Xác suất Thống kê ứng dụng trong Kinh tế Xã hội

Nguyễn Thị Nhung

Bộ môn Toán - Đại học Thăng Long

Ngày 22 tháng 8 năm 2013

Chương III

Tóm tắt và trình bày dữ liệu bằng bảng và đồ thị

Chương III

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Chương III

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Câu hỏi tình huống chương 3

Câu hỏi tình huống

Tình huống 1: Bạn là một thành viên trong văn phòng Đoàn của trường Đại học Thăng Long, bạn nhận được file về dữ liệu điểm thi tuyển sinh vào trường Đại học Thăng Long năm 2008. Làm thế nào để bạn tóm tắt được kết quả điểm thi để gửi lên trang web trường để đưa được thông tin chung cho về kết quả thi cho Ban Giám hiệu cũng như cho các thí sinh thi tuyển vào trường ĐH Thăng Long.

Câu hỏi tình huống chương 3

Câu hỏi tình huống

Tình huống 1: Bạn là một thành viên trong văn phòng Đoàn của trường Đại học Thăng Long, bạn nhận được file về dữ liệu điểm thi tuyển sinh vào trường Đại học Thăng Long năm 2008. Làm thế nào để bạn tóm tắt được kết quả điểm thi để gửi lên trang web trường để đưa được thông tin chung cho về kết quả thi cho Ban Giám hiệu cũng như cho các thí sinh thi tuyển vào trường ĐH Thăng Long.

Câu hỏi tình huống

Tình huống 2: Giả sử bạn là nhân viên của một tạp chí, Tổng biên tập nhận được bảng số liệu về tỉ lệ thời gian dùng Internet ở Việt Nam năm 2009 và yêu cầu bạn điều tra viết bài và đưa thông tin này lên báo. Bạn sẽ trình bày bảng số liệu này lên báo thế nào để bạn đọc tiếp nhận thông tin tốt.

Địa điểm dùng Internet	Tỉ lệ thời gian sử dụng
Internet cafes và WIFI Hotspot	*19%*
Điện thoại di động	*1%*
Tại nhà của bạn bè	*2%*
Công sở	*20%*
Tại nhà	*55%*
Trường và Thư viện công cộng	*3%*

Nội dung chính được giới thiệu trong chương

- Giới thiệu cách tóm tắt một tập dữ liệu định tính bằng bảng tần số.
- Giới thiệu cách tóm tắt một tập dữ liệu định lượng bằng bảng tần số.
- Giới thiệu cách phân một tập dữ liệu nhiều giá trị biểu hiện thành các tổ.
- Giới thiệu cách tóm tắt một tập dữ liệu định lượng bằng các kiểu biểu đồ: biểu đồ phân phối tần số, đa giác tần số, biểu đồ thân và lá.
- Giới thiệu cách tóm tắt một tập dữ liệu định tính bằng các kiểu biểu đồ: biểu đồ hình tròn, biểu đồ thanh, biểu đồ pareto.

Yêu cầu đối với sinh viên

- Biết cách lập bảng tần số cho tập dữ liệu định tính.
- Biết cách lập bảng tần số cho tập dữ liệu định lượng.
- Biết được khi nào cần phải phân tổ dữ liệu và các điều kiện khi tiến hành phân tổ dữ liệu.
- Biết cách minh họa phân phối của một tập dữ liệu định lượng bằng biểu đồ phân phối tần số, đa giác tần số, biểu đồ thân và lá. Phân biệt được với tập dữ liệu nào thì nên dùng biểu đồ nào cho phù hợp.
- Biết cách minh họa phân phối của một tập dữ liệu định tính bằng biểu đồ tròn, biểu đồ thanh, biểu đồ Pareto. Phân biệt được với tập dữ liệu nào thì nên dùng biểu đồ nào cho phù hợp.

Nội dung trình bày

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Đại học ThăngLong
Khương Trung, Thanh Xuân, Hà Nội
Tel.: 04.8587346 , Fax.: 04.5636775

Bảng tần số

Khái niệm

Bảng tần số là một bảng tổng hợp các biểu hiện có thể có của đặc điểm quan sát, hoặc các khoảng giá trị mà trong đó dữ liệu (định lượng) có thể rơi vào và số quan sát (tần số), tỉ lệ phần trăm chiếm (tần suất) tương ứng với mỗi biểu hiện hoặc khoảng giá trị dữ liệu.

Bảng tần số thường gồm ba cột:

- Cột đầu tiên mô tả các biểu hiện hoặc các giá trị hay khoảng giá trị được xác định cho dữ liệu;
- Cột thứ hai mô tả tần số tương ứng với các biểu hiện hay giá trị;
- Cột thứ ba mô tả các tần suất tương ứng.

Chú ý: Để thêm thông tin trong bảng tần số ta có thể thêm cột tần số tích lũy và tần suất tích lũy của các biểu hiện hoặc giá trị.

Bảng tần số

Khái niệm

Bảng tần số là một bảng tổng hợp các biểu hiện có thể có của đặc điểm quan sát, hoặc các khoảng giá trị mà trong đó dữ liệu (định lượng) có thể rơi vào và số quan sát (tần số), tỉ lệ phần trăm chiếm (tần suất) tương ứng với mỗi biểu hiện hoặc khoảng giá trị dữ liệu.

Bảng tần số thường gồm ba cột:

- Cột đầu tiên mô tả các biểu hiện hoặc các giá trị hay khoảng giá trị được xác định cho dữ liệu;
- Cột thứ hai mô tả tần số tương ứng với các biểu hiện hay giá trị;
- Cột thứ ba mô tả các tần suất tương ứng.

Chú ý: Để thêm thông tin trong bảng tần số ta có thể thêm cột tần số tích lũy và tần suất tích lũy của các biểu hiện hoặc giá trị.

Bảng tần số

Khái niệm

Bảng tần số là một bảng tổng hợp các biểu hiện có thể có của đặc điểm quan sát, hoặc các khoảng giá trị mà trong đó dữ liệu (định lượng) có thể rơi vào và số quan sát (tần số), tỉ lệ phần trăm chiếm (tần suất) tương ứng với mỗi biểu hiện hoặc khoảng giá trị dữ liệu.

Bảng tần số thường gồm ba cột:

- Cột đầu tiên mô tả các biểu hiện hoặc các giá trị hay khoảng giá trị được xác định cho dữ liệu;
- Cột thứ hai mô tả tần số tương ứng với các biểu hiện hay giá trị;
- Cột thứ ba mô tả các tần suất tương ứng.

Chú ý: Để thêm thông tin trong bảng tần số ta có thể thêm cột tần số tích lũy và tần suất tích lũy của các biểu hiện hoặc giá trị.

Bảng tần số

Khái niệm

Bảng tần số là một bảng tổng hợp các biểu hiện có thể có của đặc điểm quan sát, hoặc các khoảng giá trị mà trong đó dữ liệu (định lượng) có thể rơi vào và số quan sát (tần số), tỉ lệ phần trăm chiếm (tần suất) tương ứng với mỗi biểu hiện hoặc khoảng giá trị dữ liệu.

Bảng tần số thường gồm ba cột:

- Cột đầu tiên mô tả các biểu hiện hoặc các giá trị hay khoảng giá trị được xác định cho dữ liệu;
- Cột thứ hai mô tả tần số tương ứng với các biểu hiện hay giá trị;
- Cột thứ ba mô tả các tần suất tương ứng.

Chú ý: Để thêm thông tin trong bảng tần số ta có thể thêm cột tần số tích lũy và tần suất tích lũy của các biểu hiện hoặc giá trị.

Bảng tần số

Khái niệm

Bảng tần số là một bảng tổng hợp các biểu hiện có thể có của đặc điểm quan sát, hoặc các khoảng giá trị mà trong đó dữ liệu (định lượng) có thể rơi vào và số quan sát (tần số), tỉ lệ phần trăm chiếm (tần suất) tương ứng với mỗi biểu hiện hoặc khoảng giá trị dữ liệu.

Bảng tần số thường gồm ba cột:

- Cột đầu tiên mô tả các biểu hiện hoặc các giá trị hay khoảng giá trị được xác định cho dữ liệu;
- Cột thứ hai mô tả tần số tương ứng với các biểu hiện hay giá trị;
- Cột thứ ba mô tả các tần suất tương ứng.

Chú ý: Để thêm thông tin trong bảng tần số ta có thể thêm cột tần số tích lũy và tần suất tích lũy của các biểu hiện hoặc giá trị.

Câu hỏi tình huống

Hãy xét xem có sự khác nhau khi lập bảng tần số trong các trường hợp sau:

- *Tính số sinh viên thi vào các khối (A, B, C) trong trường;*
- *Tính số lượng tờ báo đọc trong một tuần của thành viên trong nhóm điều tra gồm 200 người;*
- *Tính số thí sinh được từng loại điểm trong kì thi tuyển sinh Toán khối A vào trường Thăng Long năm 2008.*

Nội dung trình bày

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Đại học ThăngLong
Khương Trung, Thanh Xuân, Hà Nội
Tel.: 04.8587346 , Fax.: 04.5636775

Lập bảng tần số cho dữ liệu định tính

Bảng tần số của dữ liệu định tính gồm ba cột với các thông tin sau:

- Cột đầu tiên liệt kê các biểu hiện có thể có của đối tượng theo các đặc điểm cần nghiên cứu, chẳng hạn ta có k biểu hiện;
- Cột thứ hai là tần số mỗi biểu hiện vừa liệt kê, tần số của biểu hiện thứ i kí hiệu là f_i. Nếu tổng số quan sát của tập dữ liệu là n thì ta có $n = \sum_{i=1}^{k} f_i$;
- Cột thứ ba mô tả các tần suất tương ứng với mỗi biểu hiện. Với các giả sử như trên, tần suất của biểu hiện thứ i sẽ là $\frac{f_i}{n} \times 100\%$.

Lập bảng tần số cho dữ liệu định tính

Cột tần số tích lũy và tần suất tích lũy được tính từ cột tần số và tần suất như sau:

- Tần số tích lũy của biểu hiện thứ i là $\sum_{s=1}^{i} f_s$;
- Tần suất tích lũy của biểu hiện thứ i là $\sum_{s=1}^{i} \frac{f_s}{n} \times 100\%$.

Ví dụ về bảng tần số của dữ liệu định tính

- Bảng tần số về số thí sinh dự thi các khối A, B, D vào trường Đại học Thăng Long năm 2008

Số thí sinh
Khối A
Khối B
Khối D
Tổng

Ví dụ về bảng tần số của dữ liệu định tính

- Bảng tần số về số thí sinh dự thi các khối A, B, D vào trường Đại học Thăng Long năm 2008

Số thí sinh	Tần số
Khối A	2974
Khối B	666
Khối D	1573
Tổng	5213

Ví dụ về bảng tần số của dữ liệu định tính

- Bảng tần số về số thí sinh dự thi các khối A, B, D vào trường Đại học Thăng Long năm 2008

Số thí sinh	Tần số	Tần suất (%)
Khối A	2974	57.05
Khối B	666	12.78
Khối D	1573	30.17
Tổng	5213	100

Ví dụ về bảng tần số của dữ liệu định tính

Bảng tần số về số thí sinh dự thi các khối A, B, D vào trường Đại học Thăng Long năm 2008 thêm cột tần số tích lũy và tần suất tích lũy:

Số thí sinh
Khối A
Khối B
Khối D
Tổng

Ví dụ về bảng tần số của dữ liệu định tính

Bảng tần số về số thí sinh dự thi các khối A, B, D vào trường Đại học Thăng Long năm 2008 thêm cột tần số tích lũy và tần suất tích lũy:

Số thí sinh	Tần số
Khối A	2974
Khối B	666
Khối D	1573
Tổng	5213

Ví dụ về bảng tần số của dữ liệu định tính

Bảng tần số về số thí sinh dự thi các khối A, B, D vào trường Đại học Thăng Long năm 2008 thêm cột tần số tích lũy và tần suất tích lũy:

Số thí sinh	Tần số	Tần số tích lũy
Khối A	2974	2974
Khối B	666	3640
Khối D	1573	5213
Tổng	5213	

Ví dụ về bảng tần số của dữ liệu định tính

Bảng tần số về số thí sinh dự thi các khối A, B, D vào trường Đại học Thăng Long năm 2008 thêm cột tần số tích lũy và tần suất tích lũy:

Số thí sinh	Tần số	Tần số tích lũy	Tần suất(%)
Khối A	2974	2974	57.05 %
Khối B	666	3640	12.78 %
Khối D	1573	5213	30.17 %
Tổng	5213		100 %

Ví dụ về bảng tần số của dữ liệu định tính

Bảng tần số về số thí sinh dự thi các khối A, B, D vào trường Đại học Thăng Long năm 2008 thêm cột tần số tích lũy và tần suất tích lũy:

Số thí sinh	Tần số	Tần số tích lũy	Tần suất(%)	Tần suất tích lũy(%)
Khối A	2974	2974	57.05 %	57.05 %
Khối B	666	3640	12.78 %	69.83 %
Khối D	1573	5213	30.17 %	100 %
Tổng	5213		100 %	

Lệnh trong R dùng để tính tần số, tần suất, tần số tích lũy và tần suất tích lũy

Cho `x` là một tập dữ liệu. Trong R ta có thể tính tần số, tần suất, tần số tích lũy, tần suất tích lũy của các phần tử của `x` bằng các hàm sau:

- `table(x)`: cho tần số của các phần tử trong `x`;
- `prop.table(table(x))`: cho tần suất của các phần tử trong `x`;
- `cumsum(table(x))`: cho tần số tích lũy của các phần tử trong `x`;
- `cumsum(prop.table(table(x)))`: cho tần suất tích lũy của các phần tử trong `x`;

Thực hiện tính tần số, tần suất, tần số tích lũy và tần suất tích lũy trong R

- Nhập dữ liệu về số học sinh từng khối:

```
> SoHocSinh = c(2974, 666, 1573)
```

- Tính tần suất của số học sinh của từng khối thi:

```
> prop.table(SoHocSinh)
[1] 0.5704968  0.1277575  0.3017456
```

- Tính tần số tích lũy của số học sinh cho những khối thi:

```
> cumsum(SoHocSinh)
[1] 2974  3640  5213
```

- Tính tần suất tích lũy của số học sinh cho những khối thi:

```
> cumsum(prop.table(SoHocSinh))
[1] 0.5704968  0.6982544  1.0000000
```

Nội dung trình bày

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Đại học ThăngLong
Khương Trung, Thanh Xuân, Hà Nội
Tel.: 04.8587346 , Fax.: 04.5636775

Lập bảng tần số cho dữ liệu định lượng

- Nếu dữ liệu định lượng mà đặc điểm quan tâm có ít biểu hiện thì cách lập bảng tần số giống như cách lập bảng tần số cho dữ liệu định tính, xem mỗi giá trị như là một biểu hiện.
- Nếu dữ liệu định lượng mà đặc điểm quan tâm có quá nhiều biểu hiện thì việc liệt kê từng biểu hiện làm bảng tần số dài, mất đi tác dụng tóm lược thông tin. Trong trường hợp này ta phải tiến hành phân tổ dữ liệu và lập bảng tần số cho dữ liệu đã được phân tổ.

Lập bảng tần số cho dữ liệu định lượng

- Nếu dữ liệu định lượng mà đặc điểm quan tâm có ít biểu hiện thì cách lập bảng tần số giống như cách lập bảng tần số cho dữ liệu định tính, xem mỗi giá trị như là một biểu hiện.
- Nếu dữ liệu định lượng mà đặc điểm quan tâm có quá nhiều biểu hiện thì việc liệt kê từng biểu hiện làm bảng tần số dài, mất đi tác dụng tóm lược thông tin. Trong trường hợp này ta phải tiến hành phân tổ dữ liệu và lập bảng tần số cho dữ liệu đã được phân tổ.

Ví dụ về bảng tần số của dữ liệu định lượng có ít biểu hiện

- Bảng tần số về số tờ báo đọc trong một tuần khi khảo sát 200 người dân tại một quận ở Hà Nội:

Số tờ báo
0
1
2
3
4
5

Ví dụ về bảng tần số của dữ liệu định lượng có ít biểu hiện

- Bảng tần số về số tờ báo đọc trong một tuần khi khảo sát 200 người dân tại một quận ở Hà Nội:

Số tờ báo	Tần số
0	60
1	40
2	30
3	40
4	20
5	10

Ví dụ về bảng tần số của dữ liệu định lượng có ít biểu hiện

- Bảng tần số về số tờ báo đọc trong một tuần khi khảo sát 200 người dân tại một quận ở Hà Nội:

Số tờ báo	Tần số	Tần suất (%)
0	60	30
1	40	20
2	30	15
3	40	20
4	20	10
5	10	5

Ví dụ về bảng tần số cho dữ liệu định lượng có nhiều biểu hiện

Bảng tần số về điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008:

Điểm
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
3.25
3.50
3.75
4.00

Ví dụ về bảng tần số cho dữ liệu định lượng có nhiều biểu hiện

Bảng tần số về điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008:

Điểm
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
3.25
3.50
3.75
4.00

Điểm
4.25
4.50
4.75
5.00
5.25
5.50
5.75
6.00
6.25
6.50
6.75
7.00
7.25
7.50
7.75
8.00
8.75

Ví dụ về bảng tần số cho dữ liệu định lượng có nhiều biểu hiện

Bảng tần số về điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008:

Điểm	Tần số
0.00	214
0.25	101
0.50	166
0.75	181
1.00	358
1.25	119
1.50	178
1.75	130
2.00	247
2.25	140
2.50	155
2.75	152
3.00	143
3.25	109
3.50	107
3.75	89
4.00	89

Điểm
4.25
4.50
4.75
5.00
5.25
5.50
5.75
6.00
6.25
6.50
6.75
7.00
7.25
7.50
7.75
8.00
8.75

Ví dụ về bảng tần số cho dữ liệu định lượng có nhiều biểu hiện

Bảng tần số về điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008:

Điểm	Tần số
0.00	214
0.25	101
0.50	166
0.75	181
1.00	358
1.25	119
1.50	178
1.75	130
2.00	247
2.25	140
2.50	155
2.75	152
3.00	143
3.25	109
3.50	107
3.75	89
4.00	89

Điểm	Tần số
4.25	61
4.50	63
4.75	42
5.00	43
5.25	26
5.50	18
5.75	13
6.00	9
6.25	6
6.50	4
6.75	1
7.00	5
7.25	1
7.50	1
7.75	1
8.00	1
8.75	1

Ví dụ về bảng tần số cho dữ liệu định lượng có nhiều biểu hiện

Bảng tần số về điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008:

Điểm	Tần số	Tần suất	Điểm	Tần số
0.00	214	7.2	4.25	61
0.25	101	3.4	4.50	63
0.50	166	5.6	4.75	42
0.75	181	6.0	5.00	43
1.00	358	12.0	5.25	26
1.25	119	4.0	5.50	18
1.50	178	6.0	5.75	13
1.75	130	4.4	6.00	9
2.00	247	8.3	6.25	6
2.25	140	4.7	6.50	4
2.50	155	5.2	6.75	1
2.75	152	5.1	7.00	5
3.00	143	4.8	7.25	1
3.25	109	3.7	7.50	1
3.50	107	3.6	7.75	1
3.75	89	3.0	8.00	1
4.00	89	3.0	8.75	1

Ví dụ về bảng tần số cho dữ liệu định lượng có nhiều biểu hiện

Bảng tần số về điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008:

Điểm	Tần số	Tần suất	Điểm	Tần số	Tần suất
0.00	214	7.2	4.25	61	2.0
0.25	101	3.4	4.50	63	2.1
0.50	166	5.6	4.75	42	1.4
0.75	181	6.0	5.00	43	1.5
1.00	358	12.0	5.25	26	0.9
1.25	119	4.0	5.50	18	0.6
1.50	178	6.0	5.75	13	0.4
1.75	130	4.4	6.00	9	0.3
2.00	247	8.3	6.25	6	0.2
2.25	140	4.7	6.50	4	0.1
2.50	155	5.2	6.75	1	0.03
2.75	152	5.1	7.00	5	0.2
3.00	143	4.8	7.25	1	0.03
3.25	109	3.7	7.50	1	0.03
3.50	107	3.6	7.75	1	0.03
3.75	89	3.0	8.00	1	0.03
4.00	89	3.0	8.75	1	0.03

Phân tổ dữ liệu

- Một số khái niệm trong phân tổ dữ liệu:
  - Trong mỗi một tổ, giới hạn dưới là trị số nhỏ nhất của tổ, giới hạn trên là trị số lớn nhất của tổ.
  - Khoảng cách của một tổ là hiệu giữa giới hạn trên và giới hạn dưới.
  - Phân tổ đều là cách phân chia sao cho tất cả các tổ trong bảng tần số đều có khoảng cách bằng nhau.
  - Phân tổ không đều là cách phân chia sao cho ít nhất hai tổ trong bảng tần số có khoảng cách không bằng nhau.
- Một số điều kiện khi tiến hành phân tổ:
  - Các tổ không được giao nhau, để đảm bảo một quan sát bất kì chỉ thuộc vào một tổ;
  - Các tổ được phân chia phải bao quát hết tất cả các giá trị của tập dữ liệu;
  - Mọi tổ phải chứa ít nhất một quan sát.

Phân tổ dữ liệu

- Một số khái niệm trong phân tổ dữ liệu:
  - Trong mỗi một tổ, giới hạn dưới là trị số nhỏ nhất của tổ, giới hạn trên là trị số lớn nhất của tổ.
  - Khoảng cách của một tổ là hiệu giữa giới hạn trên và giới hạn dưới.
  - Phân tổ đều là cách phân chia sao cho tất cả các tổ trong bảng tần số đều có khoảng cách bằng nhau.
  - Phân tổ không đều là cách phân chia sao cho ít nhất hai tổ trong bảng tần số có khoảng cách không bằng nhau.
- Một số điều kiện khi tiến hành phân tổ:
  - Các tổ không được giao nhau, để đảm bảo một quan sát bất kì chỉ thuộc vào một tổ;
  - Các tổ được phân chia phải bao quát hết tất cả các giá trị của tập dữ liệu;
  - Mọi tổ phải chứa ít nhất một quan sát.

Phân tổ dữ liệu

- Một số khái niệm trong phân tổ dữ liệu:
  - Trong mỗi một tổ, giới hạn dưới là trị số nhỏ nhất của tổ, giới hạn trên là trị số lớn nhất của tổ.
  - Khoảng cách của một tổ là hiệu giữa giới hạn trên và giới hạn dưới.
  - Phân tổ đều là cách phân chia sao cho tất cả các tổ trong bảng tần số đều có khoảng cách bằng nhau.
  - Phân tổ không đều là cách phân chia sao cho ít nhất hai tổ trong bảng tần số có khoảng cách không bằng nhau.
- Một số điều kiện khi tiến hành phân tổ:
  - Các tổ không được giao nhau, để đảm bảo một quan sát bất kì chỉ thuộc vào một tổ;
  - Các tổ được phân chia phải bao quát hết tất cả các giá trị của tập dữ liệu;
  - Mọi tổ phải chứa ít nhất một quan sát.

Câu hỏi

Hãy nêu các bước khi tiến hành phân tổ một tập dữ liệu.

Thủ tục phân tổ đều

Các bước của thủ tục phân tổ đều:

- Xác định số tổ cần chia k.
- Xác định khoảng cách tổ h: $h = \dfrac{x_{\max} - x_{\min}}{k}$, trong đó $x_{\max}$ và $x_{\min}$ tương ứng là giá trị nhỏ nhất và giá trị lớn nhất của tập dữ liệu.
- Xác định giới hạn dưới và giới hạn trên của các tổ: giới hạn giữa các tổ thỏa mãn các điều kiện:
  - Tổ đầu tiên phải chứa giá trị nhỏ nhất $x_{\min}$;
  - Tổ cuối cùng phải chứa giá trị lớn nhất $x_{\max}$;
  - Giới hạn trên của tổ trước phải trùng với giới hạn dưới của tổ liền sau.

Thủ tục phân tổ đều

Các bước của thủ tục phân tổ đều:

- Xác định số tổ cần chia k.
- Xác định khoảng cách tổ h: $h = \dfrac{x_{max} - x_{min}}{k}$, trong đó x_{max} và x_{min} tương ứng là giá trị nhỏ nhất và giá trị lớn nhất của tập dữ liệu.
- Xác định giới hạn dưới và giới hạn trên của các tổ: giới hạn giữa các tổ thỏa mãn các điều kiện:
  - Tổ đầu tiên phải chứa giá trị nhỏ nhất x_{min};
  - Tổ cuối cùng phải chứa giá trị lớn nhất x_{max};
  - Giới hạn trên của tổ trước phải trùng với giới hạn dưới của tổ liền sau.

Thủ tục phân tổ đều

Các bước của thủ tục phân tổ đều:

- Xác định số tổ cần chia k.
- Xác định khoảng cách tổ h: $h = \dfrac{x_{\max} - x_{\min}}{k}$, trong đó $x_{\max}$ và $x_{\min}$ tương ứng là giá trị nhỏ nhất và giá trị lớn nhất của tập dữ liệu.
- Xác định giới hạn dưới và giới hạn trên của các tổ: giới hạn giữa các tổ thỏa mãn các điều kiện:
  - Tổ đầu tiên phải chứa giá trị nhỏ nhất $x_{\min}$;
  - Tổ cuối cùng phải chứa giá trị lớn nhất $x_{\max}$;
  - Giới hạn trên của tổ trước phải trùng với giới hạn dưới của tổ liền sau.

Thủ tục phân tổ đều

- Phân chia các quan sát vào các tổ: quan sát có giá trị phù hợp với tổ nào thì thì xếp vào tổ đó, tức là quan sát có giá trị x_i được xếp vào tổ thỏa mãn điều kiện:

$$\text{Giới hạn dưới} \leqslant x_i < \text{Giới hạn trên}$$

hoặc

$$\text{Giới hạn dưới} \leqslant x_i < \text{Giới hạn trên}$$

- Dưới đây là một gợi ý trong khi phân chia thành các tổ:

$$\begin{array}{ll} \text{Tổ 1:} & [x_{\min}, x_{\min} + h) \\ \text{Tổ 2:} & [x_{\min} + h, x_{\min} + 2h) \\ \text{Tổ 3:} & [x_{\min} + 2h, x_{\min} + 3h) \\ \ldots & \ldots \\ \text{Tổ k:} & [x_{\min} + (k-1)h, x_{\min} + kh) \end{array}$$

Thủ tục phân tổ đều

- Phân chia các quan sát vào các tổ: quan sát có giá trị phù hợp với tổ nào thì thì xếp vào tổ đó, tức là quan sát có giá trị x_i được xếp vào tổ thỏa mãn điều kiện:

$$\text{Giới hạn dưới} \leqslant x_i < \text{Giới hạn trên}$$

hoặc

$$\text{Giới hạn dưới} \leqslant x_i < \text{Giới hạn trên}$$

- Dưới đây là một gợi ý trong khi phân chia thành các tổ:

$$\begin{array}{ll} \text{Tổ 1:} & [x_{\min}, x_{\min} + h) \\ \text{Tổ 2:} & [x_{\min} + h, x_{\min} + 2h) \\ \text{Tổ 3:} & [x_{\min} + 2h, x_{\min} + 3h) \\ \ldots & \ldots \\ \text{Tổ k:} & [x_{\min} + (k-1)h, x_{\min} + kh) \end{array}$$

Ví dụ về bảng tần số thu gọn

Ta có thể tóm tắt tập dữ liệu điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008 bằng cách phân thành các tổ như sau:

Khoảng điểm
$[0.0, 1.5]$
$(1.5, 3.0]$
$(3.0, 4.5]$
$(4.5, 6.0]$
$(6.0, 7.5]$
$(7.5, 9.0]$
Tổng

Ví dụ về bảng tần số thu gọn

Ta có thể tóm tắt tập dữ liệu điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008 bằng cách phân thành các tổ như sau:

Khoảng điểm	**Tần số**
$[0.0, 1.5]$	1317
$(1.5, 3.0]$	967
$(3.0, 4.5]$	518
$(4.5, 6.0]$	151
$(6.0, 7.5]$	18
$(7.5, 9.0]$	3
Tổng	2974

Ví dụ về bảng tần số thu gọn

Ta có thể tóm tắt tập dữ liệu điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008 bằng cách phân thành các tổ như sau:

Khoảng điểm	**Tần số**	**Tần suất (%)**
$[0.0, 1.5]$	1317	44.28
$(1.5, 3.0]$	967	32.51
$(3.0, 4.5]$	518	17.42
$(4.5, 6.0]$	151	5.08
$(6.0, 7.5]$	18	0.61
$(7.5, 9.0]$	3	0.10
Tổng	2974	100

Ví dụ về bảng tần số thu gọn

Bảng tần số về điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008 có thêm cột tần số tích lũy và tần suất tích lũy:

Khoảng điểm
$[0.0, 1.5]$
$(1.5, 3.0]$
$(3.0, 4.5]$
$(4.5, 6.0]$
$(6.0, 7.5]$
$(7.5, 9.0]$

Ví dụ về bảng tần số thu gọn

Bảng tần số về điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008 có thêm cột tần số tích lũy và tần suất tích lũy:

Khoảng điểm	Tần số
$[0.0, 1.5]$	1317
$(1.5, 3.0]$	967
$(3.0, 4.5]$	518
$(4.5, 6.0]$	151
$(6.0, 7.5]$	18
$(7.5, 9.0]$	3

Ví dụ về bảng tần số thu gọn

Bảng tần số về điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008 có thêm cột tần số tích lũy và tần suất tích lũy:

Khoảng điểm	Tần số	Tần số tích lũy
$[0.0, 1.5]$	1317	1317
$(1.5, 3.0]$	967	2284
$(3.0, 4.5]$	518	1802
$(4.5, 6.0]$	151	2953
$(6.0, 7.5]$	18	2971
$(7.5, 9.0]$	3	2974

Ví dụ về bảng tần số thu gọn

Bảng tần số về điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008 có thêm cột tần số tích lũy và tần suất tích lũy:

Khoảng điểm	Tần số	Tần số tích lũy	Tần suất(%)
$[0.0, 1.5]$	1317	1317	44.28 (%)
$(1.5, 3.0]$	967	2284	32.51 (%)
$(3.0, 4.5]$	518	1802	17.42 (%)
$(4.5, 6.0]$	151	2953	5.08 (%)
$(6.0, 7.5]$	18	2971	0.61 (%)
$(7.5, 9.0]$	3	2974	0.10 (%)

Ví dụ về bảng tần số thu gọn

Bảng tần số về điểm thi Toán khối A vào trường Đại học Thăng Long năm 2008 có thêm cột tần số tích lũy và tần suất tích lũy:

Khoảng điểm	Tần số	Tần số tích lũy	Tần suất(%)	Tần suất tích lũy(%)
$[0.0, 1.5]$	1317	1317	44.28 (%)	44.28 (%)
$(1.5, 3.0]$	967	2284	32.51 (%)	76.79 (%)
$(3.0, 4.5]$	518	1802	17.42 (%)	94.22 (%)
$(4.5, 6.0]$	151	2953	5.08 (%)	99.29 (%)
$(6.0, 7.5]$	18	2971	0.61 (%)	99.90 (%)
$(7.5, 9.0]$	3	2974	0.10 (%)	100 (%)

Giải quyết bài toán tình huống

- Dựa vào bảng tần số thu gọn về điểm thi Toán khối A năm 2008, hãy nhận xét về phân phối điểm thi tuyển sinh vào trường Đại học Thăng Long năm 2008.

Phân tổ dữ liệu trong R bằng hàm cut

`cut(x, breaks, labels, right = TRUE, include.lowest = FALSE)`
trong đó

- `x` véc tơ dữ liệu dạng số phân tổ;
- `breaks` véc tơ số (ít nhất hai tọa độ) gồm các điểm chia hoặc số nguyên dương (lớn hơn hoặc bằng 2) chỉ số tổ;
- `labels` nhãn các khoảng chia, mặc định các tổ dạng nửa khoảng $(a, b]$;
- `right` tham số logic, nếu `right = TRUE` khoảng chia dạng $(a, b]$, nếu `right = FALSE` khoảng chia dạng $[a, b)$;
- `include.lowest` là tham số dạng logic xem xét lấy thêm điểm chia nhỏ nhất vào tổ đầu hoặc lớn nhất vào tổ cuối.

Nội dung trình bày

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Đại học ThăngLong
Khương Trung, Thanh Xuân, Hà Nội
Tel.: 04.8587346 , Fax.: 04.5636775

Một số biểu đồ và đồ thị minh họa cho tập dữ liệu

- Một số biểu đồ cho dữ liệu định lượng:
  - Biểu đồ phân phối tần số;
  - Đa giác tần số;
  - Biểu đồ thân và lá.
- Một số biểu đồ cho dữ liệu định tính:
  - Biểu đồ tròn;
  - Biểu đồ thanh;
  - Biểu đồ Pareto.

Nội dung trình bày

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Đại học ThăngLong
Khương Trung, Thanh Xuân, Hà Nội
Tel.: 04.8587346 , Fax.: 04.5636775

Khái niệm và cách lập biểu đồ phân phối tần số

Khái niệm

Biểu đồ phân phối tần số gồm các cột đứng dùng để miêu tả phân phối tần số của tập dữ liệu định lượng.

Cách lập biểu đồ phân phối tần số:

- Phân chia tập dữ liệu thành các tổ;
- Xác định tần số trong mỗi tổ;
- Vẽ các cột đứng đặt cạnh nhau với độ rộng là khoảng cách trong mỗi tổ và độ cao tương ứng là tần số trong mỗi tổ.

Ví dụ về biểu đồ phân phối tần số

Khoảng điểm	Tần số
$[0.0, 1.5]$	1317
$(1.5, 3.0]$	967
$(3.0, 4.5]$	518
$(4.5, 6.0]$	151
$(6.0, 7.5]$	18
$(7.5, 9.0]$	3

Thông tin từ biểu đồ phân phối tần số

Dựa trên hình dáng của biểu đồ phân phối tần số ta có thể biết được:

- Mức độ tập trung tương đối của phân phối của tập dữ liệu;
- Hình dạng tương đối của phân phối của tập dữ liệu là bằng phẳng, lệch hay cân đối.

Phân phối điểm thi toán Khối A

- Dựa vào biểu đồ phân phối tần số điểm thi toán khối A, hãy nhận xét về phân phối điểm toán khối A năm 2008.

Vẽ biểu đồ phân phối tần số trong R bằng hàm hist

```
hist(x, breaks, include.lowest = TRUE,
    right = TRUE, col = , border = , labels = FALSE,
    main = "", xlim = , ylim = , xlab = "", ylab = "")
```

trong đó,

- `x` véc tơ dữ liệu vẽ biểu đồ;
- `breaks` véc tơ số gồm các điểm chia giữa các cột trong biểu đồ.
- `right` xem hàm `cut`;
- `include.lowest` xem hàm `cut`;
- `col` màu của các cột;
- `border` màu của đường biên của các cột;
- `labels` tham số điền tên trên đỉnh mỗi cột;
- `main`,`xlab`,`ylab` tên của biểu đồ, tên trục x, y;
- `xlim`,`ylim` giới hạn trên các trục.

Nội dung trình bày

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Đại học ThăngLong
Khương Trung, Thanh Xuân, Hà Nội
Tel.: 04.8587346 , Fax.: 04.5636775

Khái niệm và cách vẽ đa giác tần số

Khái niệm

Đa giác tần số là đồ thị gồm các đoạn thẳng nối các điểm với nhau dùng để miêu tả phân phối tần số của tập dữ liệu định lượng.

Cách vẽ đa giác tần số

- Phân chia tập dữ liệu thành các tổ, giả sử với các điểm chia là $x_0 < x_1 < \ldots < x_n$ và các tần số tương ứng trong mỗi tổ là $y_1, y_2, \ldots, y_n$.
- Xác định các điểm trong đồ thị:

$$A_0 = (x_0, 0), A_1 = (\frac{x_0 + x_1}{2}, y_1), \ldots A_{n-1} = (\frac{x_{n-1} + x_n}{2}, y_n), A_n = (x_n, 0).$$

- Đa giác tần số được xác định bằng cách nối lần lượt các điểm $A_0, A_1, \ldots, A_n$ lại với nhau.

Khái niệm và cách vẽ đa giác tần số

Khái niệm

Đa giác tần số là đồ thị gồm các đoạn thẳng nối các điểm với nhau dùng để miêu tả phân phối tần số của tập dữ liệu định lượng.

Cách vẽ đa giác tần số

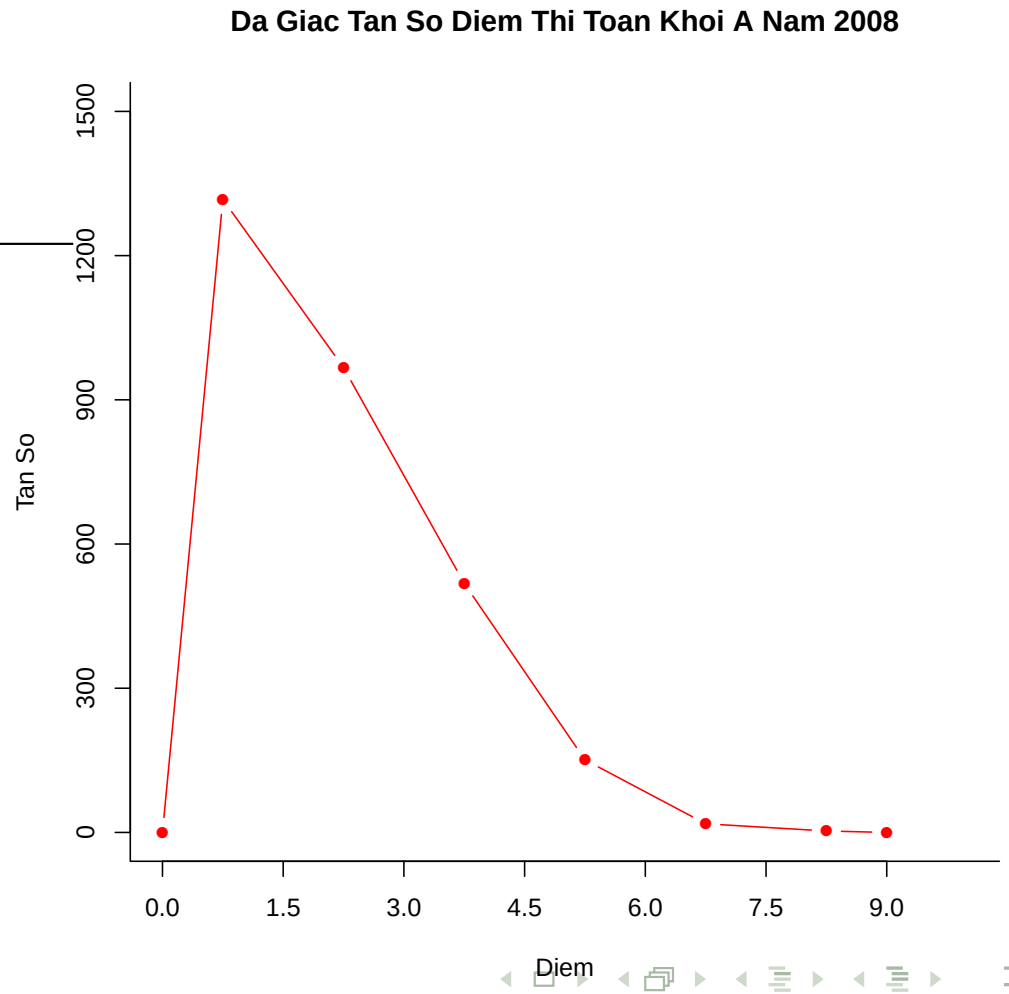
- Phân chia tập dữ liệu thành các tổ, giả sử với các điểm chia là $x_0 < x_1 < \ldots < x_n$ và các tần số tương ứng trong mỗi tổ là $y_1, y_2, \ldots, y_n$.
- Xác định các điểm trong đồ thị:

$$A_0 = (x_0, 0), A_1 = (\frac{x_0 + x_1}{2}, y_1), \ldots A_{n-1} = (\frac{x_{n-1} + x_n}{2}, y_n), A_n = (x_n, 0).$$

- Đa giác tần số được xác định bằng cách nối lần lượt các điểm $A_0, A_1, \ldots, A_n$ lại với nhau.

Ví dụ về đa giác tần số

Điểm	Hoành độ	Tung độ
A_0	0.0	0
A_1	0.75	1317
A_2	2.25	967
A_3	3.75	518
A_4	5.25	151
A_5	6.75	18
A_6	8.25	3
A_7	9.0	0



Thông tin từ đa giác tần số

Tương tự như biểu đồ phân phối tần số, hình dáng của đa giác tần số giúp ta có thể biết được:

- Mức độ tập trung tương đối của phân phối của tập dữ liệu;
- Hình dạng tương đối của phân phối của tập dữ liệu là bằng phẳng, lệch hay cân đối.

Phân phối điểm thi toán Khối A

- Dựa vào đa giác tần số điểm thi toán khối A, hãy nhận xét về phân phối điểm toán khối A năm 2008.

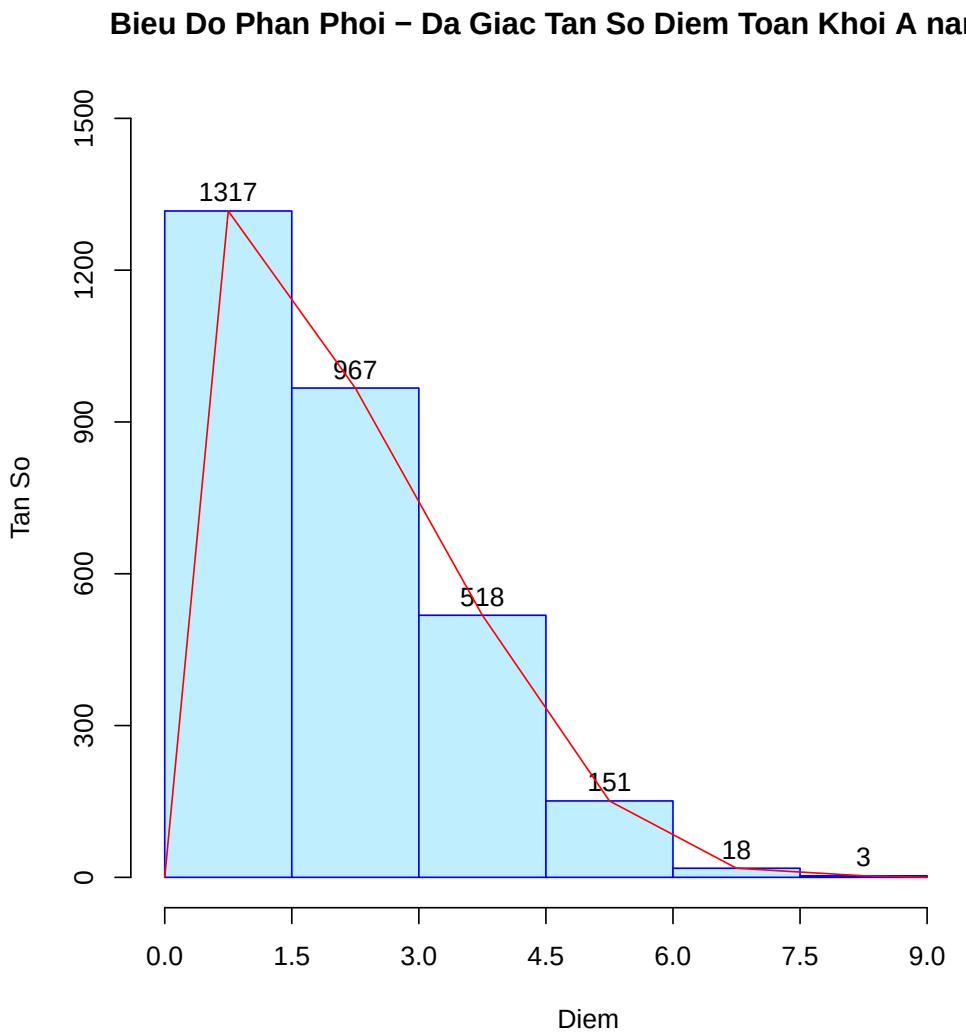
Cách đa giác tần số trong R bằng hàm plot

```
plot(x, y, type = "l"("b"),
     main = "", sub = "", xlab = "", ylab = "")
```

trong đó,

- `x,y` tương ứng là hoành độ, tung độ của các điểm trong đồ thị;
- `type="l"("b")` chỉ kiểu vẽ trong đồ thị là điểm nối bởi các đoạn thẳng;
- `main = "", sub = ""` là tham số tên chính và phụ của đồ thị;
- `xlab,ylab` là những tham số chỉ tên trục x, y;

Kết hợp biểu đồ phân phối tần số và đa giác tần số trên cùng một hình



Nội dung trình bày

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Đại học ThăngLong
Khương Trung, Thanh Xuân, Hà Nội
Tel.: 04.8587346 , Fax.: 04.5636775

Khái niệm và cách lập biểu đồ thân và lá

Khái niệm

Biểu đồ thân và lá được xây dựng bằng cách chia các chữ số trong mỗi số của tập dữ liệu định lượng thành hai phần: phần thân và phần lá để miêu tả phân phối của tập dữ liệu.

Cách lập biểu đồ phân phối tần số:

- Xác định phần thân, phần lá trong mỗi dữ liệu;
- Lập phần thân của biểu đồ;
- Xếp phần lá tương ứng với mỗi phần thân đã lập.

Khái niệm và cách lập biểu đồ thân và lá

Khái niệm

Biểu đồ thân và lá được xây dựng bằng cách chia các chữ số trong mỗi số của tập dữ liệu định lượng thành hai phần: phần thân và phần lá để miêu tả phân phối của tập dữ liệu.

Cách lập biểu đồ phân phối tần số:

- Xác định phần thân, phần lá trong mỗi dữ liệu;
- Lập phần thân của biểu đồ;
- Xếp phần lá tương ứng với mỗi phần thân đã lập.

Ví dụ về biểu đồ thân và lá điểm thi của 25 sinh viên

Điểm thi của 25 sinh viên				
86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	82	97	89

Thân

Ví dụ về biểu đồ thân và lá điểm thi của 25 sinh viên

Điểm thi của 25 sinh viên				
86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	82	97	89

Thân	Lá
2	3
3	
4	7
5	9
6	8
7	5 7 6
8	2 8 9
9	2 7 7

Ví dụ về biểu đồ thân và lá điểm thi của 25 sinh viên

Điểm thi của 25 sinh viên				
86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	82	97	89

Thân	Lá	
3		
7		
5	9	
0	7	
2	5	
1	2	
1	2	

Ví dụ về biểu đồ thân và lá điểm thi của 25 sinh viên

Điểm thi của 25 sinh viên				
86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	82	97	89

Thân	Lá		
3			
4			
5	9		
6	8		
7	5	5	
8	2	2	
9	2	7	

Ví dụ về biểu đồ thân và lá điểm thi của 25 sinh viên

Điểm thi của 25 sinh viên				
86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	82	97	89

Thân	Lá			
3				
4				
5	9			
6	7	8		
7	5	5	6	
8	2	2	3	
9	2	7	7	

Ví dụ về biểu đồ thân và lá điểm thi của 25 sinh viên

Điểm thi của 25 sinh viên				
86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	82	97	89

Thân	Lá				
3					
4					
5	9				
6	7	8			
7	5	5	6	7	
8	2	2	3	6	
9	2	7	7		

Ví dụ về biểu đồ thân và lá điểm thi của 25 sinh viên

Điểm thi của 25 sinh viên				
86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	82	97	89

Thân	Lá
3	
4	
5	9
6	7 8
7	5 5 6 7 7
8	1 2 2 3 6 8
9	1 2 7 7

Ví dụ về biểu đồ thân và lá điểm thi của 25 sinh viên

Điểm thi của 25 sinh viên				
86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	82	97	89

Thân	Lá
3	
4	
5	9
6	7 8
7	5 5 6 7 7
8	1 2 2 3 6 8 9
9	1 2 7 7

Ví dụ về biểu đồ thân và lá điểm thi của 25 sinh viên

Điểm thi của 25 sinh viên				
86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	82	97	89

Thân	Lá
3	
4	
5	9
6	7 8
7	5 5 6 7 7
8	1 2 2 3 6 8 9 9
9	1 2 7 7

Ví dụ về biểu đồ thân và lá điểm thi của 25 sinh viên

Điểm thi của 25 sinh viên				
86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	82	97	89

Thân	Lá						
2	3						
3							
4	7						
5	5	9					
6	0	7	8				
7	2	5	5	6	7	7	
8	1	2	2	3	6	8	9	9
9	1	2	7	7			

Thông tin từ biểu đồ thân và lá

Tương tự như biểu đồ phân phối tần số và đa giác tần số, hình dáng của biểu đồ thân và lá giúp ta có thể biết được:

- Mức độ tập trung tương đối của phân phối của tập dữ liệu;
- Hình dạng tương đối của phân phối của tập dữ liệu là bằng phẳng, lệch hay cân đối.

Phân phối điểm thi của 25 sinh viên

- Dựa vào biểu đồ thân và lá hãy nhận xét về phân phối điểm thi của 25 sinh viên.

Vẽ biểu đồ thân và lá trong R bằng hàm `stem()`

```
stem(x, scale = 1)
```

trong đó,

- `x` là véc tơ dữ liệu dạng số;
- `scale` là tham số điều chỉnh chiều dài của biểu đồ;

Ví dụ về vẽ biểu đồ thân và lá trên R

Biểu đồ thân và lá về dữ liệu điểm thi của 25 sinh viên

```
> DiemThi = c(86, 77, 91, 60, 55, 76, 92, 47, 88, 67,
23, 59, 72, 75, 83, 77, 68, 82, 97, 89, 81, 75, 82, 97, 89)
> stem(DiemThi)
  The decimal point is 1 digit(s) to the right of the |
  2 | 3
  4 | 759
  6 | 078255677
  8 | 122368991277
```

Ví dụ về vẽ biểu đồ thân và lá trên R

```
> DiemThi = c(86, 77, 91, 60, 55, 76, 92, 47, 88, 67,
23, 59, 72, 75, 83, 77, 68, 82, 97, 89, 81, 75, 82, 97, 89)
> stem(DiemThi, scale = 2)
  The decimal point is 1 digit(s) to the right of the |
  2 | 3
  3 |
  4 | 7
  5 | 59
  6 | 078
  7 | 2556779
  8 | 122368999
  9 | 1277
```

Nội dung trình bày

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Khái niệm và cách vẽ biểu đồ hình tròn

Khái niệm

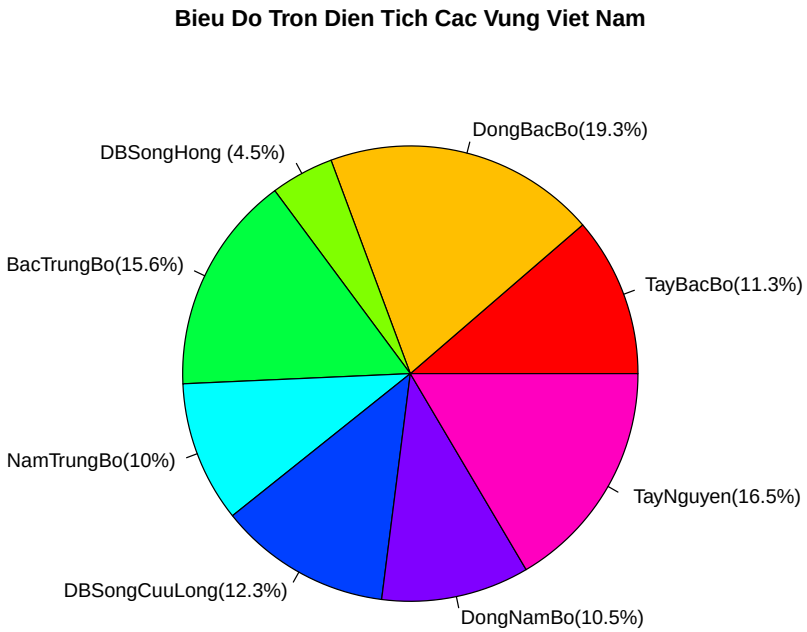
Biểu đồ hình tròn miêu tả dữ liệu định tính dưới dạng hình tròn, trong đó diện tích của toàn bộ hình tròn đại diện cho 100% *các phần tử của tập dữ liệu và diện tích các hình quạt đại diện cho phần trăm của các tập con cần biểu diễn.*

Cách vẽ biểu đồ hình tròn:

- Xác định các biểu hiện trong tập dữ liệu (định tính);
- Xác định tỉ lệ cho từng biểu hiện;
- Tính góc của hình quạt tương ứng với từng biểu hiện
- Dựa trên số đo góc cho hình quạt để vẽ phần diện tích cho mỗi biểu hiện tương ứng.

Ví dụ về biểu đồ hình tròn

Vùng	Diện tích (m^2)	Tỉ lệ (%)	Góc (độ)
ĐB Sông Hồng	14862.4	4.50	16.15
Đông Bắc Bộ	64025.2	19.3	69.59
Tây Bắc Bộ	37533.9	11.3	40.80
Bắc Trung Bộ	51551.9	15.6	56.03
Nam Trung Bộ	33166.0	10.0	36.05
Tây Nguyên	54659.6	16.5	59.41
Đông Nam Bộ	34807.8	10.5	37.83
ĐB Sông Cửu Long	40604.8	12.3	44.13



Thông tin từ biểu đồ hình tròn

Dựa trên biểu đồ hình tròn ta có thể biết được:

- Mức phân phối tương đối của mỗi biểu hiện so với toàn thể;
- Biểu hiện nào có chiếm nhiều nhất, ít nhất.

Phân phối diện tích các vùng ở Việt Nam

- Dựa vào biểu đồ tròn hãy nhận xét về phân phối diện tích của các vùng ở Việt Nam.

Vẽ biểu đồ hình tròn trong R bằng hàm pie()

```
pie(x, labels = names(x), col = NULL, border = NULL,
    lty = NULL, main = NULL)
```

trong đó

- `x` là véc tơ dạng số thể hiện giá trị của mỗi hình quạt trong biểu đồ;
- `labels` là tham số chỉ tên của những hình quạt trong biểu đồ;
- `col` là tham số chỉ màu của các hình quạt;
- `border` là tham số chỉ màu của đường danh giới giữa các hình quạt;
- `main,sub` là những tham số chỉ tiêu đề và tiêu đề phụ của biểu đồ.

Nội dung trình bày

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Khái niệm và cách vẽ biểu đồ thanh

Khái niệm

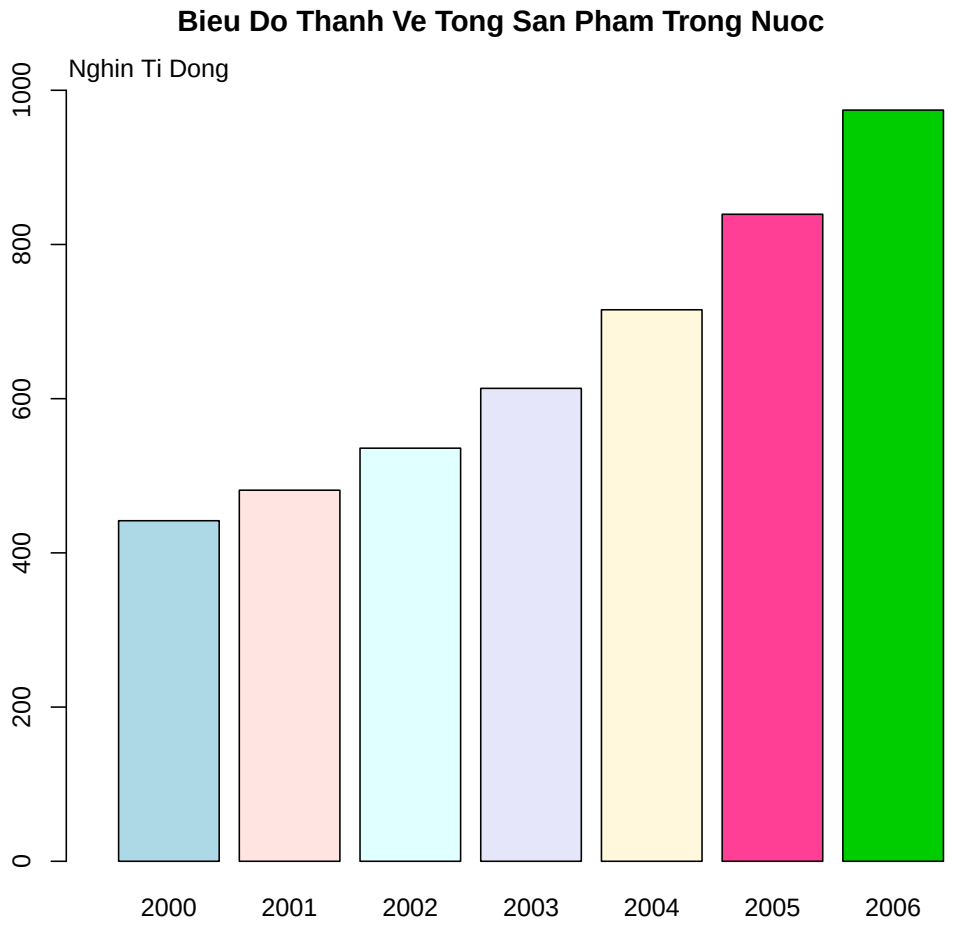
Biểu đồ thanh bao gồm thanh đứng (ngang) dùng để miêu tả phân phối tần số của tập dữ liệu định tính.

Cách vẽ biểu đồ thanh:

- Xác định các biểu hiện trong tập dữ liệu;
- Xác định giá trị cho từng biểu hiện trong tập dữ liệu;
- Vẽ các thanh đứng cho từng biểu hiện với chiều cao tương ứng với giá trị mỗi biểu hiện.

Ví dụ về biểu đồ thanh

Năm	Tổng sản phẩm
2000	441.7
2001	481.3
2002	535.8
2003	613.4
2004	715.3
2005	839.2
2005	974.3



Thông tin từ biểu đồ thanh

Dựa trên biểu đồ hình thanh ta có thể biết được:

- Sự so sánh tương đối giữa các biểu hiện với nhau;
- Biểu hiện nào có chiếm nhiều nhất, ít nhất.

Phân phối của tổng sản phẩm trong nước từ năm 2000 đến năm 2006

- Dựa vào biểu đồ thanh hãy nhận xét về phân phối của tổng sản phẩm trong nước từ năm 2000 đến năm 2006.

Vẽ biểu đồ hình thanh trong R bằng hàm barplot

`barplot(height, names.arg, col = "", border = "")`

- `height` véc tơ hoặc ma trận dữ liệu dùng để vẽ biểu đồ;
- `names.arg` tên viết dưới mỗi thanh hoặc nhóm các thanh trong biểu đồ;
- `col` màu của các thanh;
- `border` màu của đường biên của các cột.

Nội dung trình bày

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Đại học ThăngLong
Khương Trung, Thanh Xuân, Hà Nội
Tel.: 04.8587346 , Fax.: 04.5636775

Khái niệm và cách vẽ biểu đồ Pareto

Khái niệm

Biểu đồ Pareto dùng để miêu tả phân phối của tập dữ liệu định tính, gồm các thanh đứng, trong đó thông tin về các quan sát được phân loại và được đưa lên biểu đồ theo thứ tự giảm dần của các tần số, đồng thời kết hợp luôn đa giác tích lũy trên cùng biểu đồ này.

Cách lập biểu đồ Pareto

- Xác định các biểu hiện của tập dữ liệu;
- Xác định tần số các biểu hiện;
- Xếp các biểu hiện theo thứ tự tần số giảm dần và lập tần suất tích lũy theo thứ tự này;
- Vẽ các thanh đứng mô tả tần số các biểu hiện theo thứ tự tần số giảm dần;
- Vẽ đa giác tích lũy của các biểu hiện theo thứ tự tần số giảm dần

Khái niệm và cách vẽ biểu đồ Pareto

Khái niệm

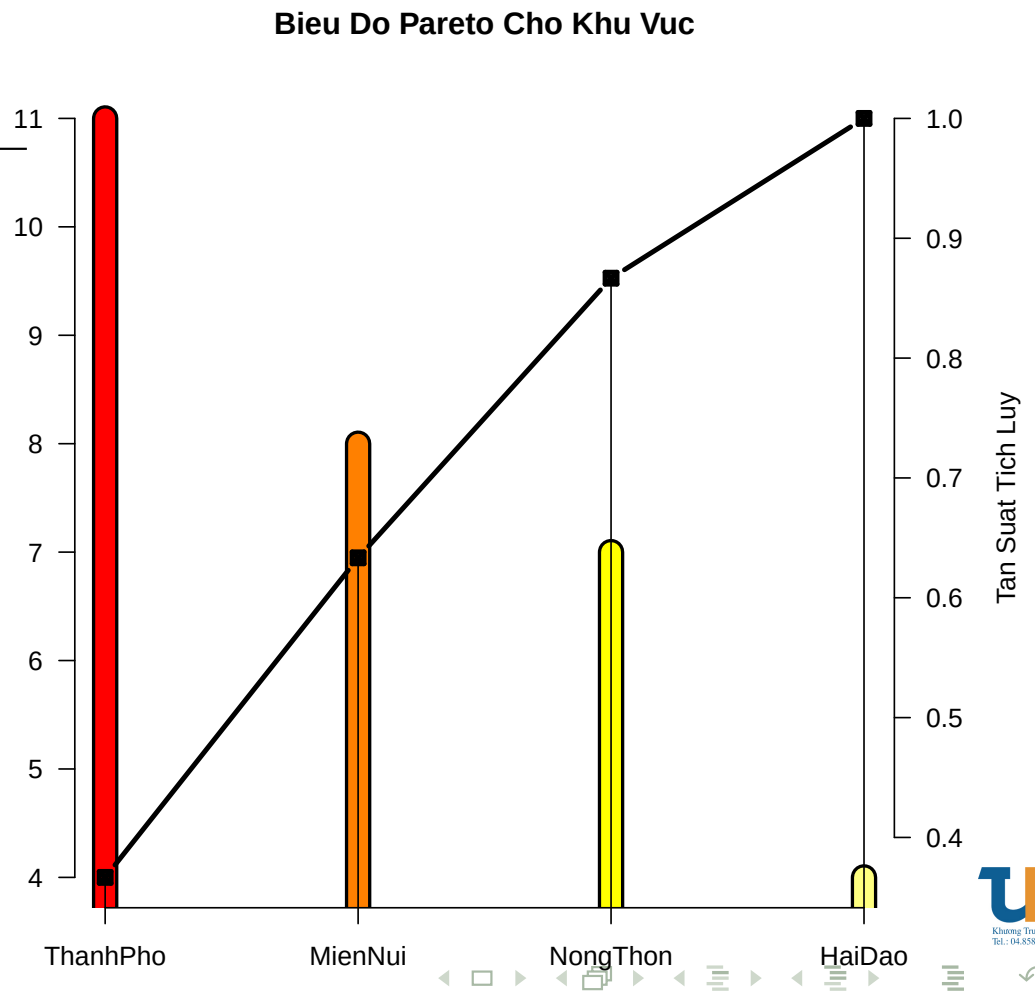
Biểu đồ Pareto dùng để miêu tả phân phối của tập dữ liệu định tính, gồm các thanh đứng, trong đó thông tin về các quan sát được phân loại và được đưa lên biểu đồ theo thứ tự giảm dần của các tần số, đồng thời kết hợp luôn đa giác tích lũy trên cùng biểu đồ này.

Cách lập biểu đồ Pareto

- Xác định các biểu hiện của tập dữ liệu;
- Xác định tần số các biểu hiện;
- Xếp các biểu hiện theo thứ tự tần số giảm dần và lập tần suất tích lũy theo thứ tự này;
- Vẽ các thanh đứng mô tả tần số các biểu hiện theo thứ tự tần số giảm dần;
- Vẽ đa giác tích lũy của các biểu hiện theo thứ tự tần số giảm dần.

Ví dụ về biểu đồ Pareto

Khu vực	Tần số	Tần suất tích lũy
Thành phố	11	0.37
Miền núi	8	0.63
Nông thôn	7	0.87
Hải đảo	4	1.00



Thông tin từ biểu đồ Pareto

Dựa trên biểu đồ Pareto ta có thể biết được:

- Sự so sánh tương đối giữa tần số của các biểu hiện với nhau: biểu hiện nào xuất hiện nhiều hơn biểu hiện nào, biểu hiện nào xuất hiện nhiều nhất, ít nhất;
- Tần suất tích lũy của các biểu hiện

Phân phối của khu vực sống trong mẫu điều tra

- Dựa vào biểu đồ Pareto hãy nhận xét về phân phối của khu vực sống trong mẫu điều tra.

Nội dung trình bày

1. Tóm tắt và trình bày dữ liệu bằng bảng tần số
   - Khái niệm bảng tần số
   - Bảng tần số cho dữ liệu định tính
   - Bảng tần số cho dữ liệu định lượng
2. Tóm tắt và trình bày dữ liệu bằng biểu đồ và đồ thị
   - Một số biểu đồ và đồ thị minh họa cho tập dữ liệu
   - Biểu đồ phân phối tần số
   - Đa giác tần số
   - Biểu đồ thân và lá
   - Biểu đồ hình tròn
   - Biểu đồ thanh
   - Biểu đồ Pareto
   - Biểu đồ tán xạ

Đại học ThăngLong
Khương Trung, Thanh Xuân, Hà Nội
Tel.: 04.8587346 , Fax.: 04.5636775

Khái niệm và cách vẽ biểu đồ tán xạ

Khái niệm

Biểu đồ tán xạ là đồ thị hai chiều bao gồm các điểm mô tả mối quan hệ giữa hai biến định lượng.

Cách lập biểu đồ tán xạ:

- Xác định các biểu hiện của biến định lượng thứ nhất: $x_1, x_2, \ldots, x_n$;
- Xác định các biểu hiện của biến định lượng thứ hai: $y_1, y_2, \ldots, y_n$;
- Xác định các cặp điểm $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$;
- Vẽ các cặp điểm này trên mặt phẳng đề các.

Khái niệm và cách vẽ biểu đồ tán xạ

Khái niệm

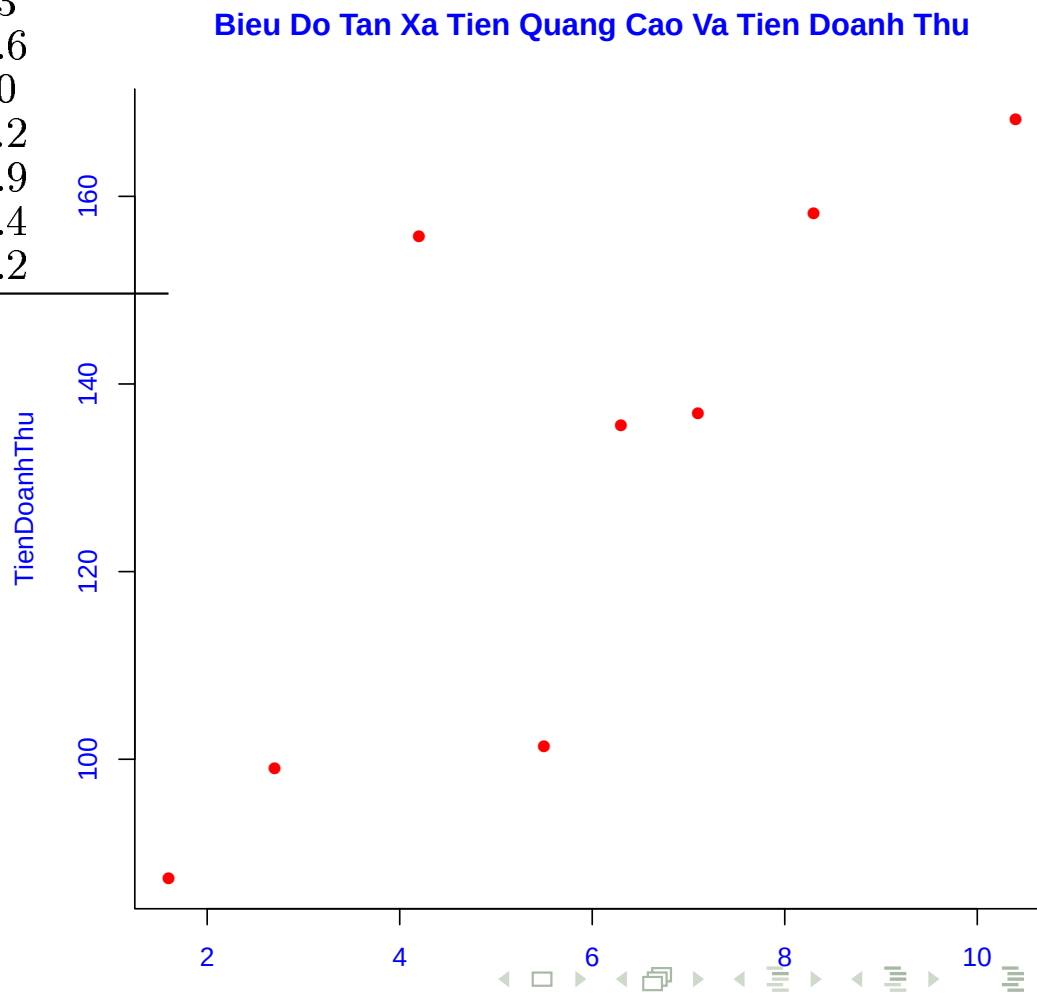
Biểu đồ tán xạ là đồ thị hai chiều bao gồm các điểm mô tả mối quan hệ giữa hai biến định lượng.

Cách lập biểu đồ tán xạ:

- Xác định các biểu hiện của biến định lượng thứ nhất: $x_1, x_2, \ldots, x_n$;
- Xác định các biểu hiện của biến định lượng thứ hai: $y_1, y_2, \ldots, y_n$;
- Xác định các cặp điểm $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$;
- Vẽ các cặp điểm này trên mặt phẳng đề các.

Ví dụ về biểu đồ tán xạ

Tiền quảng cáo (triệu đô)	Tiền doanh thu (triệu đô)
4.2	155.7
1.6	87.3
6.3	135.6
2.7	99.0
10.4	168.2
7.1	136.9
5.5	101.4
8.3	158.2



Thông tin từ biểu đồ tán xạ

Biểu đồ tán xạ cho ta những thông tin sau:

- Biểu đồ tán xạ cho ta cái nhìn tương đối về mối quan hệ giữa hai biến định lượng;
- Biểu đồ tán xạ cho ta thấy hai biến định lượng có quan hệ dạng tuyến tính hay phi tuyến.

Mối quan hệ giữa số tiền quảng cáo và tiền doanh thu

- Dựa vào biểu đồ tán xạ hãy nhận xét về mối quan hệ giữa số tiền quảng cáo và tiền doanh thu.

Vẽ biểu đồ tán xạ trên R bằng hàm plot

```
plot (x, y, type = "p", main = "", xlab = "", ylab = "",
bty = "l")
```

trong đó

- `x,y` tương ứng là véc tơ tọa độ của biến định lượng thứ nhất, thứ hai;
- `main, xlab, ylab` là tên của biểu đồ, tên biến thứ nhất, biến thứ hai;
- bty = "l" tham số cho kiểu bao quanh biểu đồ giống hai trục tọa độ đề các *oxy*.

Chương IV

Tóm tắt và trình bày dữ liệu bằng các đại lượng thống kê mô tả

Chương IV

3. Các số đo hướng tâm của tập dữ liệu
4. Các đại lượng mô tả sự phân bố của tập dữ liệu
5. Các đại lượng đo lường độ phân tán
6. Sử dụng kết hợp trung bình và độ lệch chuẩn
7. Các đại lượng miêu tả hình dáng của tập dữ liệu

Chương IV

3. Các số đo hướng tâm của tập dữ liệu
4. Các đại lượng mô tả sự phân bố của tập dữ liệu
5. Các đại lượng đo lường độ phân tán
6. Sử dụng kết hợp trung bình và độ lệch chuẩn
7. Các đại lượng miêu tả hình dáng của tập dữ liệu

Chương IV

3. Các số đo hướng tâm của tập dữ liệu
4. Các đại lượng mô tả sự phân bố của tập dữ liệu
5. Các đại lượng đo lường độ phân tán
6. Sử dụng kết hợp trung bình và độ lệch chuẩn
7. Các đại lượng miêu tả hình dáng của tập dữ liệu

Chương IV

3. Các số đo hướng tâm của tập dữ liệu
4. Các đại lượng mô tả sự phân bố của tập dữ liệu
5. Các đại lượng đo lường độ phân tán
6. Sử dụng kết hợp trung bình và độ lệch chuẩn
7. Các đại lượng miêu tả hình dáng của tập dữ liệu

Chương IV

3. Các số đo hướng tâm của tập dữ liệu
4. Các đại lượng mô tả sự phân bố của tập dữ liệu
5. Các đại lượng đo lường độ phân tán
6. Sử dụng kết hợp trung bình và độ lệch chuẩn
7. Các đại lượng miêu tả hình dáng của tập dữ liệu

Câu hỏi tình huống chương 4

Câu hỏi tình huống

Với tập dữ liệu về điểm thi tuyển sinh toán Khối A vào Đại học Thăng Long năm 2008, bạn hãy đưa những thông tin sau để Ban giám hiệu và các thí sinh hiểu được tình hình chung về kết quả của môn thi Toán:

- *Điểm trung bình môn Toán của các thí sinh là bao nhiêu?*
- *Một nửa số thí sinh có điểm môn Toán cao hơn bao nhiêu? Thấp hơn bao nhiêu?*
- *Điểm thấp nhất của nhóm thí sinh nằm trong top 10% số thí sinh có điểm thi môn Toán cao nhất là bao nhiêu?*
- *Điểm thấp nhất, cao nhất, điểm xuất hiện nhiều nhất trong tập dữ liệu điểm thi là bao nhiêu?*

Nội dung chính trong chương

- Giới thiệu những số đo hướng tâm phổ biến của một tập dữ liệu như: trung bình cộng, trung vị, mode;
- Giới thiệu những đại lượng đo độ phân bố phổ biến của một tập dữ liệu như: tứ phân vị, phân vị thứ p;
- Giới thiệu những đại lượng đo độ phân tán phổ biến của một tập dữ liệu như: khoảng biến thiên, độ trải giữa, phương sai, độ lệch chuẩn;
- Giới thiệu những qui tắc mô tả phân phối của một tập dữ liệu: qui tắc thực nghiệm, định lí Chebyshev;
- Giới thiệu những đại lượng mô tả hình dáng của tập dữ liệu: hệ số skewness và hệ sô Kurtosis.

Yêu cầu đối với sinh viên

- Biết cách tính những số đo hướng tâm phổ biến như: trung bình cộng, trung vị, mode và ý nghĩa của những giá trị này;
- Biết cách tính những đại lượng đo độ phân bố phổ biến của một tập dữ liệu như: tứ phân vị, phân vị thứ p và ý nghĩa của những đại lượng này;
- Biết cách tính những đại lượng đo độ phân tán phổ biến của một tập dữ liệu như: khoảng biến thiên, độ trải giữa, phương sai, độ lệch chuẩn và ý nghĩa của những đại lượng này;
- Biết và vận dụng được hai qui tắc mô tả phân phối của một tập dữ liệu: qui tắc thực nghiệm, định lí Chebyshev;
- Biết cách tính những đại lượng mô tả hình dáng của tập dữ liệu: hệ số skewness và hệ sô Kurtosis và ý nghĩa của những đại lượng này.

Các số đo hướng tâm của tập dữ liệu

- Trung bình cộng: trung bình cộng đơn giản, trung bình cộng có trọng số
- Trung vị
- Mode

Trung bình cộng đơn giản

Khái niệm

Trung bình cộng đơn giản được tính bằng cách cộng tất cả các giá trị quan sát của tập dữ liệu rồi chia cho số quan sát của tập dữ liệu đó.

Công thức tính trung bình cộng đơn giản:

- Trung bình cộng đơn giản $\overline{x}$ của các giá trị $x_1, x_2, \ldots, x_n$ được cho bởi công thức:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n},$$

- **Ví dụ:** Giả sử số tiền (đơn vị nghìn VND) dùng cho chi tiêu thực phẩm trong một tuần là: 120, 150, 125, 100, 180, 140, 200. Khi đó số tiền trung bình một ngày trong tuần dành cho chi tiêu thực phẩm là:

$$\overline{x} = \frac{120 + 150 + 125 + 100 + 180 + 140 + 200}{7} = 145.$$

Trung bình cộng đơn giản

Khái niệm

Trung bình cộng đơn giản được tính bằng cách cộng tất cả các giá trị quan sát của tập dữ liệu rồi chia cho số quan sát của tập dữ liệu đó.

Công thức tính trung bình cộng đơn giản:

- Trung bình cộng đơn giản $\overline{x}$ của các giá trị $x_1, x_2, \ldots, x_n$ được cho bởi công thức:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n},$$

- **Ví dụ:** Giả sử số tiền (đơn vị nghìn VND) dùng cho chi tiêu thực phẩm trong một tuần là: 120, 150, 125, 100, 180, 140, 200. Khi đó số tiền trung bình một ngày trong tuần dành cho chi tiêu thực phẩm là:

$$\overline{x} = \frac{120 + 150 + 125 + 100 + 180 + 140 + 200}{7} = 145.$$

Trung bình cộng có trọng số

Khái niệm

Trung bình cộng có trọng số được tính bằng cách cộng các tích của giá trị quan sát của tập dữ liệu với trọng số tương ứng rồi chia cho tổng các trọng số của tập dữ liệu đó.

Công thức tính trung bình cộng có trọng số

- Trung bình có trọng số $\overline{x}$ của các giá trị x_i với trọng số tương ứng w_i, $\forall i = 1, \ldots, n$ được tính bởi công thức.

$$\overline{x}_w = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_k w_k}{w_1 + w_2 + \ldots + w_k},$$

Trung bình cộng có trọng số

Khái niệm

Trung bình cộng có trọng số được tính bằng cách cộng các tích của giá trị quan sát của tập dữ liệu với trọng số tương ứng rồi chia cho tổng các trọng số của tập dữ liệu đó.

Công thức tính trung bình cộng có trọng số

- Trung bình có trọng số $\overline{x}$ của các giá trị x_i với trọng số tương ứng w_i, $\forall i = 1, \ldots, n$ được tính bởi công thức.

$$\overline{x}_w = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_k w_k}{w_1 + w_2 + \ldots + w_k},$$

Ví dụ tính trung bình cộng có trọng số

- Điểm thi kết thúc học kì của một sinh viên được cho trong bảng dưới đây:

Môn học	**Điểm**	**Số đ.v tín chỉ**
XSTK và Ứng dụng	7.0	4
Nguyên lí kế toán	5.0	2
Toán Quản lí	8.0	2
Mô hình Kinh tế	4.0	2
Thị trường chứng khoán	6.0	2

- Khi đó điểm trung bình các môn học của sinh viên trên trong kì này sẽ là:

$$\overline{x} = \frac{7 \times 4 + 5 \times 2 + 8 \times 2 + 4 \times 2 + 6 \times 2}{4 + 2 + 2 + 2 + 2} = 6.17.$$

Ví dụ tính trung bình cộng có trọng số

- Điểm thi kết thúc học kì của một sinh viên được cho trong bảng dưới đây:

Môn học	**Điểm**	**Số đ.v tín chỉ**
XSTK và Ứng dụng	7.0	4
Nguyên lí kế toán	5.0	2
Toán Quản lí	8.0	2
Mô hình Kinh tế	4.0	2
Thị trường chứng khoán	6.0	2

- Khi đó điểm trung bình các môn học của sinh viên trên trong kì này sẽ là:

$$\overline{x} = \frac{7 \times 4 + 5 \times 2 + 8 \times 2 + 4 \times 2 + 6 \times 2}{4 + 2 + 2 + 2 + 2} = 6.17.$$

Ưu điểm và nhược điểm của trung bình cộng

- Ưu điểm:
  - Trung bình cộng là một số đo hướng tâm phổ biến vì nó tác động vào mọi phần tử của tập dữ liệu;
  - Mọi tập dữ liệu đều có duy nhất một trung bình cộng;
  - Trung bình cộng là một khái niệm toán học quen thuộc, hơn nữa nó có nhiều tính chất toán học giúp cho ta có thể thực hiện được các suy diễn trong thống kê.
- Nhược điểm:
  - Trung bình cộng bị ảnh hưởng bởi các giá trị ngoại biên;
  - Trung bình cộng chỉ dùng cho dữ liệu đo bằng thang đo định lượng.

Ưu điểm và nhược điểm của trung bình cộng

- Ưu điểm:
  - Trung bình cộng là một số đo hướng tâm phổ biến vì nó tác động vào mọi phần tử của tập dữ liệu;
  - Mọi tập dữ liệu đều có duy nhất một trung bình cộng;
  - Trung bình cộng là một khái niệm toán học quen thuộc, hơn nữa nó có nhiều tính chất toán học giúp cho ta có thể thực hiện được các suy diễn trong thống kê.
- Nhược điểm:
  - Trung bình cộng bị ảnh hưởng bởi các giá trị ngoại biên;
  - Trung bình cộng chỉ dùng cho dữ liệu đo bằng thang đo định lượng.

Ưu điểm và nhược điểm của trung bình cộng

- Ưu điểm:
  - Trung bình cộng là một số đo hướng tâm phổ biến vì nó tác động vào mọi phần tử của tập dữ liệu;
  - Mọi tập dữ liệu đều có duy nhất một trung bình cộng;
  - Trung bình cộng là một khái niệm toán học quen thuộc, hơn nữa nó có nhiều tính chất toán học giúp cho ta có thể thực hiện được các suy diễn trong thống kê.
- Nhược điểm:
  - Trung bình cộng bị ảnh hưởng bởi các giá trị ngoại biên;
  - Trung bình cộng chỉ dùng cho dữ liệu đo bằng thang đo định lượng.

Ưu điểm và nhược điểm của trung bình cộng

- Ưu điểm:
  - Trung bình cộng là một số đo hướng tâm phổ biến vì nó tác động vào mọi phần tử của tập dữ liệu;
  - Mọi tập dữ liệu đều có duy nhất một trung bình cộng;
  - Trung bình cộng là một khái niệm toán học quen thuộc, hơn nữa nó có nhiều tính chất toán học giúp cho ta có thể thực hiện được các suy diễn trong thống kê.
- Nhược điểm:
  - Trung bình cộng bị ảnh hưởng bởi các giá trị ngoại biên;
  - Trung bình cộng chỉ dùng cho dữ liệu đo bằng thang đo định lượng.

Ưu điểm và nhược điểm của trung bình cộng

- Ưu điểm:
  - Trung bình cộng là một số đo hướng tâm phổ biến vì nó tác động vào mọi phần tử của tập dữ liệu;
  - Mọi tập dữ liệu đều có duy nhất một trung bình cộng;
  - Trung bình cộng là một khái niệm toán học quen thuộc, hơn nữa nó có nhiều tính chất toán học giúp cho ta có thể thực hiện được các suy diễn trong thống kê.
- Nhược điểm:
  - Trung bình cộng bị ảnh hưởng bởi các giá trị ngoại biên;
  - Trung bình cộng chỉ dùng cho dữ liệu đo bằng thang đo định lượng.

Trung vị

Khái niệm

Trung vị của tập dữ liệu đã được sắp thứ tự là giá trị mà có không quá 50% số quan sát của tập dữ liệu có giá trị nhỏ hơn trung vị và không quá 50% số quan sát của tập dữ liệu có giá trị lớn hơn trung vị.

Cách tìm trung vị của tập dữ liệu có n quan sát

- Sắp xếp lại tập dữ liệu;
- Nếu số quan sát n là số lẻ thì trung vị là quan sát ở vị trí thứ $(n+1)/2$.
- Nếu số quan sát n là số chẵn thì trung vị giá trị trung bình cộng của hai quan sát ở vị trí chính giữa của tập dữ liệu, tức là hai quan sát ở vị trí thứ $n/2$ và $(n+2)/2$.

Nhận xét: Trong nhiều trường hợp ta có thể nói có khoảng 50% số quan sát của tập dữ liệu nhỏ hơn hoặc bằng trung vị và khoảng 50% số quan sát của tập dữ liệu lớn hơn hoặc bằng trung vị.

Trung vị

Khái niệm

Trung vị của tập dữ liệu đã được sắp thứ tự là giá trị mà có không quá 50% số quan sát của tập dữ liệu có giá trị nhỏ hơn trung vị và không quá 50% số quan sát của tập dữ liệu có giá trị lớn hơn trung vị.

Cách tìm trung vị của tập dữ liệu có n quan sát

- Sắp xếp lại tập dữ liệu;
- Nếu số quan sát n là số lẻ thì trung vị là quan sát ở vị trí thứ $(n+1)/2$.
- Nếu số quan sát n là số chẵn thì trung vị giá trị trung bình cộng của hai quan sát ở vị trí chính giữa của tập dữ liệu, tức là hai quan sát ở vị trí thứ $n/2$ và $(n+2)/2$.

Nhận xét: Trong nhiều trường hợp ta có thể nói có khoảng 50% số quan sát của tập dữ liệu nhỏ hơn hoặc bằng trung vị và khoảng 50% số quan sát của tập dữ liệu lớn hơn hoặc bằng trung vị.

Trung vị

Khái niệm

Trung vị của tập dữ liệu đã được sắp thứ tự là giá trị mà có không quá 50% số quan sát của tập dữ liệu có giá trị nhỏ hơn trung vị và không quá 50% số quan sát của tập dữ liệu có giá trị lớn hơn trung vị.

Cách tìm trung vị của tập dữ liệu có n quan sát

- Sắp xếp lại tập dữ liệu;
- Nếu số quan sát n là số lẻ thì trung vị là quan sát ở vị trí thứ $(n+1)/2$.
- Nếu số quan sát n là số chẵn thì trung vị giá trị trung bình cộng của hai quan sát ở vị trí chính giữa của tập dữ liệu, tức là hai quan sát ở vị trí thứ $n/2$ và $(n+2)/2$.

Nhận xét: Trong nhiều trường hợp ta có thể nói có khoảng 50% số quan sát của tập dữ liệu nhỏ hơn hoặc bằng trung vị và khoảng 50% số quan sát của tập dữ liệu lớn hơn hoặc bằng trung vị.

Trung vị

Khái niệm

Trung vị của tập dữ liệu đã được sắp thứ tự là giá trị mà có không quá 50% số quan sát của tập dữ liệu có giá trị nhỏ hơn trung vị và không quá 50% số quan sát của tập dữ liệu có giá trị lớn hơn trung vị.

Cách tìm trung vị của tập dữ liệu có n quan sát

- Sắp xếp lại tập dữ liệu;
- Nếu số quan sát n là số lẻ thì trung vị là quan sát ở vị trí thứ $(n+1)/2$.
- Nếu số quan sát n là số chẵn thì trung vị giá trị trung bình cộng của hai quan sát ở vị trí chính giữa của tập dữ liệu, tức là hai quan sát ở vị trí thứ $n/2$ và $(n+2)/2$.

Nhận xét: Trong nhiều trường hợp ta có thể nói có khoảng 50% số quan sát của tập dữ liệu nhỏ hơn hoặc bằng trung vị và khoảng 50% số quan sát của tập dữ liệu lớn hơn hoặc bằng trung vị.

Trung vị

Khái niệm

Trung vị của tập dữ liệu đã được sắp thứ tự là giá trị mà có không quá 50% số quan sát của tập dữ liệu có giá trị nhỏ hơn trung vị và không quá 50% số quan sát của tập dữ liệu có giá trị lớn hơn trung vị.

Cách tìm trung vị của tập dữ liệu có n quan sát

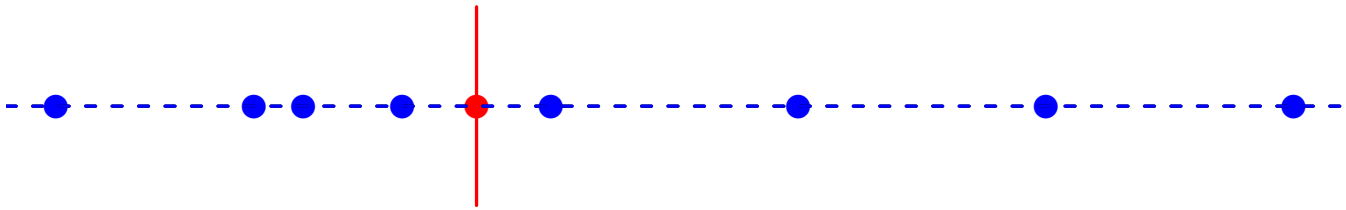
- Sắp xếp lại tập dữ liệu;
- Nếu số quan sát n là số lẻ thì trung vị là quan sát ở vị trí thứ $(n+1)/2$.
- Nếu số quan sát n là số chẵn thì trung vị giá trị trung bình cộng của hai quan sát ở vị trí chính giữa của tập dữ liệu, tức là hai quan sát ở vị trí thứ $n/2$ và $(n+2)/2$.

Nhận xét: Trong nhiều trường hợp ta có thể nói có khoảng 50% số quan sát của tập dữ liệu nhỏ hơn hoặc bằng trung vị và khoảng 50% số quan sát của tập dữ liệu lớn hơn hoặc bằng trung vị.

Ví dụ tính trung vị

Ví dụ: Hãy tìm trung vị của tập dữ liệu: $5, 9, 12, 10, 20, 15, 30, 25$ và nêu nhận xét.

Minh họa trung vị



Ưu điểm và nhược điểm của trung vị

- Ưu điểm:
  - Trung vị không bị ảnh hưởng bởi những giá trị ngoại biên;
  - Mọi tập dữ liệu đều có duy nhất một trung vị;
  - Có thể tìm trung vị cho tập dữ liệu sử dụng thang đo thứ bậc, khoảng, tỉ lệ.
- Nhược điểm:
  - Trung vị đánh đồng giữa giá trị với thứ tự của giá trị trong tập dữ liệu;
  - Để tìm trung vị ta phải sắp xếp lại thứ tự dữ liệu và việc này tốn rất nhiều thời gian.

Ưu điểm và nhược điểm của trung vị

- Ưu điểm:
  - Trung vị không bị ảnh hưởng bởi những giá trị ngoại biên;
  - Mọi tập dữ liệu đều có duy nhất một trung vị;
  - Có thể tìm trung vị cho tập dữ liệu sử dụng thang đo thứ bậc, khoảng, tỉ lệ.
- Nhược điểm:
  - Trung vị đánh đồng giữa giá trị với thứ tự của giá trị trong tập dữ liệu;
  - Để tìm trung vị ta phải sắp xếp lại thứ tự dữ liệu và việc này tốn rất nhiều thời gian.

Mode của tập dữ liệu

Khái niệm

Mode của một tập dữ liệu là giá trị xuất hiện nhiều nhất trong tập dữ liệu.

- Cách tìm mode của một tập dữ liệu:
  - Lập bảng tần số cho tập dữ liệu;
  - Tìm giá trị lớn nhất trong các tần số;
  - Tìm các giá trị của tập dữ liệu tương ứng với tần số lớn nhất; và mode của tập dữ liệu là các giá trị này.
- **Ví dụ:** Tìm mode của các tập dữ liệu sau và nhận xét.
  - $0, 1, 3, 1, 5, 2, 6, 2, 9, 2$.
  - $2, 3, 2, 5, 7, 8, 7, 15$.
  - $0, 1, 2, 3, 4, 5, 6$.

Mode của tập dữ liệu

Khái niệm

Mode của một tập dữ liệu là giá trị xuất hiện nhiều nhất trong tập dữ liệu.

- Cách tìm mode của một tập dữ liệu:
  - Lập bảng tần số cho tập dữ liệu;
  - Tìm giá trị lớn nhất trong các tần số;
  - Tìm các giá trị của tập dữ liệu tương ứng với tần số lớn nhất; và mode của tập dữ liệu là các giá trị này.
- **Ví dụ:** Tìm mode của các tập dữ liệu sau và nhận xét.
  - $0, 1, 3, 1, 5, 2, 6, 2, 9, 2$.
  - $2, 3, 2, 5, 7, 8, 7, 15$.
  - $0, 1, 2, 3, 4, 5, 6$.

Mode của tập dữ liệu

Khái niệm

Mode của một tập dữ liệu là giá trị xuất hiện nhiều nhất trong tập dữ liệu.

- Cách tìm mode của một tập dữ liệu:
  - Lập bảng tần số cho tập dữ liệu;
  - Tìm giá trị lớn nhất trong các tần số;
  - Tìm các giá trị của tập dữ liệu tương ứng với tần số lớn nhất; và mode của tập dữ liệu là các giá trị này.
- **Ví dụ:** Tìm mode của các tập dữ liệu sau và nhận xét.
  - $0, 1, 3, 1, 5, 2, 6, 2, 9, 2$.
  - $2, 3, 2, 5, 7, 8, 7, 15$.
  - $0, 1, 2, 3, 4, 5, 6$.

Ưu điểm và nhược điểm của Mode

- Ưu điểm:
  - Mode không bị ảnh hưởng bởi giá trị ngoại biên của tập dữ liệu;
  - Mode là đại lượng thống kê mô tả duy nhất có thể sử dụng cho tập dữ liệu có số đo thuộc thang đo định danh.
- Nhược điểm:
  - Mode chỉ quan tâm đến các giá trị xuất hiện nhiều nhất mà không quan tâm đến những giá trị còn lại của tập dữ liệu;
  - Mode của tập dữ liệu có thể không duy nhất, có những tập dữ liệu có nhiều mode.

Ưu điểm và nhược điểm của Mode

- Ưu điểm:
  - Mode không bị ảnh hưởng bởi giá trị ngoại biên của tập dữ liệu;
  - Mode là đại lượng thống kê mô tả duy nhất có thể sử dụng cho tập dữ liệu có số đo thuộc thang đo định danh.
- Nhược điểm:
  - Mode chỉ quan tâm đến các giá trị xuất hiện nhiều nhất mà không quan tâm đến những giá trị còn lại của tập dữ liệu;
  - Mode của tập dữ liệu có thể không duy nhất, có những tập dữ liệu có nhiều mode.

Các đại lượng mô tả sự phân bố của tập dữ liệu

- Tứ phân vị
- Phân vị thứ p

Tứ phân vị

Khái niệm

Tứ phân vị chia tập dữ liệu đã sắp xếp theo trật tự tăng dần thành bốn phần có số quan sát bằng nhau.

Cách tính tứ phân vị của tập dữ liệu có n phần tử:

- Tứ phân vị thứ nhất kí hiệu là Q_1 là giá trị của quan sát tại vị trí xác định bởi công thức $25\%(n+1)$;
- Tứ phân vị thứ hai kí hiệu là Q_2 chính là trung vị;
- Tứ phân vị thứ ba kí hiệu là Q_3 là giá trị của quan sát tại vị trí xác định bởi công thức $75\%(n+1)$;

Tứ phân vị

Khái niệm

Tứ phân vị chia tập dữ liệu đã sắp xếp theo trật tự tăng dần thành bốn phần có số quan sát bằng nhau.

Cách tính tứ phân vị của tập dữ liệu có n phần tử:

- Tứ phân vị thứ nhất kí hiệu là Q_1 là giá trị của quan sát tại vị trí xác định bởi công thức $25\%(n+1)$;
- Tứ phân vị thứ hai kí hiệu là Q_2 chính là trung vị;
- Tứ phân vị thứ ba kí hiệu là Q_3 là giá trị của quan sát tại vị trí xác định bởi công thức $75\%(n+1)$;

Ví dụ tìm tứ phân vị

Ví dụ: Tìm tứ phân vị của tập dữ liệu: $10, 13, 15, 25, 35, 40, 60, 70$ và nêu nhận xét.

Phân vị thứ p

Khái niệm

Phân vị thứ p của một tập dữ liệu đã được sắp thứ tự là giá trị chia tập dữ liệu thành hai phần, một phần không quá p% số quan sát có giá trị nhỏ hơn phân vị thứ p, phần còn lại có không quá (100 − p)% số quan sát lớn hơn phân vị thứ p.

Nhận xét: Trong nhiều trường hợp ta có thể nói khoảng $p\%$ số quan sát của tập dữ liệu nhỏ hơn hoặc bằng phân vị thứ p và khoảng $(100 - p)\%$ lớn hơn hoặc bằng phân vị thứ p

Cách tính phân vị thứ p của tập dữ liệu có n phần tử: Phân vị thứ p là giá trị có vị trí được xác định bởi công thức $i = \frac{p}{100}(n+1)$.

Phân vị thứ p

Khái niệm

Phân vị thứ p của một tập dữ liệu đã được sắp thứ tự là giá trị chia tập dữ liệu thành hai phần, một phần không quá p% số quan sát có giá trị nhỏ hơn phân vị thứ p, phần còn lại có không quá $(100-p)\%$ số quan sát lớn hơn phân vị thứ p.

Nhận xét: Trong nhiều trường hợp ta có thể nói khoảng $p\%$ số quan sát của tập dữ liệu nhỏ hơn hoặc bằng phân vị thứ p và khoảng $(100-p)\%$ lớn hơn hoặc bằng phân vị thứ p

Cách tính phân vị thứ p của tập dữ liệu có n phần tử: Phân vị thứ p là giá trị có vị trí được xác định bởi công thức $i = \frac{p}{100}(n+1)$.

Ví dụ tính phân vị thứ p

Ví dụ: Tính phân vị thứ 60 của tập dữ liệu sau và nêu nhận xét.

$$15, 32, 42, 65, 87, 92, 100, 105, 110, 120.$$

Các đại lượng mô tả độ phân tán

- Khoảng biến thiên
- Độ trải giữa
- Độ lệch tuyệt đối trung bình
- Phương sai và độ lệch chuẩn

Khoảng biến thiên

Khái niệm

Khoảng biến thiên của một tập dữ liệu là hiệu giữa giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu.

- Cách tính khoảng biến thiên R: $R = x_{\max} - x_{\min}$, trong đó $x_{\max}, x_{\min}$ là giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu.
- **Ví dụ:**
  - Khoảng biến thiên của tập dữ liệu: 1, 2, 3, 4, 5, 6, 7, 8 là $R = 8 - 1 = 7$;
  - Khoảng biến thiên của tập dữ liệu: 1, 1, 1, 1, 8, 8, 8, 8 là $R = 8 - 1 = 7$;
  - Khoảng biến thiên của tập dữ liệu: 1, 2, 3, 4, 5, 6, 7, 8, 100 là $R = 100 - 1 = 99$.
- Ưu điểm và nhược điểm: Khoảng biến thiên là số đo độ phân tán đơn giản và dễ hiểu nhưng nó chỉ phụ thuộc vào giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu rất nhạy cảm với các giá trị ngoại biên và bỏ qua cách phân bố nội bộ của tập dữ liệu.

Khoảng biến thiên

Khái niệm

Khoảng biến thiên của một tập dữ liệu là hiệu giữa giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu.

- Cách tính khoảng biến thiên R: $R = x_{\max} - x_{\min}$, trong đó $x_{\max}, x_{\min}$ là giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu.
- **Ví dụ:**
  - Khoảng biến thiên của tập dữ liệu: 1, 2, 3, 4, 5, 6, 7, 8 là $R = 8 - 1 = 7$;
  - Khoảng biến thiên của tập dữ liệu: 1, 1, 1, 1, 8, 8, 8, 8 là $R = 8 - 1 = 7$;
  - Khoảng biến thiên của tập dữ liệu: 1, 2, 3, 4, 5, 6, 7, 8, 100 là $R = 100 - 1 = 99$.
- Ưu điểm và nhược điểm: Khoảng biến thiên là số đo độ phân tán đơn giản và dễ hiểu nhưng nó chỉ phụ thuộc vào giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu rất nhạy cảm với các giá trị ngoại biên và bỏ qua cách phân bố nội bộ của tập dữ liệu.

Khoảng biến thiên

Khái niệm

Khoảng biến thiên của một tập dữ liệu là hiệu giữa giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu.

- Cách tính khoảng biến thiên R: $R = x_{\max} - x_{\min}$, trong đó $x_{\max}, x_{\min}$ là giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu.
- **Ví dụ:**
  - Khoảng biến thiên của tập dữ liệu: 1, 2, 3, 4, 5, 6, 7, 8 là $R = 8 - 1 = 7$;
  - Khoảng biến thiên của tập dữ liệu: 1, 1, 1, 1, 8, 8, 8, 8 là $R = 8 - 1 = 7$;
  - Khoảng biến thiên của tập dữ liệu: 1, 2, 3, 4, 5, 6, 7, 8, 100 là $R = 100 - 1 = 99$.
- Ưu điểm và nhược điểm: Khoảng biến thiên là số đo độ phân tán đơn giản và dễ hiểu nhưng nó chỉ phụ thuộc vào giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu rất nhạy cảm với các giá trị ngoại biên và bỏ qua cách phân bố nội bộ của tập dữ liệu.

Khoảng biến thiên

Khái niệm

Khoảng biến thiên của một tập dữ liệu là hiệu giữa giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu.

- Cách tính khoảng biến thiên R: $R = x_{\max} - x_{\min}$, trong đó $x_{\max}, x_{\min}$ là giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu.
- **Ví dụ:**
  - Khoảng biến thiên của tập dữ liệu: 1, 2, 3, 4, 5, 6, 7, 8 là $R = 8 - 1 = 7$;
  - Khoảng biến thiên của tập dữ liệu: 1, 1, 1, 1, 8, 8, 8, 8 là $R = 8 - 1 = 7$;
  - Khoảng biến thiên của tập dữ liệu: 1, 2, 3, 4, 5, 6, 7, 8, 100 là $R = 100 - 1 = 99$.
- Ưu điểm và nhược điểm: Khoảng biến thiên là số đo độ phân tán đơn giản và dễ hiểu nhưng nó chỉ phụ thuộc vào giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu rất nhạy cảm với các giá trị ngoại biên và bỏ qua cách phân bố nội bộ của tập dữ liệu.

Độ trải giữa

Khái niệm

Độ trải giữa của một tập dữ liệu là hiệu độ chênh lệch giữa tứ phân vị thứ ba và tứ phân vị thứ nhất của tập dữ liệu.

- Cách tính độ trải giữa R_Q: $R_Q = Q_3 - Q_1$, trong đó Q_3, Q_1 là tứ phân vị thứ ba và thứ nhất của tập dữ liệu.
- Ví dụ: Tập dữ liệu 10, 13, 15, 25, 35, 40, 60, 70 có $Q_1 = 13.5$ và $Q_3 = 55$, từ đó có độ trải giữa $R_Q = Q_3 - Q_1 = 55 - 13.5 = 41.5$.
- Ưu điểm và nhược điểm: Độ trải giữa không phụ thuộc vào giá trị ngoại biên nhưng đánh đồng giữa giá trị với thứ hạng của giá trị của tập dữ liệu.

Độ trải giữa

Khái niệm

Độ trải giữa của một tập dữ liệu là hiệu độ chênh lệch giữa tứ phân vị thứ ba và tứ phân vị thứ nhất của tập dữ liệu.

- Cách tính độ trải giữa R_Q: $R_Q = Q_3 - Q_1$, trong đó Q_3, Q_1 là tứ phân vị thứ ba và thứ nhất của tập dữ liệu.
- Ví dụ: Tập dữ liệu 10, 13, 15, 25, 35, 40, 60, 70 có $Q_1 = 13.5$ và $Q_3 = 55$, từ đó có độ trải giữa $R_Q = Q_3 - Q_1 = 55 - 13.5 = 41.5$.
- Ưu điểm và nhược điểm: Độ trải giữa không phụ thuộc vào giá trị ngoại biên nhưng đánh đồng giữa giá trị với thứ hạng của giá trị của tập dữ liệu.

Độ lệch tuyệt đối trung bình

Khái niệm

Độ lệch tuyệt đối trung bình của một tập dữ liệu gồm n quan sát có giá trị $x_1, x_2, \ldots, x_n$ *với trung bình* $\overline{x}$ *được cho bởi công thức:*

$$\frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}.$$

Ưu điểm và nhược điểm của độ lệch tuyệt đối trung bình:

- Ưu điểm: Độ lệch tuyệt đối trung bình căn cứ vào mọi điểm của tập dữ liệu và chỉ ra một cách trung bình mỗi điểm dữ liệu nằm cách xa trung bình bao nhiêu.
- Nhược điểm: Công thức tính độ lệch tuyệt đối trung bình dùng đến giá trị tuyệt đối nên khó thực hiện các phép biến đổi toán học.

Độ lệch tuyệt đối trung bình

Khái niệm

Độ lệch tuyệt đối trung bình của một tập dữ liệu gồm n quan sát có giá trị $x_1, x_2, \ldots, x_n$ *với trung bình* $\overline{x}$ *được cho bởi công thức:*

$$\frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}.$$

Ưu điểm và nhược điểm của độ lệch tuyệt đối trung bình:

- Ưu điểm: Độ lệch tuyệt đối trung bình căn cứ vào mọi điểm của tập dữ liệu và chỉ ra một cách trung bình mỗi điểm dữ liệu nằm cách xa trung bình bao nhiêu.
- Nhược điểm: Công thức tính độ lệch tuyệt đối trung bình dùng đến giá trị tuyệt đối nên khó thực hiện các phép biến đổi toán học.

Độ lệch tuyệt đối trung bình

Ví dụ: Cho tập dữ liệu về số tiền chi tiêu thực phẩm trong một tuần: 120, 150, 125, 100, 180, 140, 200. Trung bình của số tiền chi tiêu trong một ngày là $\overline{x} = 145$. Khi đó độ lệch tuyệt đối trung bình cho số tiền chi tiêu thực phẩm là:

$$\begin{aligned} A &= \frac{|120-145| + |150-145| + |125-145| + |100-145|}{7} \\ &\quad + \frac{|180-145| + |140-145| + |200-145|}{7} \\ &= 27.14. \end{aligned}$$

Khái niệm về phương sai và độ lệch chuẩn

Khái niệm

- *Phương sai của một tập dữ liệu tổng thể, kí hiệu là σ^2, được xác định bởi công thức: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$, ở đây μ là trung bình của tổng thể và N là số quan sát trong tổng thể.*
- *Phương sai của một tập dữ liệu mẫu, kí hiệu là s^2, được xác định bởi công thức: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$, ở đây $\overline{x}$ là trung bình của mẫu và n là số quan sát trong mẫu.*

Khái niệm về phương sai và độ lệch chuẩn

Khái niệm

- *Độ lệch chuẩn của một tập dữ liệu tổng thể, kí hiệu là σ, là căn bậc hai của phương sai của tổng thể:*

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}.$$

ii. *Độ lệch chuẩn của một tập dữ liệu mẫu, kí hiệu là s, là căn bậc hai của phương sai mẫu:*

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}.$$

Ví dụ tính phương sai và độ lệch chuẩn

Ví dụ: Cho tập dữ liệu

$$10, 15, 32, 18, 25, 65, 30, 38.$$

Tính phương sai và độ lệch chuẩn trong cả hai trường hợp tập dữ liệu dạng tổng thể và dạng mẫu.

Trung bình, Phương sai và Độ lệch chuẩn của tập dữ liệu thu gọn

Khái niệm

Cho một tập dữ liệu thu gọn (cho dưới dạng bảng tần số) với x_i là giá trị quan sát thứ i hoặc giá trị đại diện của tổ thứ i, f_i là tần số của quan sát hoặc tổ thứ i và số phân tử của tập dữ liệu $n = \sum_{i=1}^{k} f_i$. Khi đó

- *Trung bình cộng của dữ liệu thu gọn được cho bởi công thức:*

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}.$$

- *Phương sai của một tập dữ liệu tổng thể thu gọn được xác định bởi công thức:*

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \mu)^2 f_i}{\sum_{i=1}^{k} f_i}.$$

Trung bình, Phương sai và Độ lệch chuẩn của tập dữ liệu thu gọn

Khái niệm

- *Phương sai của một tập dữ liệu mẫu thu gọn được xác định bởi công thức:*

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1},$$

- *Độ lệch chuẩn của một tập dữ liệu tổng thể (mẫu) được tính bằng căn bậc hai của phương sai tổng thể (mẫu):*

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \mu)^2 f_i}{\sum_{i=1}^{k} f_i}} \qquad s = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1}}.$$

Ví dụ tính trung bình, phương sai của tập dữ liệu thu gọn

Tính điểm trung bình, phương sai, độ lệch chuẩn của điểm toán khối A dựa trên bảng tần số thu gọn sau:

Khoảng điểm	**Điểm đại diện** (x_i^*)	**Tần số** (f_i)
[0.0, 1.5]	0.75	1317
(1.5, 3.0]	2.25	967
(3.0, 4.5]	3.75	518
(4.5, 6.0]	5.25	151
(6.0, 7.5]	6.75	18
(7.5, 9.0]	8.25	3

Lời giải

Số đo hướng tâm đối với các loại thang đo

- Đối với thang đo định danh, số đo hướng tâm duy nhất là mode.
- Đối với thang đo thứ bậc, số đo hướng tâm là mode và trung vị. Trung vị cho nhiều thông tin hơn mode.
- Đối với thang đo khoảng và thang đo tỉ lệ, số đo hướng tâm có thể là mode, trung vị và trung bình cộng, trong đó trung bình cộng chứa nhiều thông tin nhất.

Số đo độ phân bố đối với các loại thang đo

- Đối với thang đo định danh không thể tính các số đo độ phân bố như tứ phân vị hay phân vị thứ p.
- Đối với thang đo thứ bậc, thang đo khoảng và thang đo tỉ lệ có thể tính các số đo độ phân bố như tứ phân vị và phân vị thứ p.

Số đo độ phân tán đối với các loại thang đo

- Đối với thang đo định danh và thang đo thứ bậc độ phân tán thống kê có thể đo bằng các tỉ lệ, tức là tần suất xuất hiện của các biểu hiện, không tính được khoảng biến thiên, độ trải giữa, phương sai và độ lệch chuẩn.
- Đối với thang đo khoảng và thang đo tỉ lệ, tất cả các đại lượng: khoảng biến thiên, độ trải giữa, độ lệch tuyệt đối trung bình, phương sai và độ lệch chuẩn đều có thể dùng để đo độ phân tán, trong đó phương sai và độ lệch chuẩn là hai số đo độ phân tán tốt nhất.

Những hàm tính các đại lượng thống kê mô tả trên R

`mean(x)`	tính trung bình cộng của véc tơ `x`
`median(x)`	tính trung vị của véc tơ `x`
`which(table(x)` `==` `max(table(x)))`	tính mode của véc tơ `x`
`summary(x)`	cho các giá trị lớn nhất, nhỏ nhất, tứ phân vị, trung bình của véc tơ `x`
`quantile(x)`	tính phân vị tùy ý của véc tơ `x`
`range(x)`	cho giá trị nhỏ nhất và lớn nhất của véc tơ `x`
`var(x)`	cho phương sai của véc tơ `x`
`sd(x)`	cho độ lệch chuẩn của véc tơ `x`

Tính trung bình và phương sai của tập dữ liệu thu gọn trên R

Để tính trung bình, phương sai, độ lệch chuẩn của điểm thi Toán khối A năm 2008 cho dưới dạng bảng tần số thu gọn (phân tổ dữ liệu) trên R ta có thể làm như sau:

- Nhập dữ liệu về Điểm và Tần số:

```
> Diem = c(0.75, 2.25, 3.75, 5.25, 6.75, 8.25)
> TanSo = c(1317, 967, 518, 151, 18, 3)
```

- Tính điểm trung bình:

```
> tb = sum(Diem*TanSo)/sum(TanSo)
> tb
[1] 2.032616
```

Tính trung bình và phương sai của tập dữ liệu thu gọn trên R

- Tính Phương sai:

  $ps = sum((Diem - tb)^2 * TanSo)/sum(TanSo)$

```
> ps
[1]     1.956863
```

- Tính Độ lệch chuẩn:

```
 dlc = sqrt(ps)
> dlc
[1]     1.398879
```

Tính trung bình và phương sai của tập dữ liệu thu gọn trên R

Hoặc cũng có thể tính như sau:

- Nhập dữ liệu về Điểm và Tần số:

```
> Diem = c(0.75, 2.25, 3.75, 5.25, 6.75, 8.25)
> TanSo = c(1317, 967, 518, 151, 18, 3)
```

- Tạo dãy điểm ban đầu:

```
> DayDiem = rep(Diem,TanSo)
```

Tính trung bình và phương sai của tập dữ liệu thu gọn trên R

- Tính Trung bình, phương sai, độ lệch chuẩn:

```
> tb = mean(DayDiem)
> tb
[1]     2.032616
 ps = var(DayDiem)
> ps
[1]     1.957521
```

- Tính Độ lệch chuẩn:

```
dlc = sd(DayDiem))
> dlc
[1]      1.399115
```

Biểu đồ hộp và râu

Khái niệm

Biểu đồ hộp và râu là một biểu đồ nghiên cứu về độ hướng tâm, độ phân tán và phân phối của một tập dữ liệu định lượng. Biểu đồ hộp và râu cho ta các thông tin về: giá trị cực đại, giá trị cực tiểu, ba tứ phân vị và các quan sát ngoại lệ của tập dữ liệu.

Để vẽ biểu đồ hộp và râu của một tập dữ liệu, ta phải tính một số đại lượng thống kê mô tả sau:

- Trung bình, trung vị của tập dữ liệu;
- Tứ phân vị thứ nhất (Q_1), tứ phân vị thứ ba (Q_3), độ trải giữa $R_Q = Q_3 - Q_1$ của tập dữ liệu;
- Giá trị nhỏ nhất, lớn nhất và các giá trị ngoại biên (nếu có) của tập dữ liệu.

Biểu đồ hộp và râu

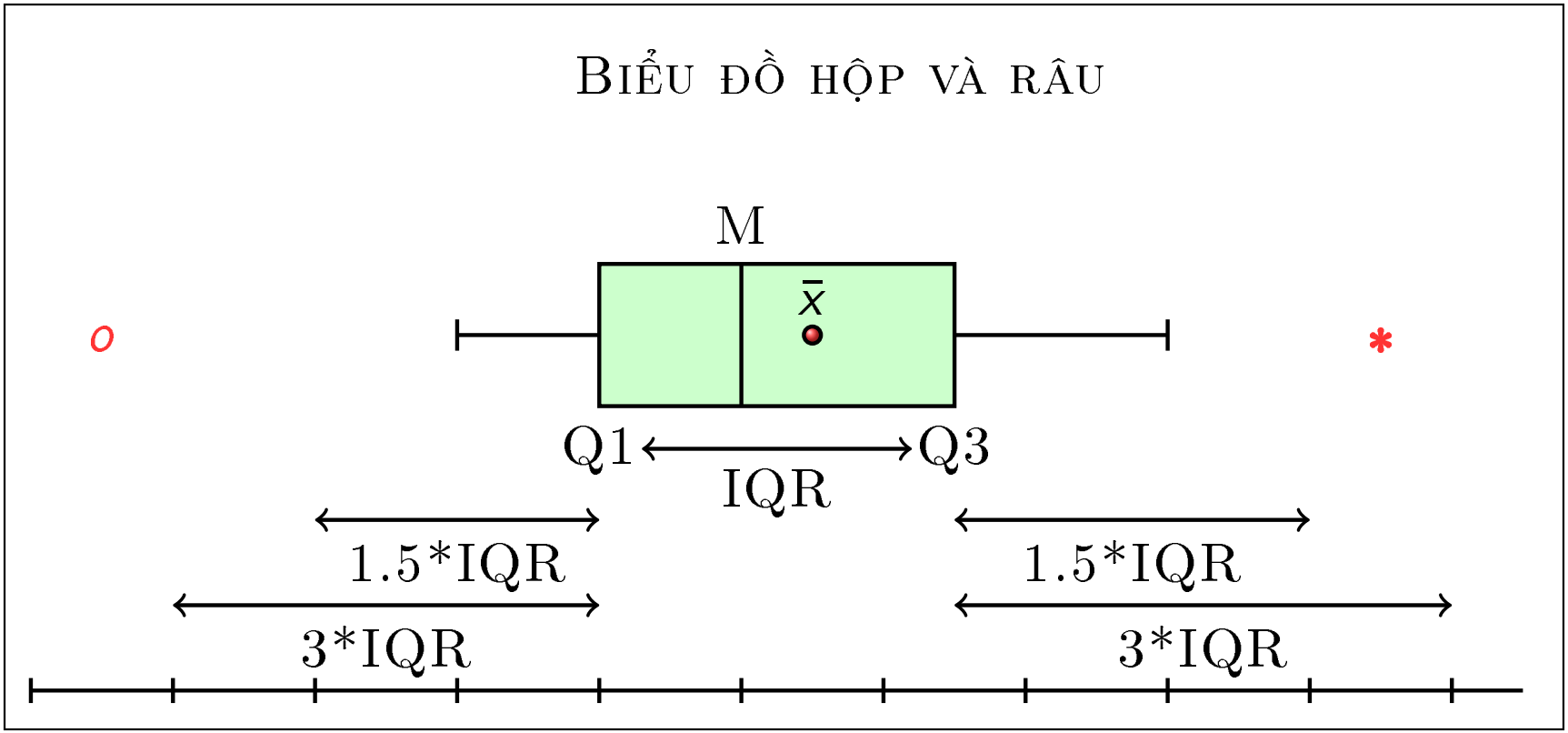
Khái niệm

Biểu đồ hộp và râu là một biểu đồ nghiên cứu về độ hướng tâm, độ phân tán và phân phối của một tập dữ liệu định lượng. Biểu đồ hộp và râu cho ta các thông tin về: giá trị cực đại, giá trị cực tiểu, ba tứ phân vị và các quan sát ngoại lệ của tập dữ liệu.

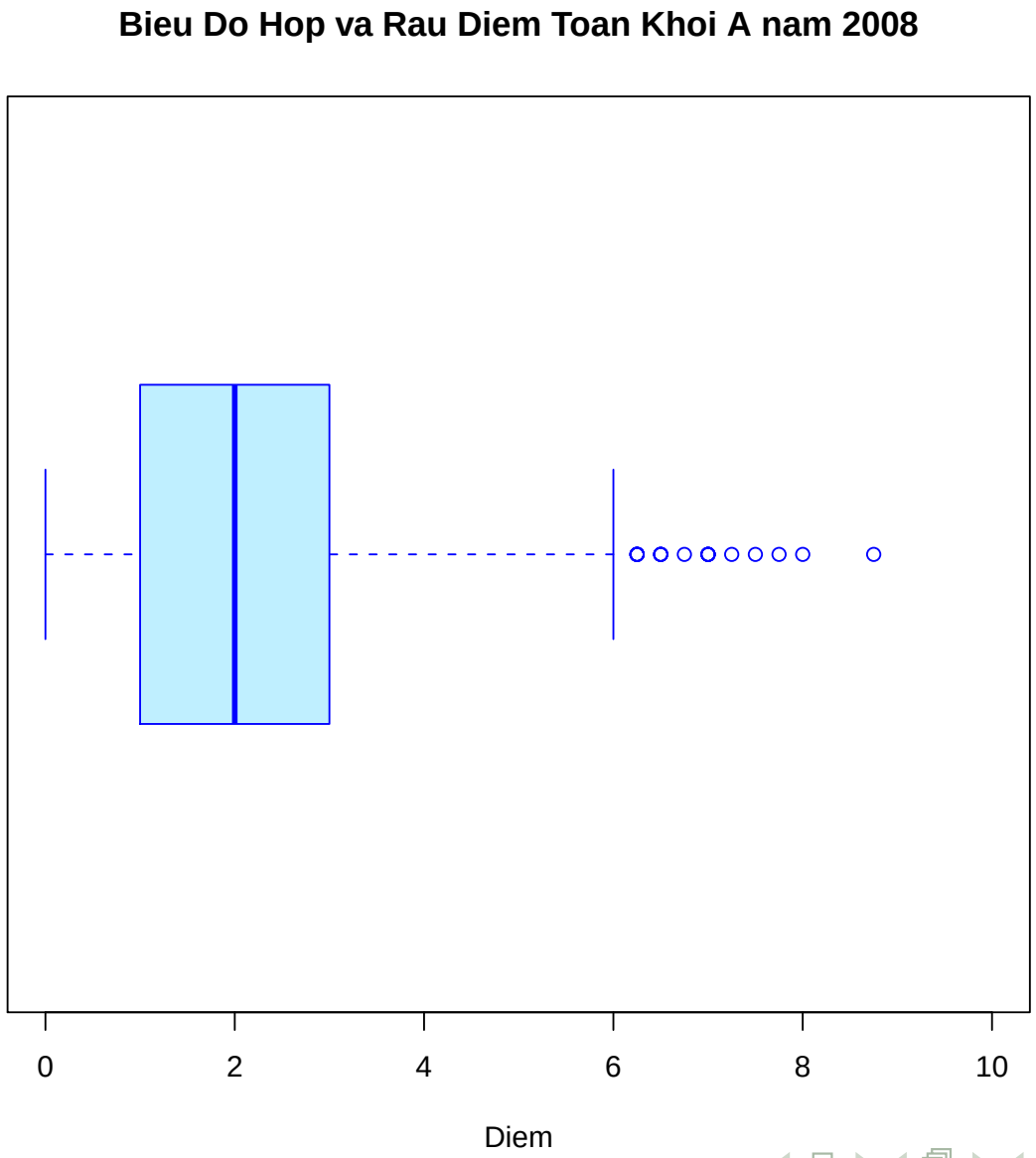
Để vẽ biểu đồ hộp và râu của một tập dữ liệu, ta phải tính một số đại lượng thống kê mô tả sau:

- Trung bình, trung vị của tập dữ liệu;
- Tứ phân vị thứ nhất (Q_1), tứ phân vị thứ ba (Q_3), độ trải giữa $R_Q = Q_3 - Q_1$ của tập dữ liệu;
- Giá trị nhỏ nhất, lớn nhất và các giá trị ngoại biên (nếu có) của tập dữ liệu.

Minh họa biểu đồ hộp và râu



Ví dụ về biểu đồ hộp và râu



Vẽ biểu đồ hộp và râu trong R bằng hàm boxplot

`boxplot(x, border = "", col = "", horizontal=FALSE)`
trong đó,

- `x` véc tơ dữ liệu số cần vẽ biểu đồ;
- `border` màu của râu, đường biên của hộp và giá trị ngoại biên;
- `col` màu của hộp;
- `horizontal` nếu `horizontal=FALSE` thì biểu đồ được vẽ đứng, nếu `horizontal=TRUE` thì biểu đồ được vẽ ngang.

Sử dụng kết hợp trung bình và độ lệch chuẩn

- Qui tắc thực nghiệm;
- Định lí Chebyshev;
- Hệ số biến thiên.

Qui tắc thực nghiệm

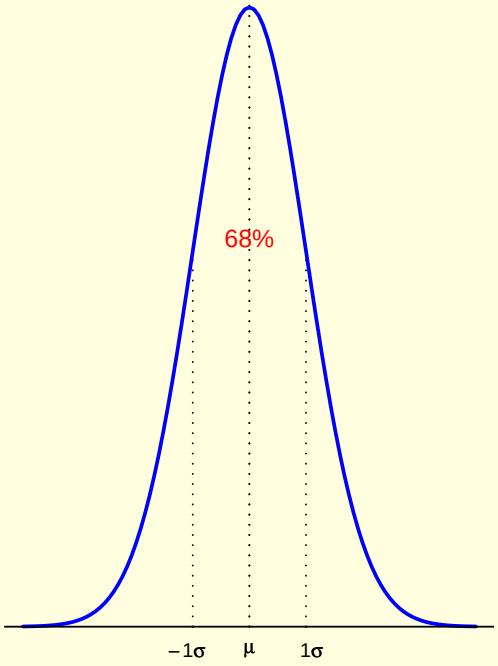
Định lí (Qui tắc thực nghiệm)

Đối với một tập dữ liệu tuân theo phân phối chuẩn, qui tắc thực nghiệm cho ta biết tỉ lệ phần trăm các giá trị nằm trong vòng một số lần độ lệch chuẩn tính từ trung bình, cụ thể ta có các kết luận sau:

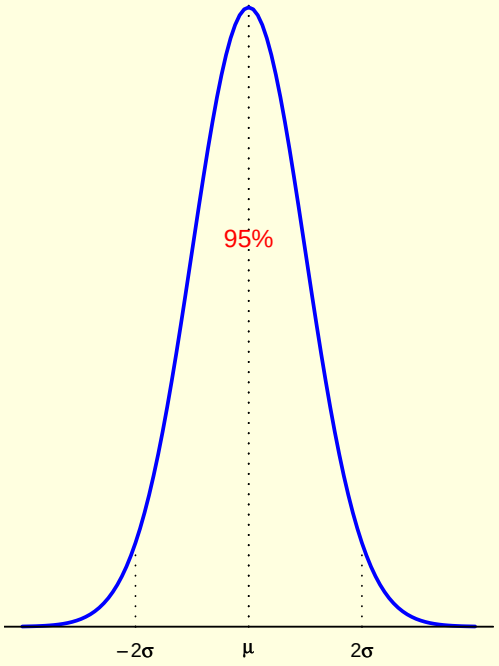
- *Có khoảng* 68% *số quan sát của tập dữ liệu tập trung trong khoảng* $[\mu - \sigma, \mu + \sigma]$.
- *Có khoảng* 95% *số quan sát của tập dữ liệu tập trung trong khoảng* $[\mu - 2\sigma, \mu + 2\sigma]$.
- *Có khoảng* 99.7% *số quan sát của tập dữ liệu tập trung trong khoảng* $[\mu - 3\sigma, \mu + 3\sigma]$.

Qui tắc thực nghiệm

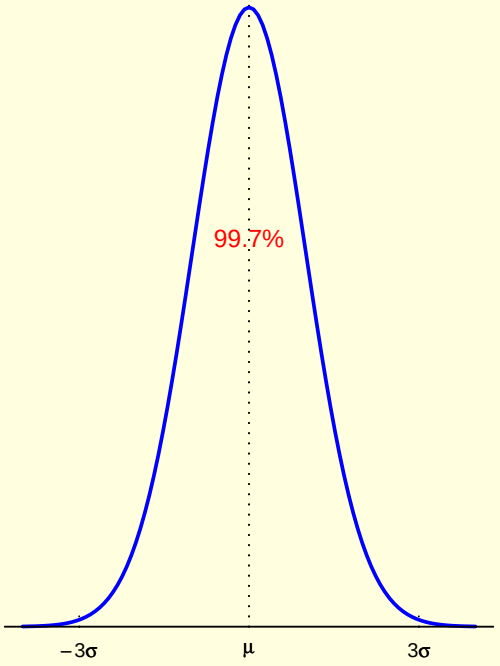
Empirical Rule for one standard Deviation



Empirical Rule for two standard Deviations



Empirical Rule for three standard Deviations



Ví dụ áp dụng qui tắc thực nghiệm

Ví dụ: Chỉ số IQ của con người được coi là tuân theo phân phối chuẩn với trung bình $\mu = 100$ và độ lệch chuẩn $\sigma = 15$. Dựa vào qui tắc thực nghiệm hãy đưa ra những nhận xét về chỉ số IQ của con người.

Định lí Chebyshev

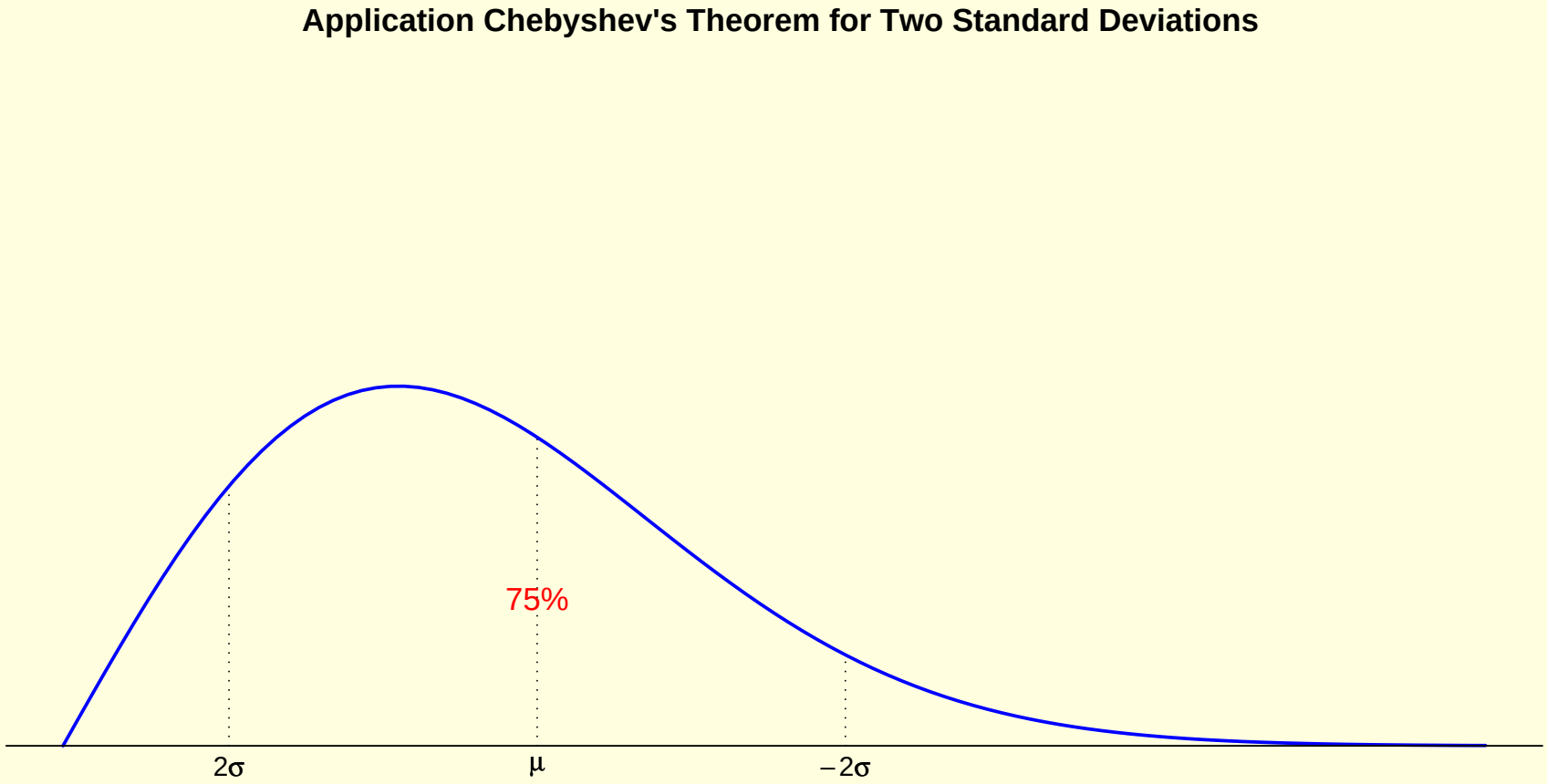
Định lí (Định lí Chebyshev)

Đối với một tập dữ liệu có phân phối tùy ý, định lí Chebyshev cho ta biết khi $k > 1$ thì trong khoảng $[\mu - k\sigma, \mu + k\sigma]$ chứa ít nhất $1 - 1/k^2$ giá trị của tập dữ liệu.

Áp dụng định lí Chebyshev cho một số giá trị k ta có những kết quả sau:

- Với $k = 2$, có ít nhất 75% số quan sát của tập dữ liệu tập trung trong khoảng $[\mu - 2\sigma, \mu + 2\sigma]$.
- Với $k = 2.5$, có ít nhất 84% số quan sát của tập dữ liệu tập trung trong khoảng $[\mu - 2.5\sigma, \mu + 2.5\sigma]$.
- Với $k = 3$, có ít nhất 89% số quan sát của tập dữ liệu tập trung trong khoảng $[\mu - 3\sigma, \mu + 3\sigma]$.

Định lí Chebyshev



Ví dụ áp dụng định lí Chebyshev

Ví dụ: Thu nhập của dân chúng một vùng có trung bình là $\mu = 50$, độ lệch chuẩn $\sigma = 10$. Dựa trên định lí Chebyshev hãy đưa ra những nhận định cho thu nhập của dân cư vùng này.

Câu hỏi tình huống chương 4

Câu hỏi tình huống

Kĩ thuật viên A hoàn thành trung bình mỗi quí 40 phân tích thí nghiệm với độ lệch chuẩn là 5, kĩ thuật viên B hoàn thành trung bình mỗi quí 160 phân tích với độ lệch chuẩn là 15. Ai tỏ ra ít biến động hơn trong sản xuất?

Hệ số biến thiên

Khái niệm

Hệ số biến thiên của một tập dữ liệu là tỉ số (tính bằng phần trăm) của độ lệch chuẩn và trung bình. Hệ số biến thiên dùng để so sánh độ biến thiên của hai tập dữ liệu khi trung bình của chúng khác nhau.

Công thức tính hệ số biến thiên CV:

- Hệ số biến thiên của dữ liệu tổng thể: $CV = \frac{\sigma}{\mu}.100\%$;
- Hệ số biến thiên của dữ liệu mẫu: $CV = \frac{s}{\overline{x}}.100\%$;

Ví dụ tính hệ số biến thiên

Ví dụ: Giả sử trong 5 tuần, giá của cổ phiếu A là: 22, 28, 20, 27, 25 và giá của cổ phiếu B là: 40, 50, 48, 55, 45. Hãy xét xem giá của cổ phiếu nào biến thiên nhiều hơn.

Lời giải:

Ta có: $\mu_A = 24.4$, $\sigma_A = 3.36$ và $\mu_B = 47.6$, $\sigma_B = 5.60$. Do giá trung bình của hai loại cổ phiếu khác nhau nên ta dùng hệ số biến thiên để đưa ra kết luận về sự biến thiên về giá của hai loại cổ phiếu.

$$CV_A = \frac{3.36}{24.4}.100\% = 13.77\%; \qquad \frac{5.60}{47.6}.100\% = 11.77\%.$$

Vì $CV_A > CV_B$ nên giá của cổ phiếu A biến thiên nhiều hơn giá của cổ phiếu B.

Các đại lượng miêu tả hình dáng của tập dữ liệu

- Hệ số bất đối xứng: Skewness.
- Hệ số đo độ nhọn Kurtosis

Skewness

Khái niệm

Skewness là đại lượng đo lường mức độ lệch của phân phối, còn được gọi là hệ số bất đối xứng.

- Hệ số Skewness S_k có thể tính bằng công thức: $S_k = \frac{\mu - M_d}{\sigma}$, trong đó μ, σ, M_d tương ứng là trung bình, độ lệch chuẩn, trung vị của một tập dữ liệu.
- Skewness và mối liên hệ với trung bình, trung vị và mode.
  - Nếu phân phối của tập dữ liệu là đối xứng thì hệ số skewness bằng không và khi đó trung bình, trung vị và mode trùng nhau;
  - Nếu phân phối của tập dữ liệu tập trung ở bên trái thì hệ số skewness dương và mode $<$ trung vị $<$ trung bình;
  - Nếu phân phối của tập dữ liệu tập trung bên phải thì hệ số skewness âm và trung bình $<$ trung vị $<$ mode.

Skewness

Khái niệm

Skewness là đại lượng đo lường mức độ lệch của phân phối, còn được gọi là hệ số bất đối xứng.

- Hệ số Skewness S_k có thể tính bằng công thức: $S_k = \frac{\mu - M_d}{\sigma}$, trong đó μ, σ, M_d tương ứng là trung bình, độ lệch chuẩn, trung vị của một tập dữ liệu.
- Skewness và mối liên hệ với trung bình, trung vị và mode.
  - Nếu phân phối của tập dữ liệu là đối xứng thì hệ số skewness bằng không và khi đó trung bình, trung vị và mode trùng nhau;
  - Nếu phân phối của tập dữ liệu tập trung ở bên trái thì hệ số skewness dương và mode $<$ trung vị $<$ trung bình;
  - Nếu phân phối của tập dữ liệu tập trung bên phải thì hệ số skewness âm và trung bình $<$ trung vị $<$ mode.

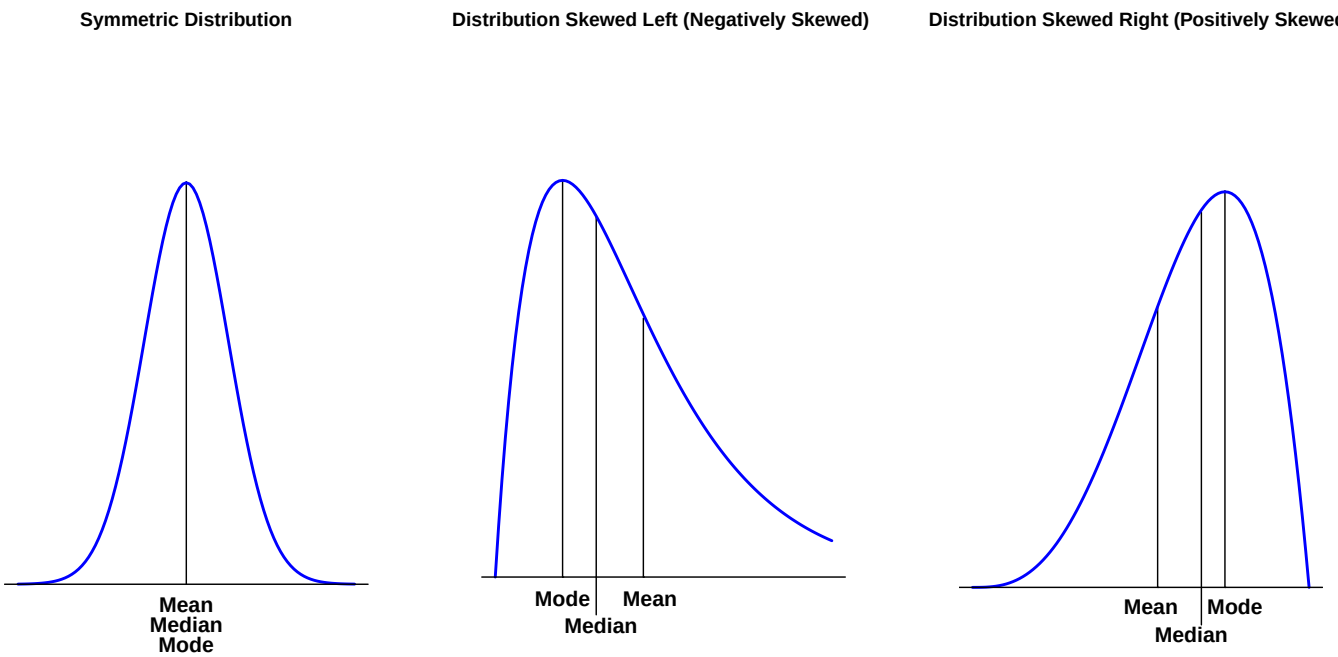
Skewness

Khái niệm

Skewness là đại lượng đo lường mức độ lệch của phân phối, còn được gọi là hệ số bất đối xứng.

- Hệ số Skewness S_k có thể tính bằng công thức: $S_k = \frac{\mu - M_d}{\sigma}$, trong đó μ, σ, M_d tương ứng là trung bình, độ lệch chuẩn, trung vị của một tập dữ liệu.
- Skewness và mối liên hệ với trung bình, trung vị và mode.
  - Nếu phân phối của tập dữ liệu là đối xứng thì hệ số skewness bằng không và khi đó trung bình, trung vị và mode trùng nhau;
  - Nếu phân phối của tập dữ liệu tập trung ở bên trái thì hệ số skewness dương và mode $<$ trung vị $<$ trung bình;
  - Nếu phân phối của tập dữ liệu tập trung bên phải thì hệ số skewness âm và trung bình $<$ trung vị $<$ mode.

Mô tả hình học hệ số Skewness



Kurtosis

Khái niệm

Kurtosis là đại lượng đo mức độ tập trung tương đối của các quan sát quanh trung tâm của nó trong mối quan hệ so sánh với hai đuôi.

Hệ số Kurtosis và hình dáng của tập dữ liệu:

- Khi phân phối tập trung ở mức độ bình thường thì hệ số Kurtosis $= 3$;
- Khi phân phối tập trung hơn mức bình thường (hình dáng của phân phối cao và nhọn với hai đuôi hẹp) thì Kurtosis > 3;
- Khi phân phối không tập trung như mức bình thường (hình dáng của phân phối phẳng và trải dài) thì Kurtosis < 3.

Kurtosis

Khái niệm

Kurtosis là đại lượng đo mức độ tập trung tương đối của các quan sát quanh trung tâm của nó trong mối quan hệ so sánh với hai đuôi.

Hệ số Kurtosis và hình dáng của tập dữ liệu:

- Khi phân phối tập trung ở mức độ bình thường thì hệ số Kurtosis $= 3$;
- Khi phân phối tập trung hơn mức bình thường (hình dáng của phân phối cao và nhọn với hai đuôi hẹp) thì Kurtosis > 3;
- Khi phân phối không tập trung như mức bình thường (hình dáng của phân phối phẳng và trải dài) thì Kurtosis < 3.

Mô tả hình học hệ số Kurtosis