

Chương

02

TỔNG HỢP VÀ TRỰC 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

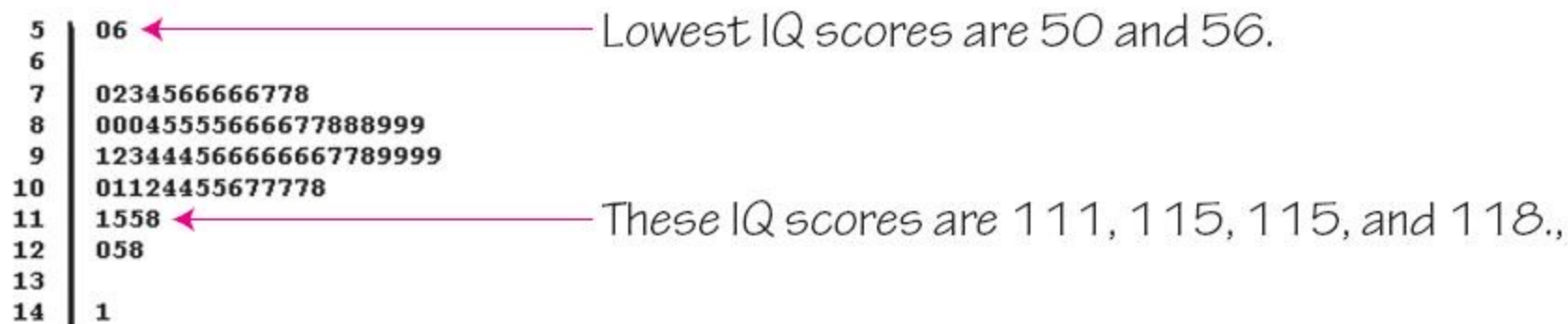
- Khi tập dữ liệu mẫu được thu thập về, thông thường chúng ta phải thực hiện tính toán, và biến đổi một chút để có thể biết được các đặc tính của chúng.
- Tuy nhiên, việc thay đổi dữ liệu cần phải thực hiện cẩn thận để tránh làm mất mát thông tin mà dữ liệu chứa đựng.
- Để có cái nhìn ban đầu về dữ liệu,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



15,16,21,23,23,26,26,30,32,41

Stem	Leaf
1	5 6
2	1 3 3 6 6
3	0 2
4	1

*how to
place "32"*

- 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

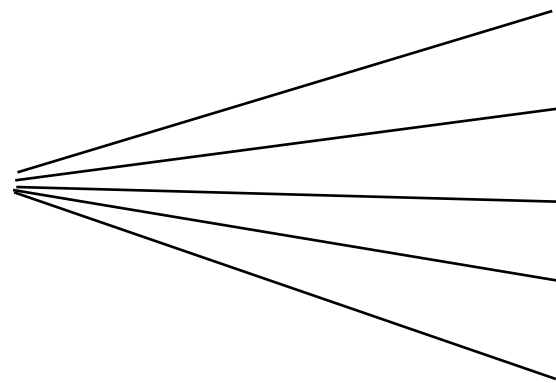
Phân phối tần số

- **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.

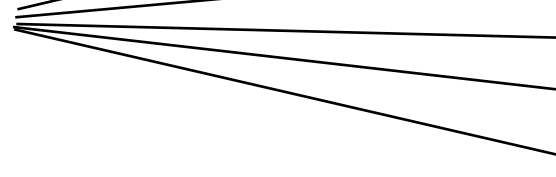


IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1

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 suất:

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.

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

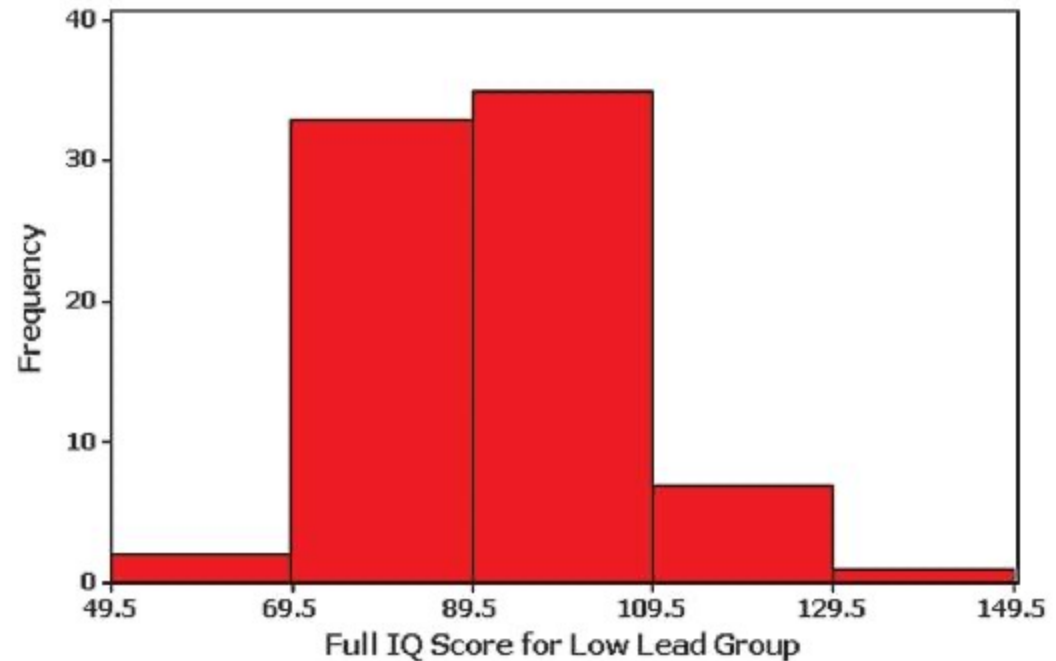
Histograms

- 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 suất của lớp
- Chiều cao của các cột tương ứng với tần suất 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



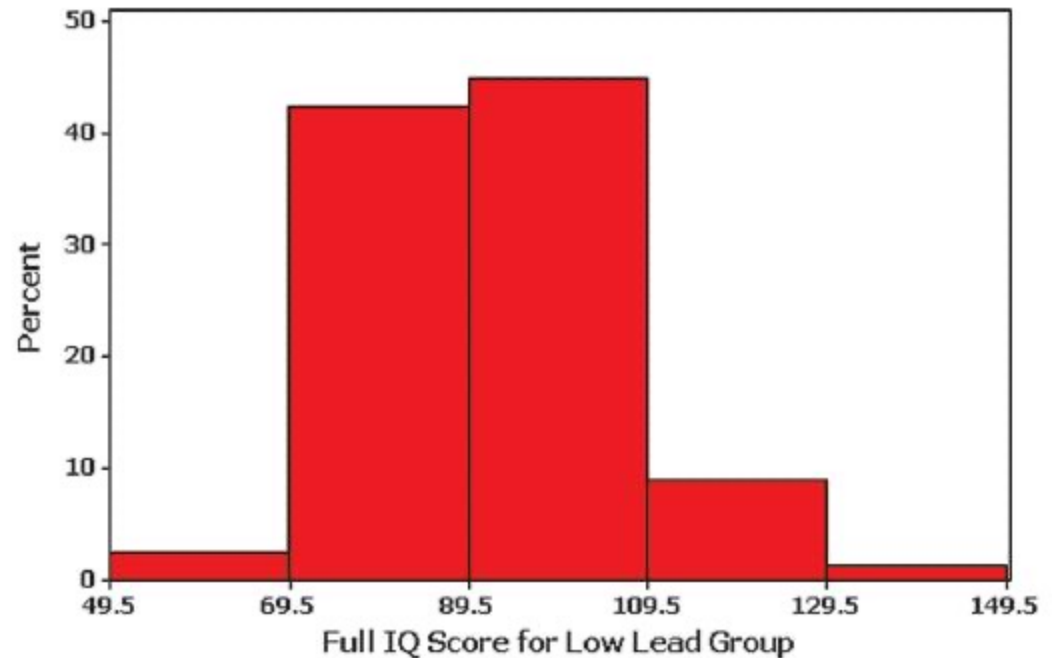
Histograms

- 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

has the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies instead of actual frequencies

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

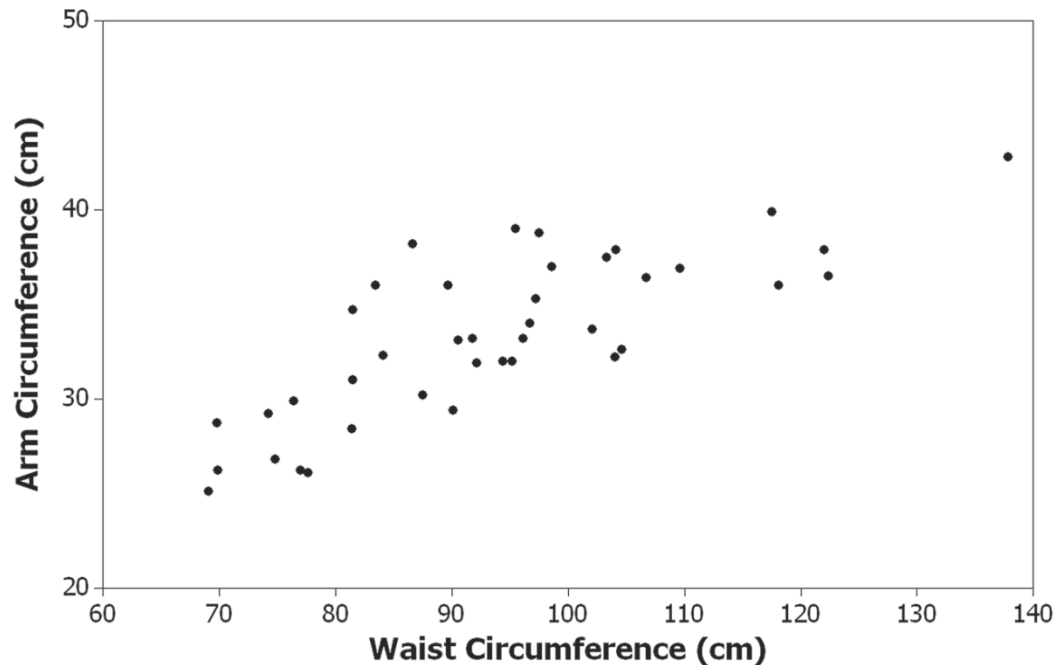


NỘI DUNG

- 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)

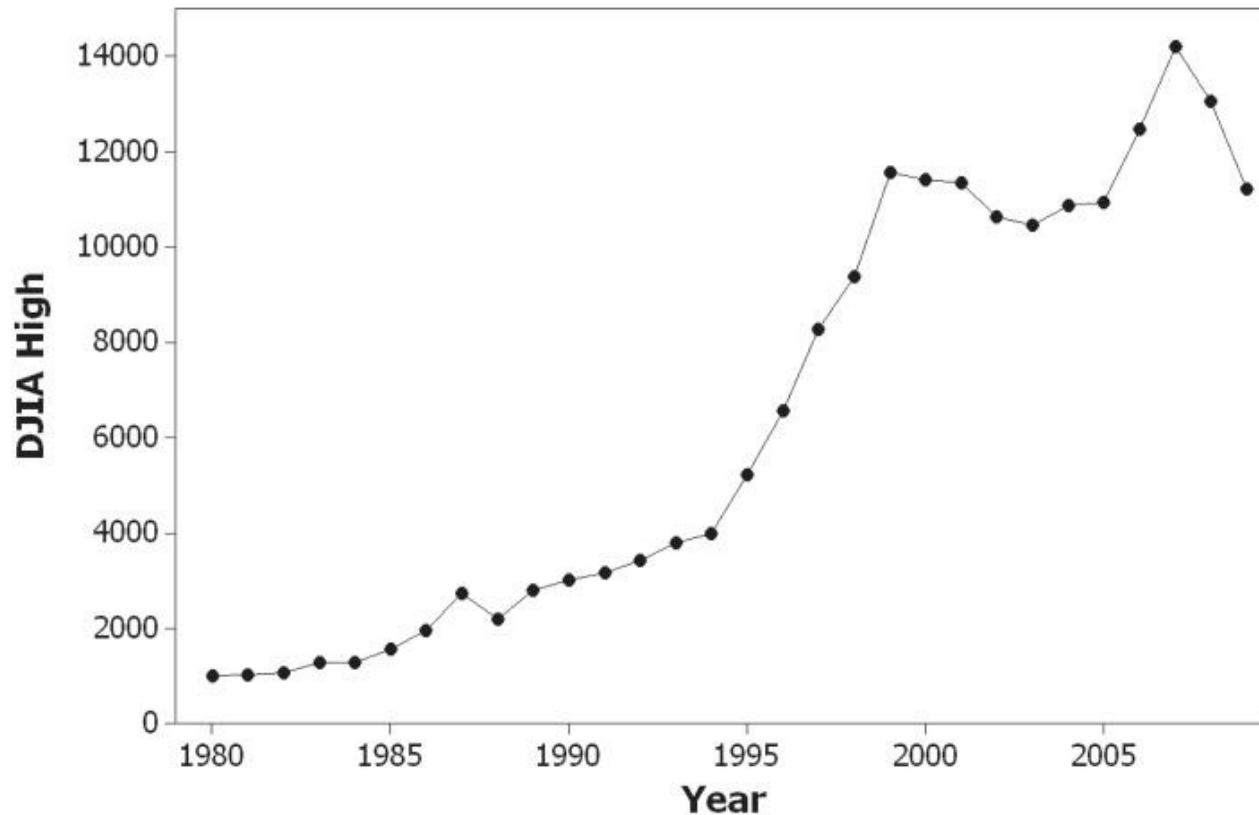
A plot of paired (x, y) quantitative data with a horizontal x-axis and a vertical y-axis. Used to determine whether there is a relationship between the two variables.



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

Time-Series Graph

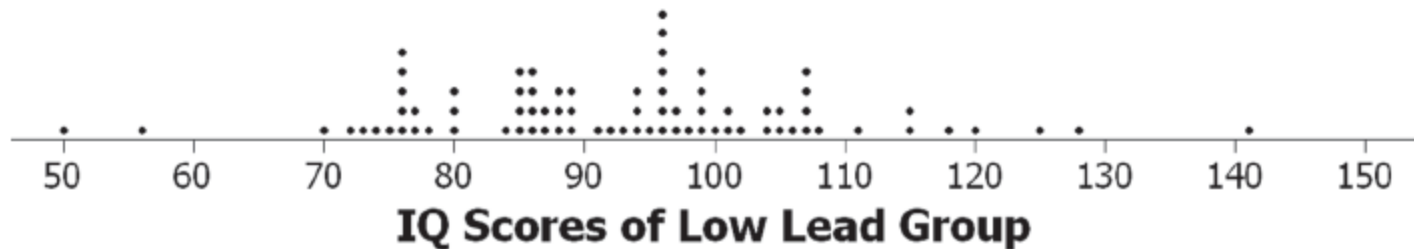
Data that have been collected at different points in time: *time-series data*



Yearly high values of the Dow Jones Industrial Average

Dotplot

Consists of a graph in which each data value is plotted as a point (or dot) along a scale of values. Dots representing equal values are stacked.

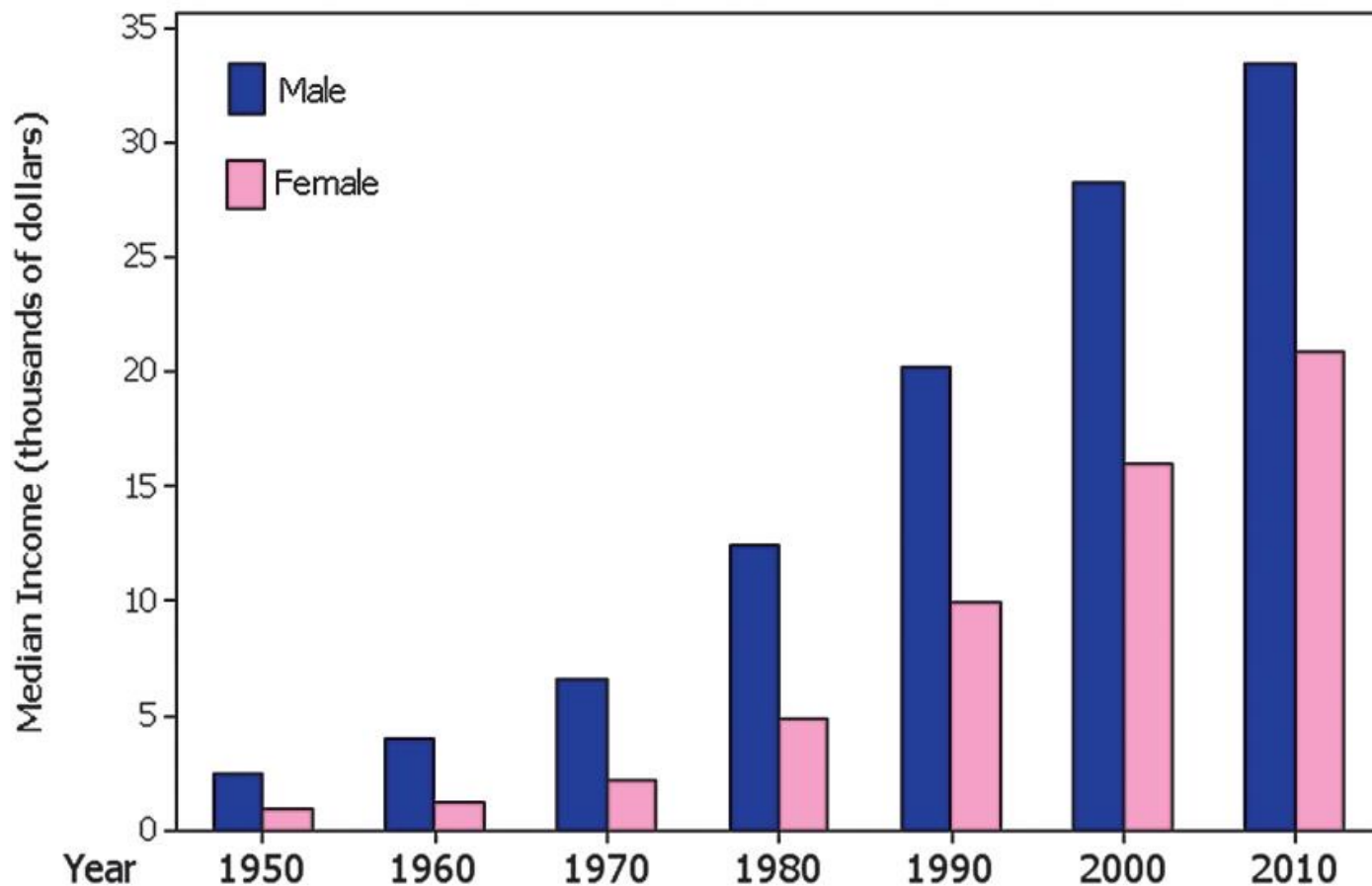


Bar Graph

Uses bars of equal width to show frequencies of categorical, or qualitative, data. Vertical scale represents frequencies or relative frequencies. Horizontal scale identifies the different categories of qualitative data.

A multiple bar graph has two or more sets of bars and is used to compare two or more data sets.

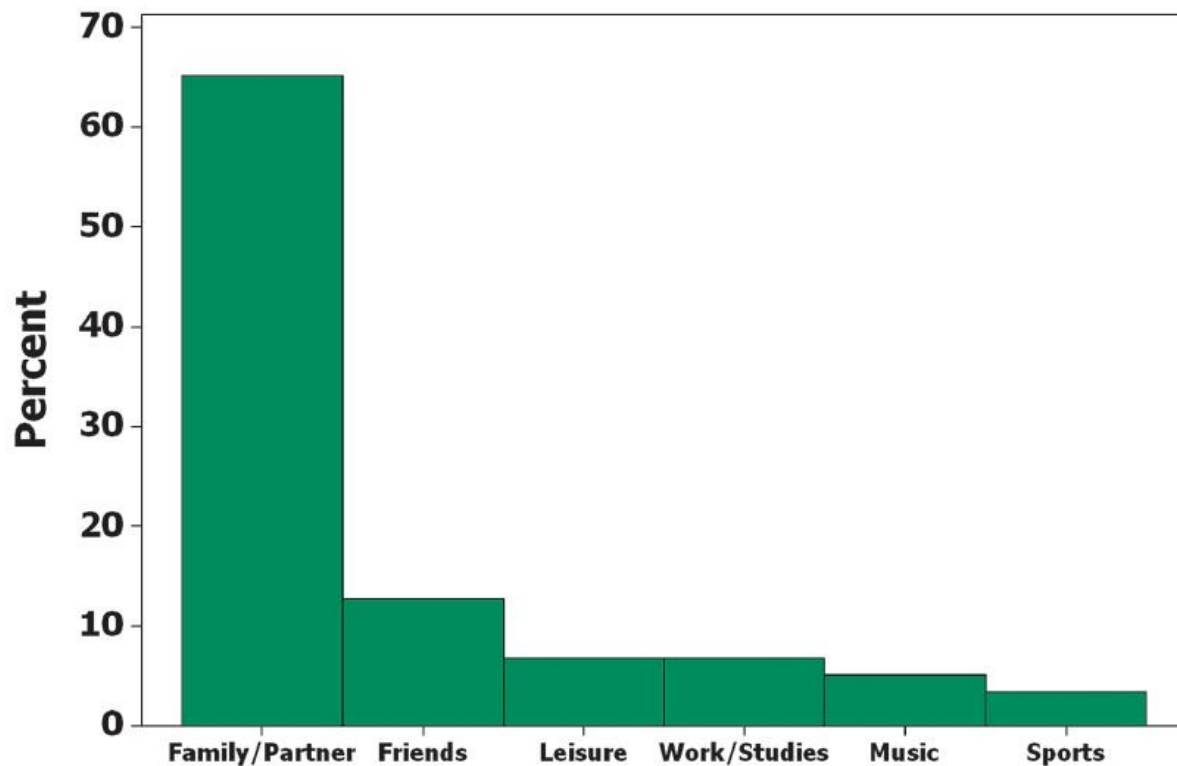
Multiple Bar Graph



Multiple Bar Graph: Median Income by Gender

Pareto Chart

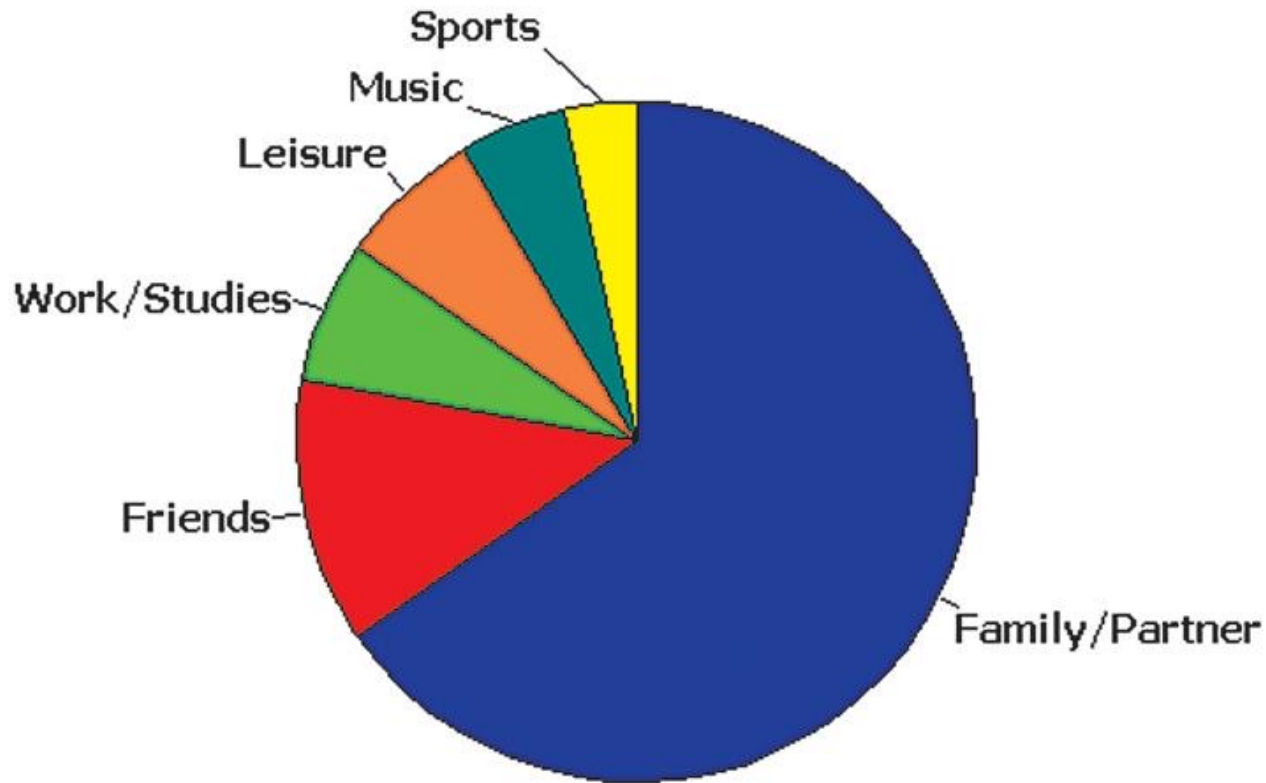
A bar graph for qualitative data, with the bars arranged in descending order according to frequencies



Pareto Chart: What Contributes Most to Happiness?

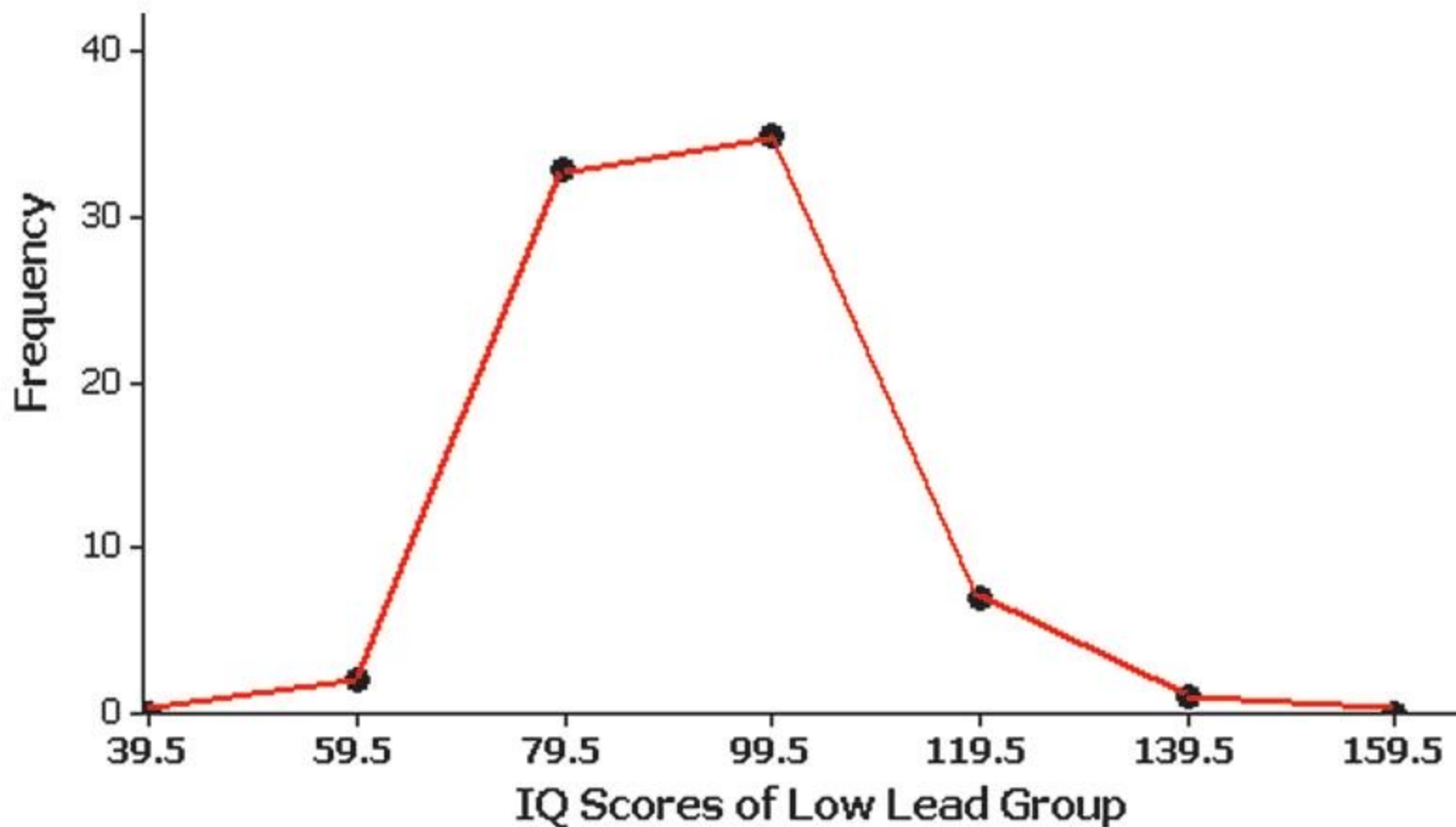
Pie Chart

A graph depicting qualitative data as slices of a circle, in which the size of each slice is proportional to frequency count



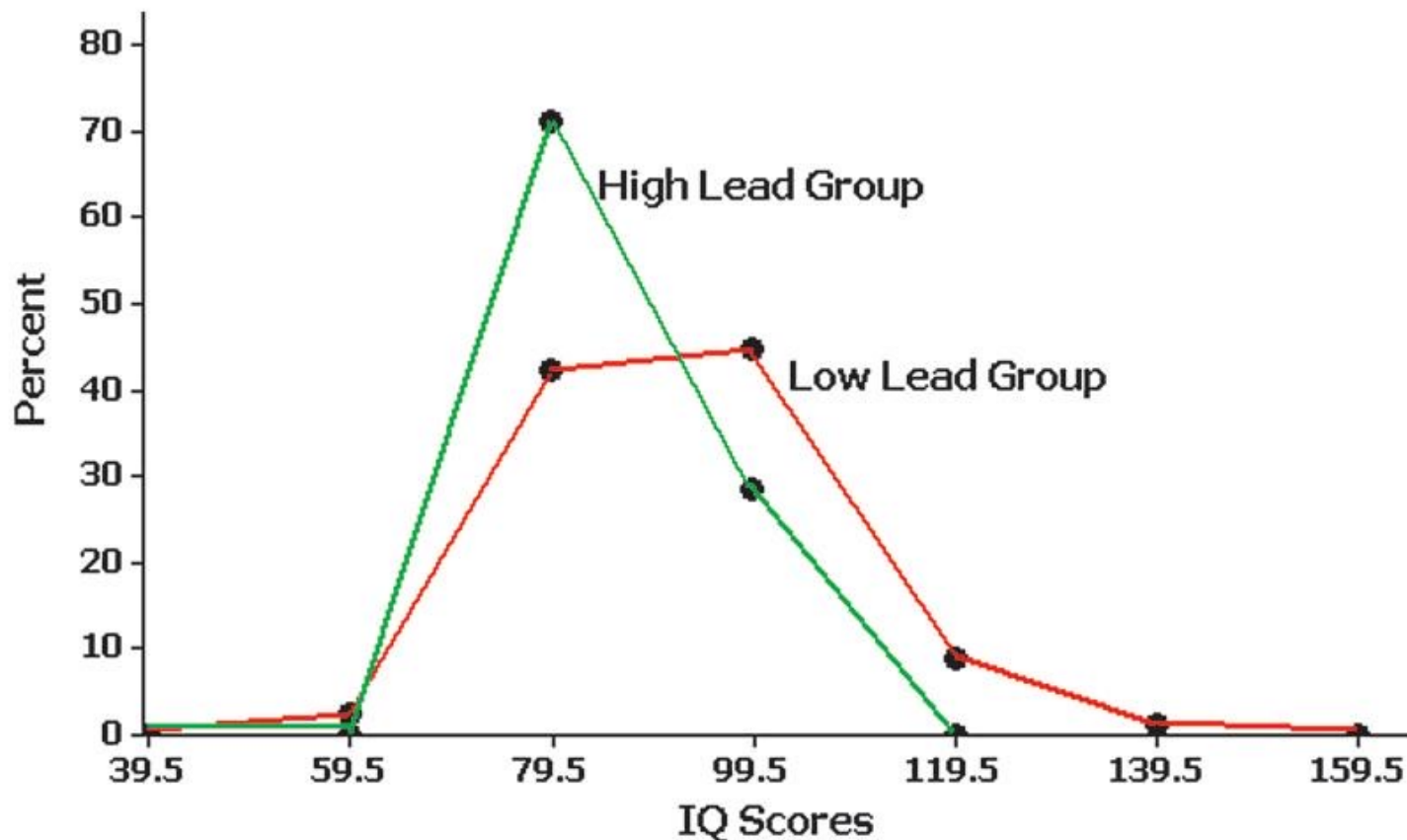
Frequency Polygon

uses line segments connected to points directly above class midpoint values.



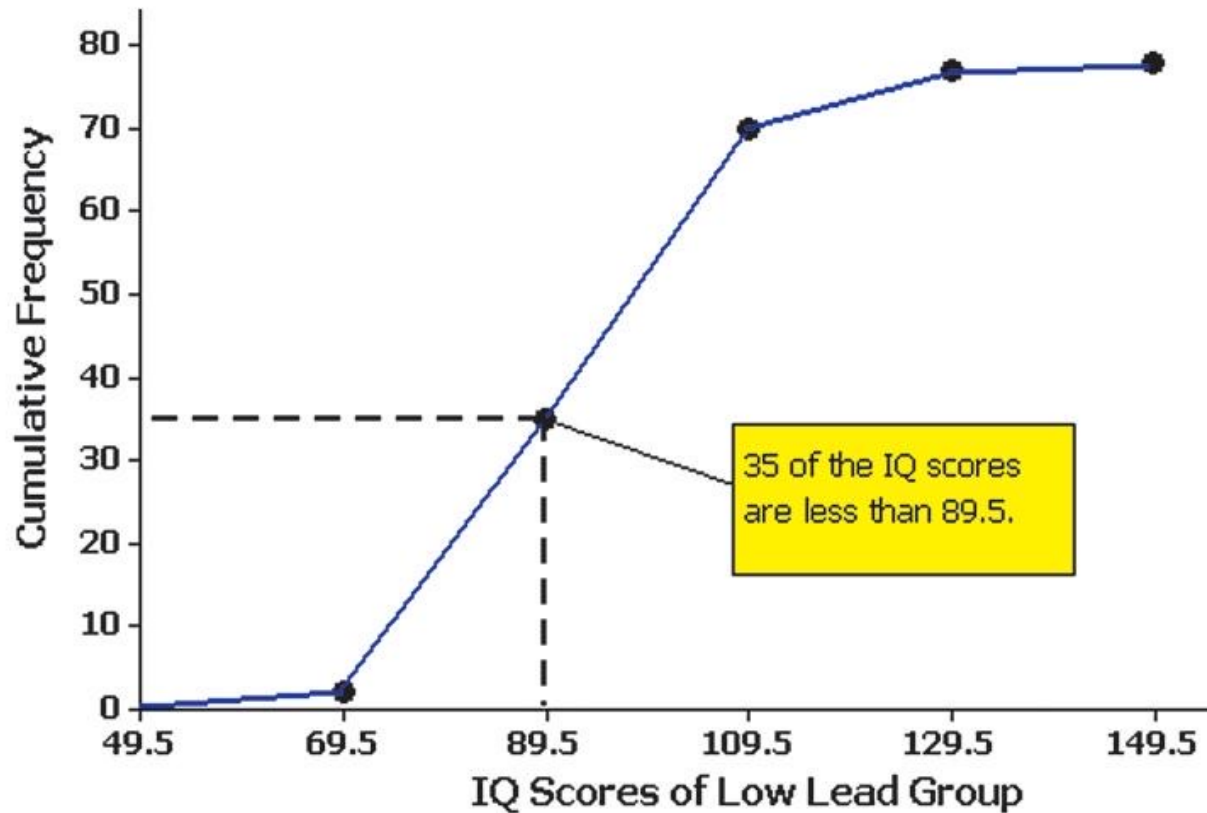
Relative Frequency Polygon

Uses relative frequencies (proportions or percentages) for the vertical scale.



Ogive

A line graph that depicts **cumulative** frequencies



BÀI TẬP

Doanh thu các cửa hàng rau UK

- Dưới đây là bảng dữ liệu doanh thu của 50 cửa hàng rau ở Anh (đơn vị ngàn bảng). Hãy vẽ các đồ thị

8.1	11.8	8.7	10.6	9.5
9.3	11.5	10.7	11.6	7.8
10.5	7.6	10.1	8.9	8.6
11.1	10.2	11.1	9.9	9.8
11.6	15.1	12.5	6.5	7.5
10.3	12.9	9.2	10.7	12.8
12.5	9.3	10.4	12.7	10.5
10.3	11.1	9.6	9.7	14.5
13.7	6.7	11.5	8.4	10.3
13.7	11.2	7.3	5.3	12.5

Doanh thu các chi nhánh của siêu thị

- Khảo sát doanh thu (bảng Anh) trong một tháng của 60 chi nhánh thuộc một siêu thị tại Anh cho bảng dữ liệu như sau:

Table 1.4 Raw data of sales revenue from a supermarket (£'000s).

15.5	7.8	12.7	15.6	14.8	8.5	11.5	13.5	8.8	9.8
10.7	16.0	9.0	9.1	13.6	14.5	8.9	11.7	11.5	14.9
15.4	16.0	16.1	13.8	9.2	13.1	15.8	13.2	12.6	10.9
12.9	9.6	12.1	15.2	11.9	10.4	10.6	13.7	14.4	13.8
9.6	12.0	11.0	10.5	12.4	11.5	11.7	14.1	11.2	12.2
12.5	10.8	10.0	11.1	10.2	11.2	14.2	11.0	12.1	12.5

Doanh thu chi nhánh theo quốc gia

- Khảo sát doanh thu (dollars) trong năm của một công ty đa quốc gia. Cho ta bảng dữ liệu như sau:

Group	Country	Sales revenues (\$)
1	Austria	522,065
2	Belgium	1,266,054
3	Finland	741,639
4	France	2,470,257
5	Germany	2,876,431
6	Italy	2,086,829
7	Netherlands	1,091,779
8	Portugal	1,161,479
9	Sweden	3,884,566
10	United Kingdom	4,432,234

THANK YOU