Lý Thuyết Về Thống Kê

Khái Quát

- Giới thiệu sơ lược khoa học thống kê.
- Mẫu và phương pháp thu thập
- Các đặc điểm mẫu.

Khoa Học Thống Kê

- Thống kê là môn khoa học bao gồm phương pháp thu thập, xử lý, tổ chức, phân tích, giải thích và trình bày dữ liệu.

Thống Kê Mô Tả Và Thống Kê Suy Luận

- **Thống kê mô tả (descriptive statistics)** là quá trình tổng hợp, sắp xếp để tạo ra đặc điểm cô đọng của một tập dữ liệu (như là mẫu quan sát).
- **Thống kê suy luận (inferential statistics)** là quá trình sử dụng lý thuyết xác suất để suy luận các đặc tính tổng quát hơn của một tập dữ liệu (dùng mẫu để suy luận ra quần thể thống kê).

Thu Thập Dữ Liệu

- Dữ liệu từ thí nghiệm (experimental data): là dữ liệu thu thập được từ các thí nghiệm (experiments) khoa học, trong đó các yếu tố ảnh hưởng có thể được kiểm soát để tìm hiểu ảnh hưởng của tác động nhân quả cần nghiên cứu.
  - Ví dụ: thu thập dữ liệu thử nghiệm vaccine Covid-19 trên người.
- Dữ liệu quan sát (observational data): là dữ liệu được thu thập mà nhà nghiên cứu không thể tác động gì lên hiện tượng tạo ra dữ liệu.
  - Ví dụ: thu thập dữ liệu tiền lương, việc làm dân cư.

Vấn Đề Thu Thập Mẫu Từ Mẫu Từ Quần Thể

- **Quần thể thống kê (statistical population)** là tập hợp tất cả các phần tử chúng ta quan tâm trong một nghiên cứu.
- **Mẫu (a sample)** là một tập hợp con của quần thể mà chúng ta cần nghiên cứu.
  - Trong thực tế vì các lý do khác nhau (như tài chính, thời gian, sự phức tạp của quá trình nghiên cứu…), chúng ta chỉ có thể lấy mẫu để nghiên cứu mà không thể điều tra cả tổng thể.
  - Một tổng thể có thể được lấy nhiều mẫu.

Phương Pháp Thu Thập Mẫu

- Các phương pháp thu thập mẫu ngẫu nhiên:
  - Simple random sampling.
  - Stratified random sampling.
  - Cluster sampling.

- Mẫu (representative) có đại diện cho quần thể hay không?

Thống Kê Mô Tả (Descriptive Statistics)

Các Dạng Dữ Liệu

- **Định tính:**
  - **Định danh (nominal):** chỉ thuần mô tả sự khác biệt, không có thứ tự so sánh.
  - **Thứ bậc (ordinal):** mô tả sự khác biệt và có thứ bậc so sánh.
- **Định lượng:**
  - **Thang đo (interval):** có thứ tự, khác biệt giữa giá trị trong thang đo có ý nghĩal; nhưng giá trị 0 không có ý nghĩa.
  - **Tỷ lệ (ratio)**: giống như thang đo và giá trị 0 có ý nghĩa.

Các Dạng Dữ Liệu: Ví Dụ

Định tính không thứ bậc (Nominal)	Định tính có thứ bậc (Ordinal)	Thang đo (Interval)	Tỷ lệ (Ratio)
Nam, nữ	Trung bình, Khá, Giỏi, Xuất sắc	Nhiệt độ	Giá sản phẩm
Kinh, Tày, Nùng, Dao…	5 sao, 4 sao, 3 sao, 2 sao, 1 sao.	Chỉ số thông minh IQ	Trọng lượng
Mobiphone, Viettel, Vinaphone	Rất đồng tính, Đồng tình, Trung tính, Không Đồng Tình, Rất Không Đồng Tình.	Độ pH	Lợi nhuận doanh nghiệp

Các Đặc Điểm Của Mẫu (1)

- **Bảng tần số**: biểu diễn tần số hoặc tỷ lệ tương đối của từng giá trị quan sát hoặc khoảng giá trị quan sát

X	x_1	x_2	x_3	x_3	x_4
Tần số	n_1	n_2	n_3	n_4	n_5

- **Biểu đồ tần số**: biển diễn bảng tần số dưới dạng biểu đồ.

Các Đặc Điểm Của Mẫu (2)

- **Trung bình mẫu (sample mean):**

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- **Phương sai và độ lệch chuẩn mẫu (sample variance and sample standard deviation):**

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}; \quad \hat{\sigma} = S$$

- **Hệ số biến thiên (coefficient of variation):**

$$CV = \left(\frac{S}{\bar{X}}\right) * 100\%$$

Ví Dụ:

- Dữ liệu về năng suất lúa (tạ/ha) của 10 hộ dân như sau:

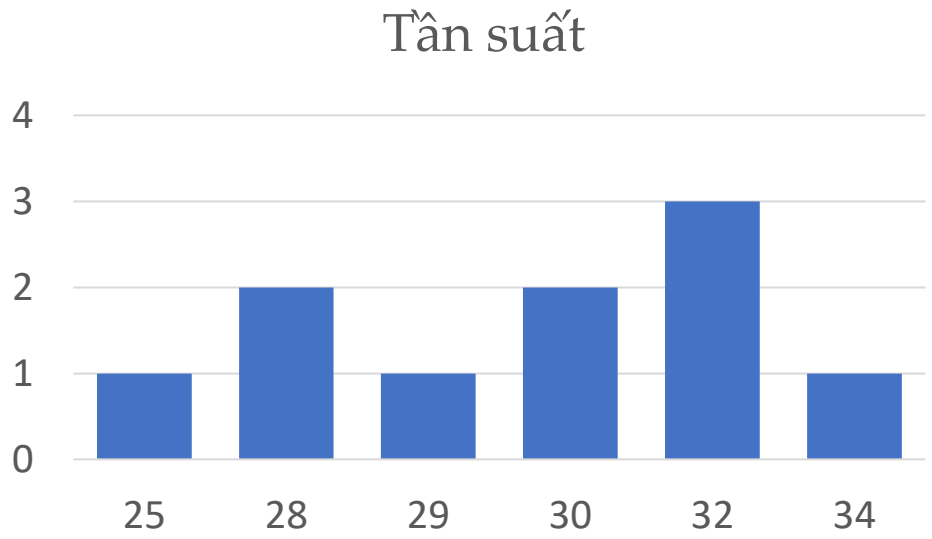
$$30, 32, 29, 30, 34, 32, 28, 32, 28, 25$$

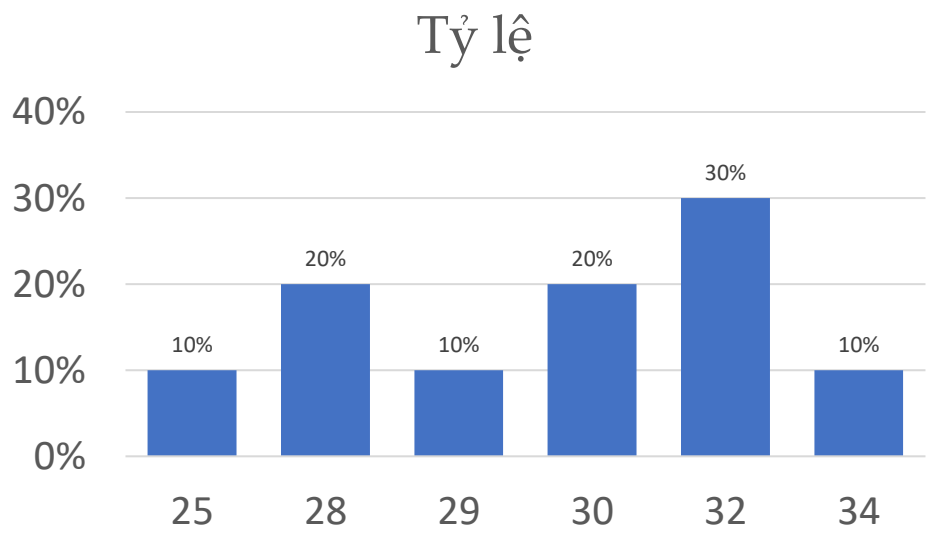
- Hãy tính:
  - Bảng tần số.
  - Trung bình mẫu, phương sai và độ lệch chuẩn mẫu.

Ví Dụ:

Năng suất (tạ/ha)	25	28	29	30	32	34
Tần suất	1	2	1	2	3	1
Tỷ lệ	0.1	0.2	0.1	0.2	0.3	0.1

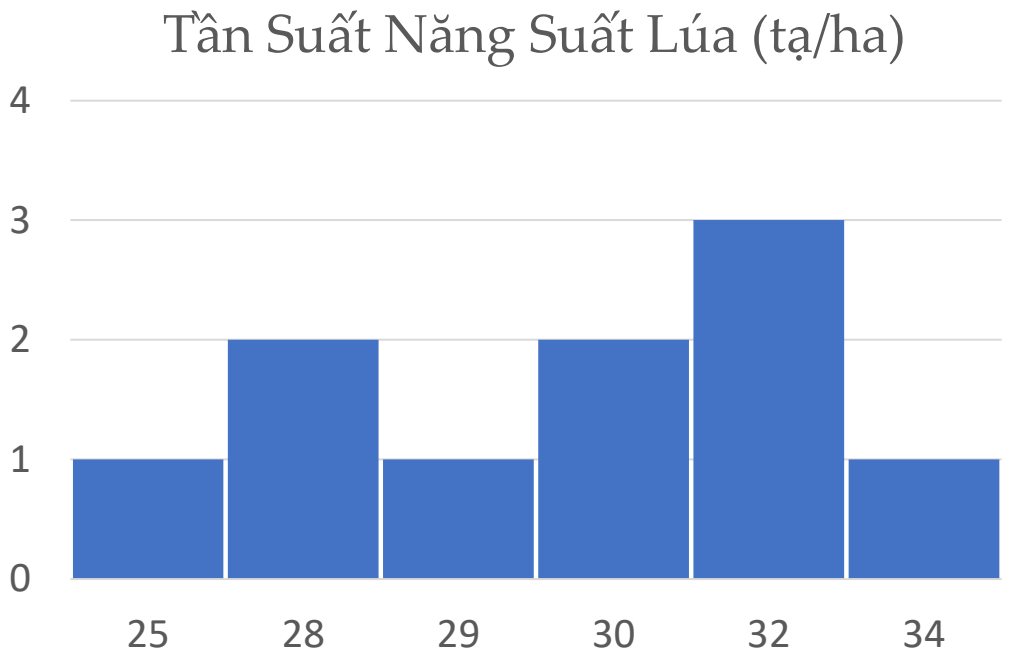
Trung bình mẫu	30
Phương sai mẫu	6.89
Độ lệch chuẩn mẫu	2.62

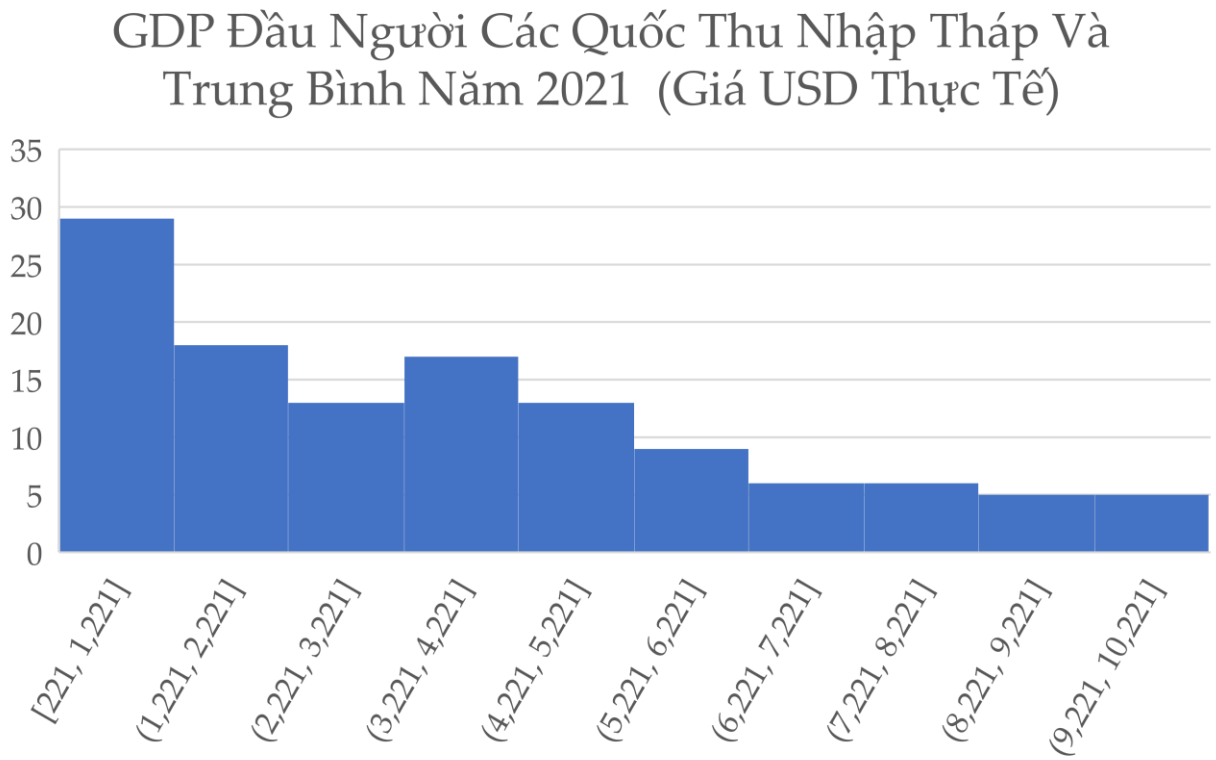




Biểu Đồ Histogram

- Histogram là biểu đồ thể hiện tần suất/tỷ lệ dưới dạng hình các cột hình chữ nhật.





Các Đặc Điểm Của Mẫu (3)

- Giá trị lớn nhất.
- Giá trị nhỏ nhất.
- Trung vị (median).
- Số yếu vị (mode).
- Khoảng cách giữa giá trị lớn nhất và nhỏ nhất (range).

Ví Dụ:

- Dữ liệu về năng suất lúa (tạ/ha) của 10 hộ dân như sau:

$$30, 32, 29, 30, 34, 32, 28, 32, 28, 25$$

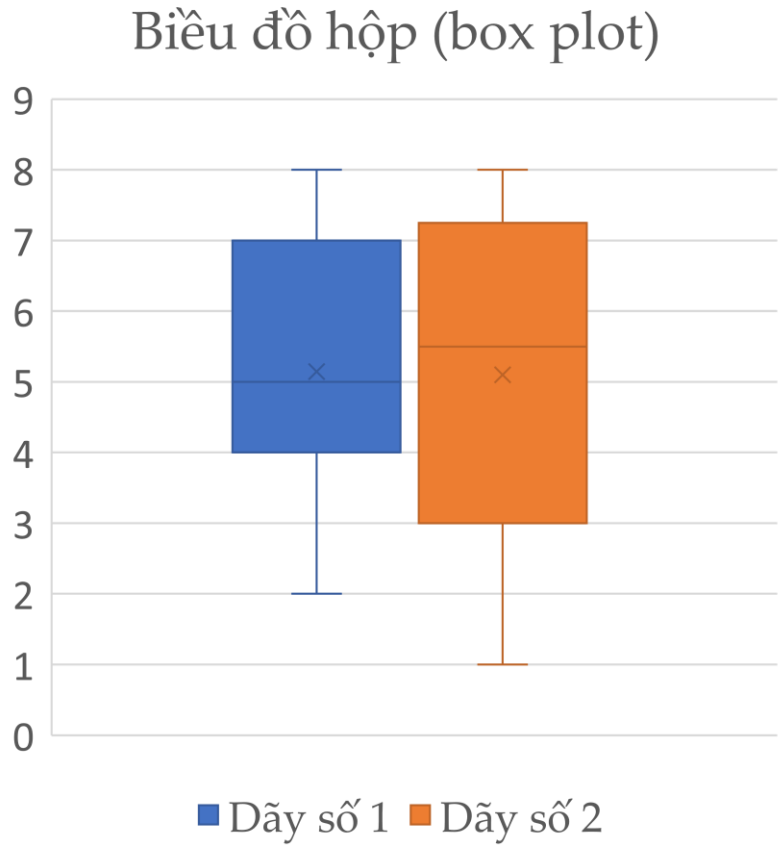
- Hãy tính:
  - Median, mode and range?

Các Đặc Trưng Của Mẫu (3)

- **Điểm tứ phân vị (quartile)**: chia dữ liệu thành bốn phần có tỷ suất phân bố bằng nhau:
  - Điểm tứ phân vị thứ nhất Q1 (first quartile): 25% số lượng giá trị nhỏ hơn Q1.
  - Điểm tứ phân vị thứ nhì Q2 (second quartile): đúng bằng trung vị (median), 50% số lượng giá trị nằm giữa Q2.
  - Điểm tứ phân vị thứ ba Q3 (third quartile): 75% số lượng giá trị nhỏ hơn Q3.
- Khoảng cách giữa Q1 và Q3 gọi là **độ trải giữa (interquartile range IQR).**

Ví Dụ: Tính Q_1, Q_2, Q_3 cho dãy số

- Cho dãy số thứ nhất: $2, 4, 4, 5, 6, 7, 8$



Ví Dụ: Tính Q_1, Q_2, Q_3 cho dãy số

- Cho dãy số thứ nhất: $2, 4, 4, 5, 6, 7, 8$

  $2, 4, 4, \mathbf{5}, 6, 7, 8$

  $2, \mathbf{4}, 4, 5, 6, \mathbf{7}, 8$

  $Q_1 = 4, Q_2 = 5, Q_3 = 7$

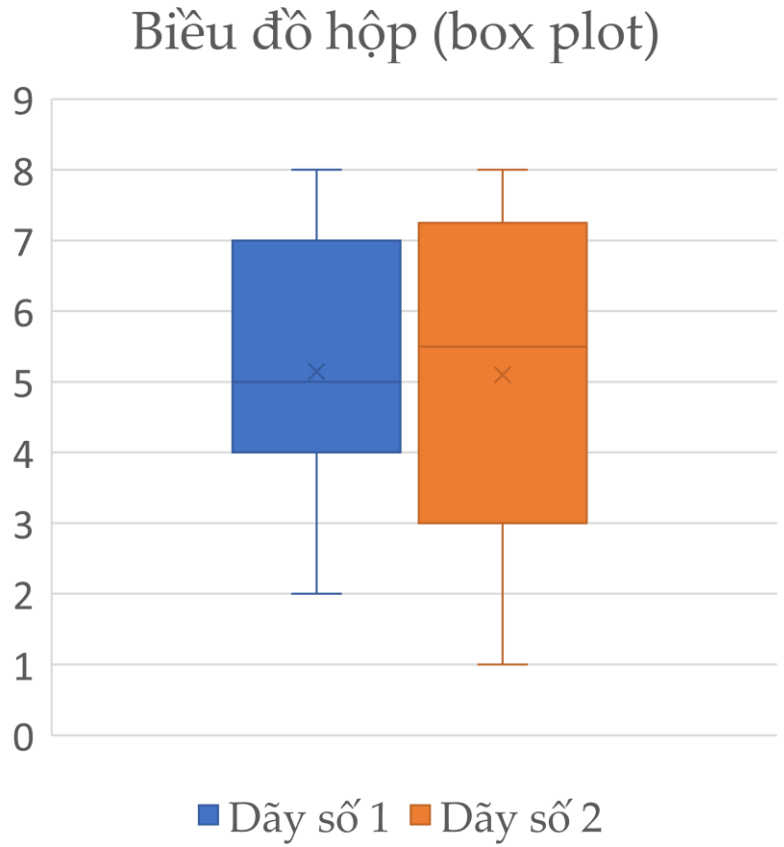
- Cho dãy số thứ hai: $1, 3, 3, 4, 5, 6, 6, 7, 8, 8$

  $1, 3, 3, 4, 5, 6, 6, 7, 8, 8$

  $Q_2 = (5/6) = 5.5$

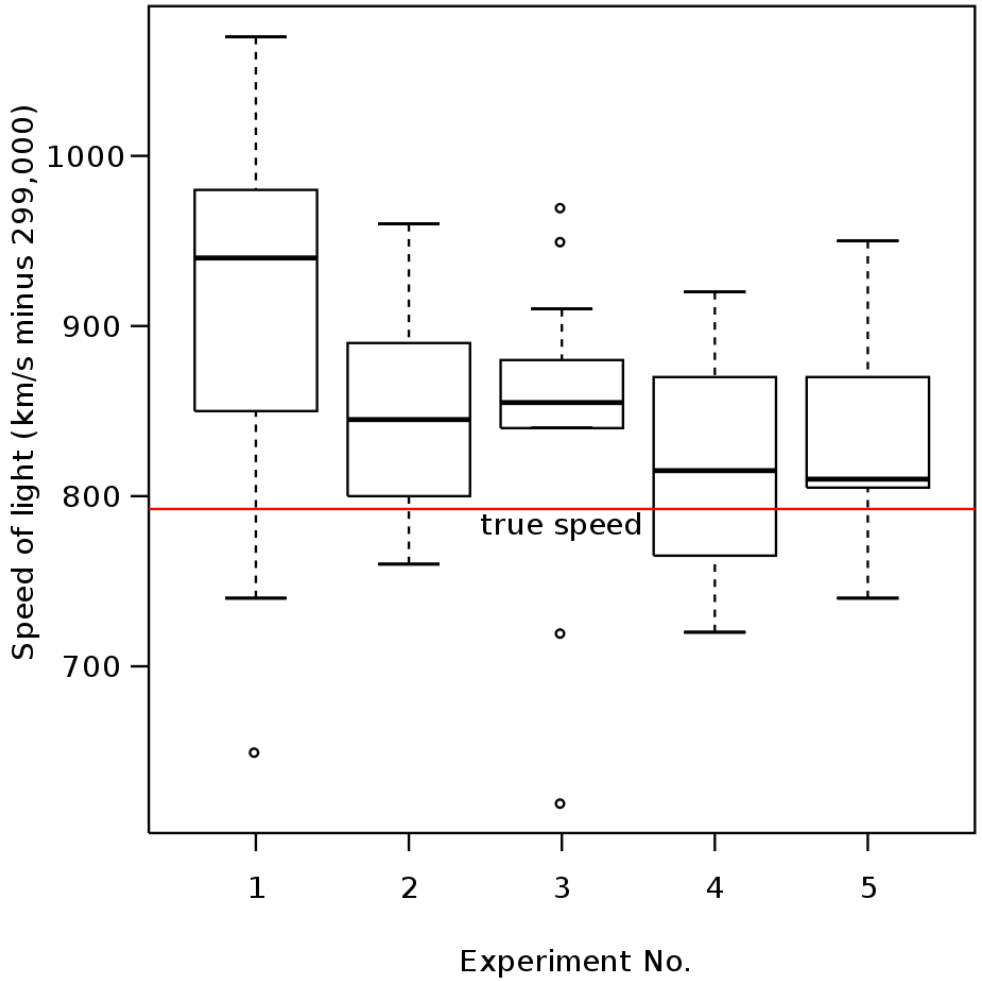
  $1, 3, \mathbf{3}, 4, 5, | \, 6, 6, \mathbf{7}, 8, 8$

  $Q_1 = 3, Q_3 = 7$



Biểu Đồ Hộp (Box Plot)

- Biểu đồ hộp (box and whisker plot or box plot) thể hiện 5 thông tin:
  - Giá trị lớn nhất Q_0 (max).
  - Giá trị nhỏ nhất Q_4 (min).
  - Trung vị (median).
  - Tứ vị phân đầu tiên Q_1 (first quartile).
  - Tứ vị phân thứ ba Q_3 (third quartile).
- Thông tin độ trải giữa IQR có thể được tính: $\text{IQR} = Q_3 - Q_1$.



Source: https://en.wikipedia.org/wiki/Box_plot

So Sánh Các Dạng Dữ Liệu

Dạng dữ liệu	Phép toán	Đo lường tính trung tâm	Đo lường sự phân tán
Định danh (nominal)	• So sánh ngang bằng (=, ≠)	• Mode	• Không có
Thứ bậc (ordinal)	• So sánh ngang bằng (=, ≠) • So sánh hơn kém (<, >)	• Mode • Trung vị (median)	• Khoảng cách (range) • IQR
Thang đo (interval)	• So sánh ngang bằng (=, ≠) • So sánh hơn kém (<, >) • Cộng trừ (+, −)	• Mode • Trung vị (median) • Giá trị trung bình	• Khoảng cách (range) • IQR • Phương sai, độ lệch chuẩn
Tỷ lệ (ratio)	• So sánh ngang bằng (=, ≠) • So sánh hơn kém (<, >) • Cộng trừ (+, −) • Nhân chia (×,÷)	• Mode • Trung vị (median) • Giá trị trung bình cộng • Giá trị trung bình nhân	• Khoảng cách (range) • IQR • Phương sai, độ lệch chuẩn • Hệ số biến thiên (CV)

Các Đặc Điểm Của Mẫu (4)

- **Hiệp phương sai mẫu (sample variance):**

$$\widehat{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

- **Hệ số tương quan mẫu (sample correlation):**

$$\hat{\rho}_{XY} = \frac{\widehat{cov}(X,Y)}{S_X S_Y}$$

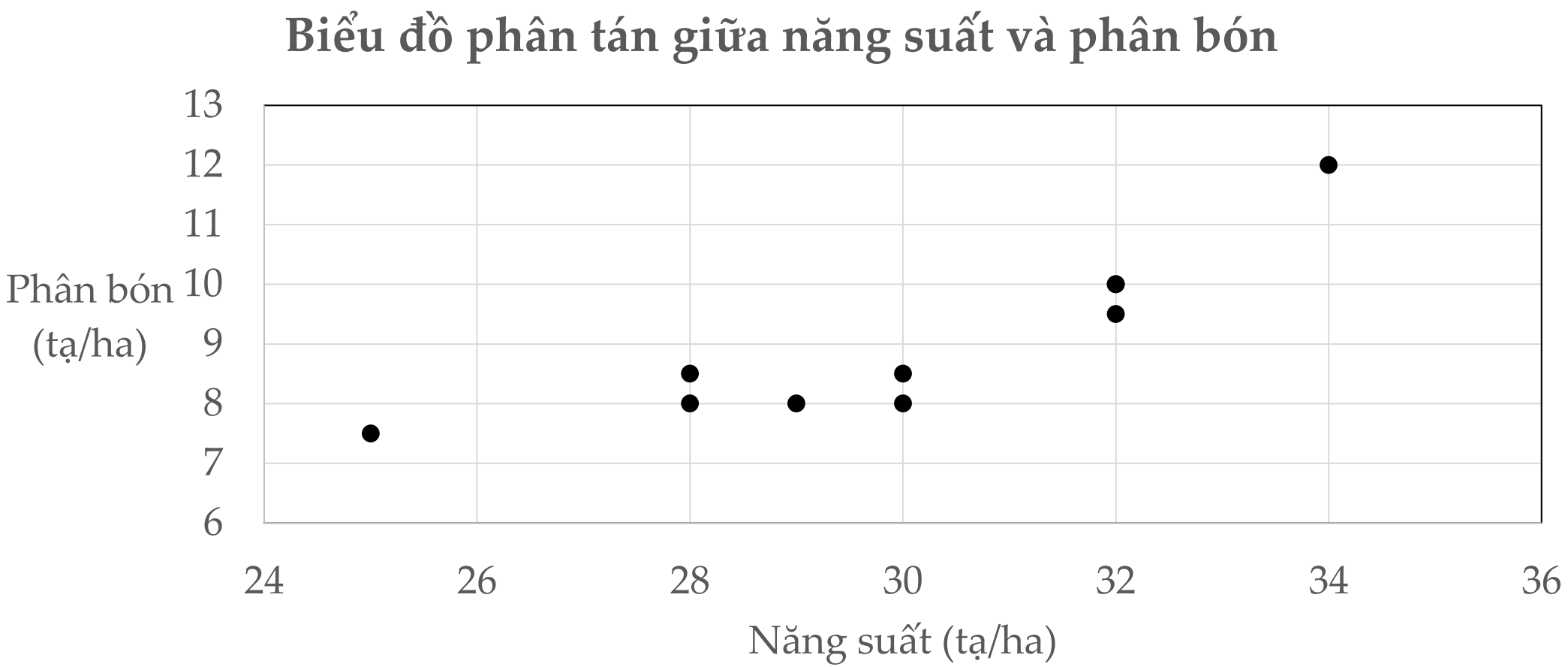
Ví dụ:

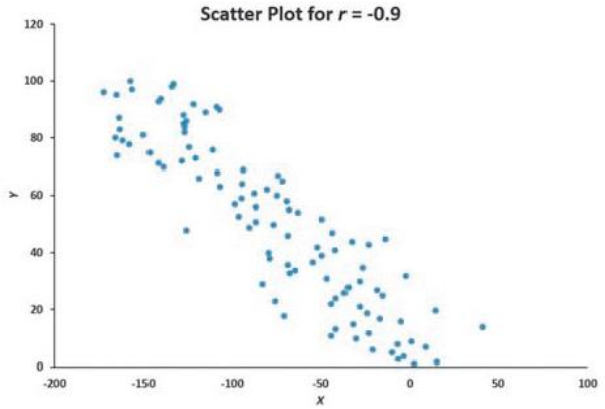
- Dữ liệu về năng suất lúa (tạ/ha) và lượng phân bón sử dụng trung bình (tạ/ha) của 10 hộ dân như sau:

Năng suất	30	32	29	30	34	32	28	32	28	25
Phân bón	8.5	9.5	8	8	12	10	8	10	8.5	7.5

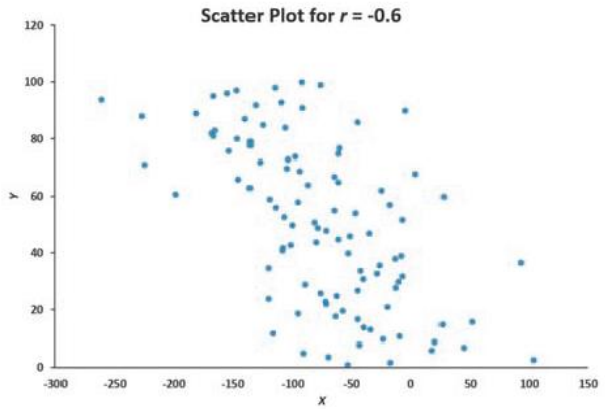
- Hãy tính:
  - Hiệp phương sai mẫu.
  - Hệ số tương quan mẫu.

Biểu Đồ Phân Tán (Scatter Plot)

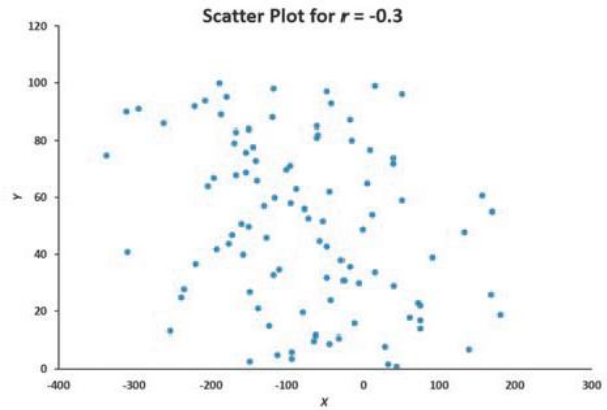




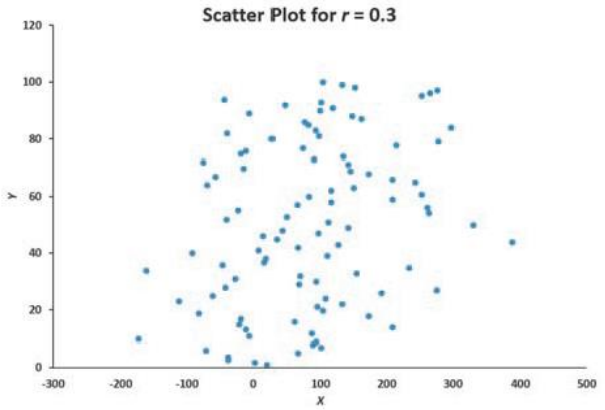
Panel A



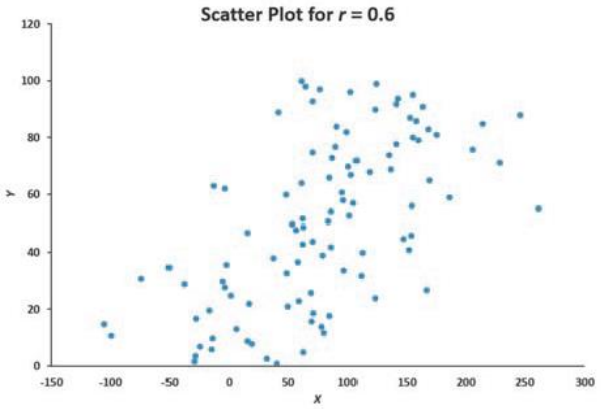
Panel B



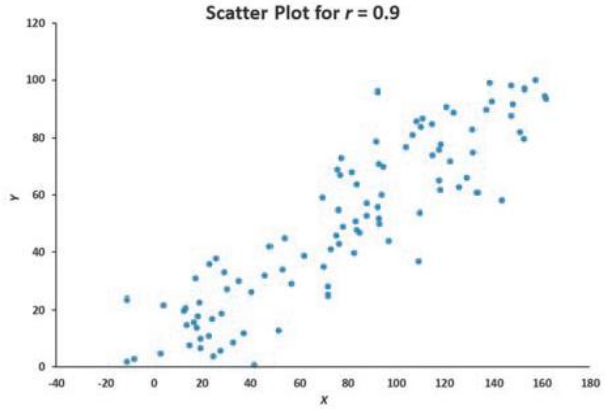
Panel C



Panel D



Panel E

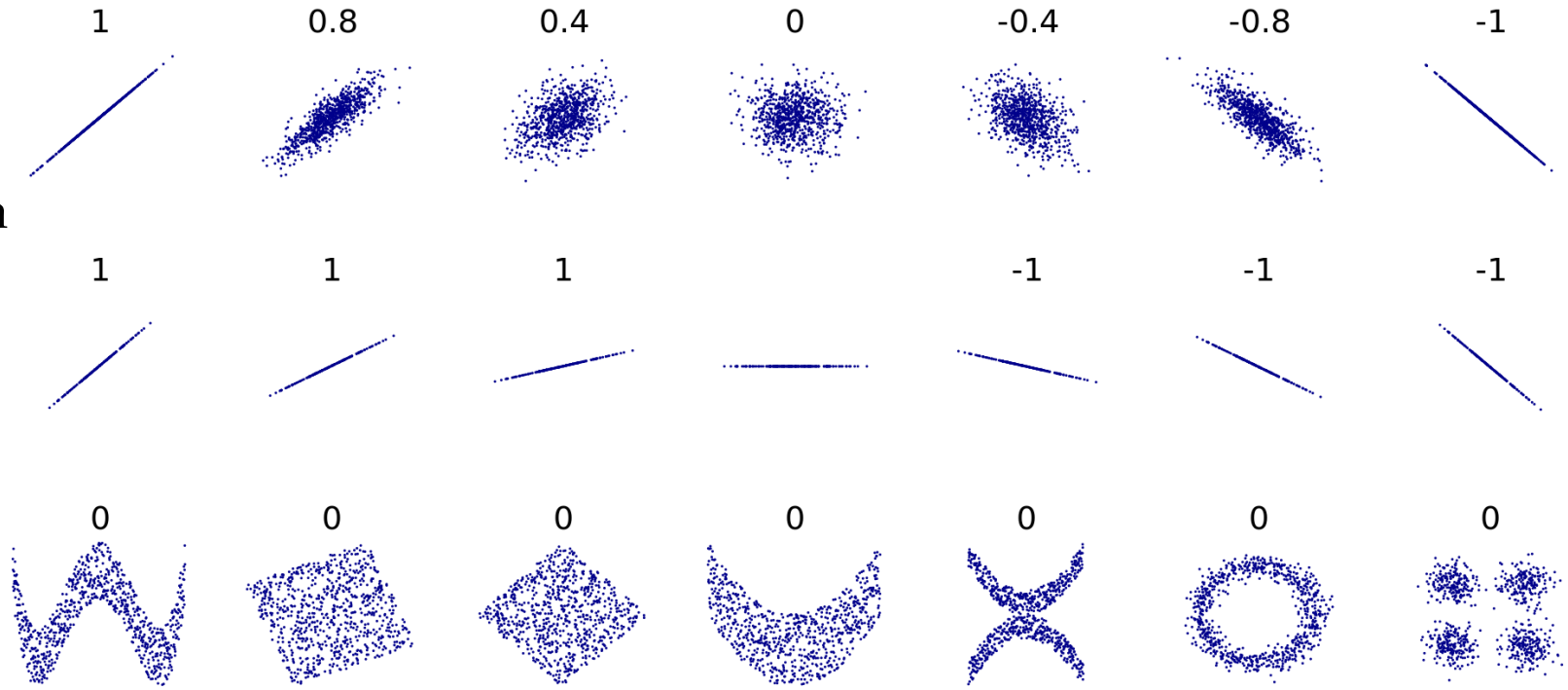


Panel F

Ghi chú:

Hệ số tương quan mẫu (các số trên đầu từng biểu đồ scatter plot) thể hiện:

- Dòng 1: hướng (âm hay dương) của mối quan hệ tuyến tính của hai biến.
- Dòng 2: không thể hiện đúng mối quan hệ tuyến tính.
- Dòng 3: không phản ánh một chút nào mối quan hệ phi tuyến tính giữa hai biến.



Sự Khác Biệt Giữa Các Đặc Điểm Của Mẫu Và Phân Phối Xác Suất

- Phân phối xác suất so với tần suất từ mẫu thu thập?
- Giá trị kì vọng so với trung bình mẫu?
- Trừu tượng và quan sát được?