Thống kê Mô tả

Khoa Công nghệ Thông tin và Truyền thông

Nội dung

Giới thiệu về Thống kê

Thống kê là gì?

- Trong thực tế, ta thường gặp các số liệu thống kê và thông tin về thống kê. Ví dụ:
  - thông tin khảo sát khách hàng,
  - thông tin tiếp thị, quảng cáo,
  - thăm dò bỏ phiếu trong các cuộc tranh cử.

Thống kê là gì?

- Trong thực tế, ta thường gặp các số liệu thống kê và thông tin về thống kê. Ví dụ:
  - thông tin khảo sát khách hàng,
  - thông tin tiếp thị, quảng cáo,
  - thăm dò bỏ phiếu trong các cuộc tranh cử.
- Làm thế nào để hiểu được tất cả các số liệu này?

Một số vấn đề thực tế

Vấn đề

Làm thế nào để biết được thu nhập trung bình của người dân ở một địa phương?

Một số vấn đề thực tế

Vấn đề

Làm thế nào để biết được thu nhập trung bình của người dân ở một địa phương?

- Một nhà điều tra thử tính thu nhập trung bình của 100 người dân ở địa phương này thì được kết quả là 180USD/tháng. Có thể nói gì về thu nhập của người dân ở toàn địa phương đó?

Một số vấn đề thực tế

Vấn đề

Làm thế nào để biết được thu nhập trung bình của người dân ở một địa phương?

- Một nhà điều tra thử tính thu nhập trung bình của 100 người dân ở địa phương này thì được kết quả là 180USD/tháng. Có thể nói gì về thu nhập của người dân ở toàn địa phương đó?
- Nhà quản lý nói rằng thu nhập trung bình của người dân ở toàn địa phương đó là 200 USD/tháng. Nhận định của nhà quản lý có đúng không?

Một số vấn đề thực tế (tiếp)

Vấn đề

Làm thế nào để so sánh hiệu quả của phác đồ điều trị mới cho một bệnh nào đó?

- Điều trị 250 bệnh nhân theo phác đồ điều trị cũ thấy có 190 người khỏi bệnh. Điều trị 100 bệnh nhân theo phác đồ điều trị mới thấy có 80 người khỏi bệnh. Hỏi phác đồ điều trị mới có thực sự hiệu quả hơn phác đồ cũ?

Một số vấn đề thực tế (tiếp)

Vấn đề

Con của bạn sẽ cao bao nhiêu cm?

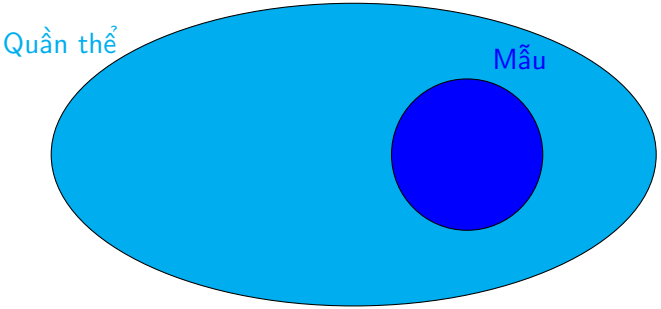
- Chiều cao của con có bị ảnh hưởng bởi chiều cao của bố hay chiều cao của mẹ?
- Chiều cao của bố/mẹ và con liên quan như thế nào với nhau?
- Nếu biết chiều cao của bố/mẹ thì có thể dự đoán được chiều cao của con không?

Một số khái niệm của thống kê

- Quần thể (population): Là tập hợp tất cả các đối tượng mà ta cần nghiên cứu. Quần thể thường rất lớn, đôi khi là vô hạn.
- Mẫu (sample): Là tập hợp một số phần tử đại diện lấy từ quần thể mà ta chọn để tiến hành nghiên cứu.
  Số phần tử của một mẫu được gọi là cỡ mẫu, ký hiệu là n.

Một số khái niệm của thống kê

- Quần thể (population): Là tập hợp tất cả các đối tượng mà ta cần nghiên cứu. Quần thể thường rất lớn, đôi khi là vô hạn.
- Mẫu (sample): Là tập hợp một số phần tử đại diện lấy từ quần thể mà ta chọn để tiến hành nghiên cứu.
  Số phần tử của một mẫu được gọi là cỡ mẫu, ký hiệu là n.



Thống kê là gì?

Để nghiên cứu các tính chất của một quần thể ta có thể:

- khảo sát toàn bộ các phần tử của quần thể, hoặc
- khảo sát một bộ phận của quần thể đó, sau đó tìm cách rút ra kết luận dựa trên dữ liệu quan sát được.

Thống kê là gì?

Để nghiên cứu các tính chất của một quần thể ta có thể:

- khảo sát toàn bộ các phần tử của quần thể, hoặc
- khảo sát một bộ phận của quần thể đó, sau đó tìm cách rút ra kết luận dựa trên dữ liệu quan sát được.

Định nghĩa 1

Thống kê là một quá trình sử dụng phương pháp khoa học để trả lời các câu hỏi và đưa ra quyết định. Quá trình đó bao gồm thiết kế các nghiên cứu, thu thập, xử lý, biểu diễn, phân tích mẫu số liệu thu thập được từ một quần thể để rút ra được các kết luận có độ tin cậy cho toàn bộ quần thể đó.

Xác định dữ liệu cần thu thập

- Xác định rõ dữ liệu nào cần thu thập, thứ tự ưu tiên của các dữ liệu này. Nếu không sẽ mất rất nhiều thời gian và chi phí cho những dữ liệu ít quan trọng hay không liên quan đến vấn đề cần nghiên cứu.
- Xác định số các đơn vị điều tra (cỡ mẫu).

Dữ liệu sơ cấp và thứ cấp

Dữ liệu sơ cấp là dữ liệu thu thập trực tiếp, ban đầu từ đối tượng nghiên cứu.

- Ưu điểm: đáp ứng tốt nhu cầu nghiên cứu.
- Nhược điểm: tốn kém nhiều về thời gian và chi phí.
- Phương pháp thu thập: thực nghiệm, khảo sát qua điện thoại, thư hỏi, quan sát trực tiếp và phỏng vấn cá nhân.

Dữ liệu sơ cấp và thứ cấp

Dữ liệu sơ cấp là dữ liệu thu thập trực tiếp, ban đầu từ đối tượng nghiên cứu.

- Ưu điểm: đáp ứng tốt nhu cầu nghiên cứu.
- Nhược điểm: tốn kém nhiều về thời gian và chi phí.
- Phương pháp thu thập: thực nghiệm, khảo sát qua điện thoại, thư hỏi, quan sát trực tiếp và phỏng vấn cá nhân.

Dữ liệu thứ cấp là dữ liệu đã qua tổng hợp, xử lý.

- Ưu điểm: thu thập nhanh, ít tốn kém chi phí.
- Nhược điểm: đôi khi ít chi tiết và không đáp ứng đúng nhu cầu nghiên cứu.
- Nguồn cung cấp: số liệu nội bộ, số liệu từ cơ quan thống kê nhà nước, cơ quan chính phủ, báo, tạp chí, các tổ chức, hiệp hội, viện nghiên cứu,...

Lấy mẫu hoàn lại và không hoàn lại

- Lấy mẫu ngẫu nhiên có hoàn lại: lần lượt lấy ngẫu nhiên từ quần thể ra một phần tử, thu thập các thông tin cần thiết từ phần tử đó rồi trả nó trở lại quần thể trước khi lấy tiếp lần sau.

Lấy mẫu hoàn lại và không hoàn lại

- Lấy mẫu ngẫu nhiên có hoàn lại: lần lượt lấy ngẫu nhiên từ quần thể ra một phần tử, thu thập các thông tin cần thiết từ phần tử đó rồi trả nó trở lại quần thể trước khi lấy tiếp lần sau.
- Lấy mẫu ngẫu nhiên không hoàn lại: tương tự như trên nhưng khác ở chỗ các phần tử đã lấy ra sẽ không được chọn lại ở lần sau.

Quy tắc lấy mẫu

Căn cứ vào dữ liệu của mẫu mà ta thu thập được, để có thể đưa ra những kết luận đủ chính xác về dấu hiệu nghiên cứu trong quần thể thì trước hết mẫu được chọn phải *mang tính đại diện cho quần thể.*

- Mỗi phần tử được lấy vào mẫu một cách hoàn toàn ngẫu nhiên, tức là mọi phần tử của quần thể đều có thể được lấy vào mẫu với khả năng như nhau.
- Các phần tử của mẫu được chọn lần lượt, độc lập với nhau và có hoàn lại từ quần thể.

Quy tắc lấy mẫu

Căn cứ vào dữ liệu của mẫu mà ta thu thập được, để có thể đưa ra những kết luận đủ chính xác về dấu hiệu nghiên cứu trong quần thể thì trước hết mẫu được chọn phải *mang tính đại diện cho quần thể.*

- Mỗi phần tử được lấy vào mẫu một cách hoàn toàn ngẫu nhiên, tức là mọi phần tử của quần thể đều có thể được lấy vào mẫu với khả năng như nhau.
- Các phần tử của mẫu được chọn lần lượt, độc lập với nhau và có hoàn lại từ quần thể.

Nhận xét

Khi kích thước của quần thể khá lớn còn kích thước của mẫu lại nhỏ thì phương thức lấy mẫu hoàn lại và lấy mẫu không hoàn lại cho ta kết quả sai lệch không đáng kể.

Các phương pháp lấy mẫu

- Mẫu giản đơn: là mẫu được chọn trực tiếp từ danh sách đã được đánh số của quần thể. Từ quần thể kích thước m, ta rút ra mẫu n phần tử bằng cách bốc thăm, chọn số ngẫu nhiên từ bảng hoặc sinh số ngẫu nhiên từ máy tính.

Các phương pháp lấy mẫu

- Mẫu giản đơn: là mẫu được chọn trực tiếp từ danh sách đã được đánh số của quần thể. Từ quần thể kích thước m, ta rút ra mẫu n phần tử bằng cách bốc thăm, chọn số ngẫu nhiên từ bảng hoặc sinh số ngẫu nhiên từ máy tính.
- Mẫu phân tầng: quần thể được chia thành nhóm và mỗi nhóm được lấy mẫu giản đơn.

Các phương pháp lấy mẫu

- Mẫu giản đơn: là mẫu được chọn trực tiếp từ danh sách đã được đánh số của quần thể. Từ quần thể kích thước m, ta rút ra mẫu n phần tử bằng cách bốc thăm, chọn số ngẫu nhiên từ bảng hoặc sinh số ngẫu nhiên từ máy tính.
- Mẫu phân tầng: quần thể được chia thành nhóm và mỗi nhóm được lấy mẫu giản đơn.
- Lấy mẫu cụm: quần thể được chia thành nhiều cụm. Đầu tiên chọn ngẫu nhiên một số cụm, sau đó lại chọn ngẫu nhiên phần tử từ các cụm được chọn bằng phương pháp lấy mẫu giản đơn.

Các phương pháp lấy mẫu

- Mẫu giản đơn: là mẫu được chọn trực tiếp từ danh sách đã được đánh số của quần thể. Từ quần thể kích thước m, ta rút ra mẫu n phần tử bằng cách bốc thăm, chọn số ngẫu nhiên từ bảng hoặc sinh số ngẫu nhiên từ máy tính.
- Mẫu phân tầng: quần thể được chia thành nhóm và mỗi nhóm được lấy mẫu giản đơn.
- Lấy mẫu cụm: quần thể được chia thành nhiều cụm. Đầu tiên chọn ngẫu nhiên một số cụm, sau đó lại chọn ngẫu nhiên phần tử từ các cụm được chọn bằng phương pháp lấy mẫu giản đơn.
- Mẫu hệ thống: đánh số các phần tử của quần thể từ 1 đến N. Chọn ngẫu nhiên 1 phần tử trong k phần tử đầu tiên ($k < N$), từ phần tử này cứ cách k phần tử của quần thể lại lấy ra một phần tử cho vào mẫu.

Các phương pháp lấy mẫu

- Mẫu giản đơn: là mẫu được chọn trực tiếp từ danh sách đã được đánh số của quần thể. Từ quần thể kích thước m, ta rút ra mẫu n phần tử bằng cách bốc thăm, chọn số ngẫu nhiên từ bảng hoặc sinh số ngẫu nhiên từ máy tính.
- Mẫu phân tầng: quần thể được chia thành nhóm và mỗi nhóm được lấy mẫu giản đơn.
- Lấy mẫu cụm: quần thể được chia thành nhiều cụm. Đầu tiên chọn ngẫu nhiên một số cụm, sau đó lại chọn ngẫu nhiên phần tử từ các cụm được chọn bằng phương pháp lấy mẫu giản đơn.
- Mẫu hệ thống: đánh số các phần tử của quần thể từ 1 đến N. Chọn ngẫu nhiên 1 phần tử trong k phần tử đầu tiên ($k < N$), từ phần tử này cứ cách k phần tử của quần thể lại lấy ra một phần tử cho vào mẫu.
- Lấy mẫu nhiều tầng: kết hợp nhiều phương pháp.

Trình bày một mẫu số liệu

Các phương pháp trình bày dữ liệu

(*a*) Đối với dữ liệu định tính

- Bảng tần số, tần suất, tần số tích lũy, tần suất tích lũy.
- Biểu đồ hình cột, hình tròn.

(*b*) Đối với dữ liệu định lượng

- Biểu đồ thân-lá.
- Bảng tần số, tần suất, tần số tích lũy, tần suất tích lũy.
- Biểu đồ hình cột, hình tròn.

Bảng tần số

- Tần số (frequency) là số lần biến số nhận một giá trị nào đó.
- Tỉ lệ (proportion) là tần số được diễn tả một cách tương đối, được tính bằng cách lấy tần số chia cho tổng số quan sát.
- Tỉ lệ phần trăm (percentage) là tỉ lệ được nhân lên cho 100. Tỉ lệ và tỉ lệ phần trăm được gọi là tần số tương đối (relative frequencies) hay tần suất.
- Bảng tần số/tần suất (frequency table) là bảng liệt kê các giá trị (hoặc khoảng giá trị) của một biến và tần số/tần suất của chúng.

Ví dụ

Ví dụ 1

Năm 2016, báo Tuổi trẻ Online thực hiện khảo sát về bình chọn Quốc hoa Việt Nam, kết quả thu được như sau:

Quốc hoa	**Số lượt bình chọn**	**Tỉ lệ** (%)
Hoa sen	67008	49.6
Cây tre	47288	35
Hoa mai	15850	11.7
Đề xuất khác	4951	3.7
Tổng	135097	100

Áp dụng

Bài tập

Thống kê sinh viên của một khóa ở một trường đại học ta có bảng số liệu sau:

Ngành học	Tần số (số sinh viên)	Tỉ lệ (%)
Quản trị Kinh doanh		
Điện tử và Viễn thông	350	
Công nghệ Thông tin		20
Tổng	1000	100

Bảng tần số (tiếp)

(a) Trường hợp dữ liệu có ít giá trị.

Ví dụ 2

Kết quả khảo sát môn Toán của HS khối 12 trường THPT X được cho bởi:

Điểm thi	Tần số (số học sinh)	Tỉ lệ (%)
3	3	3.75
4	12	15
5	15	18.75
6	20	25
7	16	20
8	8	10
9	4	5
10	2	2.5
Tổng	80	100

Bảng tần số (tiếp)

(b) Trường hợp dữ liệu có nhiều giá trị.

- Nếu dữ liệu có nhiều giá trị khác nhau, khoảng cách giữa các giá trị không đồng đều hoặc các giá trị khác nhau rất ít thì ta sẽ biểu diễn chúng dưới dạng khoảng.
- Ví dụ: khảo sát 1200 người trong độ tuổi lao động (từ 18 đến 60 tuổi), nếu lập bảng như ở ví dụ trên thì sẽ rất dài, làm mất đi tác dụng tóm lược thông tin. Do đó, ta thường phân thành các nhóm, chẳng hạn: từ 18 đến 21 tuổi, từ 21 đến 30 tuổi, từ 31 đến 40 tuổi, từ 41 đến 50 tuổi, từ 51 đến 60 tuổi.

Bảng tần số (tiếp)

(b) Trường hợp dữ liệu có nhiều giá trị.

- Nếu dữ liệu có nhiều giá trị khác nhau, khoảng cách giữa các giá trị không đồng đều hoặc các giá trị khác nhau rất ít thì ta sẽ biểu diễn chúng dưới dạng khoảng.
- Ví dụ: khảo sát 1200 người trong độ tuổi lao động (từ 18 đến 60 tuổi), nếu lập bảng như ở ví dụ trên thì sẽ rất dài, làm mất đi tác dụng tóm lược thông tin. Do đó, ta thường phân thành các nhóm, chẳng hạn: từ 18 đến 21 tuổi, từ 21 đến 30 tuổi, từ 31 đến 40 tuổi, từ 41 đến 50 tuổi, từ 51 đến 60 tuổi.
- Chú ý:
  - Số khoảng tối ưu là $\sqrt{n}$.
  - Độ dài mỗi khoảng xấp xỉ $h = \frac{x_{max} - x_{min}}{\sqrt{n}}$.

Ví dụ

Năng suất (tạ/ha) của một loại cây thu hoạch được tại 40 khu vực canh tác như sau:

153	154	156	157	158	159	159	160	160	160
161	161	161	162	162	162	163	163	163	164
164	164	165	165	166	166	167	167	168	168
170	171	172	173	174	175	176	177	178	179

Hãy lập bảng tần số của số liệu trên theo mẫu:

Năng suất	Tần số	Tần suất (%)
Tổng	40	100

Ví dụ (tiếp)

- Số khoảng tối ưu là $\sqrt{40} \approx 6$.
- Độ dài mỗi khoảng xấp xỉ $h = \frac{x_{max} - x_{min}}{\sqrt{n}} = \frac{179-153}{\sqrt{40}} \approx 4$.

Ví dụ (tiếp)

- Số khoảng tối ưu là $\sqrt{40} \approx 6$.
- Độ dài mỗi khoảng xấp xỉ $h = \frac{x_{max} - x_{min}}{\sqrt{n}} = \frac{179-153}{\sqrt{40}} \approx 4$.

Ta có bảng tần số sau.

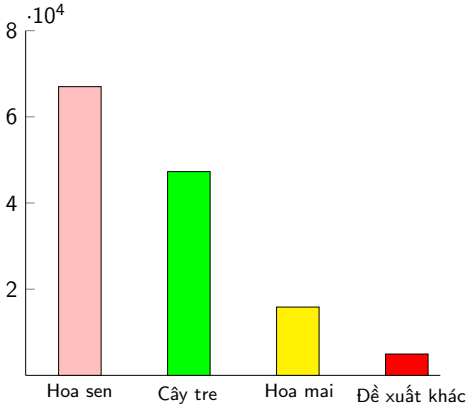
Năng suất	Tần số	Tần suất (%)
152-157	4	10
157-161	9	22.5
161-165	11	27.5
165-169	6	15
169-173	4	10
173-180	6	15
Tổng	40	100

Biểu đồ tần số

Biểu đồ tần số là cách biểu diễn trực quan bảng tần số của số liệu. Để xây dựng một biểu đồ tần số, ta thực hiện các bước như sau.

- Bước 1: gắn nhãn các mốc của từng khoảng trên một thang nằm ngang.
- Bước 2: đánh dấu và dán nhãn thang thẳng đứng theo tần số.
- Bước 3: trên mỗi khoảng, vẽ một hình chữ nhật có chiều cao bằng với tần số tương ứng với khoảng đó.

Ví dụ

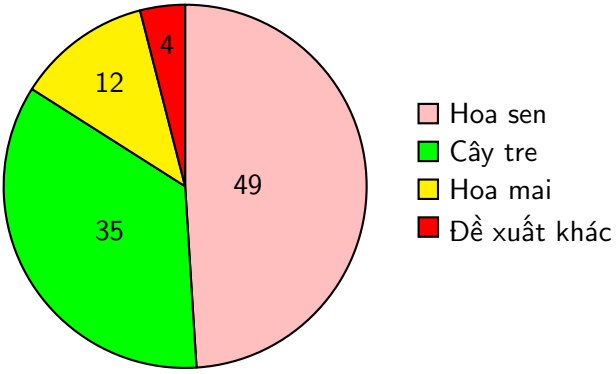


Hình 1: Số lượt bình chọn Quốc hoa.

Biểu đồ tần suất

Biểu đồ tần suất là cách biểu diễn trực quan bảng tần suất của số liệu. Biểu đồ tần suất thường có hình tròn, mỗi hình quạt tương ứng với một biến số hay khoảng biến số. Chú ý: Độ lớn góc ở tâm của hình quạt $=$ tỉ lệ $\times 360^\circ$.

Ví dụ



Hình 2: Số lượt bình chọn Quốc hoa.

Hàm phân phối mẫu

Hàm phân phối mẫu

Định nghĩa 2

Cho mẫu ngẫu nhiên $(X_1, \ldots, X_n)$ từ phân phối $F(x)$. Hàm số

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i < x), \quad x \in \mathbb{R},$$

với $\mathbb{I}(\cdot)$ là hàm chỉ dấu, được gọi là hàm phân phối mẫu.

Tính chất

Một số tính chất của hàm phân phối mẫu

- $0 \leq F_n(x) \leq 1$.
- $F_n(x)$ là hàm không giảm theo x.
- $F_n(x)$ là hàm liên tục trái và có giới hạn phải.
- $\lim_{n \to -\infty} F_n(x) = 0$, $\lim_{n \to \infty} F_n(x) = 1$.

Các đặc trưng của mẫu

Trung bình mẫu

Định nghĩa 3 (Trung bình mẫu - Sample mean)

Cho mẫu các giá trị của biến X với kích thước n gồm $\{x_1, \ldots, x_n\}$. *Khi đó, trung bình mẫu, ký hiệu bởi* $\bar{x}$, *được cho bởi công thức*

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

Trung bình mẫu

- Nếu mẫu dữ liệu được biểu diễn dưới dạng một bảng tần số:

Giá trị	x_1	$\cdots$	x_k
Tần số	n_1	$\cdots$	n_k

thì,

$$\bar{x} = \frac{x_1 n_1 + \cdots + x_k n_k}{n}.$$

Trung bình mẫu

- Nếu mẫu dữ liệu được biểu diễn dưới dạng một bảng tần số:

Giá trị	x_1	$\cdots$	x_k
Tần số	n_1	$\cdots$	n_k

thì,

$$\bar{x} = \frac{x_1 n_1 + \cdots + x_k n_k}{n}.$$

- Nếu mẫu dữ liệu được biểu diễn dưới dạng một bảng tần số theo nhóm:

Giá trị	(a_1, a_2)	(a_2, a_3)	$\cdots$	(a_k, a_{k+1})
Tần số	n_1	n_2	$\cdots$	n_k

Gọi $x_i = \frac{a_i + a_{i+1}}{2}$ là giá trị đại diện cho khoảng (a_i, a_{i+1}). Khi đó,

$$\bar{x} = \frac{x_1 n_1 + \cdots + x_k n_k}{n}.$$

Ý nghĩa

- Số trung bình mẫu được dùng làm đại diện cho các số liệu của mẫu. Nó là một số đặc trưng quan trọng của mẫu số liệu.
- Ví dụ: nếu biết điểm trung bình môn Toán của lớp A là 6.5, của lớp C là 7.5 thì ta có thể cho rằng sinh viên lớp C đạt điểm cao hơn sinh viên lớp A.

Ý nghĩa

- Số trung bình mẫu được dùng làm đại diện cho các số liệu của mẫu. Nó là một số đặc trưng quan trọng của mẫu số liệu.
- Ví dụ: nếu biết điểm trung bình môn Toán của lớp A là 6.5, của lớp C là 7.5 thì ta có thể cho rằng sinh viên lớp C đạt điểm cao hơn sinh viên lớp A.
- Khi các số liệu trong mẫu có sự chênh lệch rất lớn đối với nhau thì số trung bình mẫu chưa đại diện tốt cho các số liệu trong mẫu. Khi đó, ta dùng một số đặc trưng khác thích hợp hơn, được gọi là *trung vị mẫu*.

Trung vị

Định nghĩa 4 (Trung vị - Median)

Trung vị của một mẫu số liệu, ký hiệu bởi M_e, là một số có tính chất: số các giá trị của mẫu không vượt quá M_e thì bằng số các giá trị của mẫu không nhỏ hơn M_e.

Giả sử mẫu số liệu $\{x_1, \ldots, x_n\}$ là một mẫu dữ liệu ta thu thập được. Ta sắp xếp mẫu dữ liệu theo thứ tự không giảm:

$$x_1^* \leq \cdots \leq x_n^*.$$

- Nếu $n = 2k$, thì $M_e = \frac{1}{2}(x_k^* + x_{k+1}^*)$.
- Nếu $n = 2k + 1$, thì $M_e = x_{k+1}^*$.

Ví dụ

Ví dụ 3

Để xác định hiệu năng của một bộ vi xử lý, người ta đo thời gian chạy (theo giây) của CPU với $n = 30$ nhiệm vụ ngẫu nhiên, kết quả cho bởi

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

Ta ước tính thời gian chạy trung bình μ bởi

$$\overline{X} = \frac{70 + 36 + \cdots + 56 + 19}{30} = 48.2333.$$

Chúng ta có thể kết luận, thời gian trung bình để CPU xử lý một nhiệm vụ bất kỳ là $48.2333s$.

Ví dụ (tiếp)

Để tìm trung vị, chúng ta thực hiện

- Sắp xếp các số liệu trên theo thứ tự tăng dần, *i.e.*,

9	15	19	22	24	25	30	34	35	35
36	36	37	38	42	43	46	48	54	55
56	56	59	62	69	70	82	82	89	139

Ví dụ (tiếp)

Để tìm trung vị, chúng ta thực hiện

- Sắp xếp các số liệu trên theo thứ tự tăng dần, *i.e.*,

9	15	19	22	24	25	30	34	35	35
36	36	37	38	42	43	46	48	54	55
56	56	59	62	69	70	82	82	89	139

- Do $n = 30$ là một số chẵn, ta tìm giá trị nhỏ thứ $n/2 = 15$ và giá trị nhỏ thứ $n/2 + 1 = 16$. Hai giá trị này là 42 và 43. Từ đó có kết luận $M_e = \frac{42+43}{2} = 42.5$.

Mốt

Định nghĩa 5 (Mốt - Mode)

Mốt là giá trị của mẫu số liệu có tần số xuất hiện lớn nhất. Mốt được ký hiệu là M_o.

Tứ phân vị

Định nghĩa 6 (Tứ phân vị - Quartile)

Tứ phân vị là các giá trị chia mẫu số liệu đã sắp xếp theo thứ tự từ nhỏ đến lớn thành bốn phần, mỗi phần đều chứa 25% giá trị. Các giá trị này được ký hiệu là Q_1, Q_2, Q_3.

- *Q_1 được gọi là tứ phân vị thứ nhất hay tứ phân vị dưới.*
- *Q_2 chính là trung vị.*
- *Q_3 được gọi là tứ phân vị thứ ba hay tứ phân vị trên.*

Các bước xác định Tứ phân vị

Giả sử mẫu số liệu $\{x_1, \ldots, x_n\}$ là một mẫu dữ liệu thu thập được. Ta sắp xếp mẫu dữ liệu theo thứ tự không giảm:

$$x_1^* \leq \cdots \leq x_n^*.$$

- Bước 1: Tìm trung vị của mẫu số liệu, giá trị này là Q_2.
- Bước 2: Tìm trung vị của nửa số liệu bên trái Q_2 (không bao gồm Q_2 nếu n lẻ). Giá trị này là Q_1.
- Bước 3: Tìm trung vị của nửa số liệu bên phải Q_2 (không bao gồm Q_2 nếu n lẻ). Giá trị này là Q_3.

Các bước xác định Tứ phân vị

Giả sử mẫu số liệu $\{x_1, \ldots, x_n\}$ là một mẫu dữ liệu thu thập được. Ta sắp xếp mẫu dữ liệu theo thứ tự không giảm:

$$x_1^* \leq \cdots \leq x_n^*.$$

- Bước 1: Tìm trung vị của mẫu số liệu, giá trị này là Q_2.
- Bước 2: Tìm trung vị của nửa số liệu bên trái Q_2 (không bao gồm Q_2 nếu n lẻ). Giá trị này là Q_1.
- Bước 3: Tìm trung vị của nửa số liệu bên phải Q_2 (không bao gồm Q_2 nếu n lẻ). Giá trị này là Q_3.

Nhận xét

Tìm các tứ phân vị thực chất là lặp lại việc tìm trung vị ba lần.

Phương sai & Độ lệch mẫu

(a) Phương sai

- *Phương sai tổng thể*

$$S_n^2(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

- *Phương sai mẫu*

$$s_n^2(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

(b) Độ lệch mẫu

$$\sigma_X = \sqrt{s_n^2(X)}.$$

Với bảng tần số

- Nếu mẫu dữ liệu được biểu diễn dưới dạng bảng tần số

Giá trị	x_1	$\cdots$	x_k
Tần số	n_1	$\cdots$	n_k

thì

$$s_n^2(X) = \frac{1}{n-1} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2.$$

Với bảng tần số

- Nếu mẫu dữ liệu được biểu diễn dưới dạng bảng tần số

Giá trị	x_1	$\cdots$	x_k
Tần số	n_1	$\cdots$	n_k

thì

$$s_n^2(X) = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2.$$

- Nếu mẫu dữ liệu được biểu diễn dưới dạng một bảng tần số theo nhóm:

Giá trị	$a_1 - a_2$	$a_2 - a_3$	$\cdots$	$a_k - a_{k+1}$
Tần số	n_1	n_2	$\cdots$	n_k

Gọi $x_i = \frac{a_i + a_{i+1}}{2}$ là giá trị đại diện cho khoảng (a_i, a_{i+1}). Khi đó,

$$s_n^2(X) = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2.$$

Ý nghĩa của Phương sai và Độ lệch mẫu

- Phương sai là trung bình cộng của bình phương khoảng cách từ mỗi số liệu tới số trung bình mẫu.
- Phương sai và độ lệch mẫu đo mức độ phân tán của các số liệu trong mẫu quanh số trung bình mẫu.
- Phương sai và độ lệch mẫu càng lớn thì độ phân tán càng lớn.

Các số đặc trưng khác

- *Phạm vi mẫu* (Range) là hiệu số $x_n^* - x_1^*$.
- *Khoảng tứ phân vị* (Interquartile range) là hiệu số $IQR = Q_3 - Q_1$.

Ví dụ minh họa

Ví dụ 4

Theo dõi điểm Toán của 10 học sinh lớp A, kết quả thu được như sau:

$$10 \quad 9 \quad 5 \quad 6 \quad 1 \quad 5 \quad 7 \quad 9 \quad 5 \quad 6$$

Các số đặc trưng của mẫu dữ liệu mà ta thu được là:

- *Cỡ mẫu:* $n = 10$
- *Trung bình mẫu:* $\bar{x} = 6.3$
- *Mốt:* $M_o = 5$
- *Trung vị mẫu:* $M_e = 6$
- *Tứ phân vị:* $Q_1 = 5; Q_2 = 6; Q_3 = 9.$

Áp dụng

Bài tập

Trong một bài kiểm tra, sinh viên phải trả lời 40 câu hỏi trắc nghiệm. Kết quả được thống kê ở bảng sau:

Số câu đúng	$26-30$	$31-35$	$36-40$
Số sinh viên	12	24	4

Hãy tìm $\bar{x}, s_{40}^2(X)$ *của mẫu số liệu trên.*

Phát hiện Dữ liệu bất thường

Phát hiện Dữ liệu bất thường

Định nghĩa 7

Giá trị x_i được gọi là giá trị bất thường (outlier) của mẫu dữ liệu nếu:

$$x_i \notin (Q_1 - 1.5\Delta_Q, Q_3 + 1.5\Delta_Q),$$

trong đó Q_1 là tứ phân vị thứ nhất, Q_3 là tứ phân vị thứ ba và $\Delta_Q = Q_3 - Q_1$ là khoảng tứ phân vị.

Các bước xác định giá trị bất thường

- **Bước 1.** Tính khoảng tứ phân vị $\Delta_Q = Q_3 - Q_1$.
- **Bước 2.** Xác định các cận dưới $Q_1 - 1.5\Delta_Q$ và cận trên $Q_3 + 1.5\Delta_Q$.
- **Bước 3.** Xác định tất cả các giá trị $x \notin (Q_1 - 1.5\Delta_Q, Q_3 + 1.5\Delta_Q)$. Đây là các giá trị bất thường.

Ví dụ

Cho mẫu số liệu sau

43 37 50 51 58 105 52 45 45 10

(*i*) Tính Q_1, Q_3, Δ_Q.

(*ii*) Xác định tất cả các giá trị bất thường.

Giới thiệu về Phần mềm R

Phần mềm R

- Phát minh bởi Ross Ihaka và Robert Gentleman năm 1993.

Hình 3: Ross Ihaka và Robert Gentleman (Nguồn: Wiki).

Phần mềm R

- Phát minh bởi Ross Ihaka và Robert Gentleman năm 1993.

Hình 3: Ross Ihaka và Robert Gentleman (Nguồn: Wiki).

- Phần mềm có phiên bản miễn phí.
  - Phiên bản cho Windows: https://cran.r-project.org/bin/windows/base/
  - Phiên bản cho Mac: https://cran.r-project.org/bin/macosx/

Nhập dữ liệu

Để nhập dữ liệu, ta dùng hàm c. Giả sử chúng ta muốn nhập số lỗi đánh máy trên một số trang của slides này

2 3 0 3 1 0 0 1

Để nhập dữ liệu này vào R, chúng ta thực hiện câu lệnh:

```
> typos = c(2, 3, 0, 3, 1, 0, 0, 1)
> typos
[1] 2 3 0 3 1 0 0 1
```

trong đó, dấu "=" thể hiện phép gán mẫu dữ liệu này cho biến typos.

Một số Hàm thống kê

- Để tìm giá trị trung bình, ta dùng lệnh

```
> mean(typos)
[1] 1.25
```

Một số Hàm thống kê

- Để tìm giá trị trung bình, ta dùng lệnh

```
> mean(typos)
[1] 1.25
```

- Để tìm trung vị, phương sai, độ lệch chuẩn, ta dùng các lệnh

```
> median(typos)
[1] 1
> var(typos)
[1] 1.642857
> sd(typos)
[1] 1.28174
```

Một số Hàm thống kê (tiếp)

- Để tìm các tứ phân vị và khoảng tứ phân vị, ta dùng các câu lệnh

```
> quantile(typos)
0%   25%  50%  75%  100%
0.00 0.00 1.00 2.25 3.00
> IQR(typos)
[1] 2.25
```

Cài đặt một Gói lệnh

Để cài đặt một gói lệnh, ta dùng câu lệnh

```
> install.packages("packages_name")
```

Cài đặt một Gói lệnh

Để cài đặt một gói lệnh, ta dùng câu lệnh

```
> install.packages("packages_name")
```

E.g., câu lệnh sau đây cho phép cài đặt package DescTools

```
> install.packages("DescTools")
```

Tìm Mốt

Để tìm Mốt của mẫu số liệu ta dùng package DescTools

```
> library(DescTools)
> Mode(typos)
[1] 0
attr(,"freq")
[1] 3
```

Đồ thị cột

Để vẽ một đồ thị cột (bar chart) đơn giản, ta dùng cú pháp

```
barplot(V, xlab, ylab, main, names.arg, col)
```

Đồ thị cột

Để vẽ một đồ thị cột (bar chart) đơn giản, ta dùng cú pháp

```
barplot(V, xlab, ylab, main, names.arg, col)
```

trong đó,

- `V` là véc-tơ hay ma trận chứa các giá trị số được thể hiện trong đồ thị.
- `xlab` là ký hiệu cho trục x.
- `ylab` là ký hiệu cho trục y.
- `main` là tên của đồ thị.
- `names.arg` là véc-tơ gồm tên của các cột trong đồ thị.
- `col` là véc-tơ gồm các màu cho các cột. Nếu véc-tơ này chỉ có một tọa độ, tất cả các cột sẽ có cùng một màu.

Đồ thị cột

Để vẽ đồ thị cột cho Ví dụ 1, ta có thể thực hiện như sau

```
> d = c(67008, 47288, 15850, 4951)
> colors = c("pink","green","yellow","red")
> names = c("Hoa sen","Cây tre","Hoa mai","Đề xuất khác")
> vote  =  "Số lượt bình chọn Quốc hoa"
> barplot(d, main = vote, names.arg = names, col = colors)
```

Đồ thị quạt

Để vẽ đồ thị quạt (pie chart) cho Ví dụ 1, ta có thể thực hiện như sau

```
> d = c(67008, 47288, 15850, 4951)
> colors = c("pink","green","yellow","red")
> names = c("Hoa sen","Cây tre","Hoa mai","Đề xuất khác")
> vote  =  "Số lượt bình chọn Quốc hoa"
> pie(d, labels = names, main = vote, col = colors)
```

Đồ thị quạt

Để vẽ đồ thị quạt (pie chart) cho Ví dụ 1, ta có thể thực hiện như sau

```
> d = c(67008, 47288, 15850, 4951)
> colors = c("pink","green","yellow","red")
> names = c("Hoa sen","Cây tre","Hoa mai","Đề xuất khác")
> vote  =  "Số lượt bình chọn Quốc hoa"
> pie(d, labels = names, main = vote, col = colors)
```

Để thêm %, ta có thể thực hiện các câu lệnh sau

```
> pct = round(d/sum(d)*100)
> names = paste(names, "\n", pct, "%", sept = "")
> pie(d, labels = names, main = vote, col = colors)
```

End of Chapter 3