

TỔNG HỢP VÀ TRƯC QUAN HÓA DỮ LIỆU

NỘI DUNG

- Một số đặc tính của dữ liệu
- Đồ thị Stem & Leaf
- Phân phối tần số
- Histograms
- Các dạng đồ thị khác

NỘI DUNG

- **Một số đặc tính của dữ liệu**
- Đồ thị Stem & Leaf
- Phân phối tần số
- Histograms
- Các dạng đồ thị khác

Một số đặc tính của dữ liệu

- **Độ tập trung** (*central tendency*): thể hiện vị trí mà phần lớn tập dữ liệu tập trung
- **Độ phân tán** (*variation*): thể hiện sự phân tán của các giá trị dữ liệu
- **Phân phối** (*phân bố*): hình dạng của dữ liệu khi sắp xếp theo giá trị.
- **Giá trị ngoại lệ** (*outliers*): các giá trị nằm cách xa so với hầu hết các giá trị khác trong tập dữ liệu.
- **Thời gian** (*time*): sự thay đổi đặc tính của dữ liệu theo thời gian

NỘI DUNG

- Một số đặc tính của dữ liệu
- Đồ thị Stem & Leaf
- Phân phối tần số
- Histograms
- Các dạng đồ thị khác

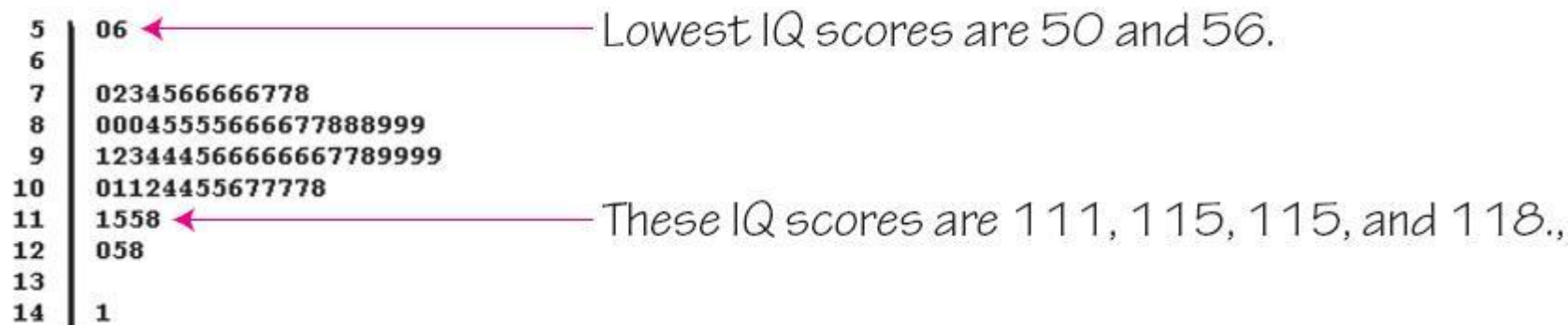
- Khi tập dữ liệu mẫu được thu thập về, thông thường chúng ta phải thực hiện tính toán, và biến đổi một chút để có thể biết được các đặc tính của chúng.
- Tuy nhiên, việc thay đổi dữ liệu cần phải thực hiện cẩn thận để tránh làm mất mát thông tin mà dữ liệu chứa đựng.
- Để có cái nhìn ban đầu về dữ liệu, mà không làm thay đổi chúng, ta có thể sử dụng đồ thị **stem & leaf**

NỘI DUNG

- Một số đặc tính của dữ liệu
- **Đồ thị Stem & Leaf**
- Phân phối tần số
- Histograms
- Các dạng đồ thị khác

Đồ thị Stem & Leaf

- Đồ thị Stem & Leaf biểu diễn dữ liệu định lượng bằng cách tách giá trị dữ liệu thành hai phần: phần thân/**the stem** (chẳng hạn chữ số trái nhất), và phần lá/**the leaf** (chẳng hạn chữ số ngoài cùng bên phải)



- Ngoài ra, để hiểu các đặc tính của dữ liệu, chúng ta có thể tổ chức và tổng hợp để xây dựng bảng phân phối tần số của dữ liệu.

NỘI DUNG

- Một số đặc tính của dữ liệu
- Đồ thị Stem & Leaf
- **Phân phối tần số**
- Histograms
- Các dạng đồ thị khác

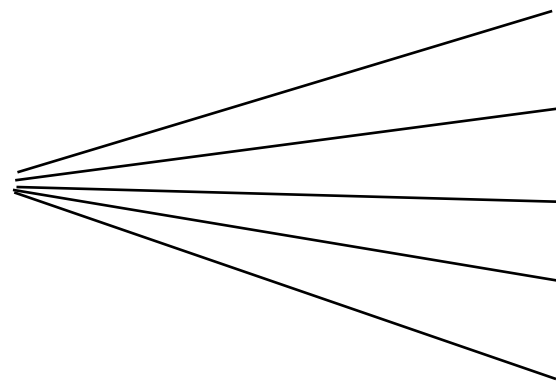
Phân phối tần số

- **Phân phối tần số** (frequency table): dùng để hiển thị phân vùng của các lớp của dữ liệu bằng cách liệt kê tất cả các lớp dữ liệu và số lần xuất hiện (tần số) tương ứng

IQ Scores of Low Lead Group

Lower Class Limits

are the smallest numbers that can actually belong to different classes.

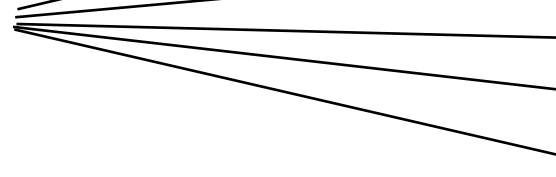


IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1

IQ Scores of Low Lead Group

Upper Class Limits

are the largest numbers that can actually belong to different classes.



IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1

IQ Scores of Low Lead Group

Class Boundaries

are the numbers used to separate classes, but without the gaps created by class limits.

49.5

69.5

89.5

109.5

129.5

149.5

IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1

IQ Scores of Low Lead Group

Class Midpoints

are the values in the middle of the classes and can be found by adding the lower class limit to the upper class limit and dividing the sum by 2.

59.5

79.5

99.5

119.5

139.5

IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1

IQ Scores of Low Lead Group

Class Width

is the difference between two consecutive lower class limits or two consecutive lower class boundaries.

20

20

20

20

20

IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1

Phân phối tần số

➤ Lý do sử dụng bảng phân phối tần số:

1. Có thể tổng hợp được tập dữ liệu lớn
2. Có thể phân tích tính tự nhiên của dữ liệu
3. Có cơ sở để xây dựng các đồ thị khác

Phân phối tần số

➤ Cách xây dựng một bảng phân phối tần số:

1. Xác định số lớp (thông thường từ 5-20)
2. Tính độ rộng của lớp

$$\text{class width} \approx \frac{(\text{maximum value}) - (\text{minimum value})}{\text{number of classes}}$$

3. Chọn giá trị bắt đầu (giá trị nhỏ nhất hoặc một giá trị thuận lợi nào đó)
4. Tính toán các lớp sử dụng cận dưới và độ rộng của lớp
5. Liệt kê các lớp theo hàng dọc
6. Điền các giá trị tần số.

Ví dụ:

Sử dụng dữ liệu nhịp tim của nữ trong bảng sau, xây dựng bảng tần số cho dữ liệu trên, sử dụng 7 lớp.

Females																			
76	72	88	60	72	68	80	64	68	68	80	76	68	72	96	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64

Giải:

Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1

Phân phối tần số tương đối

- Giống như phân phối tần số, nhưng tần số của lớp được thay bằng tỷ lệ của lớp so với toàn bộ dữ liệu

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$


$$\text{percentage frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}} \times 100\%$$

Relative Frequency Distribution

IQ Score	Frequency	Relative Frequency
50-69	2	2.6%
70-89	33	42.3%
90-109	35	44.9%
110-129	7	9.0%
130-149	1	1.3%

Cumulative Frequency Distribution

IQ Score	Frequency	Cumulative Frequency
50-69	2	2
70-89	33	35
90-109	35	70
110-129	7	77
130-149	1	78



Cumulative Frequencies

- Sau khi tính toán được bảng phân phối tần số, ta dùng **histogram** để phân tích hình dạng của phân phối.

NỘI DUNG

- Một số đặc tính của dữ liệu
- Đồ thị Stem & Leaf
- Phân phối tần số
- **Histograms**
- Các dạng đồ thị khác

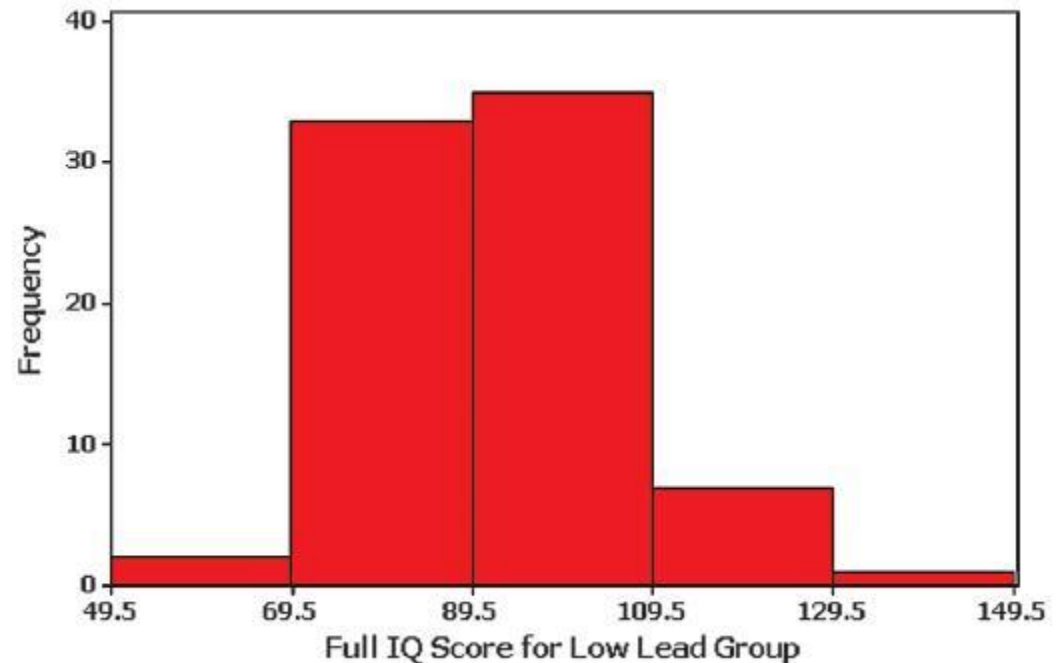
Histograms

- Histograms: là đồ thị gồm các cột có độ rộng bằng như nhau nằm cạnh nhau.
- Trục hoành thể hiện giá trị của lớp
- Trục tung thể hiện tần số của lớp
- Chiều cao của các cột tương ứng với tần số của lớp

Example

IQ scores from children with low levels of lead.

IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1



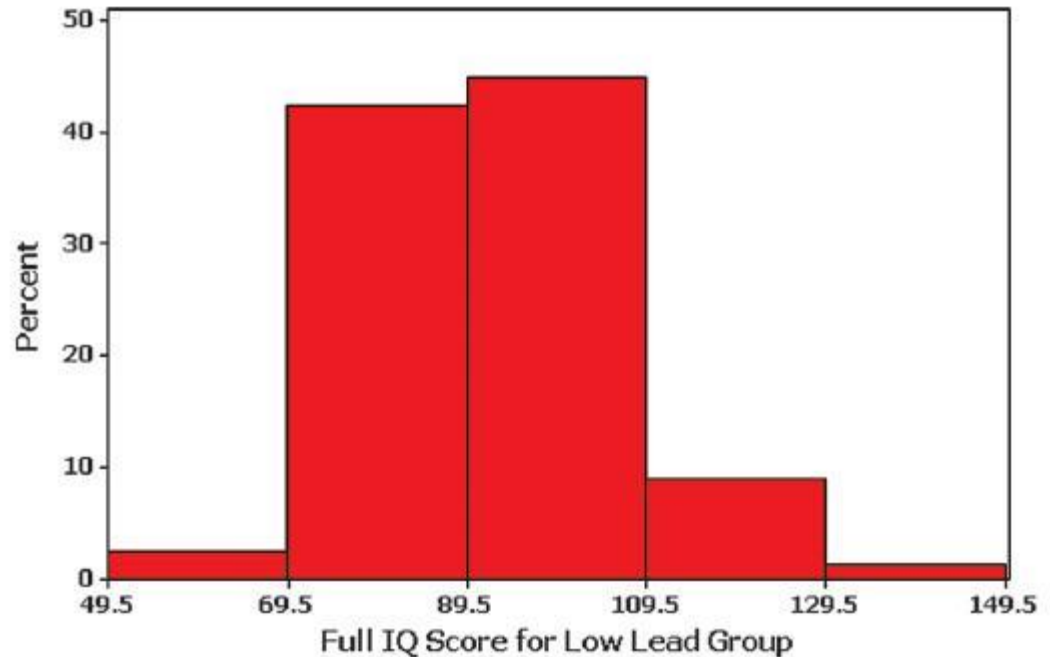
Histograms

- Hiểu một cách đơn giản: histogram là hình vẽ của bảng phân phối tần số.
- Histograms có thể được vẽ bằng các phần mềm.

Relative Frequency Histogram

Có hình dạng: trục hoành giống như histogram, nhưng trục tung được đánh dấu bằng tần số tương đối (tần suất) thay vì tần số

IQ Score	Relative Frequency
50-69	2.6%
70-89	42.3%
90-109	44.9%
110-129	9.0%
130-149	1.3%

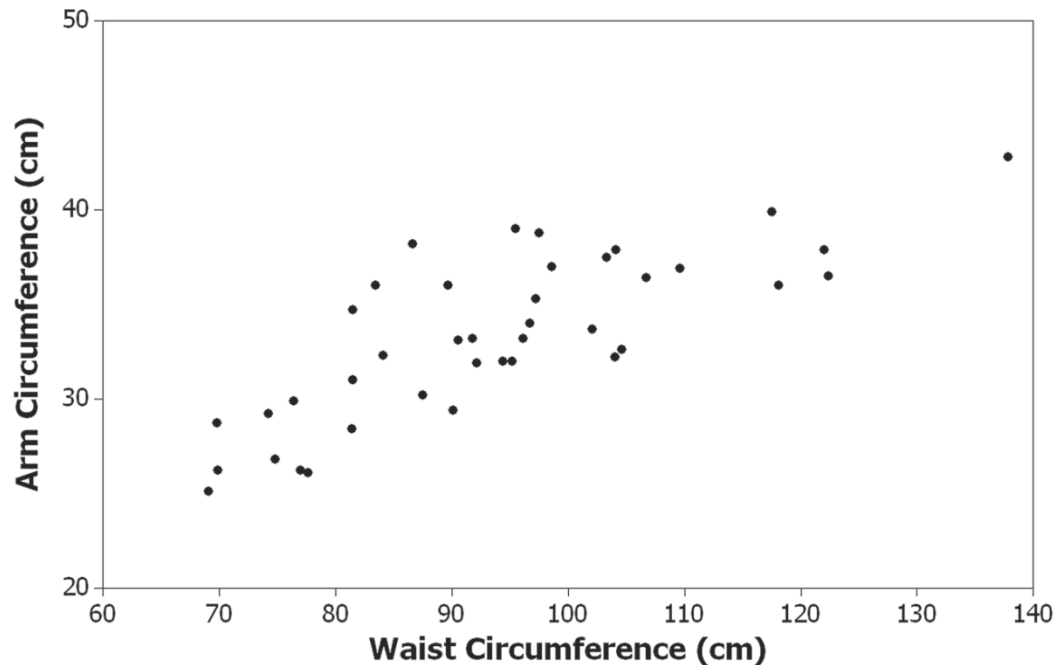


NỘI DUNG

- Một số đặc tính của dữ liệu
- Đồ thị Stem & Leaf
- Phân phối tần số
- Histograms
- **Các dạng đồ thị khác**

Scatterplot (or Scatter Diagram)

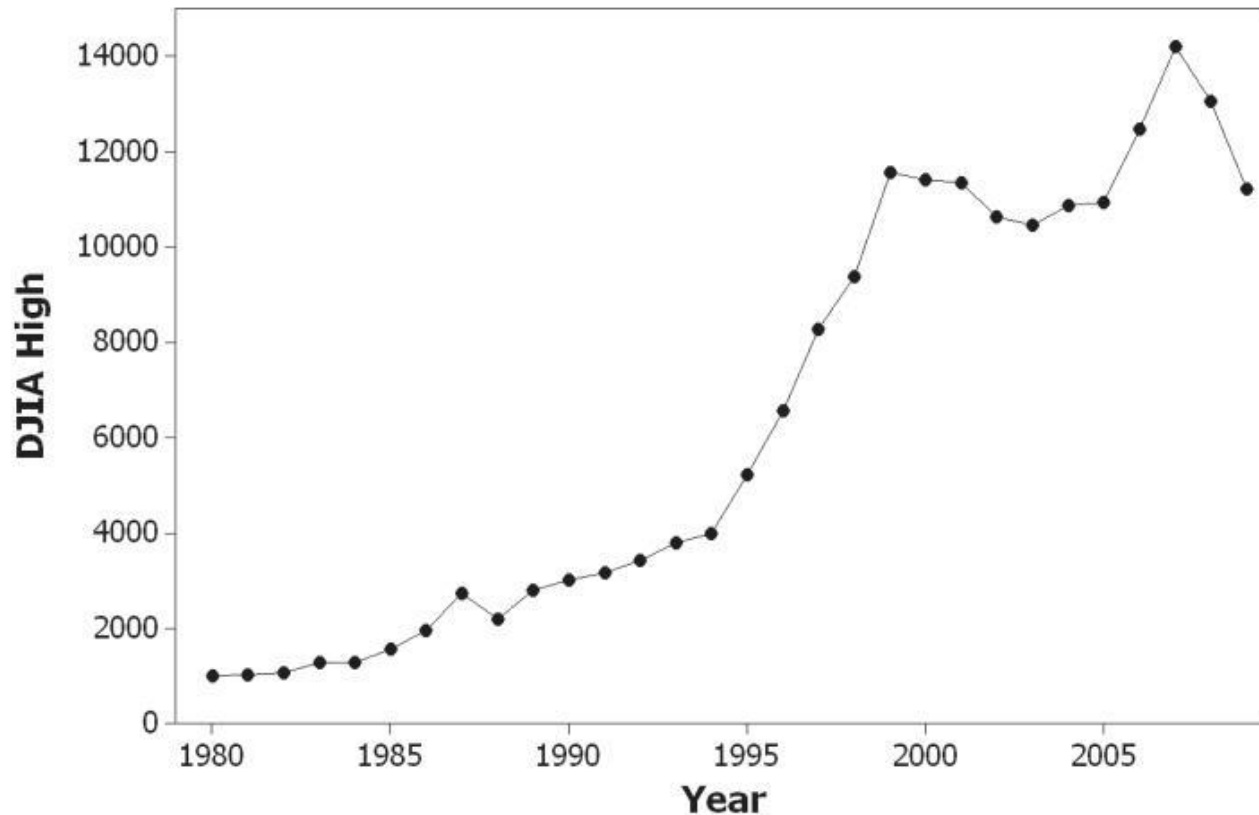
Biểu đồ của dữ liệu định lượng (x, y) với trục x nằm ngang và trục y thẳng đứng. Được sử dụng để xác định xem có mối quan hệ giữa hai biến hay không.



Randomly selected males – the pattern suggests there is a relationship.

Time-Series Graph

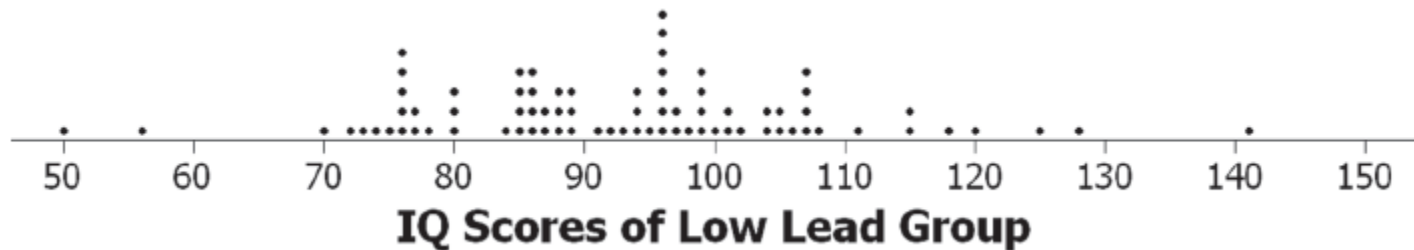
Dữ liệu đã được thu thập ở các thời điểm khác nhau: dữ liệu chuỗi thời gian



Yearly high values of the Dow Jones Industrial Average

Dotplot

Bao gồm một biểu đồ trong đó mỗi giá trị dữ liệu được vẽ dưới dạng một điểm (hoặc chấm) dọc theo thang giá trị. Các dấu chấm biểu thị các giá trị bằng nhau được xếp chồng lên nhau.

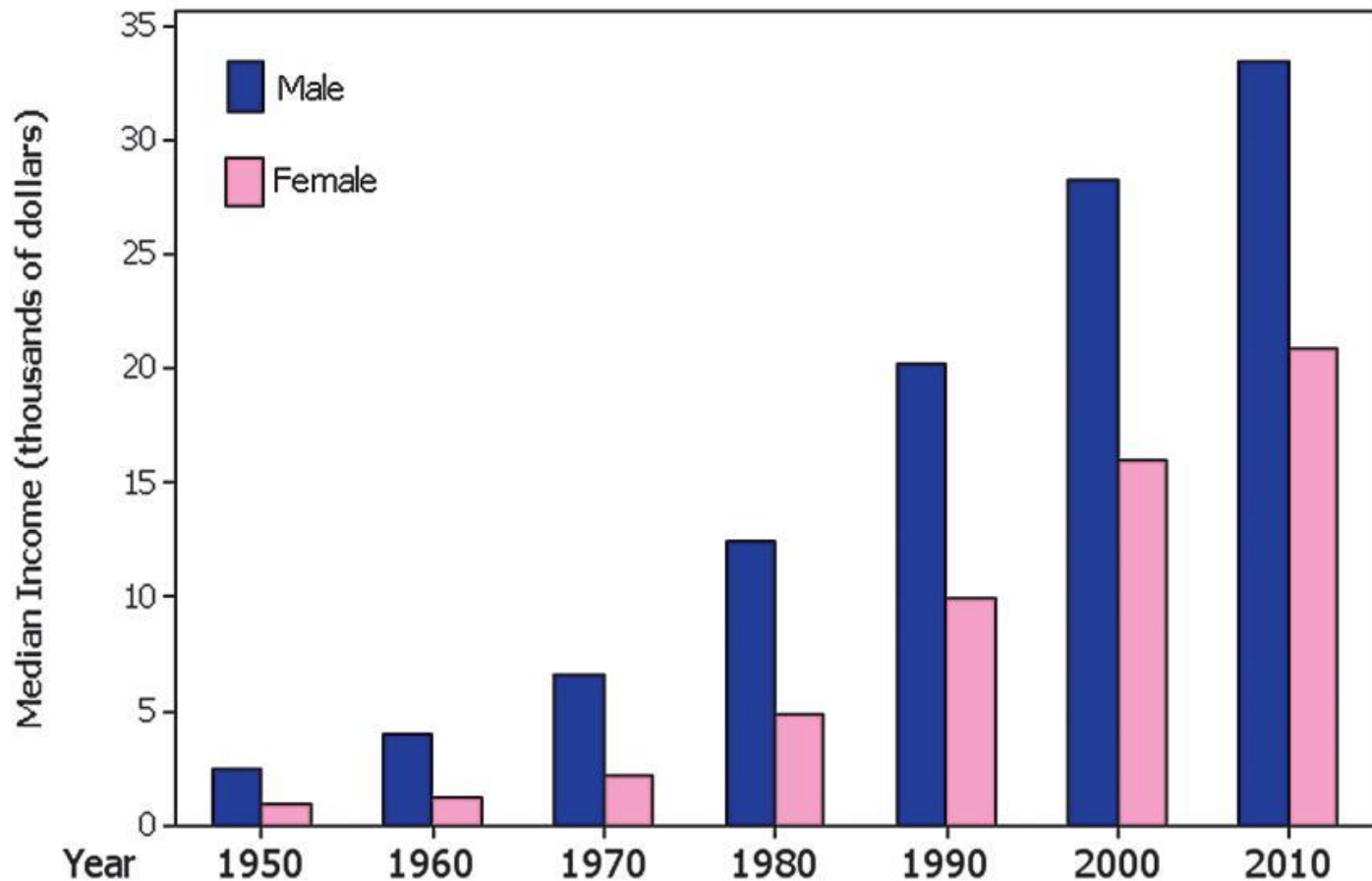


Bar Graph

Sử dụng các thanh có chiều rộng bằng nhau để hiển thị tần số của dữ liệu phân loại hoặc định tính. Thang đo theo trục tung biểu thị tần số hoặc tần số tương đối. Thang đo theo trục hoành xác định các loại dữ liệu định tính khác nhau.

Một biểu đồ thanh có hai hoặc nhiều tập hợp thanh và được sử dụng để so sánh hai hoặc nhiều tập dữ liệu.

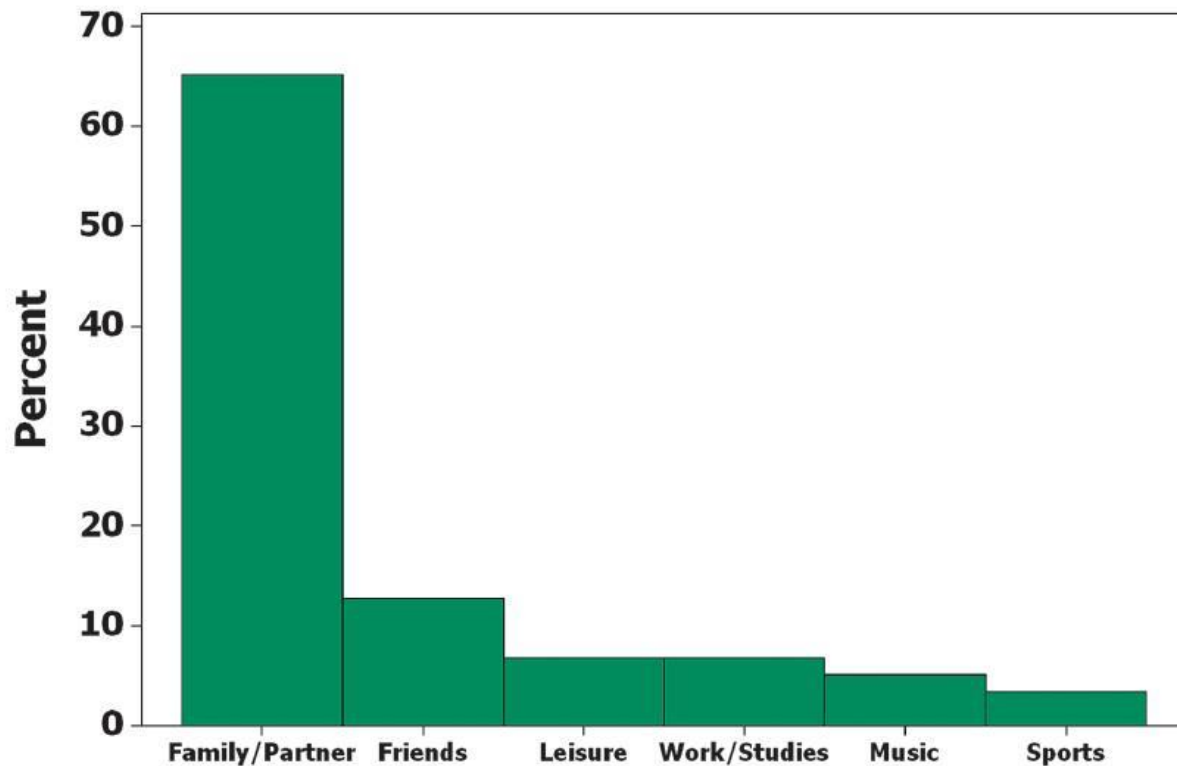
Multiple Bar Graph



Multiple Bar Graph: Median Income by Gender

Pareto Chart

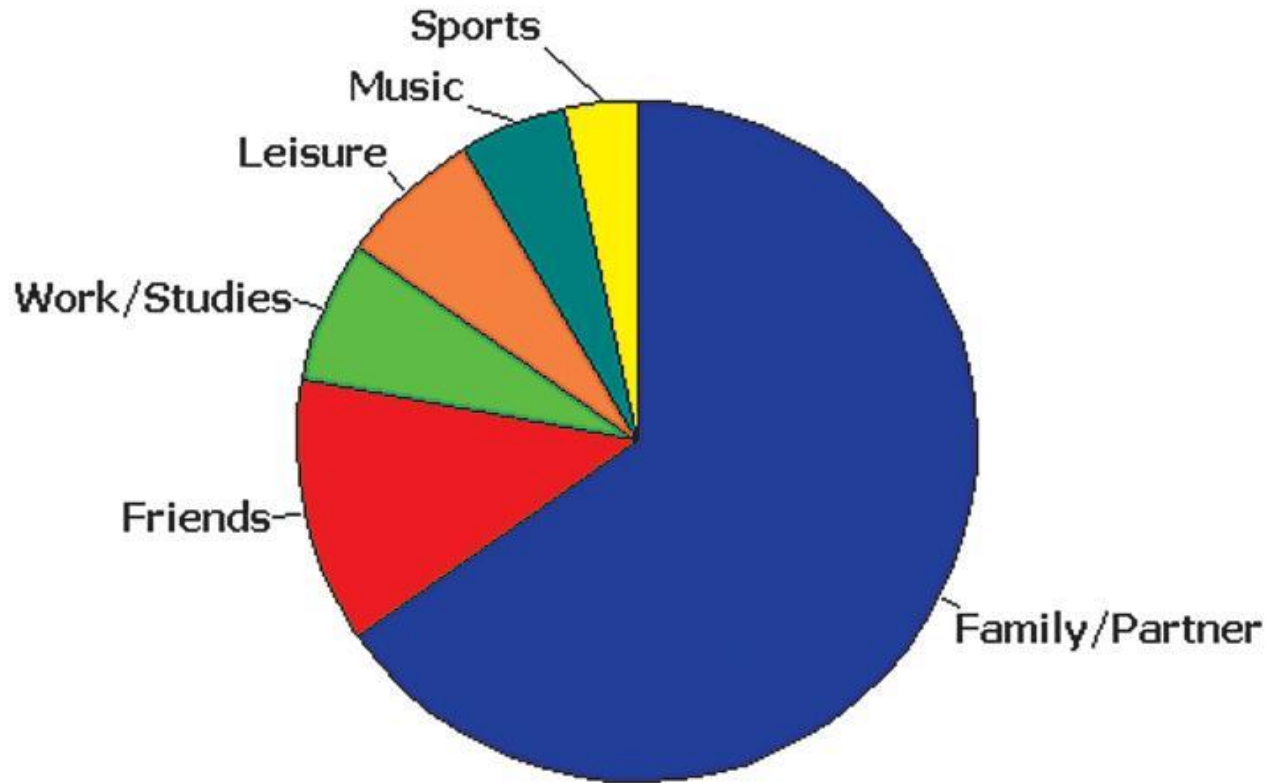
Biểu đồ thanh cho dữ liệu định tính, với các thanh được sắp xếp theo thứ tự giảm dần theo tần số



Pareto Chart: What Contributes Most to Happiness?

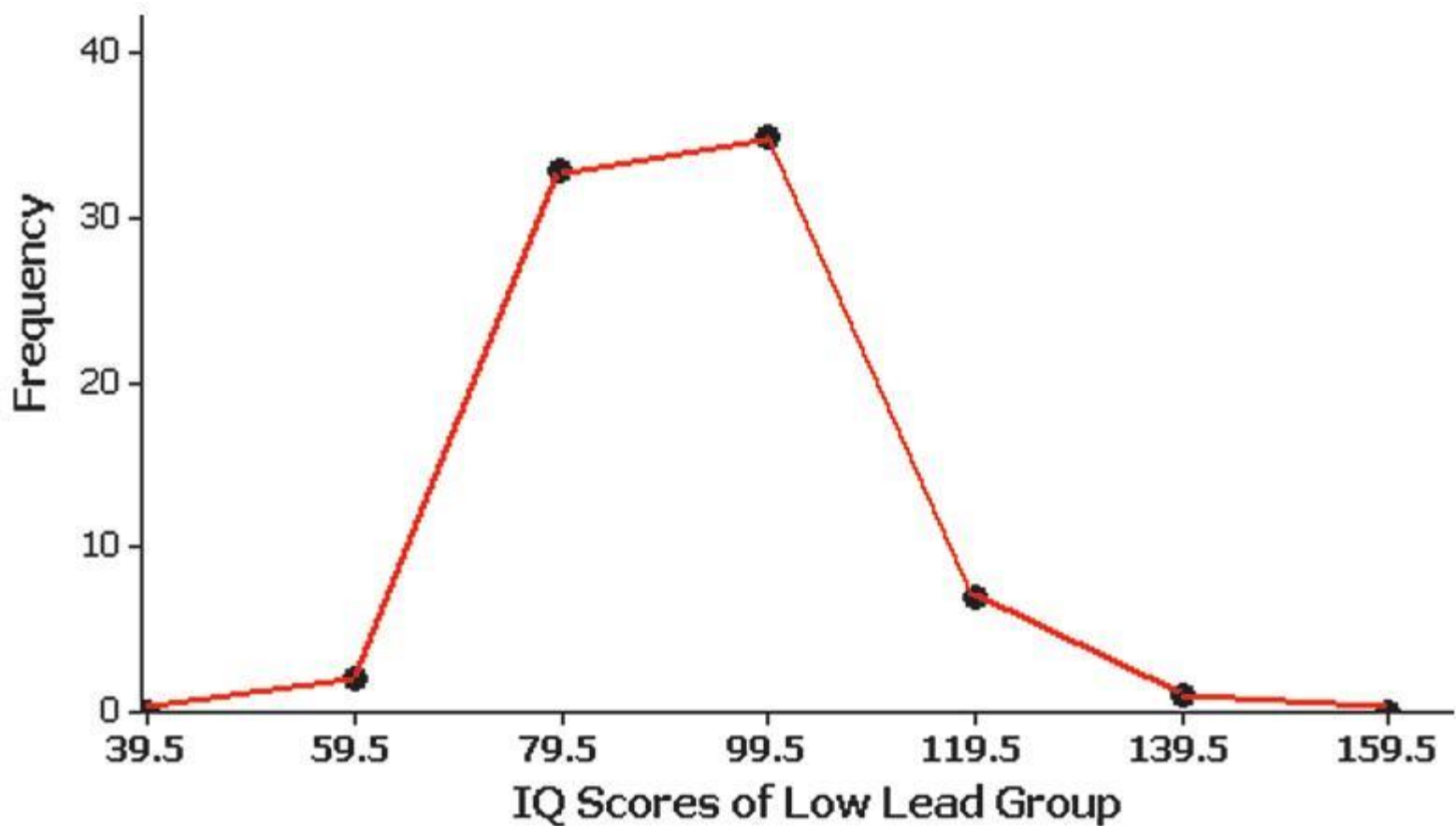
Pie Chart

Biểu đồ mô tả dữ liệu định tính dưới dạng các phần của một vòng tròn, trong đó kích thước của mỗi phần là giá trị tần suất (tỉ lệ phần trăm của tần số)



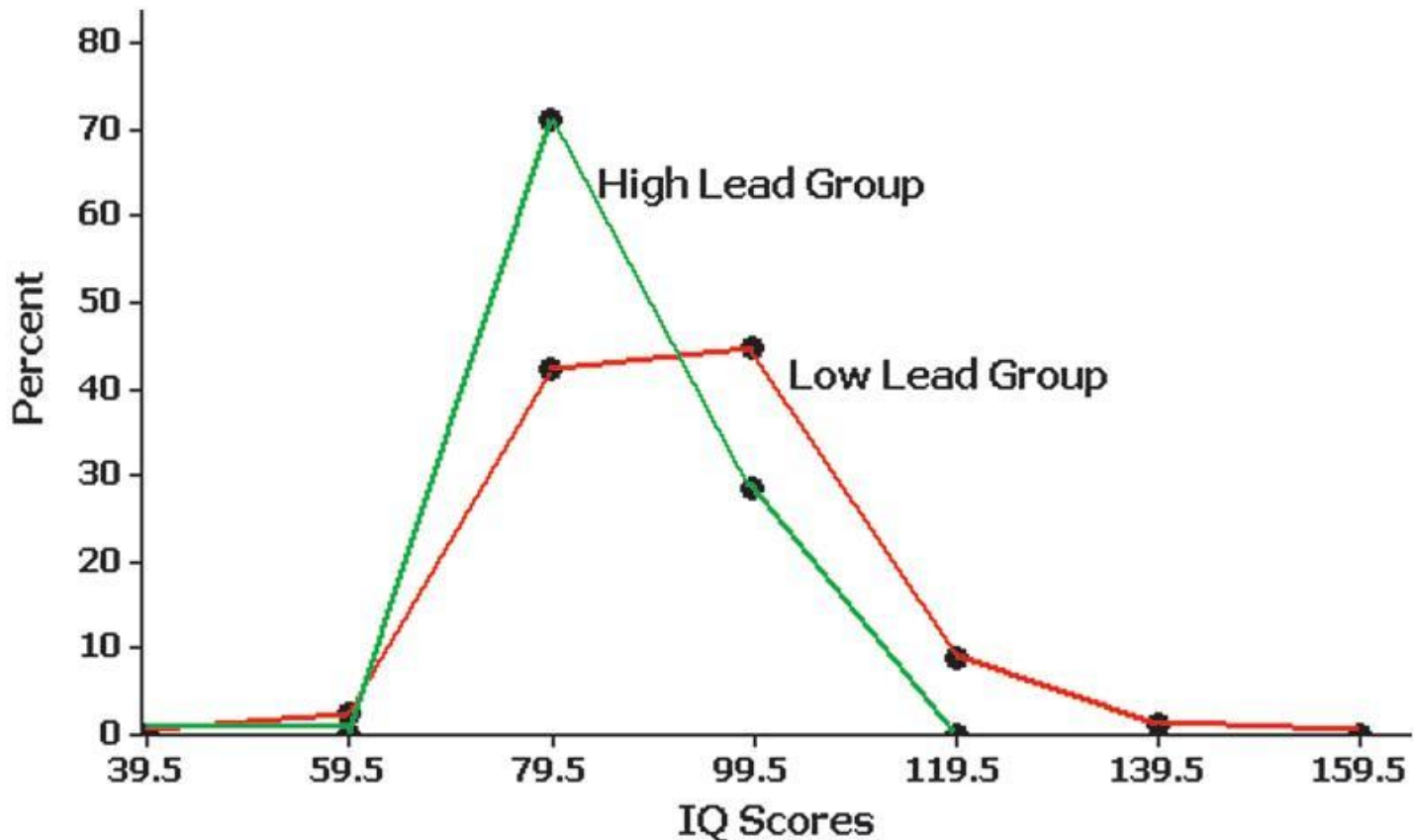
Frequency Polygon

Sử dụng các đoạn thẳng được kết nối với các điểm ngay trên giá trị điểm giữa (midpoint) của lớp.



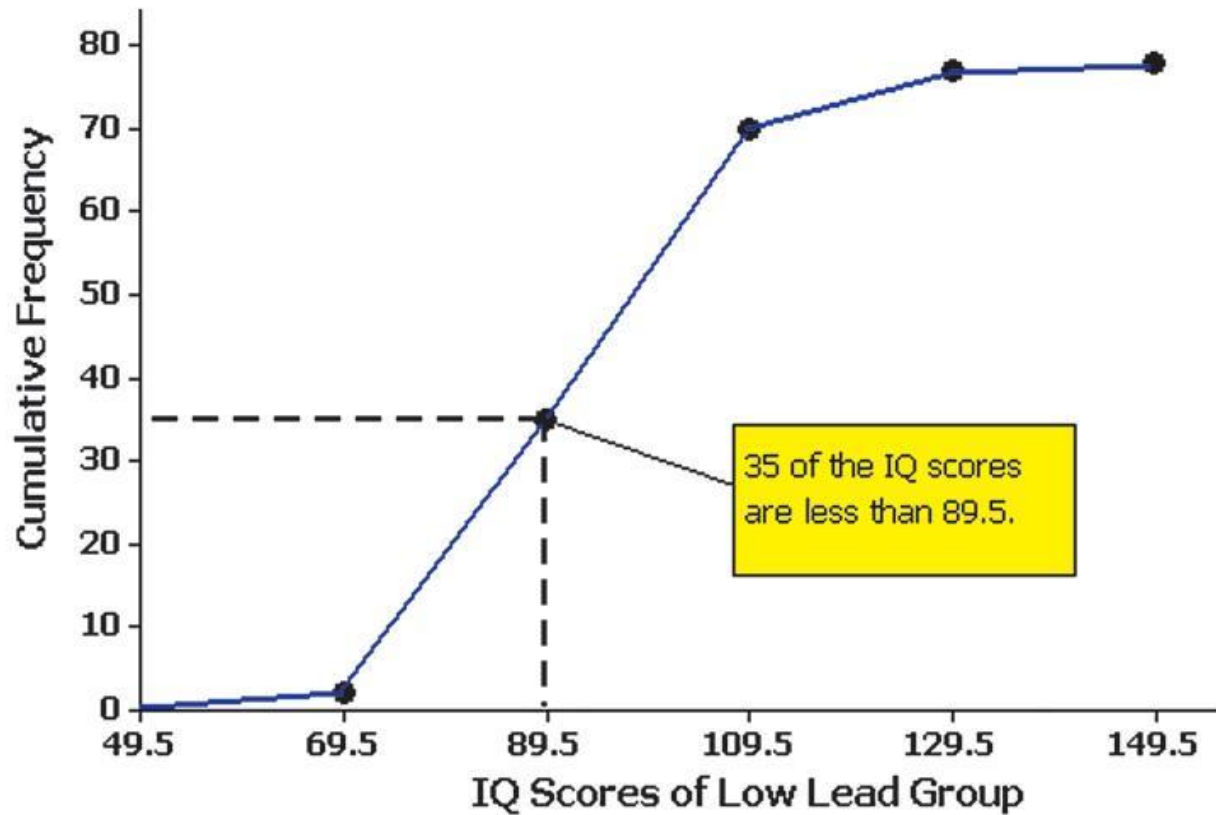
Relative Frequency Polygon

Sử dụng tần số tương đối (tỷ lệ hoặc phần trăm) cho trục tung.



Ogive

Biểu đồ đường mô tả tần số tích lũy



Cám ơn