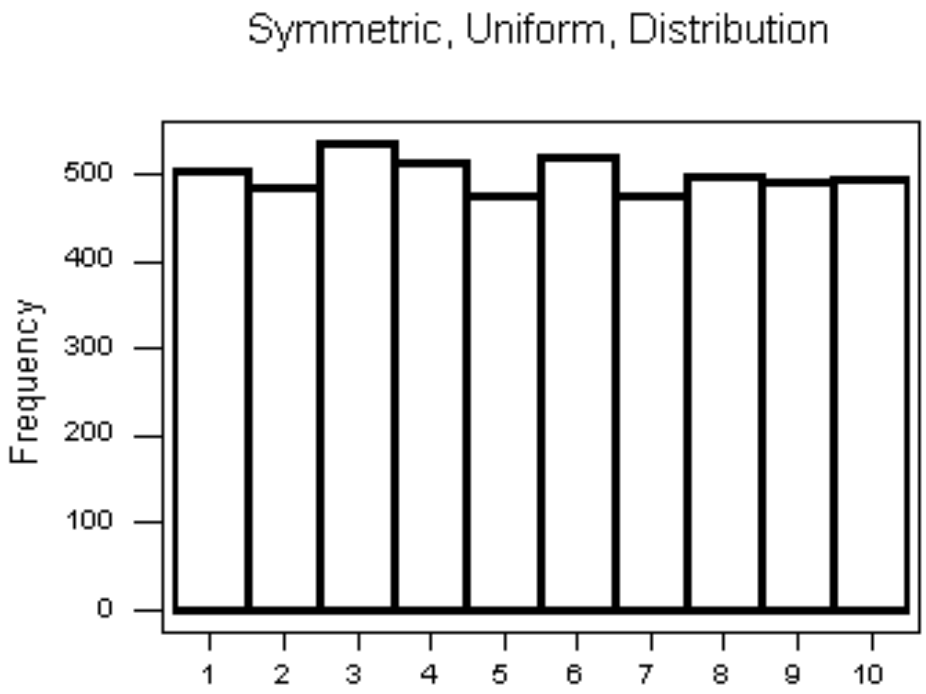


Phân bố đối xứng: symmetric, unimodal

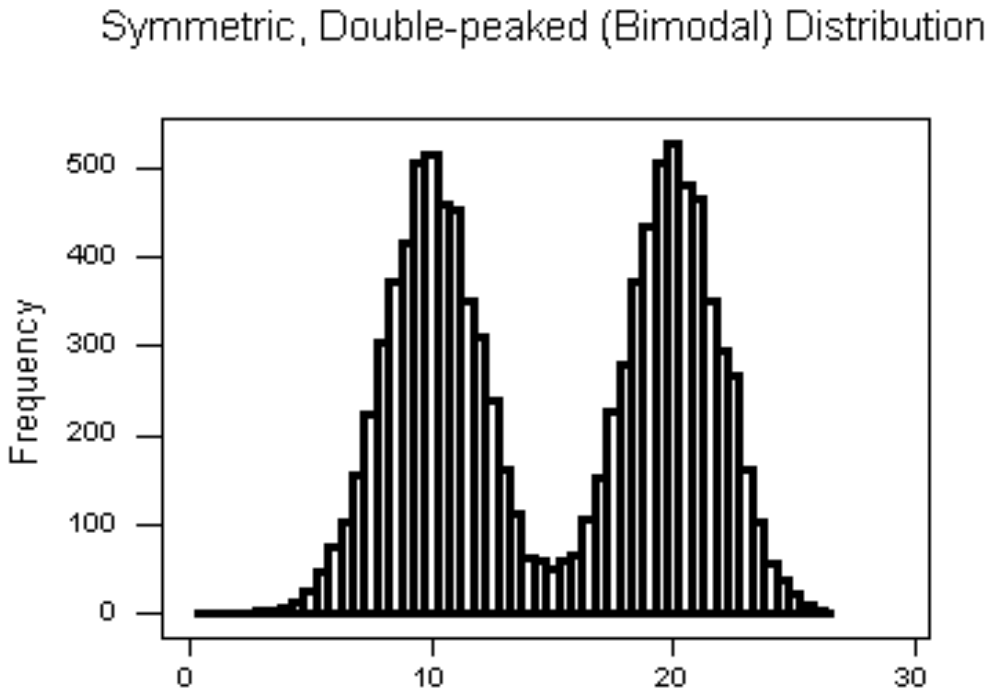
- Hình dạng (shape) của phân bố giúp ta mô tả độ lệch (skewness) như và dạng thức (modality) của dữ liệu



Phân bố đối xứng: symmetric, uniform

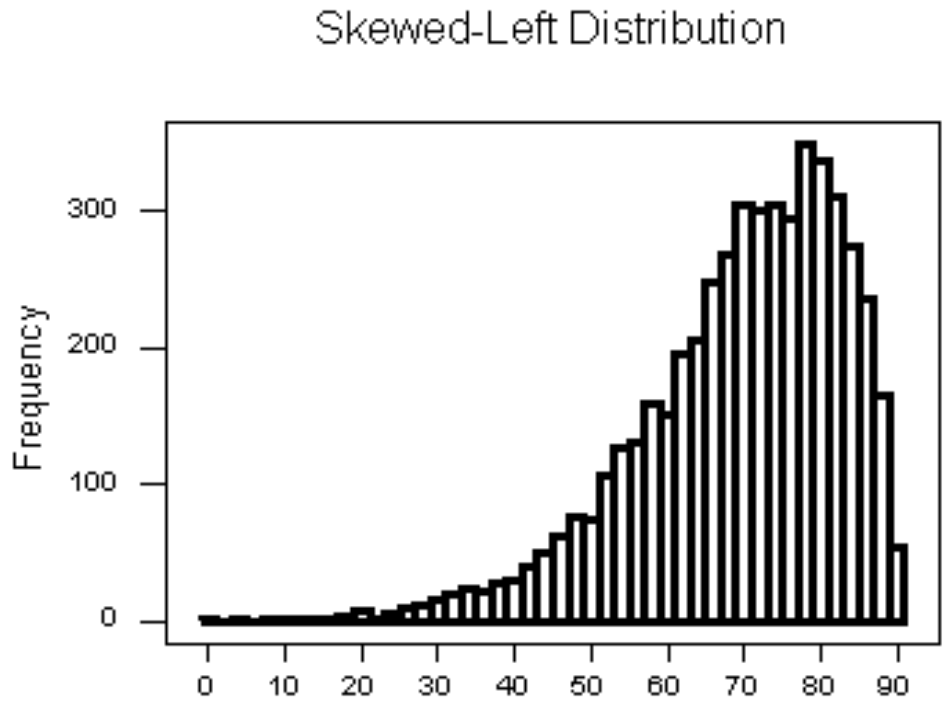


Phân bố đối xứng: symmetric, bimodal

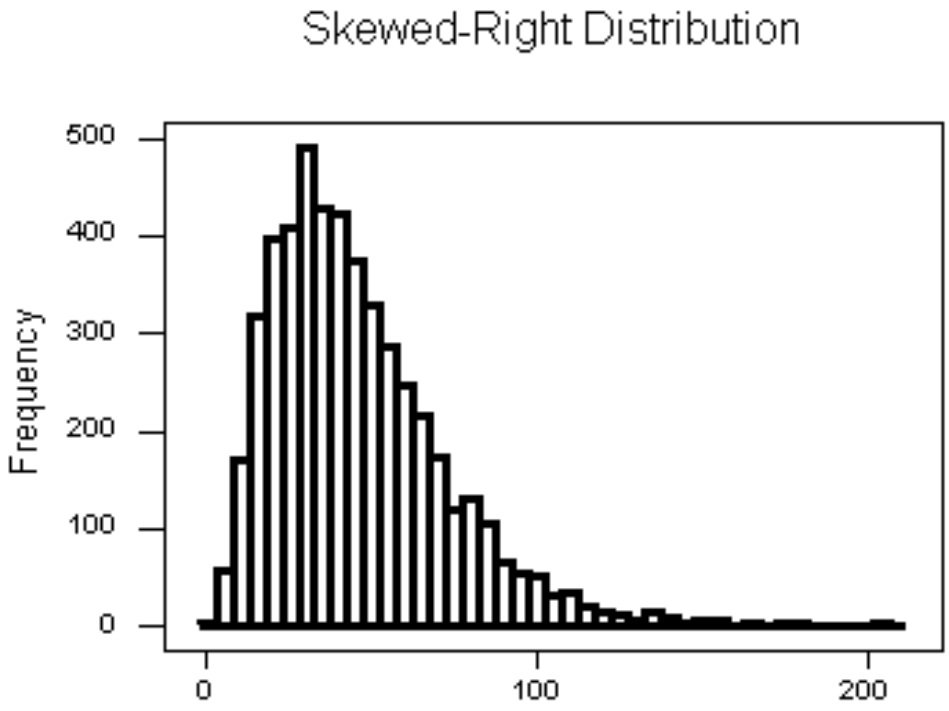


- Nếu dữ liệu có nhiều hơn hai mode, ta nói phân bố của nó là multimodal.

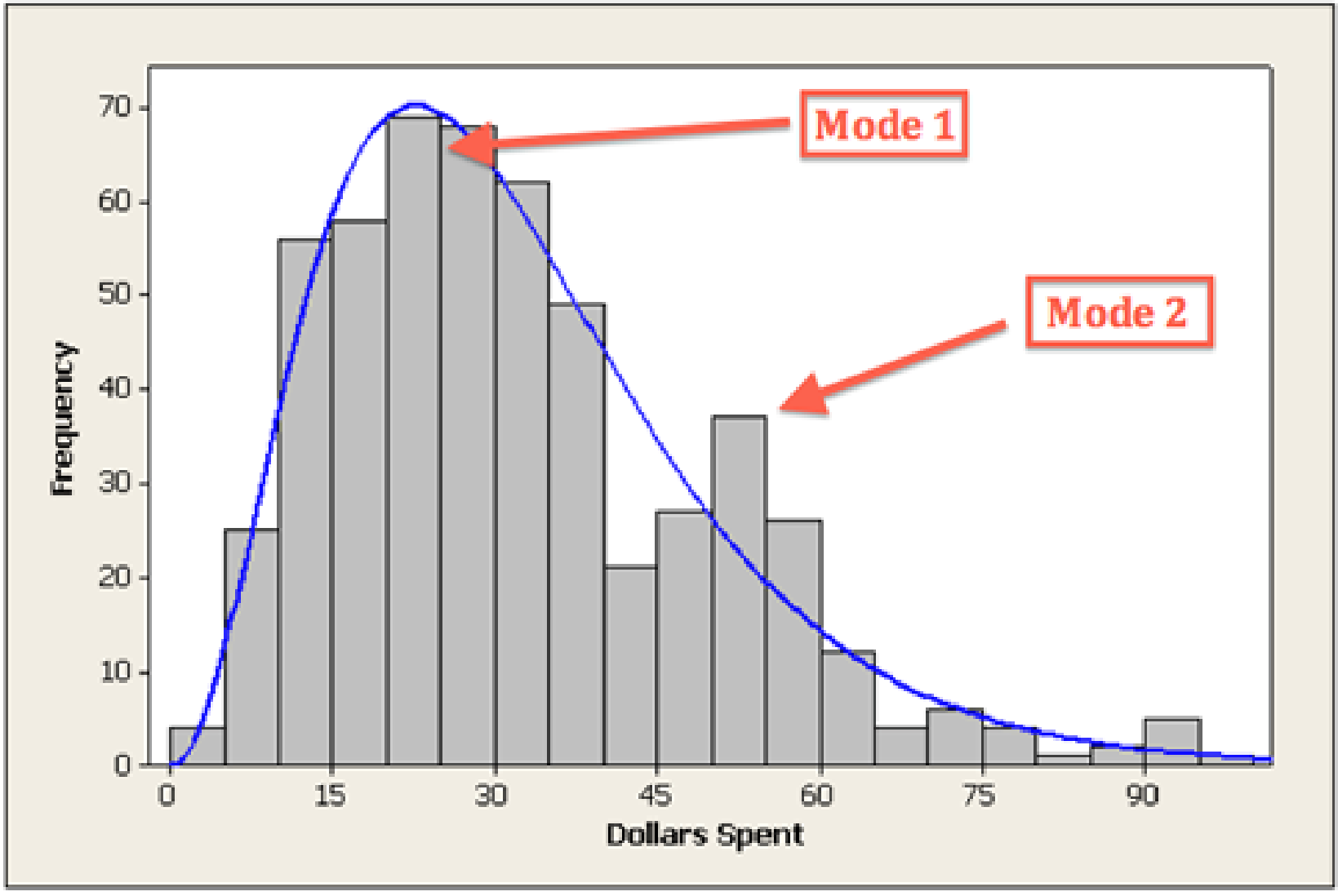
Phân bố lệch: skewed left



Phân bố lệch: skewed right



Phân bố lệch: skewed right, bimodel



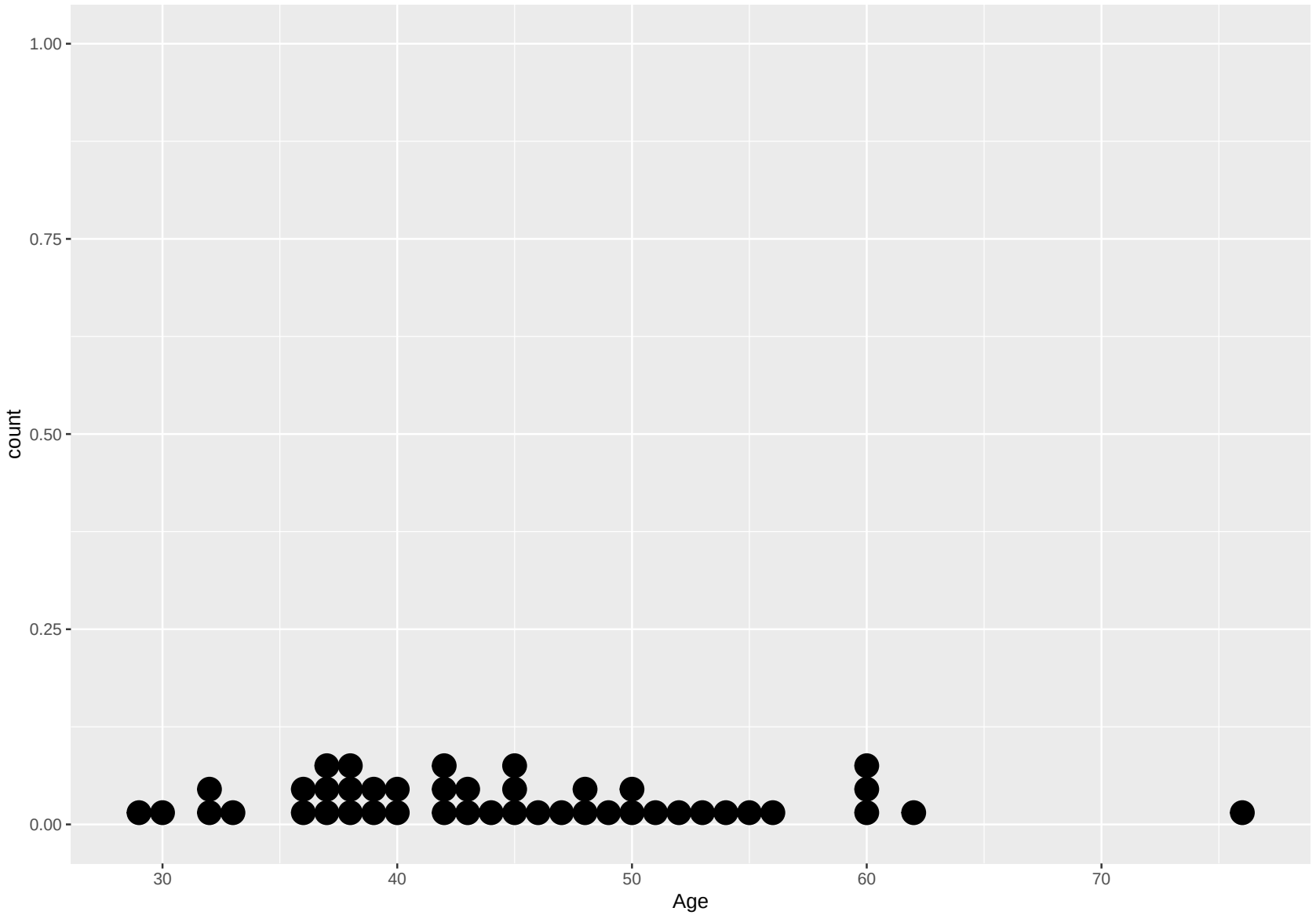
Biểu đồ stemplot

- 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.

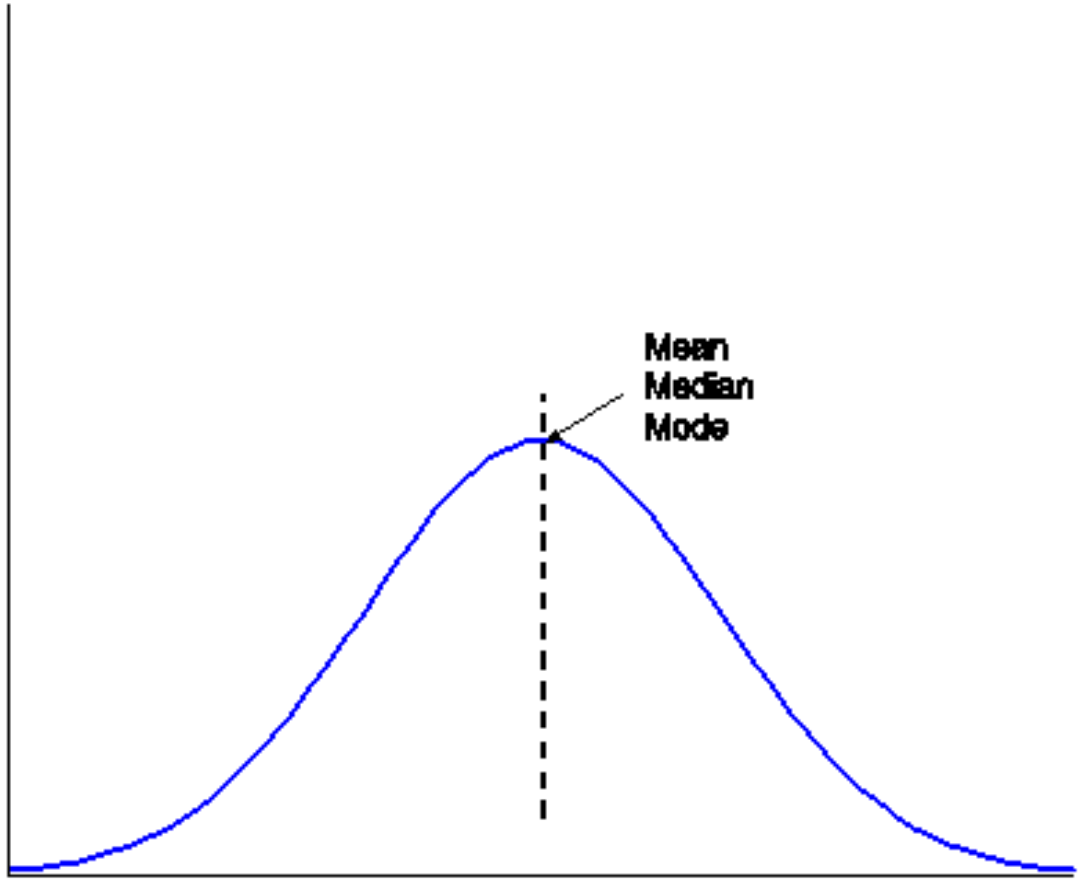
```
##  [1] 43 42 48 49 56 38 60 30 40 42 37 76 39 53 45 36 62
## [24] 37 38 32 45 60 46 40 36 47 29 43 37 38 45 50 48 60
```

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

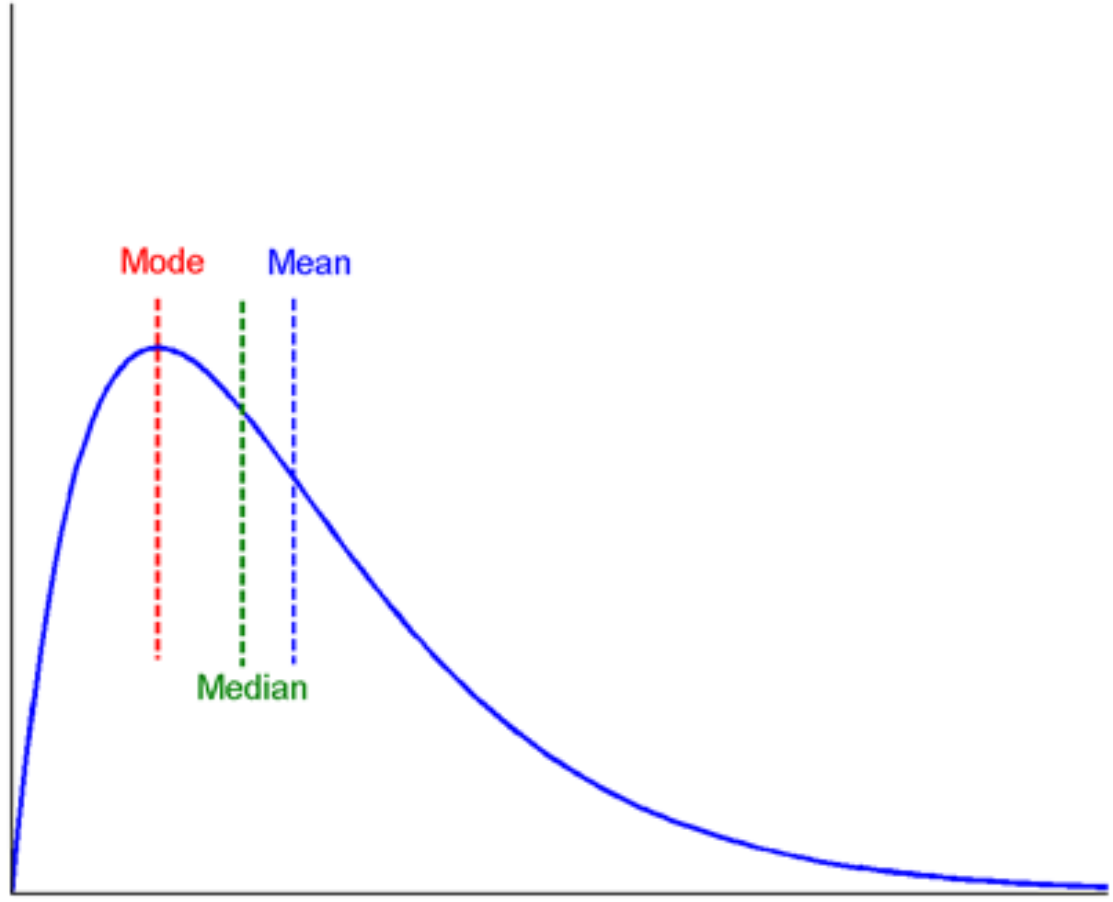
```
## [1] "modes:  42, 38, 60, 37, 45"
```

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

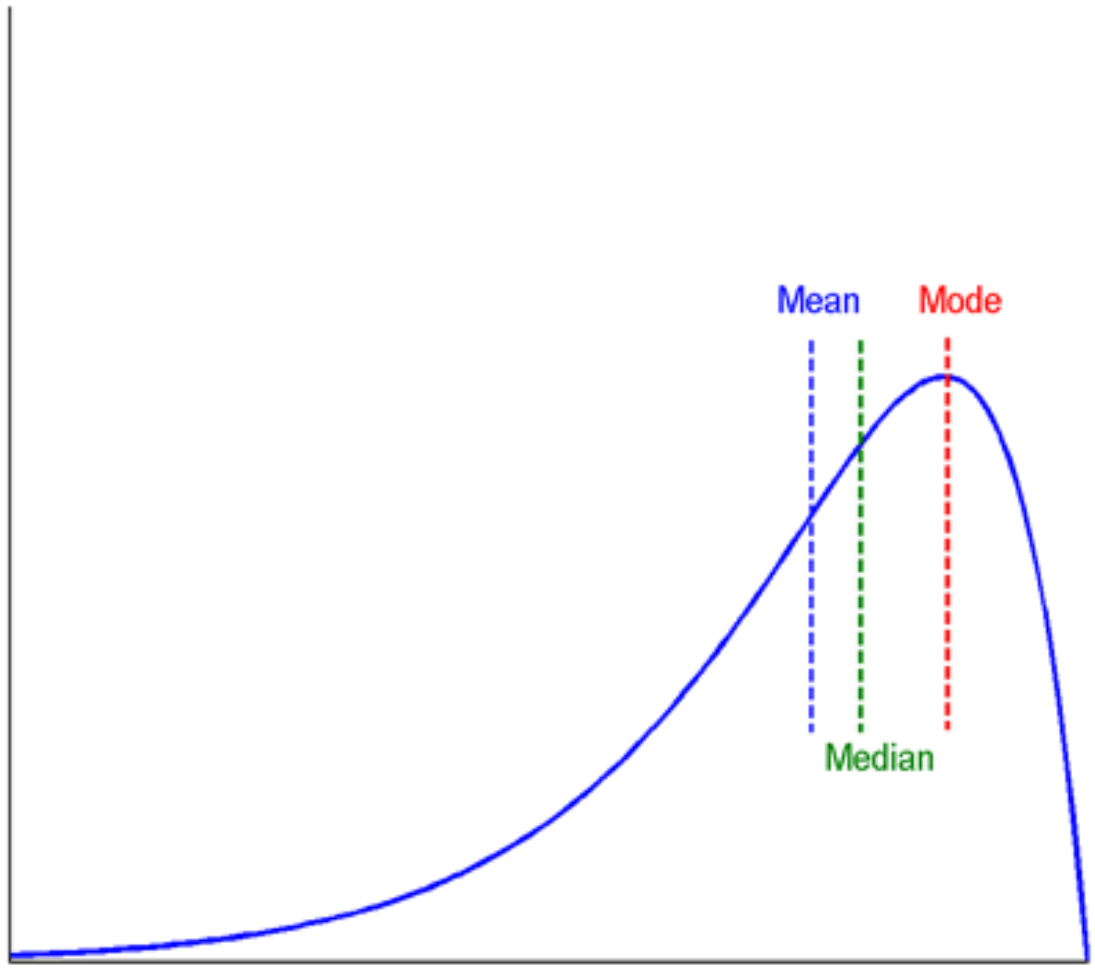
So sánh mean, mode, median: symmetric distribution



So sánh mean, mode, median: skewed right distribution

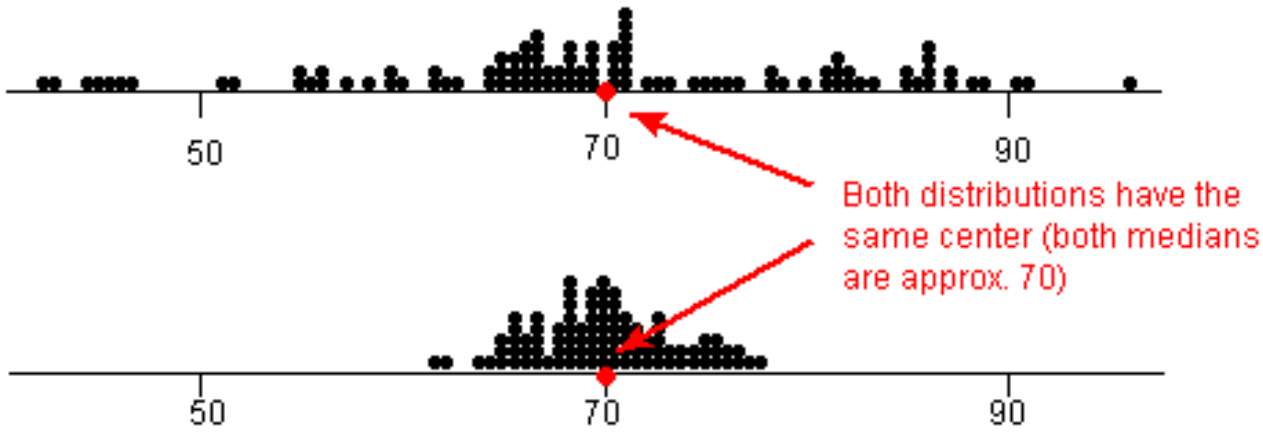


So sánh mean, mode, median: skewed left distribution



Mức độ phân tán

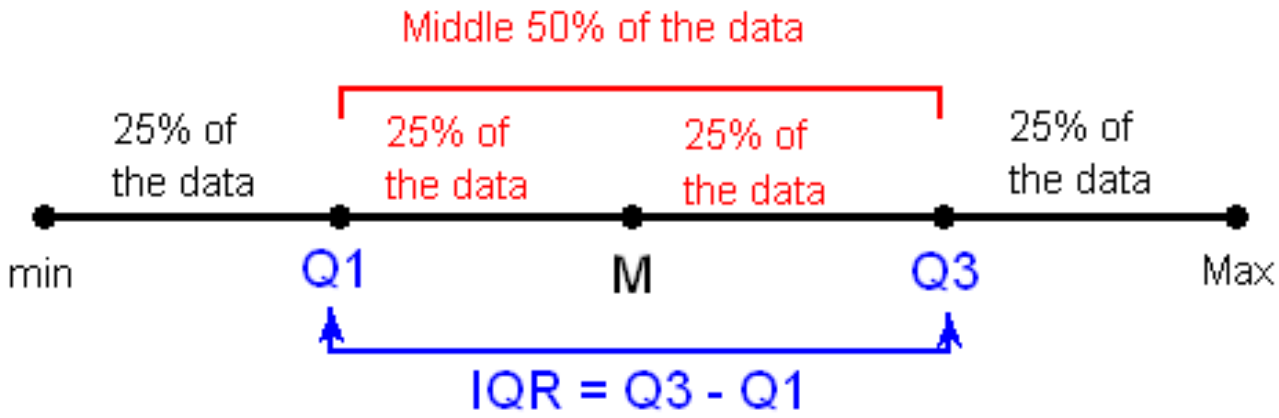
- Các giá trị trung tâm không đủ để đại diện cho một phân bố
  - Hai phân bố khác nhau có thể có các giá trị trung tâm giống nhau



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

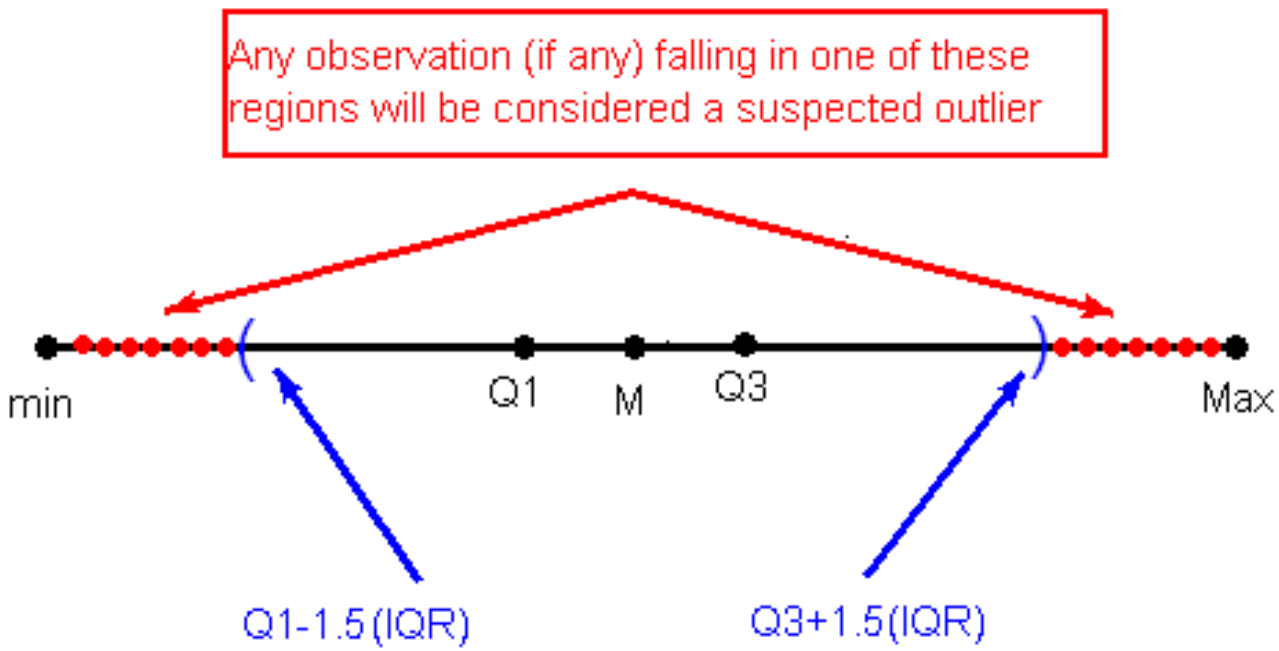
Range và inter-quartile range (IQR)

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

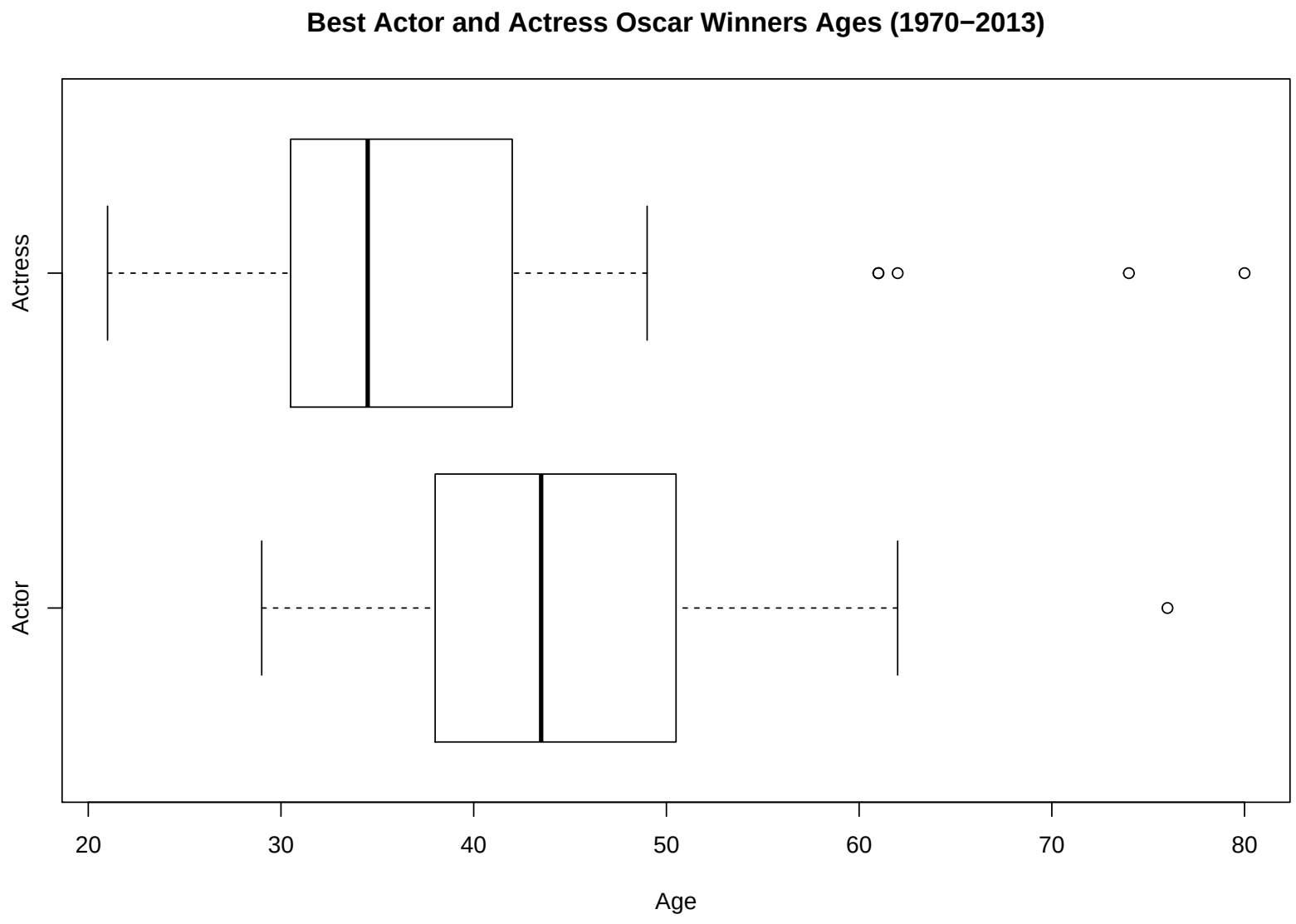


Phát hiện oulier dùng IQR

- 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
## [24] 37 38 32 45 60 46 40 36 47 29 43 37 38 45 50 48 60
```

```
## [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"
```

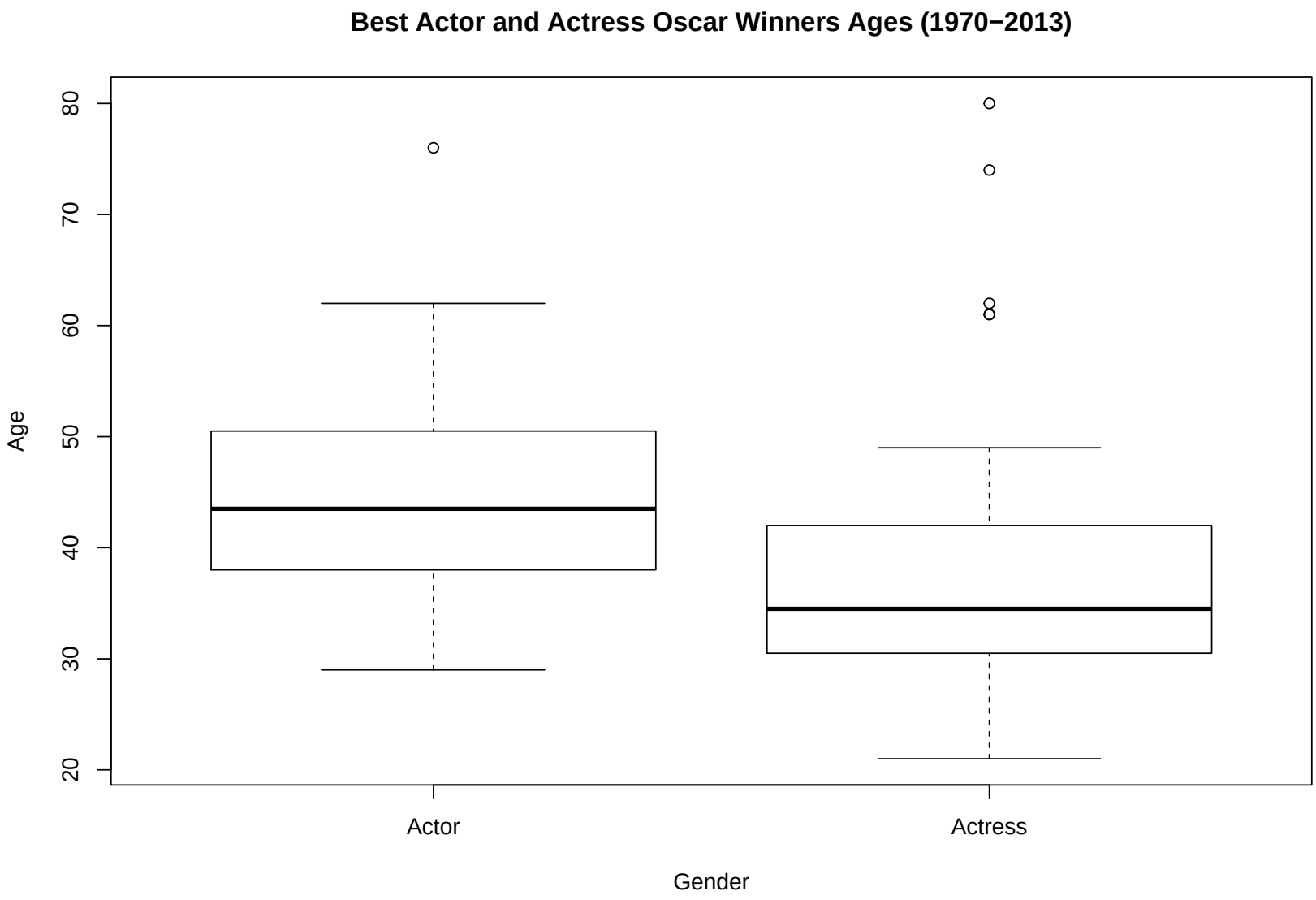
```
## [1] "var = 95"
```

Phân tích mối quan hệ giữa hai biến

- Khi phân tích mối quan hệ giữa hai biến, ta thường phân biệt vai trò của chúng.
  - Biến giải thích (explanatory/independent variable)
  - Biến phụ thuộc (response/dependent variable)
- Có 4 loại mối quan hệ giữa hai biến

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

$C \to Q$

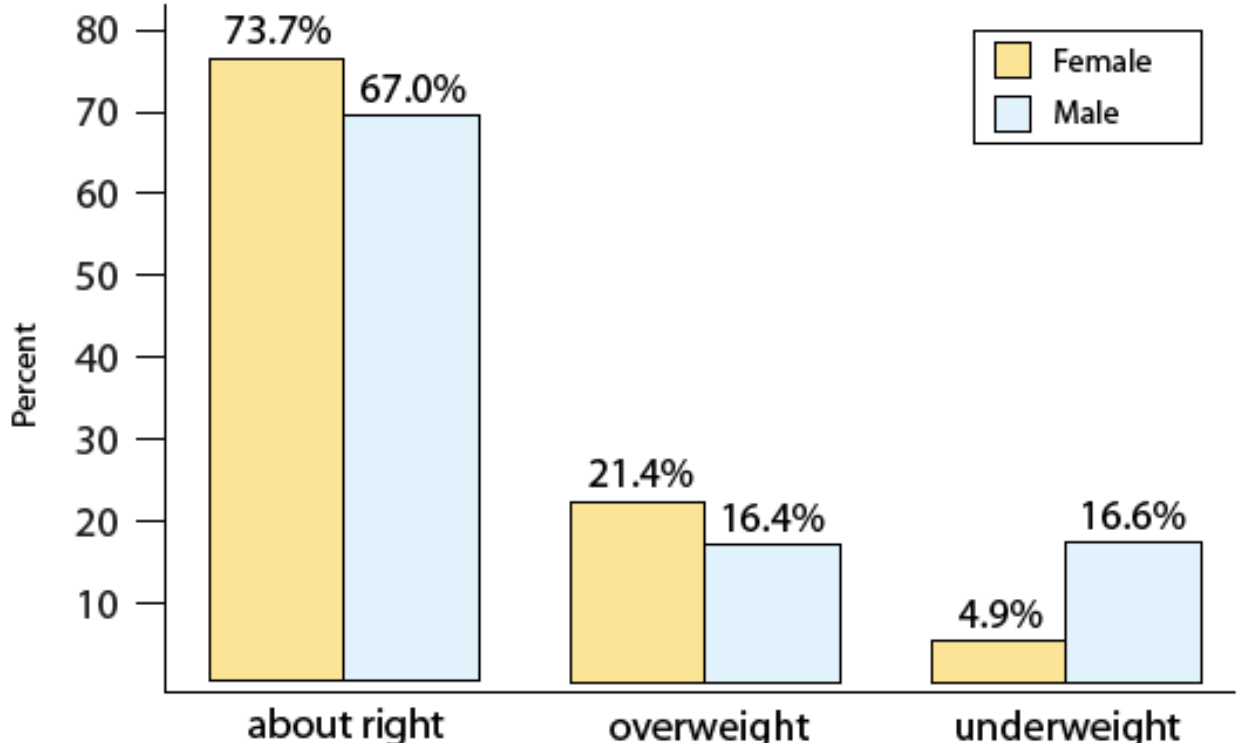

Best Actor and Actress Oscar Winners Ages (1970−2013)
Age
20
30
40
50
60
70
80
Actor
Actress
Gender

$C \to C$

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

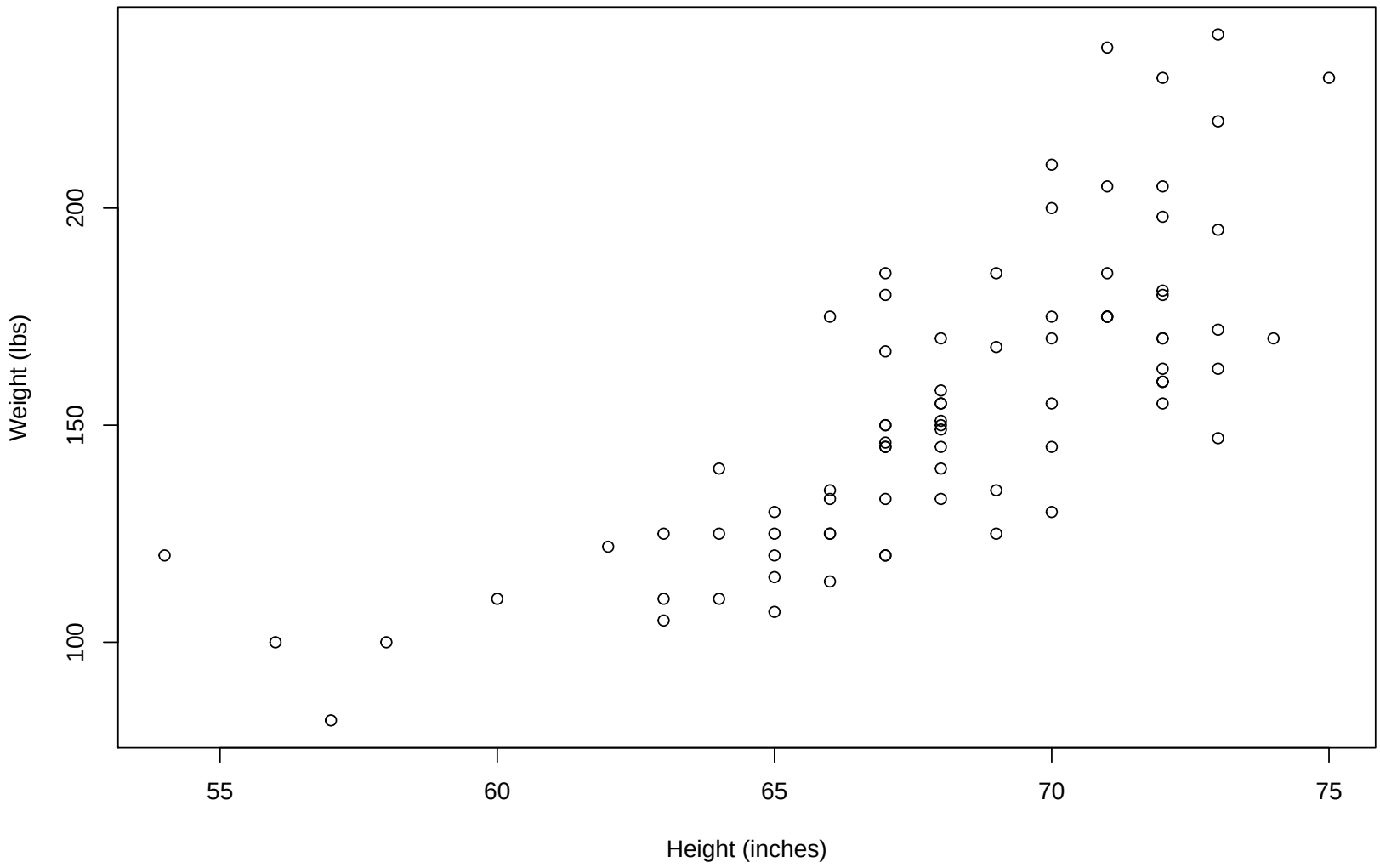
$C \to C$

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

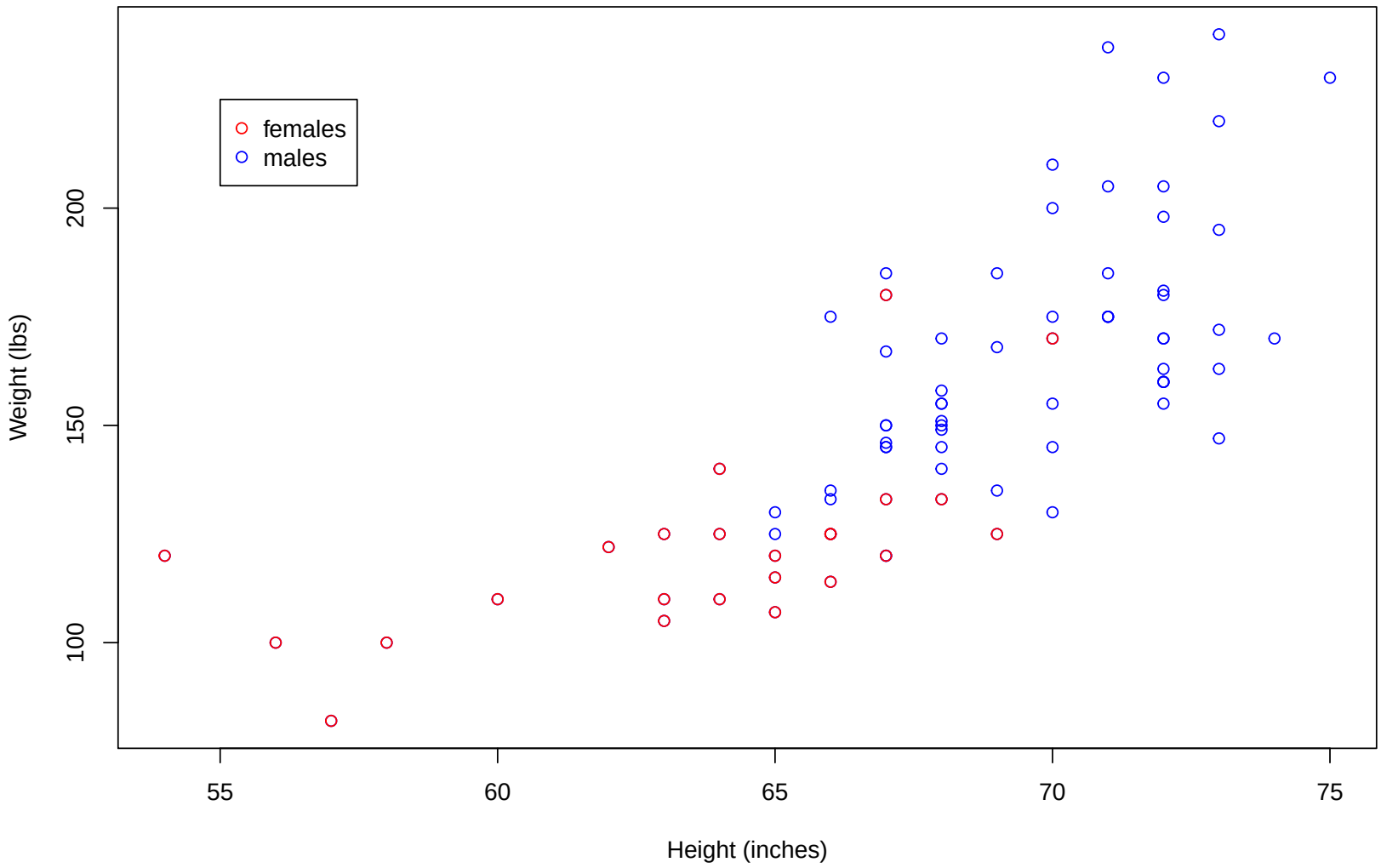


$Q \rightarrow Q$

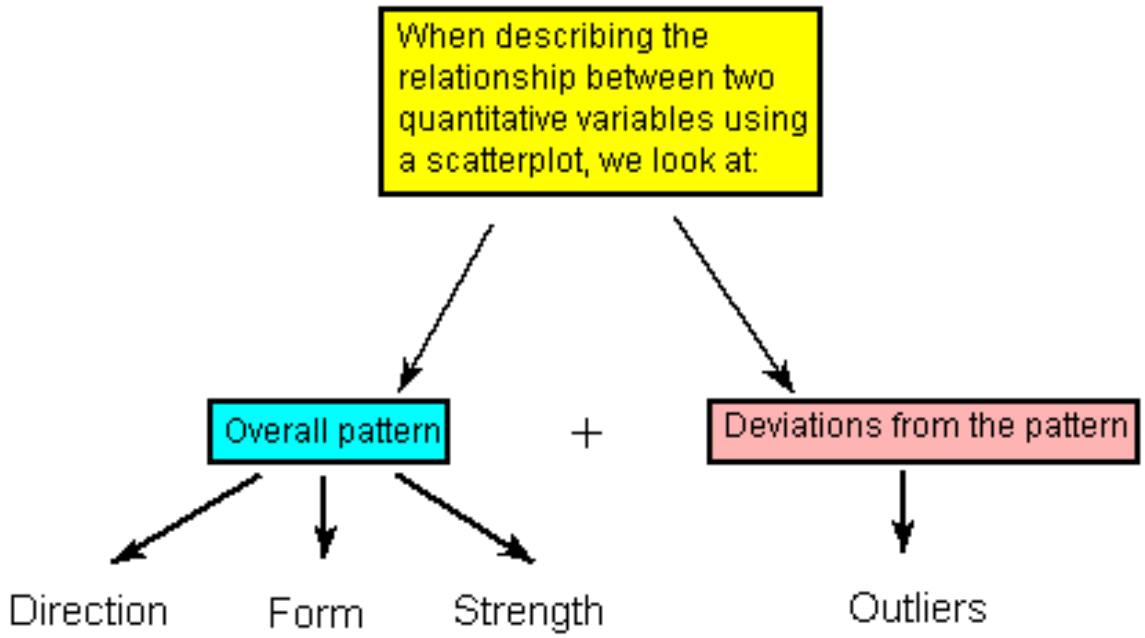
- Biểu đồ scatterplot



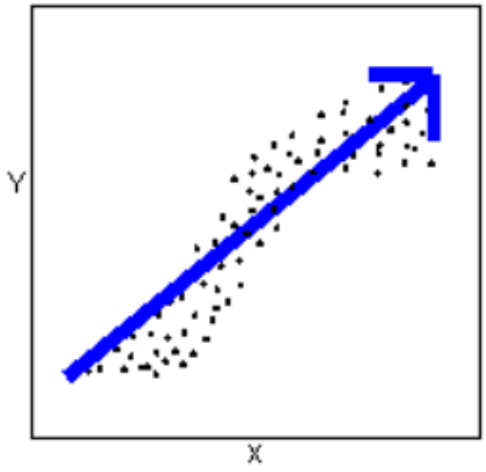
$Q \rightarrow Q$



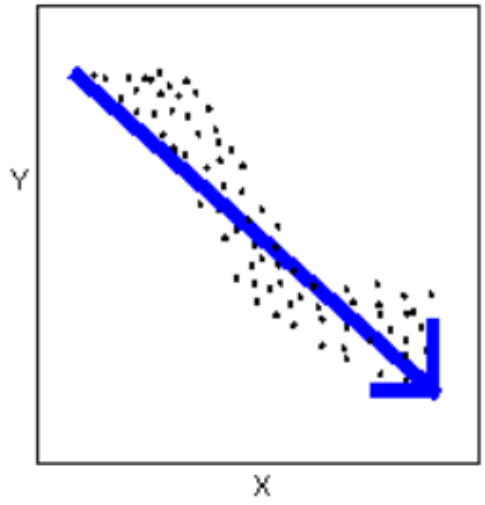
Hiểu scatterplot



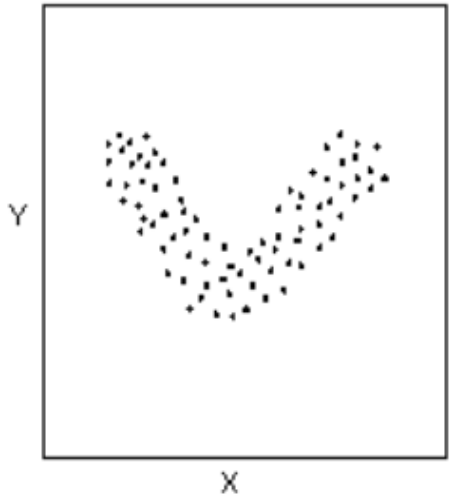
Direction của relationship



Positive relationship

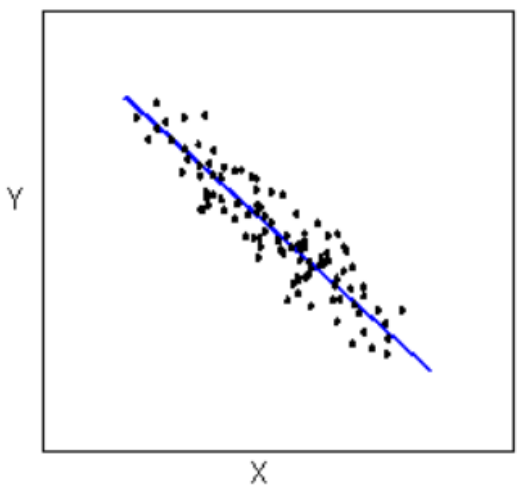


Negative relationship

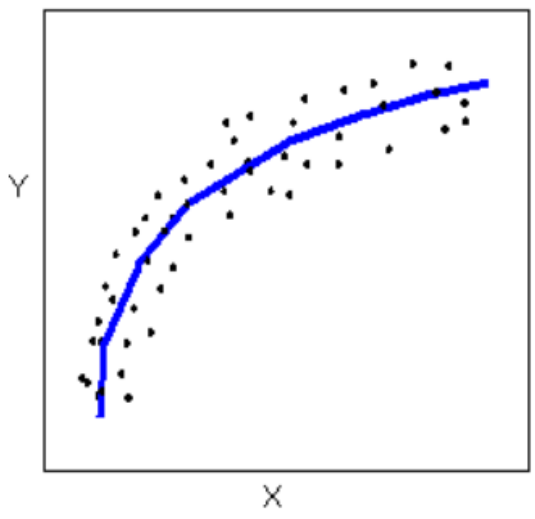


Neither positive nor negative

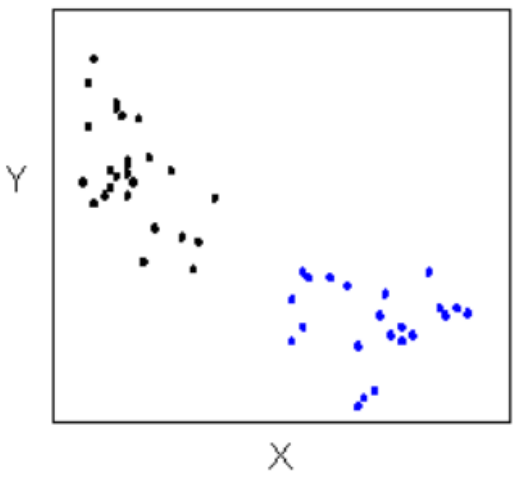
Form của relationship



linear

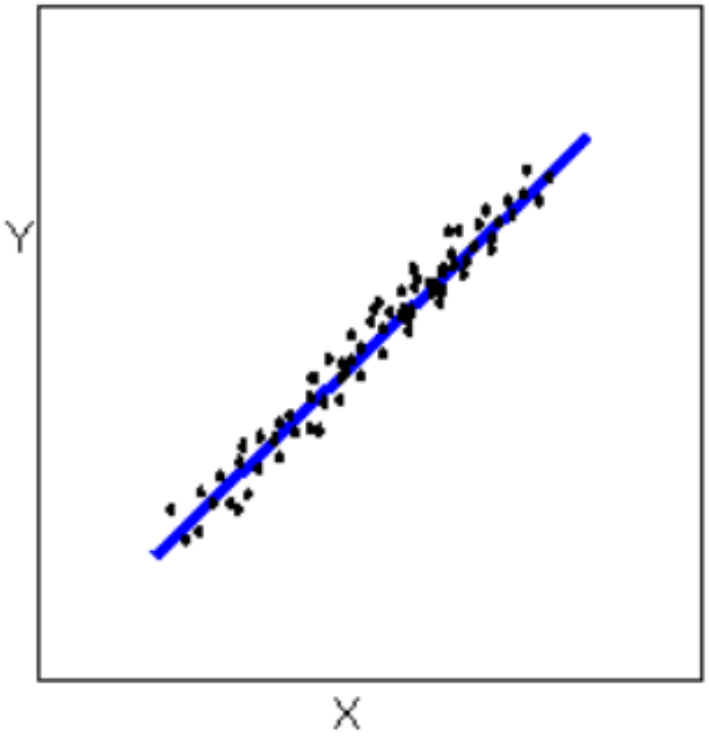


curvilinear

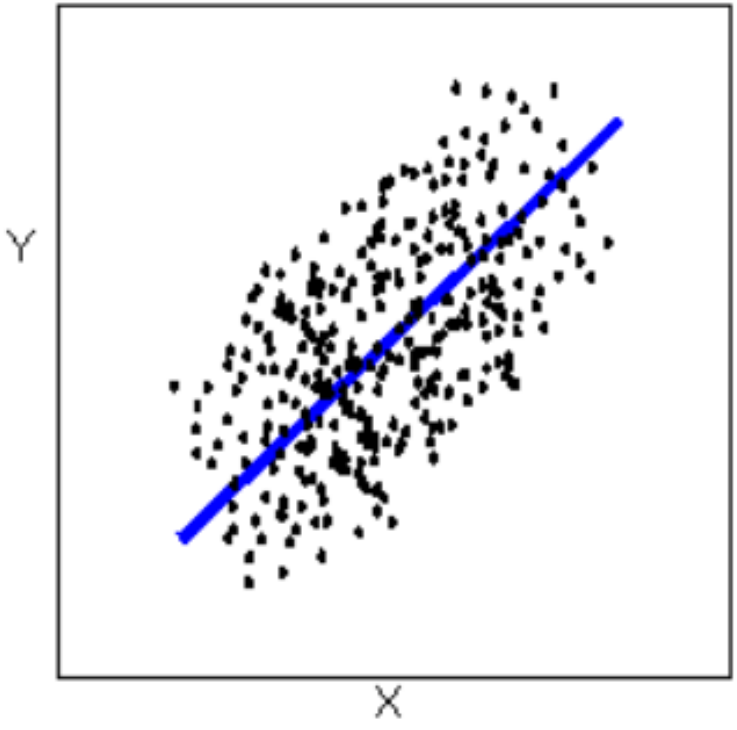


others

Strength của relationship



strong relationship



weaker relationship

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

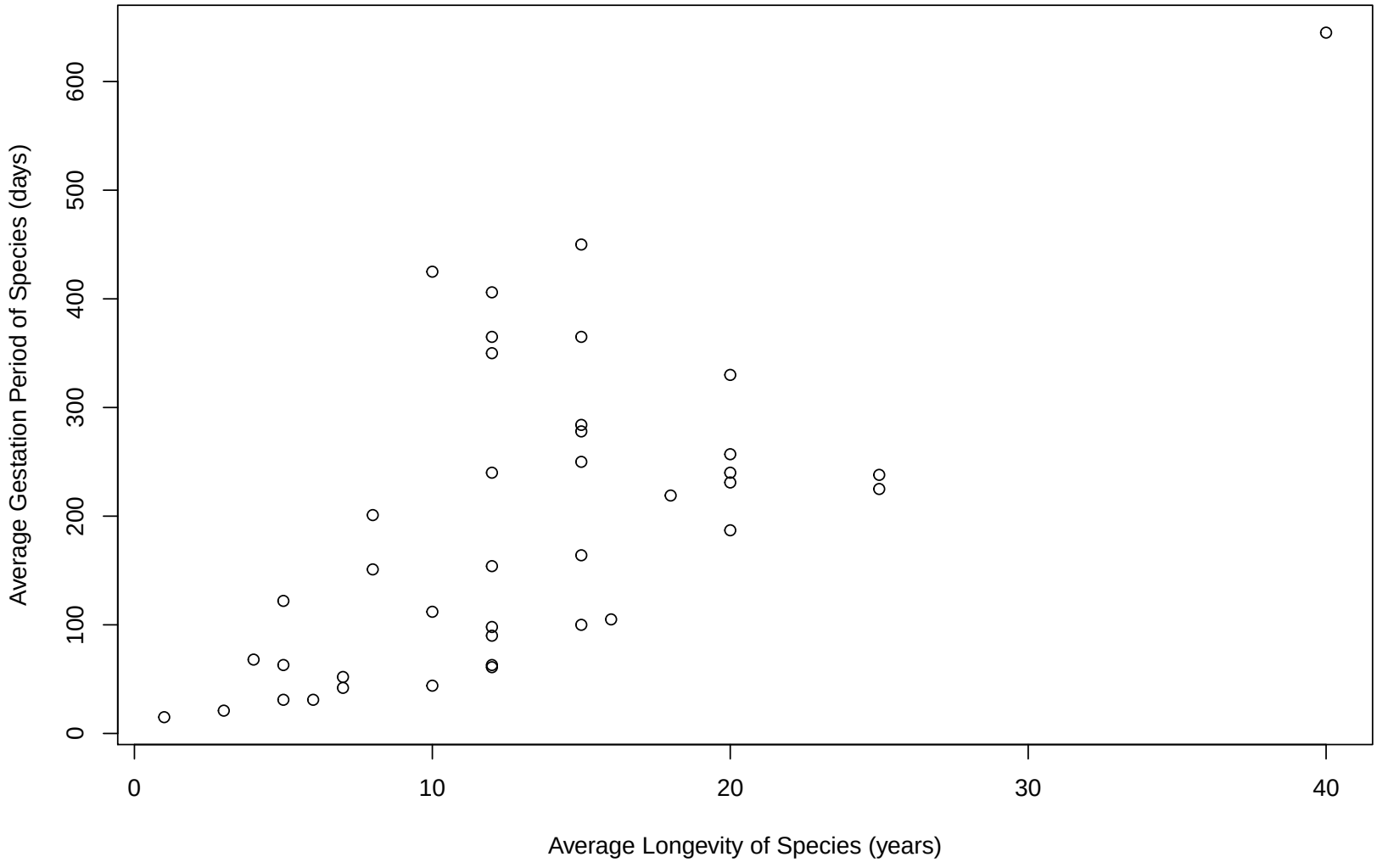
Sự tương quan (correlation)

- 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



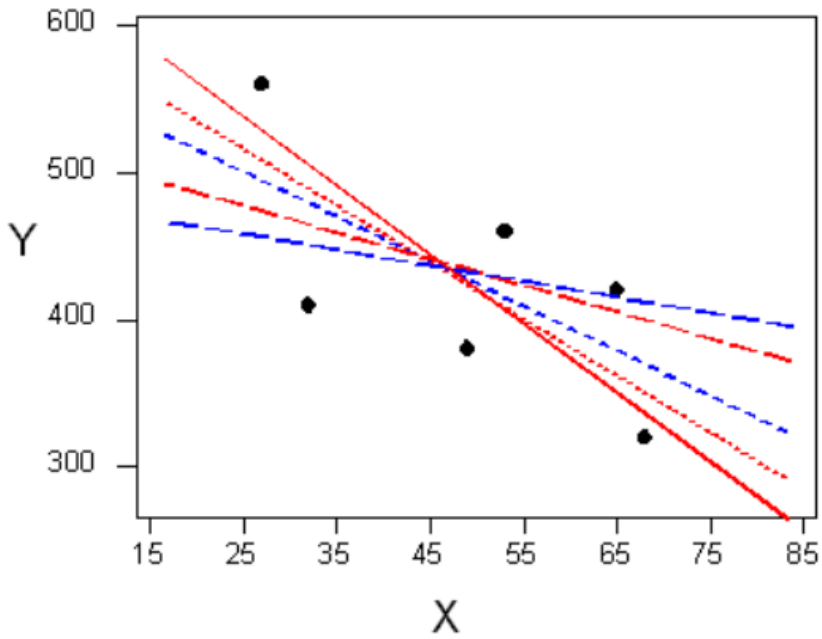

```
## [1] 0.6632397
```

Hồi quy (regression)

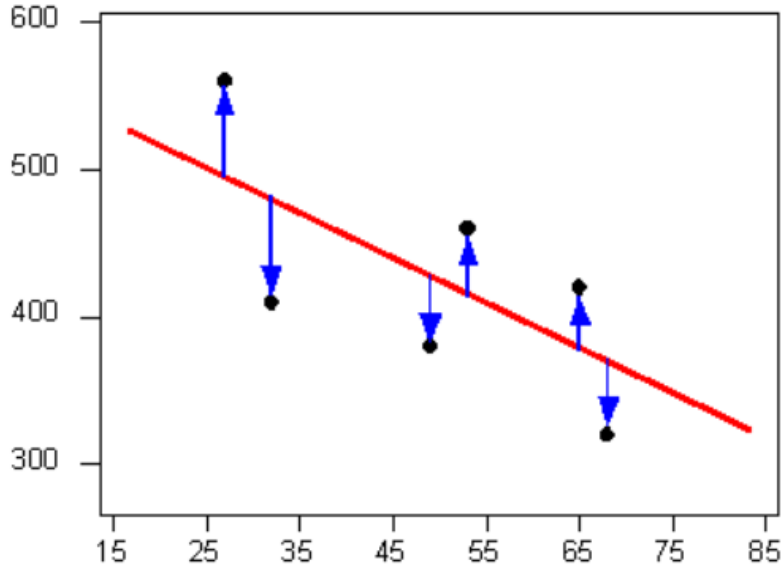
- Sự tương quan không mô tả hết mối quan hệ tuyến tính của hai biến định lượng
  - Nó chỉ mô tả strength và direction của quan hệ
- Ta thường muốn hiểu rõ hơn biến này ảnh hưởng đến biến kia như thế nào
  - Ta muốn dự đoán (predict) giá trị của response variable với giá trị của explanatory variable cho 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)



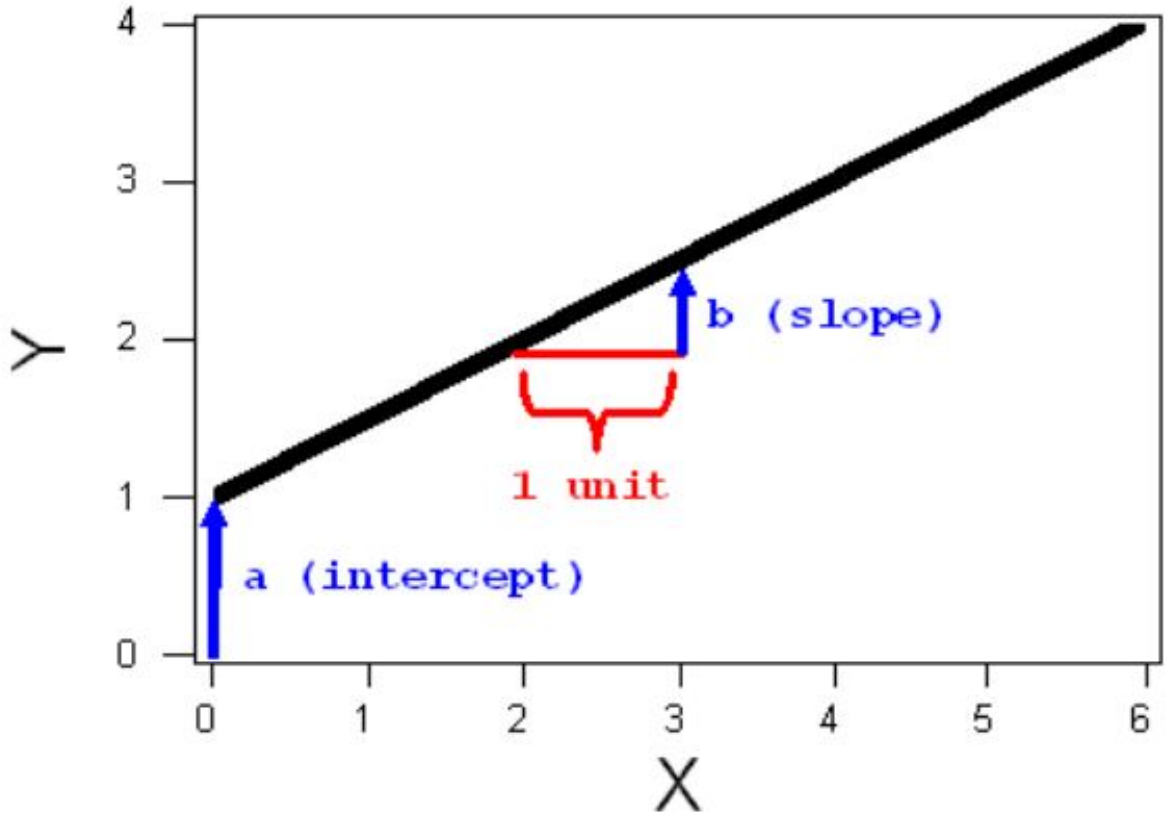
many candidates



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:

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

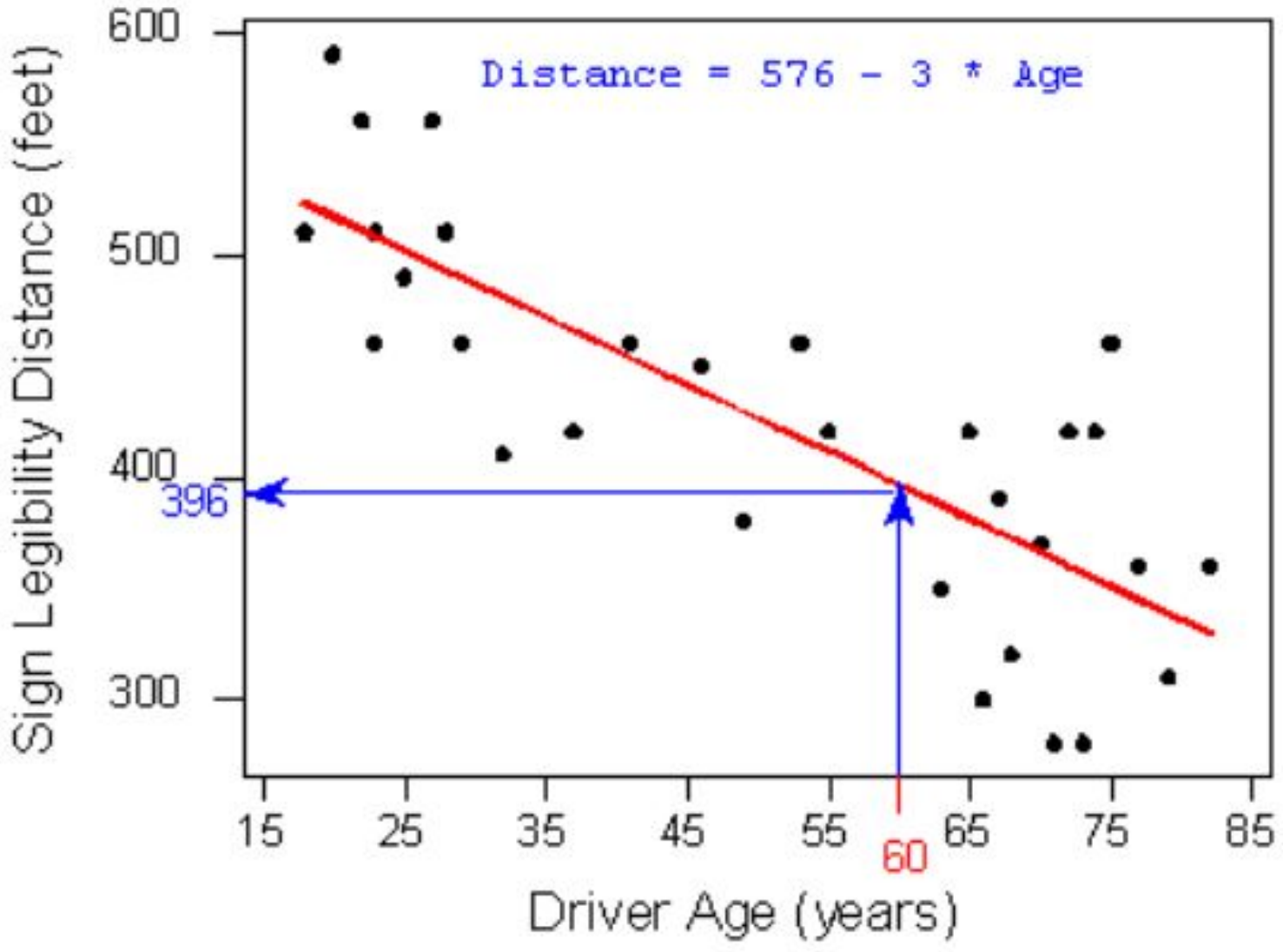
$$a = \overline{Y} - b\overline{X}$$

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

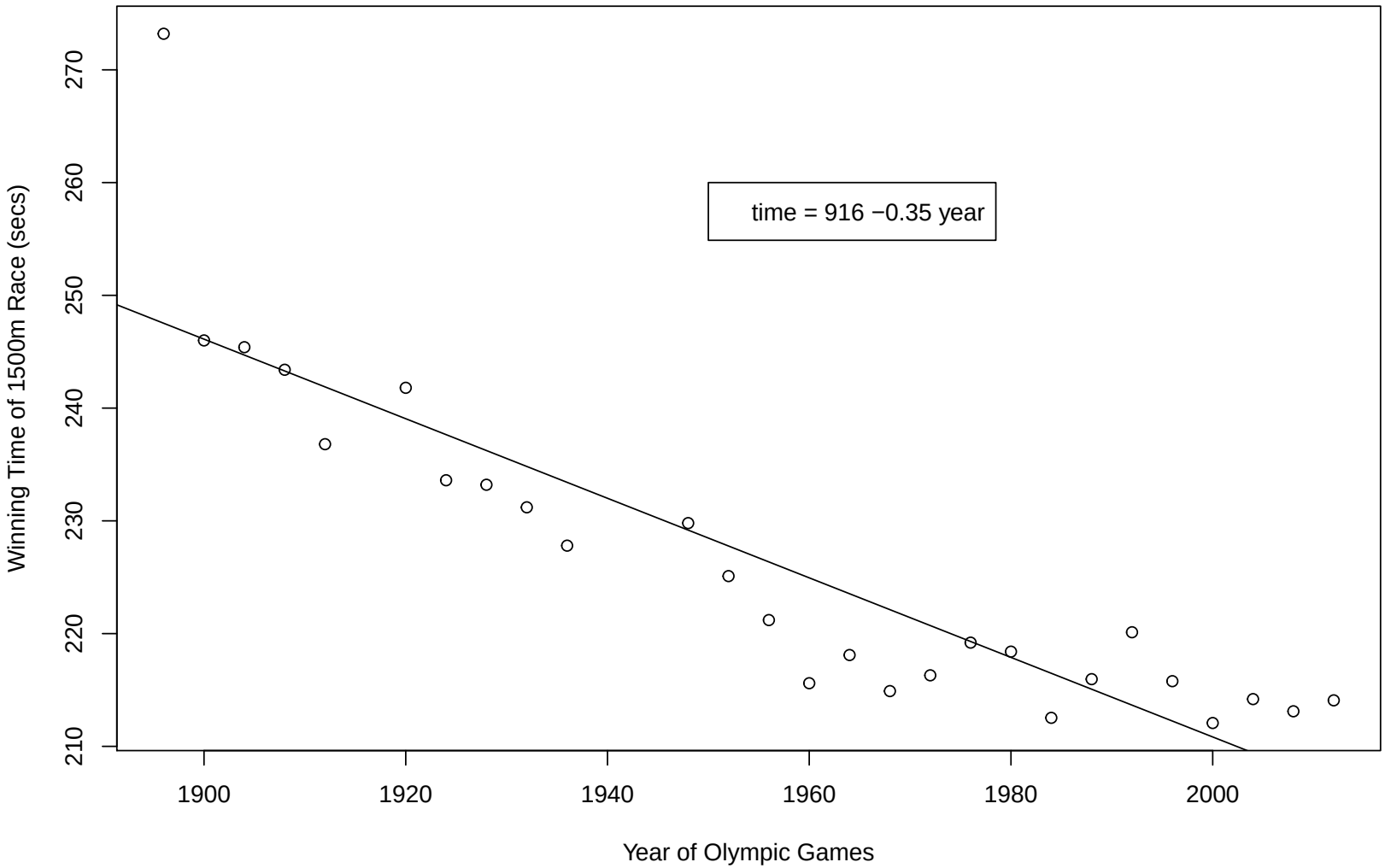
Dự đoán (prediction)

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



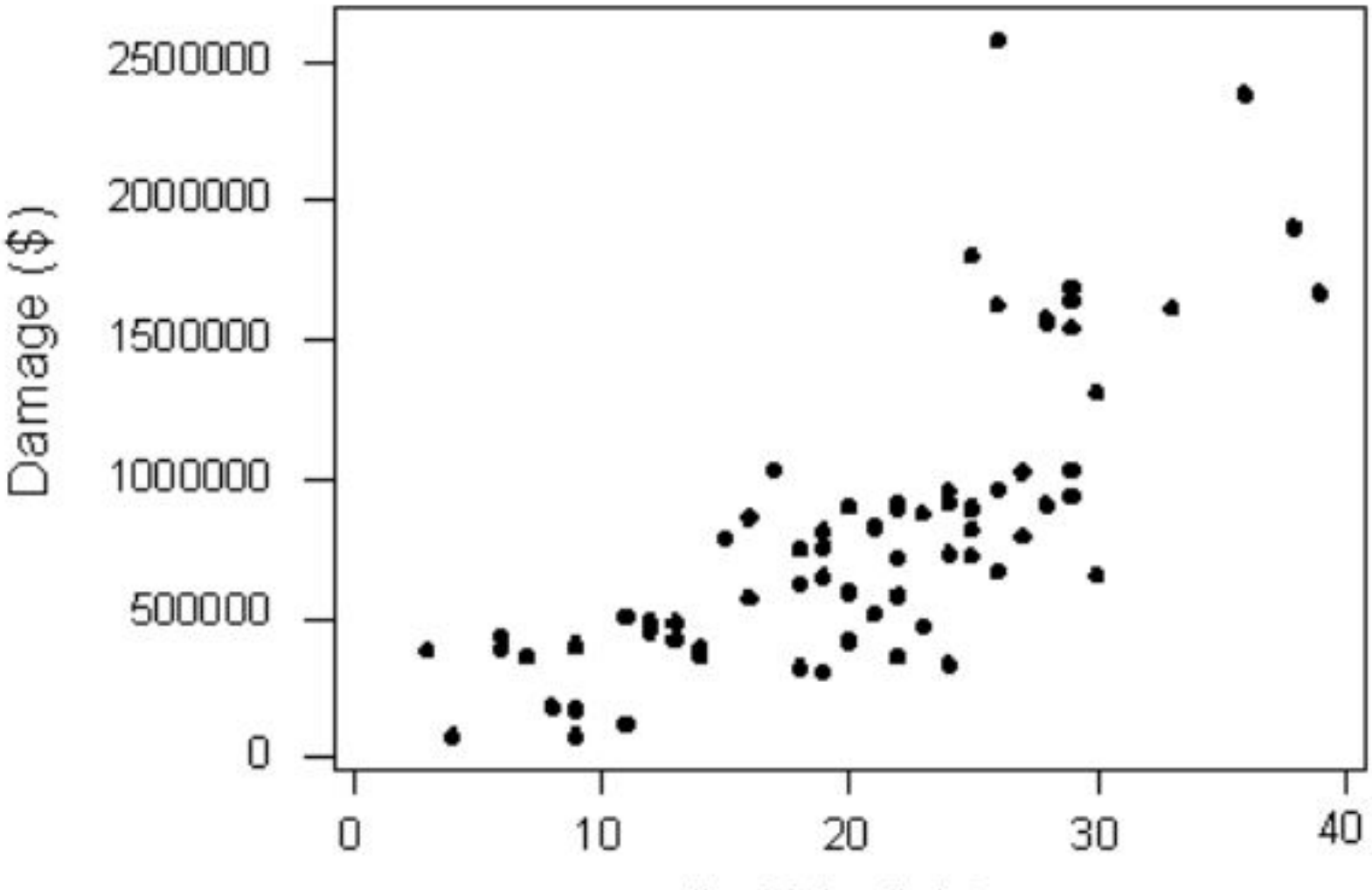
Least squares regression line

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

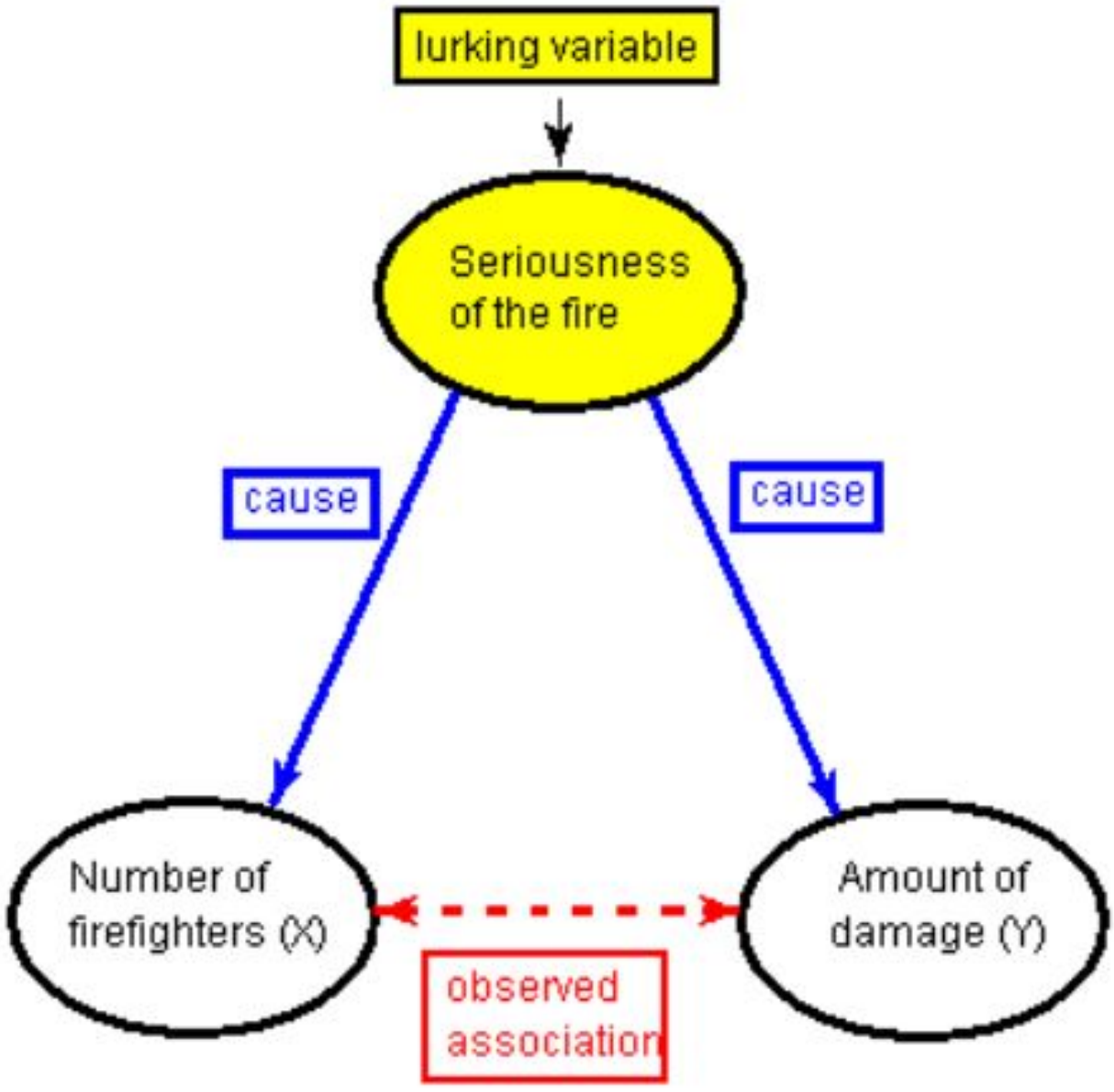


Quan hệ nhân quả (causation)

- 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.



Sự tương quan không suy ra quan hệ nhân quả



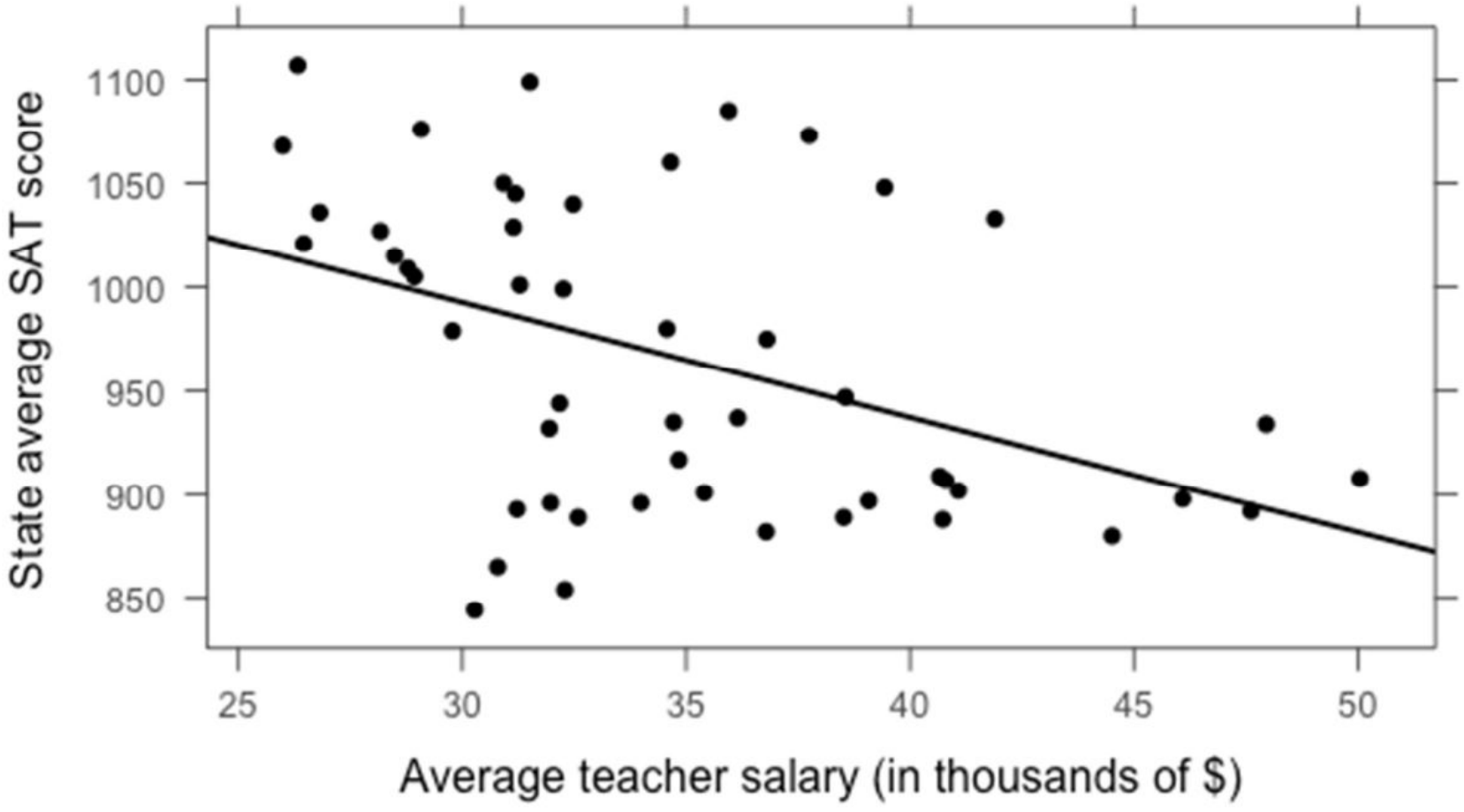
Simpson's Paradox

SMOKER	Alive	Dead
No	502 (68.6%)	230 (31.4%)
Yes	443 (76.1%)	139 (23.9%)

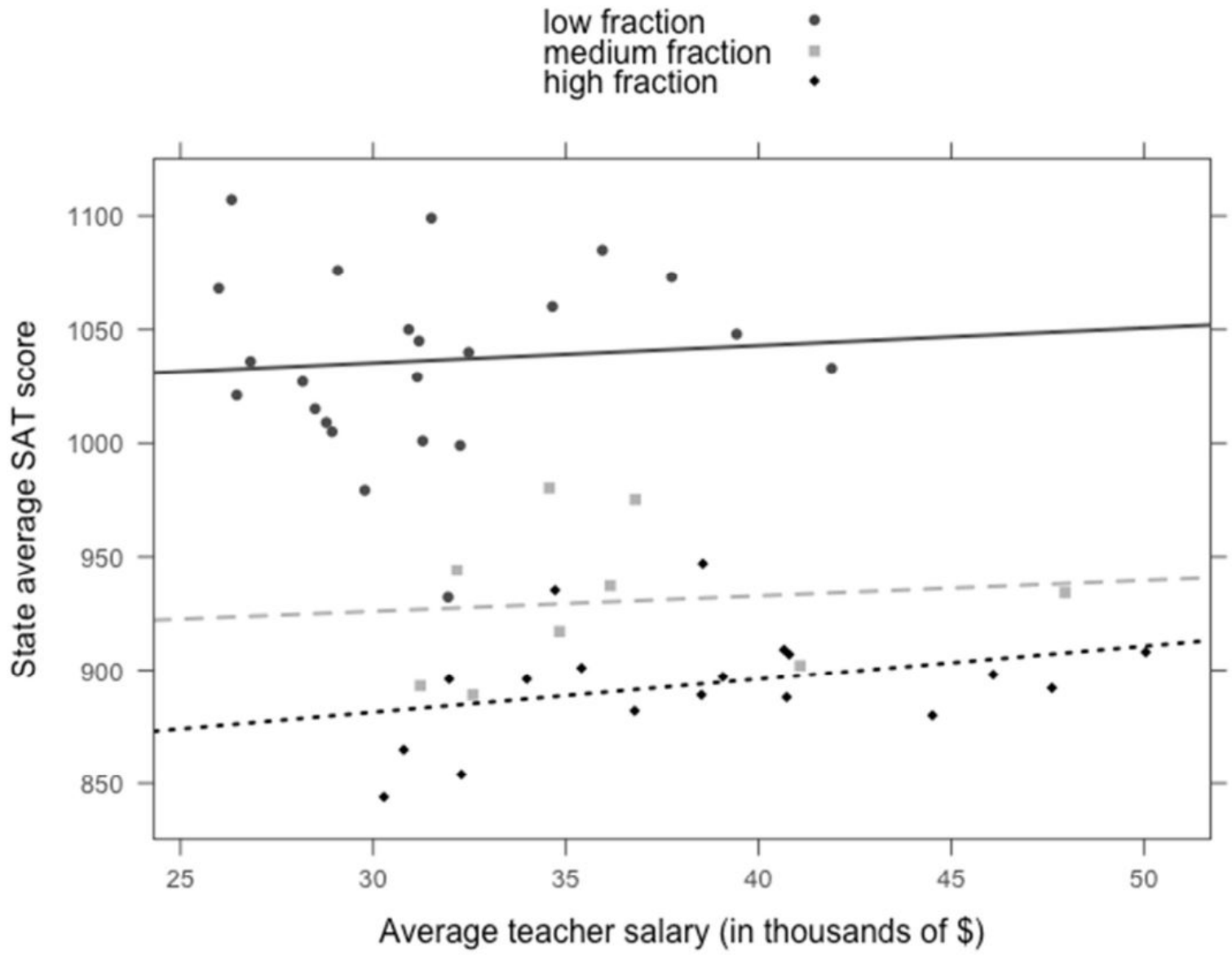
Baseline age	SMOKER	Alive	Dead
18-64	No	474 (87.9%)	65 (12.1%)
18-64	Yes	437 (82.1%)	95 (17.9%)
65+	No	28 (14.5%)	165 (85.5%)
65+	Yes	6 (12.0%)	44 (88.0%)

Age roup	Non-smoker	Smoker
18-64	539 (50.3%)	532 (49.7%)
65+	193 (79.4%)	50 (20.6%)

Simpson's Paradox



Simpson's Paradox



Tổng kết

- 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

- Phân tích mối quan hệ của hai biến

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

Tổng kết

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