B5. Descriptive Statistics

Bổ sung thêm cho bài giảng

2019

Nội dung bổ sung

1. **Descriptive statistics**
2. **Measures of central tendency**
3. **Mesures of dispersion**
4. **Measures of shape**

1. Descriptive statistics

- **Dữ liệu** (*data*)

- **Thông tin** (*information*)

- **Tri thức** (*knowledge*)

1. Descriptive statistics (tt.)

- **Analytics**
  - descriptive
  - predictive
  - prescriptive

1. Descriptive statistics (tt.)

- **Hình thức mô tả (tóm tắt) dữ liệu**
  - bảng
  - biểu đồ, đồ họa
  - số: vị trí, độ phân tán, hình dáng, mối liên hệ

1. Descriptive statistics (tt.)

- **Primary data**
  - Primary data is data that is collected by a researcher from first-hand sources, using methods like surveys, interviews, or experiments. It is collected with the research project in mind, directly from primary sources
- **Secondary data**
  - Secondary data is data gathered from studies, surveys, or experiments that have been run by other people or for other research.

https://www.statisticshowto.datasciencecentral.com/primary-data-secondary/ (09/2019)

1. Descriptive statistics (tt.)

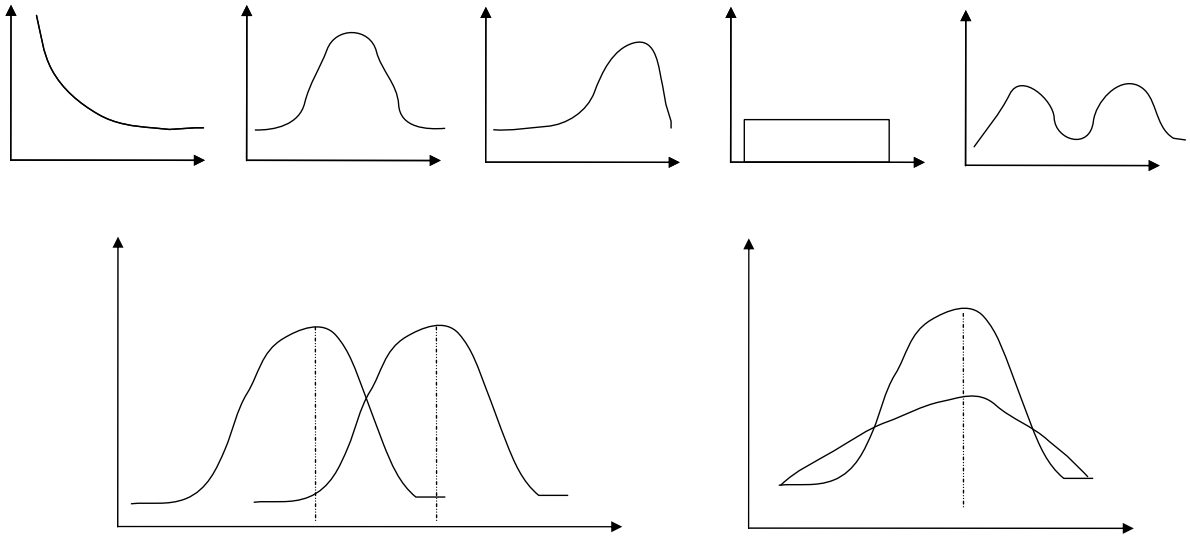
- **Tập dữ liệu** (*data set*)
  - rời rạc (discrete data): có thể "đếm được" (counted)
  - liên tục (continuous data): có thể "đo lường được" (measured) trên một thang đo (scale) → đơn vị, thứ nguyên → chia nhỏ
  - phân nhóm (grouped data): class intervals
  - nominal, ordinal, numerical
- **Phân phối** ([*frequency*] *distribution*)

1. Descriptive statistics (tt.)

- **Thuật ngữ**
  - *population*: tổng thể/quần thể (các đối tượng)
  - *sample*: mẫu (tập con của quần thể đang được khảo sát)

1. Descriptive statistics (tt.)

Central Tendency, Variability, Shape



Different Central Tendency

Different Variability

Nội dung bổ sung

1. Descriptive statistics
2. **Measures of central tendency**
3. Mesures of dispersion
4. Measures of shape

2. Measures of central tendency

- Các đại lượng thể hiện vị trí "trung tâm", điểm đặc trưng
  - mean
  - median
  - mode

2. Measures of central tendency (tt.)

- Đại lượng trung bình (*mean*)

$$\mu_{discrete} \rightarrow \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad n = |X|$$

$$\mu_{continuous} \rightarrow \bar{x} = \int_{-\infty}^{+\infty} x.f(x)$$

F(x) = Pr(X ≤ x): *cumulative distribution function*

f(x) = F'(x): *probability density function*, giới hạn ∀x: f(x) > 0

- (continuous, discrete) numeric data → arithmetic mean
- dễ bị tác động bởi outliers (đột biến) và skewed distributions

2. Measures of central tendency (tt.)

❑ **Một số định nghĩa khác**

$$\mu_{Geometric} \rightarrow \bar{x} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

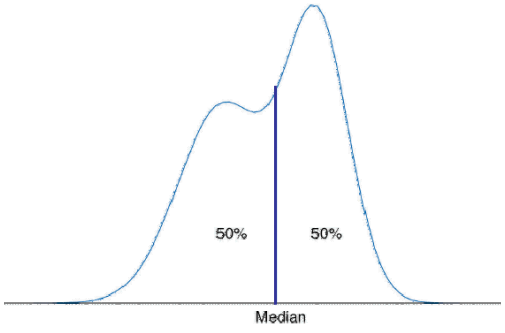
$$\mu_{Harmonic} \rightarrow \bar{x} = n \left(\sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}$$

- Trung bình lọc (*trimmed mean*): loại bỏ x% giá trị nhỏ nhất và x% giá trị lớn nhất → lọc bỏ bớt các giá trị đột biến

❑ **Lưu ý: phân biệt các ký hiệu** $\mu, \bar{x}$

2. Measures of central tendency (tt.)

❑ **Đại lượng trung vị** (*median*)

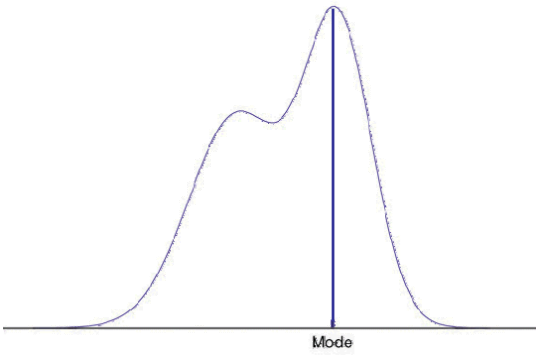


- phần tử $\tilde{x}$ tại điểm giữa của một dãy sắp xếp tăng dần
- |X| = 2k → lấy median là giá trị trung bình của 2 phần tử ở giữa
- không áp dụng được cho nominal data
- ít bị tác động bởi outliers và skewed distributions
- thích hợp cho bộ dữ liệu bất đối xứng
- X = { 5, 7, 8, 8, 8, 8 } ?

2. Measures of central tendency (tt.)

- **Đại lượng yếu vị** (*mode*)
  - tần số xuất hiện lớn nhất

$$\mu o \rightarrow \hat{x} = \arg\max_{x_i \in X} f(x_i)$$



  - có thể áp dụng cho numerical và non-numerical data
  - có thể nhiều modes cho 1 distribution (bi-modal, multi-modal)
  - dữ liệu phân phối đều ? dữ liệu liên tục ?
  - grouped data: mode của groups → mid point của group_mode
    X = { (lần chạy thứ i, t.gian chạy t_i) } ⇒ phụ thuộc phân nhóm

2. Measures of central tendency (tt.)

- **Ví dụ: Thu thập dữ liệu nhiệt độ (12g00) mỗi ngày**
  - mean: sự thay đổi nhiệt độ giữa các tháng trong năm
  - mode: những phương án thích nghi
  - median: những nơi có thay đổi đột biến, bất thường về nhiệt độ

Nội dung bổ sung

1. Descriptive statistics
2. Measures of central tendency
3. **Mesures of dispersion**
4. Measures of shape

3. Measures of dispersion

- **Measures of dispersion**

  2 lô trái cây cùng kích thước trung bình → ưa thích như nhau ?

  - mức độ phân tán, dàn trải (spread) của dữ liệu
  - mức độ biến thiên (variability) của dữ liệu
  - range, percentile, quartile, interquartile range, variance, standard deviation

3. Measures of dispersion (tt)

- **Khoảng biến thiên** (*range*)

$$R = x_{\max} - x_{\min}$$

  - độ lệch giữa max và min (KHÔNG phải các giá trị max và min)
  - "độ rộng" (độ dàn trải) của tập dữ liệu
  - khi có outliers ở 2 đầu mút ?

3. Measures of dispersion (tt)

- **[Bách] phân vị** (*percentile*)

  Phân vị thứ p là 1 giá trị:

  - có ít nhất p% các quan sát có giá trị ≤ p
  - có ít nhất (100 – p)% các quan sát có giá trị ≥ p

3. Measures of dispersion (tt)

- **[Bách] phân vị** (*percentile*)

  Các bước tính giá trị phân vị thứ p

  B1. Sắp xếp n quan sát theo thứ tự tăng dần.

  B2. Tính chỉ số i

$$i = \frac{p.n}{100}$$

  B3. Nếu i KHÔNG phải là số nguyên thì làm tròn số i và số nguyên tiếp theo sau là vị trí của phân vị thứ p

  Nếu i LÀ số nguyên thì phân vị thứ p là trung bình của hai giá trị ở vị trí thứ i và (i + 1)

3. Measures of dispersion (tt)

- **[Bách] phân vị** (*percentile*)

  3310, 3355, 3450, 3480, 3480, 3490, 3520, 3540, 3550, 3650, 3730, 3925

  Phân vị thứ 85: i = 85.12/100 = 10.2

  Vì i không phải số nguyên nên làm tròn 11

  P85 = D[11] = 3730

  Phân vị thứ 50: i = 50.12/100 = 6

  Vì i là số nguyên nên lấy trung bình D[6], D[7]

  P50 = (3490 + 3520) / 2 = 3505 (= median !)

3. Measures of dispersion (tt)

- **Tứ phân vị** (*quartile*)
  - |Q1| = |Q2| = |Q3| = |Q4| = 25%

    Q1 = P25

    Q2 = P50 (median)

    Q3 = P75
  - hạn chế tác động của những outliers ở 2 đầu mút
  - quan tâm đến 50% hai bên median (Q2 và Q3)

3. Measures of dispersion (tt)

- **Độ trải giữa** (*InterQuartile Range* – *IQR*)

  IQR = (Q3 – Q1)
  - độ biến thiên của 50% quan sát ở giữa
  - hữu dụng khi so sánh 2 tập dữ liệu (~ means, ~ medians)

3. Measures of dispersion (tt)

❑ **Phương sai** (*variance*) **và độ lệch chuẩn** (*standard deviation*)

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \sigma^2 \qquad s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- sự phân bố của các phần tử xung quanh giá trị trung bình
- phân phối chuẩn (normal distribution) >> skewed distribution
- tác động của outliers → gia tăng kích thước tập dữ liệu

3. Measures of dispersion (tt)

❑ **Hệ số biến thiên** (*coefficient of variation*)

$$CV = \frac{\sigma}{\mu}$$

- so sánh mức độ phân tán của các quần thể/mẫu có sự khác nhau về trung bình và phương sai

3. Measures of dispersion (tt)

- **VD: Cho mẫu dữ liệu**

  D = { 27, 25, 20, 15, 30, 34, 28, 25 }

  S =

  P20 =

  P25 =

  P65 =

  P75 =

3. Measures of dispersion (tt)

- **VD: Danh sách độ tuổi nhân viên của công ty**

  D = { 18, 54, 20, 46, 25, 48, 53, 27, 26, 37, 40, 36, 42, 25, 27, 33, 28, 40, 45, 25 }

  a. Tính mean và mode.

  b. Tính Q1 và Q3.

  c. Giải thích ý nghĩa của P32.

3. Measures of dispersion (tt)

- **VD: Cho mẫu dữ liệu**

  D = { 27, 25, 20, 15, 30, 34, 28, 25 }

  S =

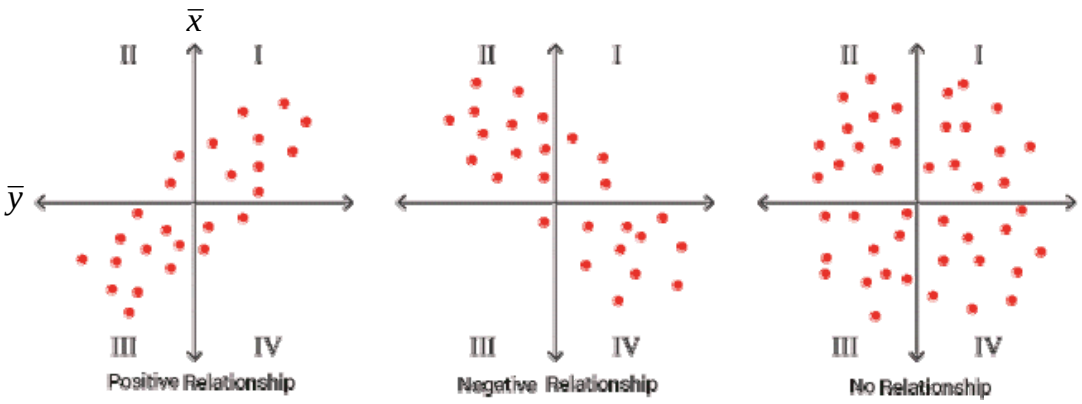
  a. Khoảng biến thiên:

  b. Độ trải giữa:

  c. Phương sai:

3. Measures of dispersion (tt)

- **Hiệp phương sai** (*covariance*)
  - mối quan hệ giữa 2 yếu tố (biến): $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
  - dữ liệu định lượng (hạn chế với rating scales)



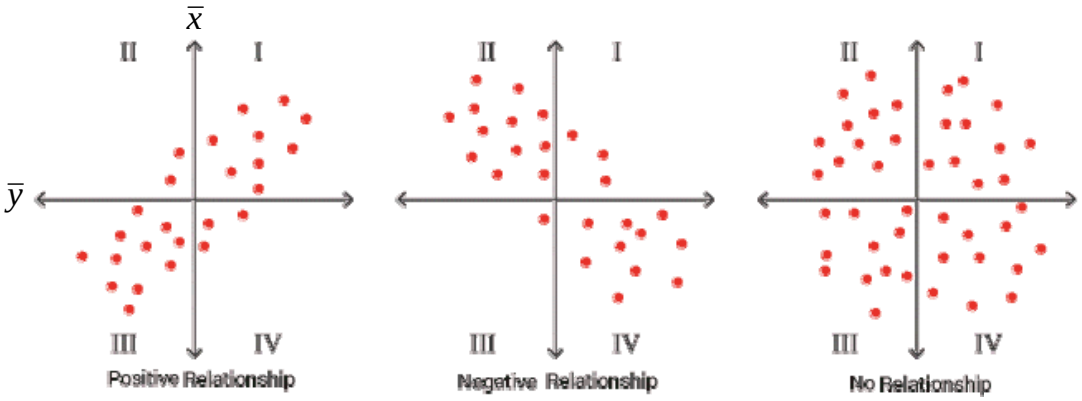
3. Measures of dispersion (tt)

❑ **Hiệp phương sai** (*covariance*)

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$s_{xy} = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$



3. Measures of dispersion (tt)

❑ **Hiệp phương sai** (*covariance*)

- hạn chế do sử dụng σ để đo lường cường độ mối quan hệ

  Mối quan hệ giữa chiều cao (cm) và cân nặng (kg)

  Thay đổi đơn vị: chiều cao (inch) và cân nặng (lb)

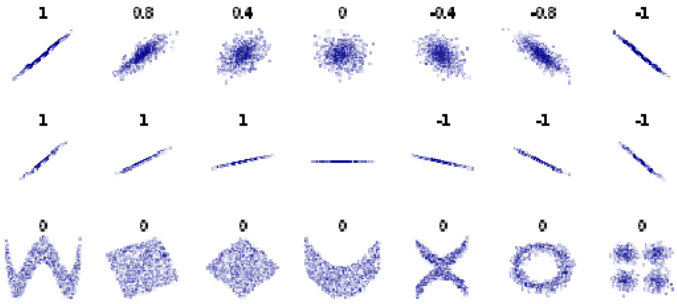
  ⇒ cường độ bị thay đổi nhưng thực tế mối quan hệ không đổi

3. Measures of dispersion (tt)

❑ Hệ số tương quan (*Pearson correlation*)

$$correlation(x,y) = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2.(y_i - \bar{y})^2}}$$

- x, y độc lập ⇒ correlation(x, y) = 0; điều ngược lại không đúng



www.wikipedia.org (09/2019)

3. Measures of dispersion (tt)

❑ So sánh covariance và correlation

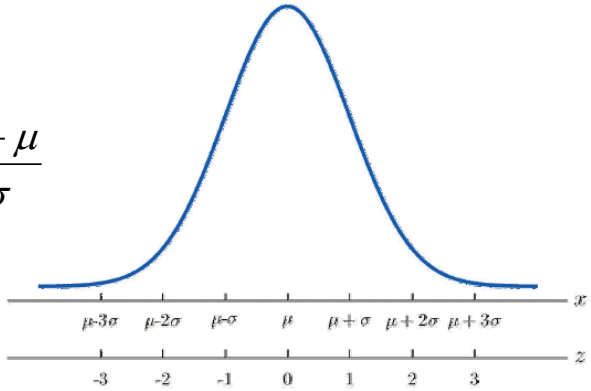
$$correlation(x,y) = \frac{\mathrm{cov}(x,y)}{s_x.s_y}$$

- cov(x, y) ∈ ℝ: đơn vị tính bằng tích của hai đơn vị tính x và y
- cov(x, y) càng lớn ⇒ mối quan hệ càng chặt nhưng đơn vị tính của x và y khác nhau cho nên cov(x, y) không thể hiện đúng mức độ phụ thuộc giữa x và y
- correlation(x, y) ∈ [–1, 1]: giá trị đã được chuẩn hóa
- correlation(x, y) thể hiện mức độ phụ thuộc giữa x và y

3. Measures of dispersion (tt)

❑ **Giá trị z** (*z-score*, *z-value*, *standard score*, *normal score*,)

$$z_{score} = \frac{x - \bar{x}}{s} \qquad z_{score} = \frac{x - \mu}{\sigma}$$



- vị trí tương đối của các giá trị trong tập dữ liệu (so với mean)
- độ lệch (chuẩn hóa) bao nhiêu lần so với độ lệch chuẩn
- xác định một giá trị có phải là outlier hay không

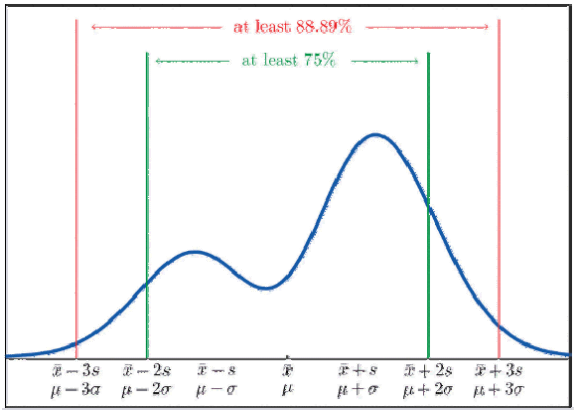
3. Measures of dispersion (tt)

❑ **Bất đẳng thức Chebyshev**

$$P(z\sigma \le (|X - \mu|) \le \frac{1}{z^2}$$

❑ **Định lý Chebyshev**

- Tối thiểu $\left(1 - \frac{1}{z^2}\right)$ quan sát nằm trong [μ – zσ, μ + zσ], với z > 1

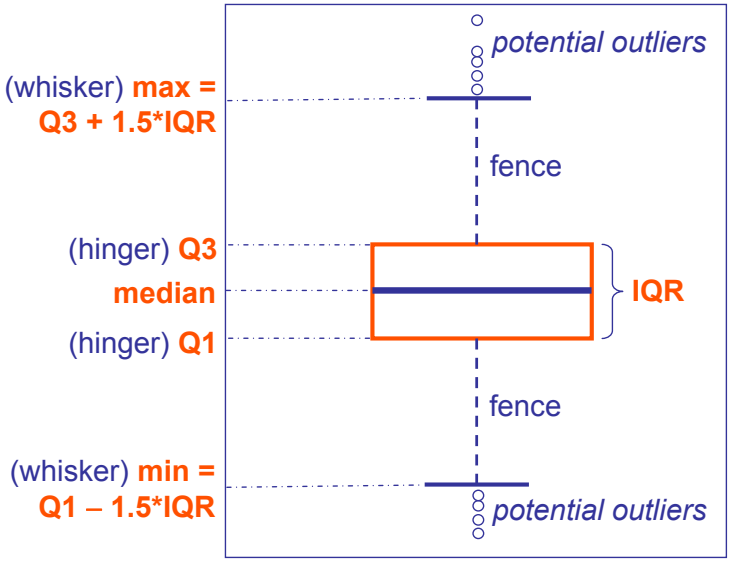


z = 3: chứa hầu hết các quan sát

Loại bỏ các quan sát có |z| > 3

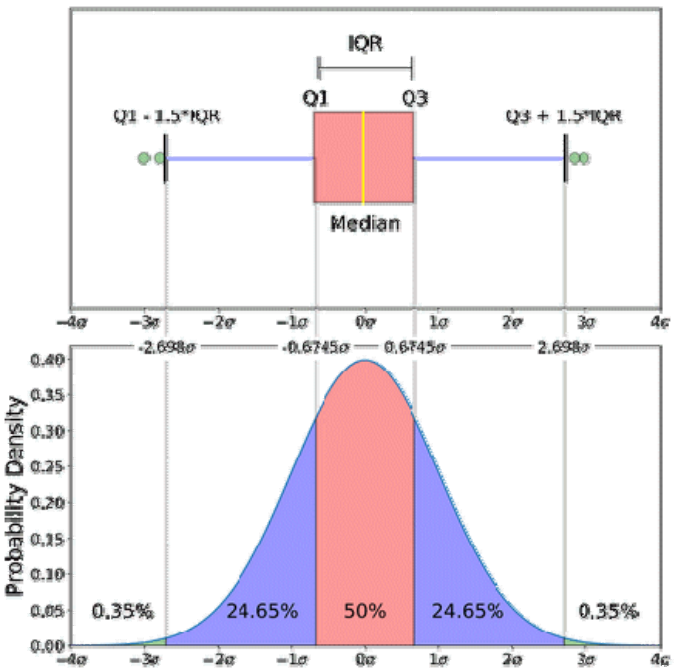
3. Measures of dispersion (tt)

- Mô tả dữ liệu liên tục bằng **box plot** (Tukey, 1977)



3. Measures of dispersion (tt)

- Đối sánh boxplot và phân phối chuẩn



towardsdatascience.com (09/2019)

3. Measures of dispersion (tt)

- **VD: Cho mẫu gồm các quan sát (x_i, y_i)**

  { (6, 6), (11, 9), (15, 6), (21, 17), (27, 12) }

  a. Hiệp phương sai mẫu:

  b. Mối quan hệ giữa x và y:

  c. Hệ số tương quan mẫu:

3. Measures of dispersion (tt)

- **VD: Cho mẫu dữ liệu có mean = 30, phương sai = 5. Áp dụng định lý Chebyshev để xác định tỷ lệ % (tối thiểu) các quan sát nằm trong khoảng giá trị:**

  a. [20, 40]

  b. [15, 45]

3. Measures of dispersion (tt)

- **VD: Cho mẫu dữ liệu như sau:**

  D = { 236, 1710, 1351, 825, 7450, 316, 4135, 1333, 1584, 387, 991, 3396, 170, 1428, 1688 }

  a. Giá trị trung bình = Trung vị =

  b. Q1 = Q3 =

  c. Khoảng biến thiên R = IQR =

  d. Phương sai =