



B5. Descriptive Statistics

Bổ sung thêm cho bài giảng

2020

Nội dung bổ sung



1. Thống kê mô tả
2. Các đại lượng về trung tâm
3. Các đại lượng về độ phân tán
4. Các đại lượng về hình dáng phân phối

1. Thống kê mô tả



❑ Dữ liệu (*data*)

❑ Thông tin (*information*)

❑ Tri thức (*knowledge*)

1. Thống kê mô tả (tt.)



❑ Analytics

- descriptive
- predictive
- prescriptive



1. Thống kê mô tả (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. Thống kê mô tả (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. Thống kê mô tả (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*)

tần số $f(x_i)$: số lần x_i xuất hiện trong tập dữ liệu (mẫu, quần thể)



1. Thống kê mô tả (tt.)

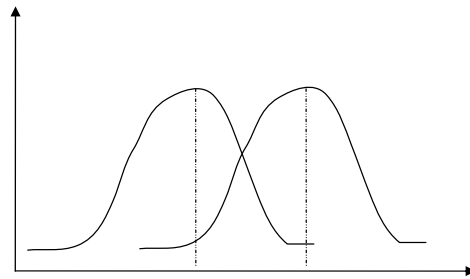
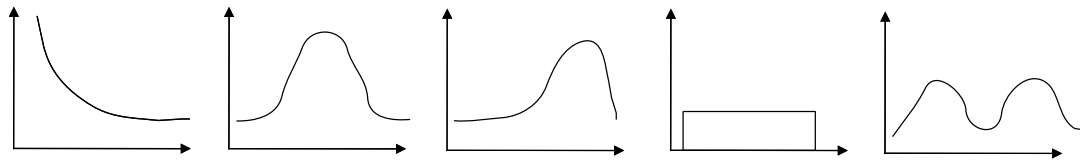
□ Thuật ngữ

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

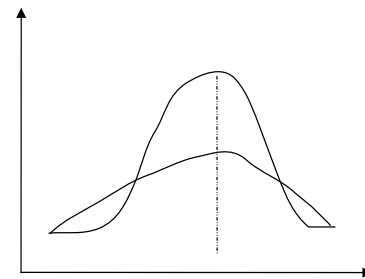
1. Thống kê mô tả (tt.)



Central Tendency, Variability, Shape



Different Central Tendency



Different Variability



Nội dung bổ sung



1. Thống kê mô tả
2. Các đại lượng về trung tâm
3. Các đại lượng về độ phân tán
4. Các đại lượng về hình dáng phân phối

2. Các đại lượng về trung tâm



□ Một giá trị số thể hiện vị trí/xu thế “trung tâm” của tập dữ liệu
→ điểm đặc trưng, điển hình, đại diện

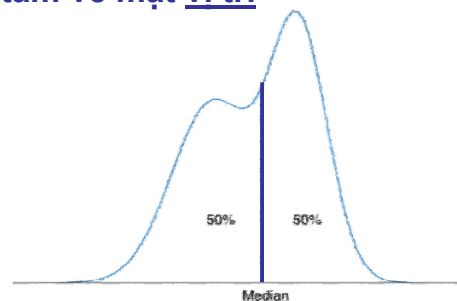
- median
- Pythagorean means
- mode



2. Các đại lượng về trung tâm (tt.)



□ Đại lượng trung vị (*median*): trung tâm về mặt vị trí



- phần tử \tilde{x} tại điểm giữa của một dãy sắp xếp (tăng hoặc giảm)
- $|X| = 2k \rightarrow$ median là giá trị trung bình của 2 phần tử ở giữa
- ít bị tác động bởi outliers
- thích hợp cho tập dữ liệu bất đối xứng
- $X = \{ 5, 7, 8, 8, 8, 8 \}$?





2. Các đại lượng về trung tâm (tt.)

□ Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

Trung bình cộng (*arithmetic mean*)

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

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

$F(x) = \Pr(X \leq x)$: *Cumulative Distribution Function*

$f(x) = F'(x)$: *Probability Density Function*, giới hạn $\forall x: f(x) > 0$

- dễ bị tác động bởi outliers và các phân phối bất đối xứng
- trung bình lọc (*trimmed mean*): loại bỏ x% giá trị nhỏ nhất và x% giá trị lớn nhất



2. Các đại lượng về trung tâm (tt.)

□ Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

VD: Đánh giá môn học ở các ĐH Mỹ: A(4), B(3), C(2), D(1), F(0).

Một SV có 9 tín chỉ A, 15 tín chỉ B, 33 tín chỉ C và 3 tín chỉ D.

Tính điểm trung bình của sinh viên.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n w_i x_i \quad n = |X| \quad \text{weighted mean}$$

$$\bar{x} = \frac{1}{(9+15+33+3)} [9*4 + 15*3 + 33*2 + 3*1] = \frac{150}{60} = 2.5$$



2. Các đại lượng về trung tâm (tt.)

□ **Đại lượng trung bình (*mean*): trung tâm về mặt giá trị**

Trung bình cộng (*arithmetic mean*)

VD:

Nhà $\xrightarrow{30\text{km/h}}$ Cơ quan $\xrightarrow{10\text{km/h}}$ Nhà. Vận tốc trung bình ?

Nhà $\xrightarrow{30\text{km/h}}$ Cơ quan $\xrightarrow{35\text{km/h}}$ Nhà hàng. Vận tốc trung bình ?

- thời gian di chuyển cho mỗi đoạn !
- đơn vị và thứ nguyên



2. Các đại lượng về trung tâm (tt.)

□ **Đại lượng trung bình (*mean*): trung tâm về mặt giá trị**

VD: Tỷ lệ nhiễm bệnh trên 3 mẫu

$$\frac{20}{20} = 100\% \quad \frac{10}{60} \approx 17\% \quad \frac{60}{70} \approx 86\%$$

Tỷ lệ nhiễm bệnh trung bình ?

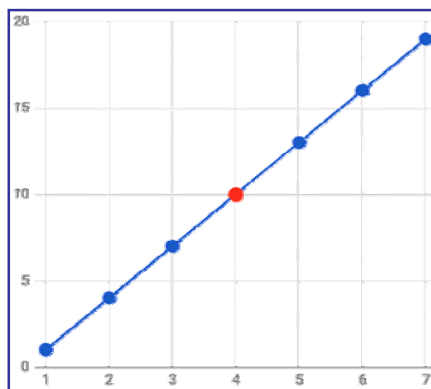




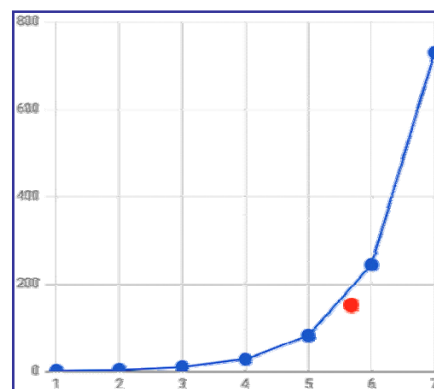
2. Các đại lượng về trung tâm (tt.)

□ Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

VD: $A = \{1, 4, 7, 10, 13, 16, 19\}$ $B = \{1, 3, 9, 27, 81, 243, 729\}$



$$\bar{A} = 10 = \tilde{A}$$



$$27 = \tilde{B} \ll \bar{B} \approx 156$$



2. Các đại lượng về trung tâm (tt.)

□ Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

VD: Số vốn 1 tỷ, lãi suất các năm: 2%, 5%, 7%, 8%, 10%.

Số tiền sau 5 năm ?

$$V_0 = 10^9$$

$$V_1 = V_0 * (1 + r_1); V_2 = V_1 * (1 + r_2) = V_0 * (1 + r_1) * (1 + r_2); \dots$$

$$V_5 = V_0 * (1 + r_1) * (1 + r_2) * (1 + r_3) * (1 + r_4) * (1 + r_5) =$$

$$= 1,361,412,360$$

Trong khi đó: $\bar{r} = (.02 + .05 + .07 + .08 + .10) / 5$

$$V_5^* = 10^9 * (1 + \bar{r})^5 = 1,363,666,440 \neq V_5$$



2. Các đại lượng về trung tâm (tt.)

□ Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

Trung bình nhân (*geometric mean*): $x_i > 0$ (tỷ lệ %)

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

VD:

Cách 1 (các mẫu riêng rẽ):

$$\bar{x}_1 = \left(\frac{20}{20} * \frac{10}{60} * \frac{60}{70} \right)^{\frac{1}{3}} \approx 0.52$$

Cách 2 (gộp chung các mẫu):

$$\bar{p} = (20 * 10 * 60)^{\frac{1}{3}} = 22.89 \quad \bar{n} = (20 * 60 * 70)^{\frac{1}{3}} = 43.80$$

$$\bar{x}_2 = \frac{\bar{p}}{\bar{n}} = \frac{22.89}{43.80} \approx 0.52 = \bar{x}_1$$

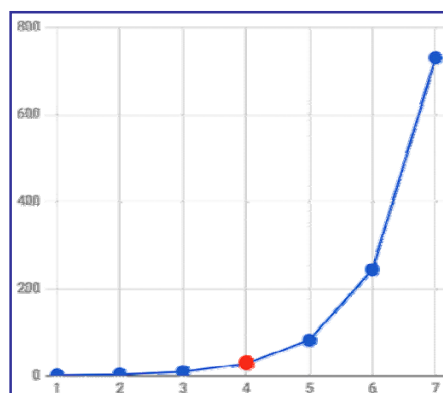


2. Các đại lượng về trung tâm (tt.)

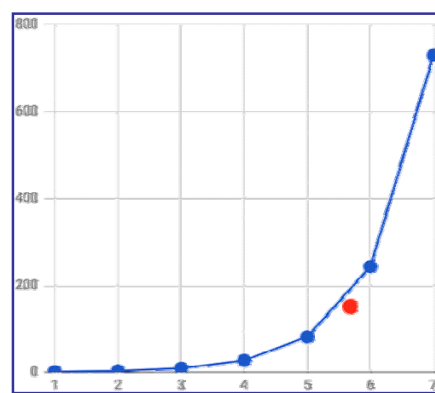
□ Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

Trung bình nhân (*geometric mean*):

VD: B = { 1, 3, 9, 27, 81, 243, 729 }



$$\tilde{B} = \bar{B}_{geo^*}$$



$$\tilde{B} << \bar{B}_{ari+}$$



2. Các đại lượng về trung tâm (tt.)

□ Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

Trung bình nhân (*geometric mean*):

VD: Vốn 1 tỷ với lãi suất các năm: 2%, 5%, 7%, 8%, 10%.

Số tiền sau 5 năm ?

$$V_5 = 1,361,412,360$$

$$\text{Ta có: } \bar{r}_{geo*} = (1.02 * 1.05 * 1.07 * 1.08 * 1.10)^{1/5}$$

$$V_5^{**} = 10^9 * (\bar{r}_{geo*})^5 = 1,361,412,360 = V_5$$



2. Các đại lượng về trung tâm (tt.)

□ Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

Trung bình nhân (*geometric mean*):

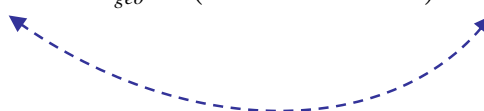
- mức độ biến thiên lớn của dữ liệu:

$$A = \{ 4, 4, 5, 5, 6, 6 \}$$

$$B = \{ 2, 2, 4, 6, 8, 8 \}$$

$$\bar{A}_{ari+} = \frac{1}{6}(4+4+5+5+6+6) = 5 \quad \bar{B}_{ari+} = \frac{1}{6}(2+2+4+6+8+8) = 5$$

$$\bar{A}_{geo*} = (4*4*5*5*6*6)^{\frac{1}{6}} \approx 4.93 \quad \bar{B}_{geo*} = (2*2*4*6*8*8)^{\frac{1}{6}} \approx 4.28$$





2. Các đại lượng về trung tâm (tt.)

□ Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

Trung bình nhân (*geometric mean*):

- nhiều nguồn dữ liệu khác thang đo → không cần chuẩn hóa

U_1 (thang điểm 5): $\text{rating}(U_1, A) = 4.5$; $\text{rating}(U_1, B) = 3.5$

U_2 (thang điểm 100): $\text{rating}(U_2, A) = 70$; $\text{rating}(U_2, B) = 80$

Chuẩn hóa: $r_{1A} = 0.90$ $r_{1B} = 0.70$ $\Rightarrow \bar{A} = 0.80$ $\bar{B} = 0.75$
 $r_{2A} = 0.70$ $r_{2B} = 0.80$ **A tốt hơn B**

Trung bình cộng: $\bar{A}_{ari+} = \frac{(4.5 + 70)}{2} = 37.25$ $\bar{B}_{ari+} = \frac{(3.5 + 80)}{2} = 41.75$

B tốt hơn A !?

Trung bình nhân: $\bar{A}_{geo*} = \sqrt{(4.5 * 70)} \approx 17.75$ $\bar{B}_{geo*} = \sqrt{(3.5 * 80)} \approx 16.73$



2. Các đại lượng về trung tâm (tt.)

□ Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

Trung bình nhân (*geometric mean*):

- xử lý tín hiệu
- xử lý ảnh
- ...

https://en.wikipedia.org/wiki/Geometric_mean



2. Các đại lượng về trung tâm (tt.)

□ Đại lượng trung bình (mean): trung tâm về mặt giá trị

Trung bình điều hòa (*harmonic mean*): $x_i > 0$, là những tỷ số giữa 2 đại lượng khác đơn vị (vận tốc, gia tốc, ...)

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

- nghịch đảo của trung bình cộng của tập dữ liệu: $\left\{ \frac{1}{x_i} \right\}_{i=1}^n$

- trung bình các tỷ số dựa trên cùng tử số, khác mẫu số

VD: vận tốc TB (km/h) $\rightarrow \{x_i\}$ gồm những vận tốc đo được trong những khoảng thời gian khác nhau



2. Các đại lượng về trung tâm (tt.)

□ Đại lượng trung bình (mean): trung tâm về mặt giá trị

Trung bình điều hòa (*harmonic mean*):

VD: Nhà $\xrightarrow{30km/h}$ Cơ quan $\xrightarrow{10km/h}$ Nhà. $d = 5km$. Vận tốc TB ?

$$\bar{v}_{ari+} = \frac{(30+10)}{2} = 20km/h$$

Trong khi đó, thời gian đi (t_1) và về (t_2) không giống nhau:

$$t_1 = \frac{5km}{30km/h} = \frac{1}{6}h = 10 \text{ min} \quad t_2 = \frac{5km}{10km/h} = \frac{1}{2}h = 30 \text{ min}$$

$$t = (t_1 + t_2) = 40 \text{ min}$$

$$\bar{v}_{weighted_mean} = \left(30 \cdot \frac{10}{40} + 10 \cdot \frac{30}{40} \right) = 15km/h$$

$$\bar{v}_{har/} = 2 \cdot \left(\frac{1}{30} + \frac{1}{10} \right)^{-1} = 2 \cdot \frac{30}{4} = 15km/h$$



2. Các đại lượng về trung tâm (tt.)

- Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

Trung bình điều hòa có trọng số (*weighted harmonic mean*):

$$\mu_{\text{Weighted_Harmonic}} \rightarrow \bar{x} = \frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n \frac{w_i}{x_i}}$$



2. Các đại lượng về trung tâm (tt.)

- Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

VD: Ngôn ngữ Python

```
Hàm: statistics.geometric_mean()    # Python 3.8
      statistics.harmonic_mean()
      scipy.stats.gmean()
      scipy.stats.hmean()
```

2. Các đại lượng về trung tâm (tt.)



□ Đại lượng trung bình (*mean*): trung tâm về mặt giá trị

Trung bình điều hòa có trọng số (*weighted harmonic mean*):

- xử lý hợp kim
- P/E (Price-Earning)
- ...

https://en.wikipedia.org/wiki/Harmonic_mean

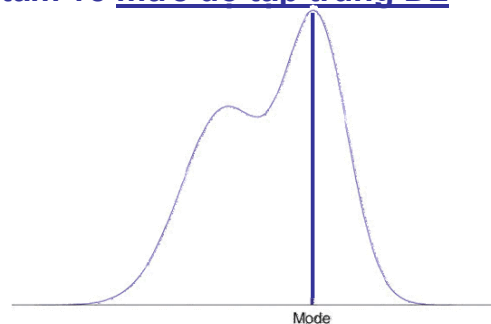


2. Các đại lượng về trung tâm (tt.)



□ Đại lượng yếu vị (*mode*): trung tâm về mức độ tập trung DL

$$\mu_0 \rightarrow \hat{x} = \arg \max_{x_i \in X} f(x_i)$$



- 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 \rightarrow mid point của group_mode
 $X = \{ (\text{lần chạy thứ } i, \text{t.gian chạy } t_i) \} \Rightarrow$ phụ thuộc phân nhóm



2. Các đại lượng về trung tâm (tt.)



□ Ví dụ: Thu thập dữ liệu nhiệt độ (12g00) mỗi ngày

- *mean*: sự thay đổi nhiệt độ giữa các tháng trong năm
- *mode*: phương án thích nghi
- *median*: những nơi có thay đổi đột biến, bất thường về nhiệt độ



Nội dung bổ sung



1. Thống kê mô tả
2. Các đại lượng về trung tâm
3. Các đại lượng về độ phân tán
4. Các đại lượng về hình dáng phân phối

3. Các đại lượng về độ phân tán



□ Đánh giá 2 tập dữ liệu có cùng mean, median, mode

- mức độ biến thiên (variability) của dữ liệu
- mức độ phân tán, dàn trải (spread) của dữ liệu
- xu thế phân tán khỏi giá trị trung tâm (điển hình)
- *range, percentile, quartile, interquartile range, variance, standard deviation*



96

3. Các đại lượng về độ phân tán (tt.)



□ Khoảng biến thiên (*range*): ước lượng ban đầu về độ phân tán

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

- 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 ?
- tập dữ liệu bất đối xứng (bị “lệch”) ?



97

3. Các đại lượng về độ phân tán (tt.)



□ [Bách] phân vị (*percentile*): $p \in (0, 100) \rightarrow$ ngưỡng giá trị v_p

- có ít nhất $p\%$ các quan sát có giá trị $\leq v_p$
- có ít nhất $(100 - p)\%$ các quan sát có giá trị $\geq v_p$

3. Các đại lượng về độ phân tán (tt.)



□ [Bách] phân vị (*percentile*): $p \in (0, 100) \rightarrow$ ngưỡng giá trị v_p

Các bước tính giá trị (v_p) của 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 chỉ số i KHÔNG phải là số nguyên thì làm tròn chỉ số i thành số nguyên tiếp theo và sẽ là vị trí của phân vị thứ p

Nếu chỉ số i LÀ số nguyên thì giá trị v_p sẽ là trung bình cộng của hai giá trị ở vị trí thứ i và $(i + 1)$

3. Các đại lượng về độ phân tán (tt.)



□ [Bách] phân vị (*percentile*): $p \in (0, 100) \rightarrow$ ngưỡng giá trị v_p

VD: Cho mẫu dữ liệu

{ 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

$P_{85} = D[11] = 3730$

Phân vị thứ 50: $i = (50 * 12) / 100 = 6$

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

$P_{50} = (3490 + 3520) / 2 = 3505$ (= median !)

3. Các đại lượng về độ phân tán (tt.)



□ [Bách] phân vị (*percentile*): $p \in (0, 100) \rightarrow$ ngưỡng giá trị v_p

VD: Cho mẫu dữ liệu

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

$S = \{ 15, 20, 25, 25, 27, 28, 30, 34 \}$

$P_{20} =$

$P_{25} =$

$P_{65} =$

$P_{75} =$

3. Các đại lượng về độ phân tán (tt.)



□ [Bách] phân vị (*percentile*): $p \in (0, 100) \rightarrow$ ngưỡng giá trị v_p

VD:

Dữ liệu: $D = \{ 8, 9, 12, 14.5, 15.5, 17, 18 \}$

P35 ?

- Excel: $\text{PERCENTILE}(D, .35) = 12.25$
- R: $\text{QUANTILE}(D, c(.35)) = 12.25$

3. Các đại lượng về độ phân tán (tt.)



□ [Bách] phân vị (*percentile*): $p \in (0, 100) \rightarrow$ ngưỡng giá trị v_p

VD: Ngôn ngữ Python

Hàm: `numpy.percentile()`

- chỉ số:

$$i = \frac{p(n-1)}{100} + 1$$

- nội suy (*interpolation*): linear, lower, upper, midpoint, nearest



3. Các đại lượng về độ phân tán (tt.)



□ Tứ phân vị (*quartile*)

- $|Q1| = |Q2| = |Q3| = |Q4| = 25\%$

$$Q1 = P25$$

$$Q2 = P50 \text{ (median)}$$

$$Q3 = P75$$

- hạn chế tác động của những outliers ở 2 đầu mút

3. Các đại lượng về độ phân tán (tt.)



□ Tứ phân vị (*quartile*)

VD: Danh sách độ tuổi nhân viên của công ty

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

a. Tính mean và mode.

$$\text{mean} = 695 / 20 = 34.75; \text{mode} = 25$$

b. Tính Q1 và Q3.

$$Q1 = (S[5] + S[6]) / 2 = 25.5; Q3 = (S[15] + S[16]) / 2 = 43.5$$

c. Giải thích ý nghĩa của P32.

3. Các đại lượng về độ phân tán (tt.)



□ Khoảng 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. Các đại lượng về độ phân tán (tt.)



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

Quần thể:
$$Var(X) = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2 = \sigma^2$$

Mẫu:
$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{(n-1)} \sum_{i=1}^n x_i^2 - \frac{n}{(n-1)} \bar{x}^2$$

- sự phân bố của các quan sát 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. Các đại lượng về độ phân tán (tt.)



□ **VD:** Cho mẫu gồm các giá trị

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

S =

a. Khoảng biến thiên:

b. Khoảng trải giữa:

c. Phương sai:

3. Các đại lượng về độ phân tán (tt.)



□ **VD:** Mẫu dữ liệu về thời gian chữa trị Covid-19:

10-14 ngày: 4

15-19 ngày: 8

20-24 ngày: 5

25-29 ngày: 2

30-34 ngày: 1

Tính trung bình và phương sai của thời gian chữa trị Covid-19.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n f_i M_i = \frac{380}{20} = 19 \quad s^2 = \frac{1}{(n-1)} \sum_{i=1}^n f_i (M_i - \bar{x})^2 = \frac{570}{19} = 30$$

3. Các đại lượng về độ phân tán (tt.)



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

- Nếu tập dữ liệu có tính đối xứng:

$$s = \frac{(x_{\max} - x_{\min})}{\sqrt{n}} \quad n < 12$$

$$s = \frac{(x_{\max} - x_{\min})}{4} \quad 20 < n < 40$$

$$s = \frac{(x_{\max} - x_{\min})}{5} \quad n \approx 100$$

$$s = \frac{(x_{\max} - x_{\min})}{6} \quad 400 < n$$

3. Các đại lượng về độ phân tán (tt.)



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

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

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

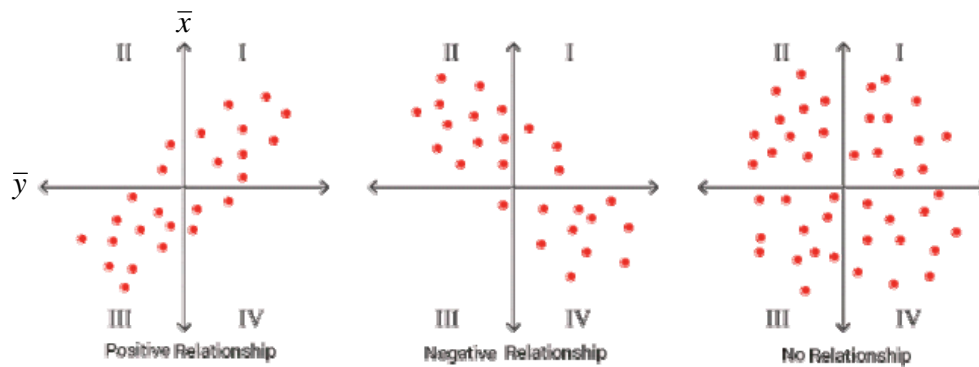


3. Các đại lượng về độ phân tán (tt.)



□ Hiệp phương sai (*covariance*)

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



3. Các đại lượng về độ phân tán (tt.)

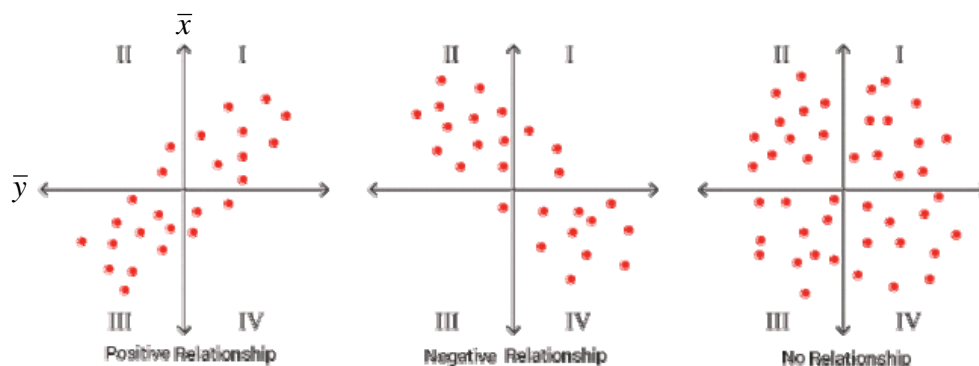


□ Hiệp phương sai (*covariance*)

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

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

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



3. Các đại lượng về độ phân tán (tt.)



□ Hiệp phương sai (covariance)

- hạn chế khi dùng để đo cường độ của mối quan hệ

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

$$H_{\text{cm}} = \{ 162, 165, 168 \} \quad W_{\text{kg}} = \{ 50, 56, 60 \}$$

$$\text{covariance}(H_{\text{cm}}, W_{\text{kg}}) = 15.00$$

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

$$H_{\text{inch}} = H_{\text{cm}} / 2.54 \quad W_{\text{lb}} = W_{\text{kg}} \times 2.20$$

$$\text{covariance}(H_{\text{inch}}, W_{\text{lb}}) = 12.99$$

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

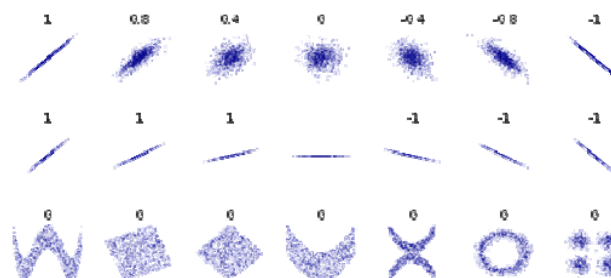
3. Các đại lượng về độ phân tán (tt.)



□ Hệ số tương quan (Pearson correlation)

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

- x, y độc lập ⇒ $\text{correlation}(x, y) = 0$; điều ngược lại không đúng



3. Các đại lượng về độ phân tán (tt.)



□ Hệ số tương quan (Pearson correlation)

$$H_{\text{cm}} = \{ 162, 165, 168 \}$$

$$W_{\text{kg}} = \{ 50, 56, 60 \}$$

$$\text{correlation}(H_{\text{cm}}, W_{\text{kg}}) = 0.9933$$

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

$$H_{\text{inch}} = H_{\text{cm}} / 2.54$$

$$W_{\text{lb}} = W_{\text{kg}} \times 2.20$$

$$\text{correlation}(H_{\text{inch}}, W_{\text{lb}}) = 0.9933$$

3. Các đại lượng về độ phân tán (tt.)



□ Hệ số tương quan (Pearson correlation)

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. Các đại lượng về độ phân tán (tt.)



□ So sánh covariance và correlation

$$\text{correlation}(x, y) = \frac{\text{cov}(x, y)}{s_x \cdot s_y}$$

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

3. Các đại lượng về độ phân tán (tt.)



□ Phân tích dữ liệu

VD: Mẫu dữ liệu về số ngày giao hàng của 2 công ty:

$A = \{ 11, 10, 9, 10, 11, 11, 10, 11, 10, 10 \}$

$B = \{ 8, 10, 13, 7, 10, 11, 10, 7, 15, 12 \}$

Phân tích thời gian giao hàng của 2 công ty (tùy mục tiêu).



3. Các đại lượng về độ phân tán (tt.)



□ Giá trị bất thường (*outliers*): quá xa giá trị trung tâm

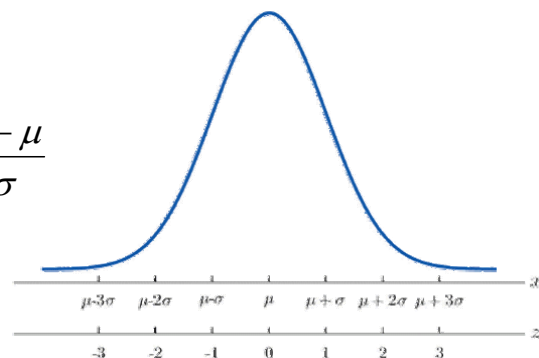
- lỗi thu thập, xử lý dữ liệu
- hiện tượng bất thường: quan tâm hay không quan tâm ?

3. Các đại lượng về độ phân tán (tt.)



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

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



- 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. Các đại lượng về độ phân tán (tt.)

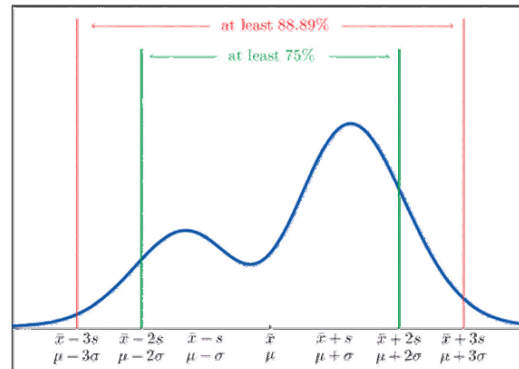


❑ Bất đẳng thức **Chebyshev**

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

❑ Định lý Chebyshev

$\forall z > 1$, có tối thiểu $\left(1 - \frac{1}{z^2}\right)$ quan sát nằm trong $[\mu - z\sigma, \mu + z\sigma]$,



$z = 3$: chứa hầu hết các quan sát

Loại bỏ các quan sát có $|z| > 3$

3. Các đại lượng về độ phân tán (tt.)



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

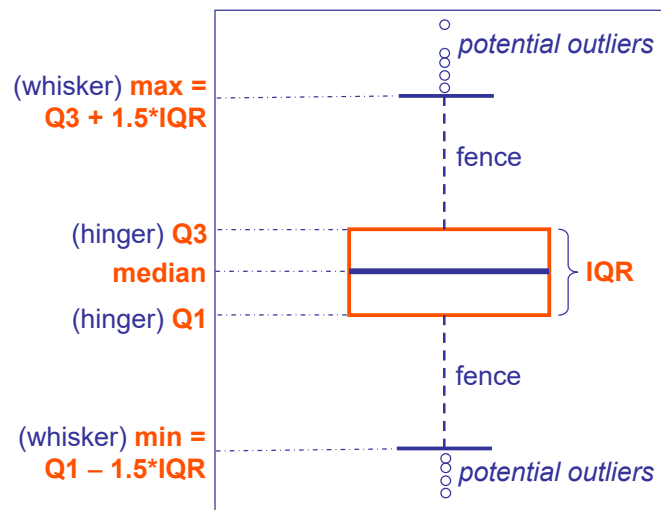
a. [15, 45]

b. [15, 40]

3. Các đại lượng về độ phân tán (tt.)



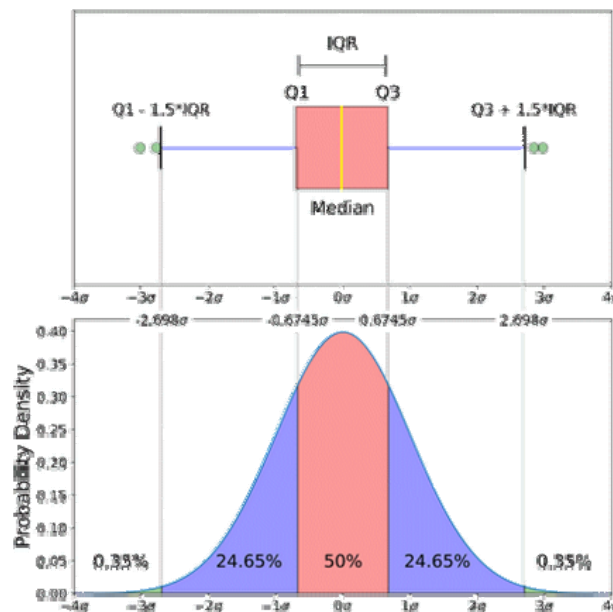
- Mô tả dữ liệu liên tục bằng **box plot** (Tukey, 1977)



3. Các đại lượng về độ phân tán (tt.)



- Đối sánh boxplot và phân phối chuẩn



Nội dung bổ sung



1. Thống kê mô tả
2. Các đại lượng về trung tâm
3. Các đại lượng về độ phân tán
4. Các đại lượng về hình dáng phân phối

4. Các đại lượng về hình dáng PP



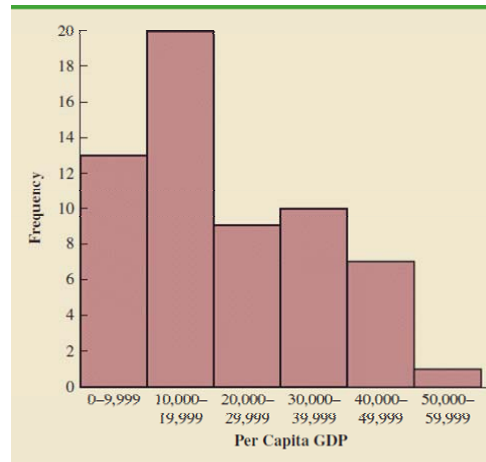
- ☐ Kiểu dáng của phân phối (Bài 6)
 - skewness
 - kurtosis

4. Các đại lượng về hình dáng PP (tt.)



□ Biểu đồ phân phối (*histogram*)

- thể hiện tần số (số lần xuất hiện) của mỗi giá trị
- không có ngăn cách giữa các nhóm như biểu đồ thanh (bar)



[Anderson+]

4. Các đại lượng về hình dáng PP (tt.)



□ Hệ số bất đối xứng, độ nghiêng (*skewness*)

$$S_{Pearson_1} = \frac{(\bar{x} - \tilde{x})}{s} \quad S_{Pearson_2} = \frac{3(\bar{x} - \tilde{x})}{s} \quad S_{Bowley} = \frac{(Q_3 + Q_1 - 2Q_2)}{(Q_3 - Q_1)}$$

$$S_k = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})^3}{s^3} \quad S_k = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \frac{(x_i - \bar{x})^3}{s^3}$$

$$S_{above} = \left| \sum_{x_i > \bar{x}} (x_i - \bar{x})^3 \right| \quad S_{below} = \left| \sum_{x_i < \bar{x}} (x_i - \bar{x})^3 \right| \quad S_k = \frac{1}{ns^3} (S_{above} - S_{below})$$

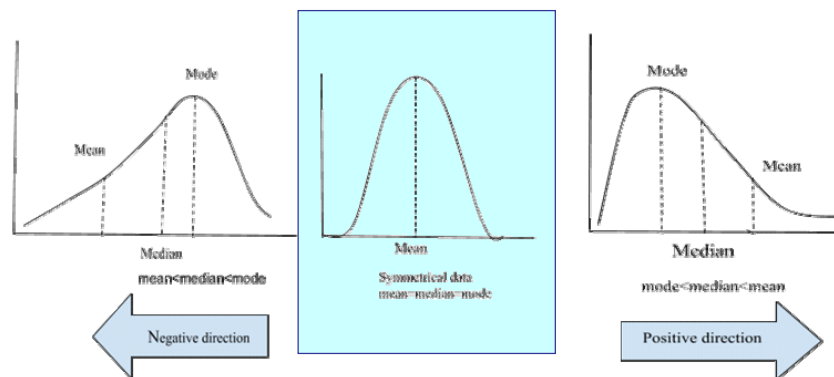
- mức độ đối xứng, bất đối xứng của một phân phối
- tỉ lệ giữa 2 “đuôi” ở 2 phía
- mối quan hệ giữa mean, median và mode: (min, max) outliers đẩy mean (mode) lệch khỏi median đi về 1 phía

4. Các đại lượng về hình dáng PP (tt.)



□ Hệ số bất đối xứng, độ nghiêng (skewness)

- skewness = 0 (*symmetrical*): mean = median = mode

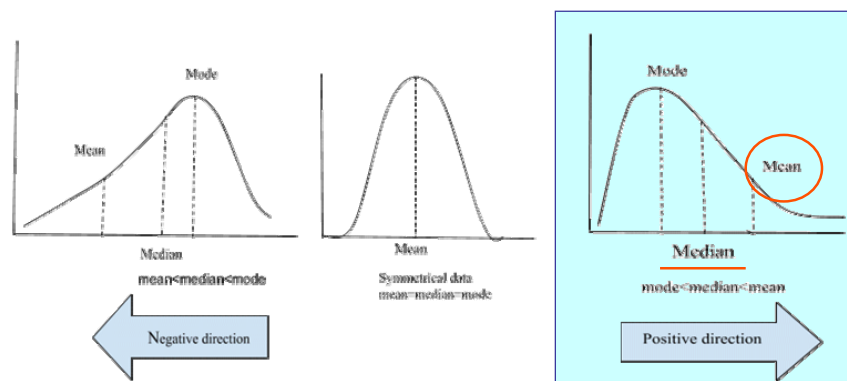


4. Các đại lượng về hình dáng PP (tt.)



□ Hệ số bất đối xứng, độ nghiêng (skewness)

- mean > median (*positive skewness*): đuôi bên PHẢI dài hơn đuôi bên trái; giá trị LỚN (outliers) đẩy mean về phía CUỐI (VD: thu nhập cá nhân → chỉ 1 số ít người thu nhập quá cao)

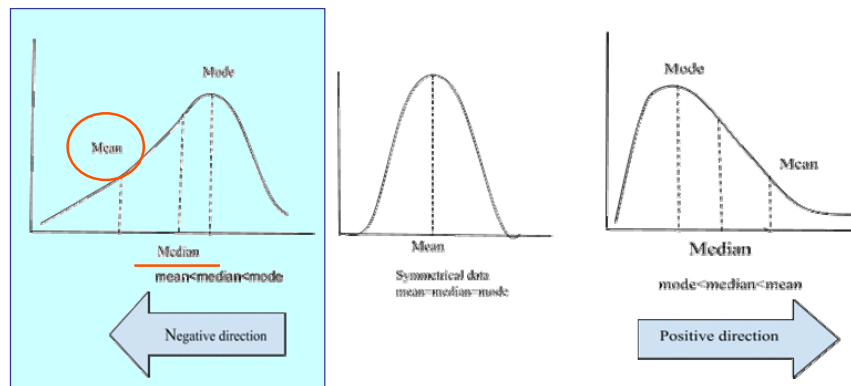


4. Các đại lượng về hình dáng PP (tt.)



□ Hệ số bất đối xứng, độ nghiêng (*skewness*)

- $\text{mean} < \text{median}$ (*negative skewness*): đuôi bên TRÁI dài hơn đuôi bên phải; giá trị NHỎ (outliers) đẩy mean về phía ĐẦU



4. Các đại lượng về hình dáng PP (tt.)



□ Độ nhọn (*kurtosis*)

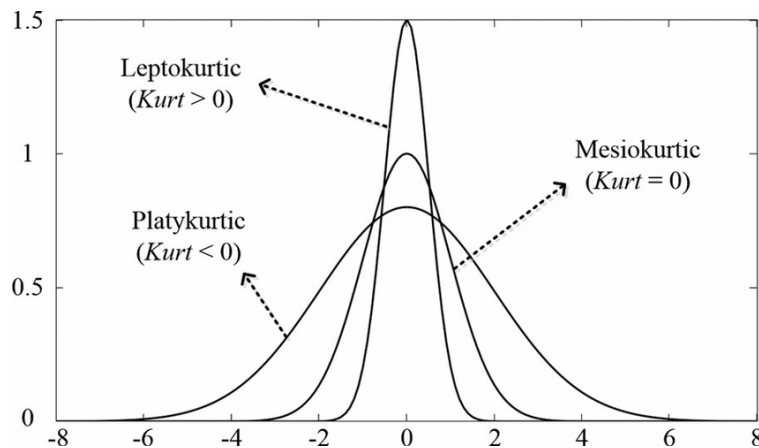
$$kurtosis = \frac{1}{|X|} \sum_{x_i \in X} \frac{(x_i - \bar{x})^4}{s^4}$$

- độ cao phần “trung tâm” hay độ “dày” phần đuôi của phân phối
- thường được so sánh với phân phối chuẩn ($kurtosis = 3$)
→ đại lượng *excess kurtosis*: $Kurt = (kurtosis - 3)$

4. Các đại lượng về hình dáng PP (tt.)



□ Độ nhọn (*kurtosis*)



Tài liệu tham khảo



Anderson et al., *Statistics for Business and Economics*, Cengage, 2016.

Nguyễn Đình Thúc và các tác giả, *Thống kê máy tính*, NXB Khoa học và kỹ thuật, 2010.

Illowsky et al., *Introductory Statistics*, OpenStax, 2017.

Pishro-Nik H., *Introduction to Probability, Statistics, and Random Processes*, Kappa Research LLC, 2014.

Schmitz Andy, *Introductory Statistics*, Saylor Academy, (https://saylordotorg.github.io/text_introductory-statistics/index.html, 09/2019).