

# Descriptive Statistics and Data Visualization

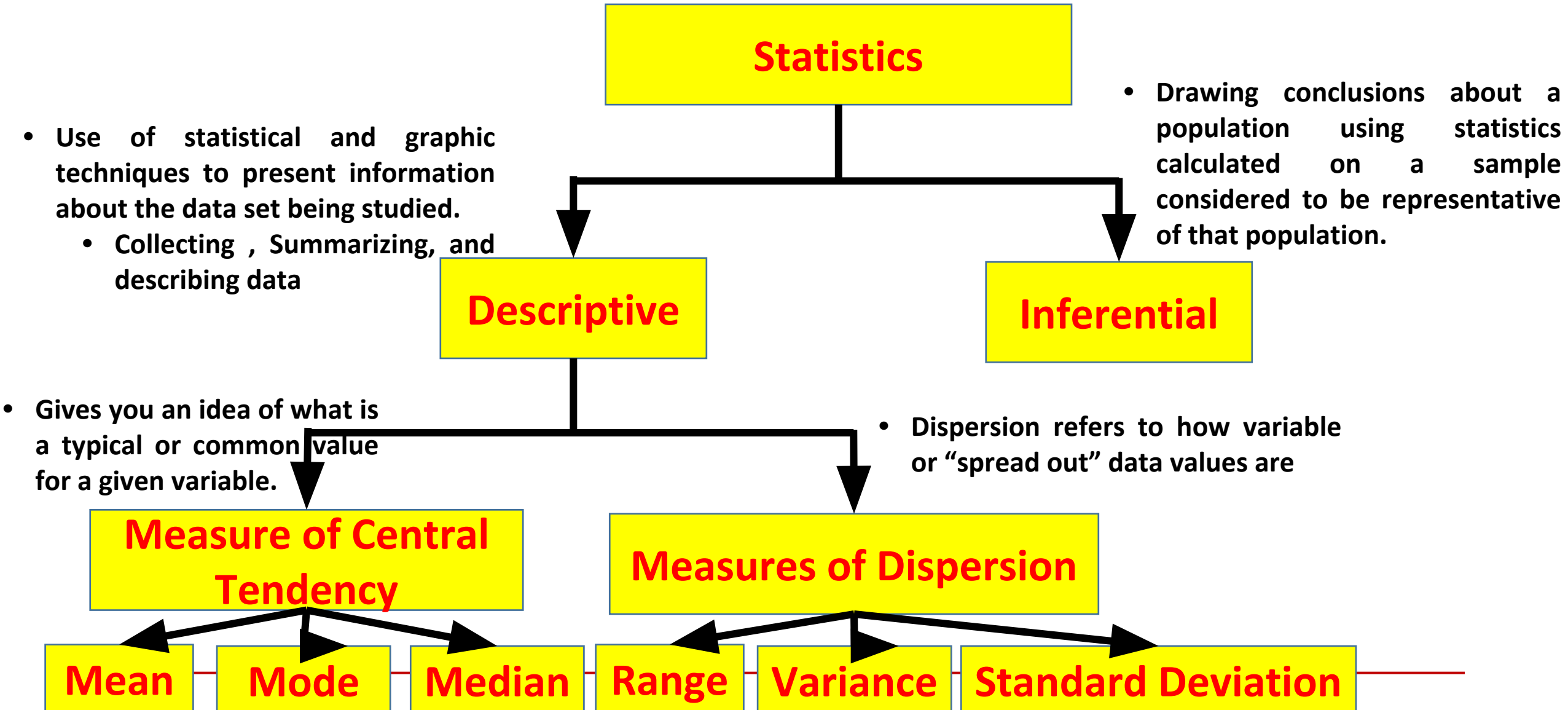
# Statistics

---

- Many studies generate large numbers of data points
  - How to make sense of all that data?
    - Statistics is used to *summarize* the data, to provide a better understanding of overall tendencies within the distributions of scores.
      - helps in summarizing the results
      - helps us recognize underlying trends and tendencies in the data
      - helps in communicating the results to others
-

# Types of Statistics

---



# Descriptive statistics

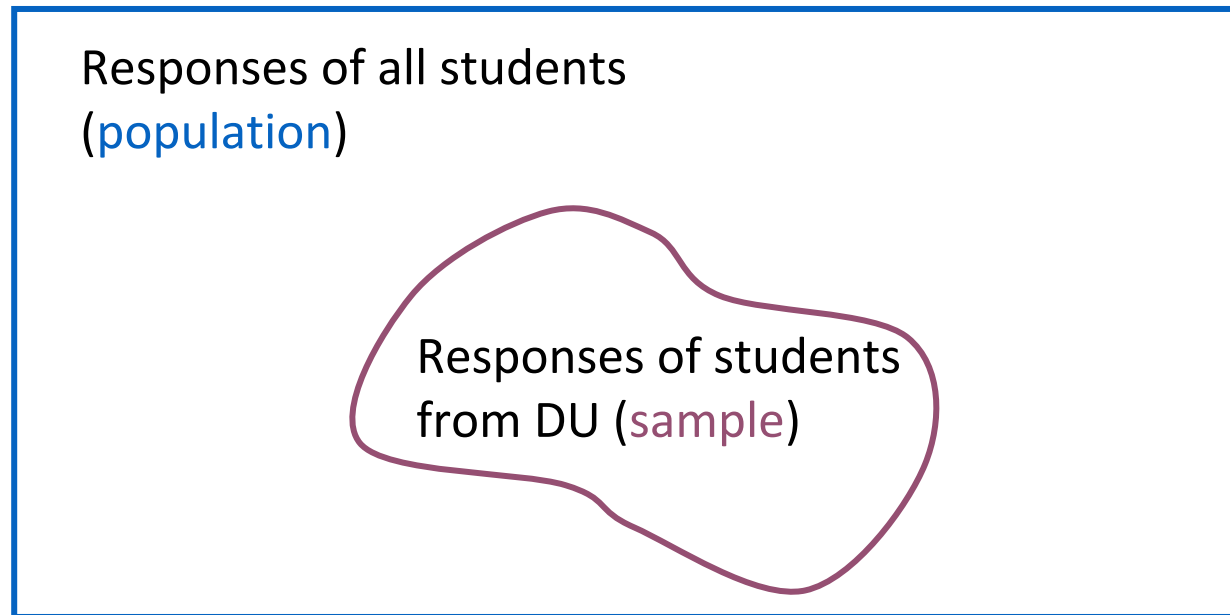
---

- If we wanted to characterize the students in this class, we would find that they are:
    - Young
    - Fit
    - Male
  - How young?
  - How fit is this class?
  - What is the distribution of males and females?
  - Goal:
    - To visualize data, understand the patterns, and make quick statements about the system's behavior
    - To understand relations among variables
-

# Populations & Samples

---

- In a survey, 250 college students were asked if they study regularly. 35 of the students said yes.



# Frequency Distributions

---

- A **frequency distribution** is a table that shows classes or intervals of data with a count of the number in each class.
- The frequency  **$f$**  of a class is the number of data points in the class.

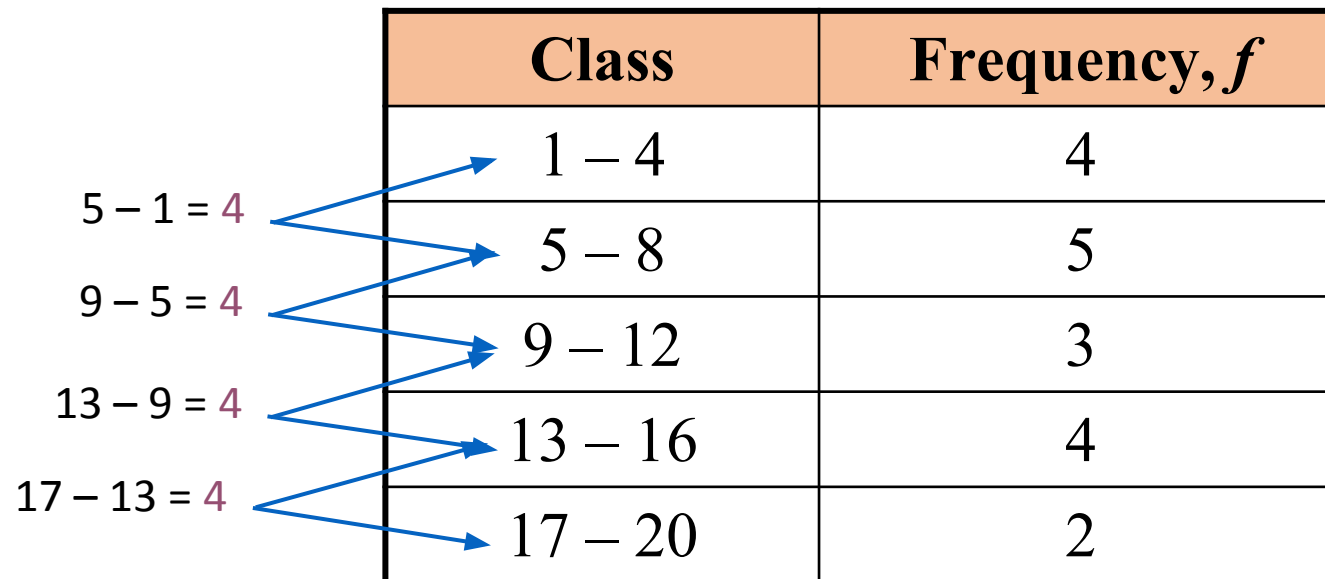
The diagram shows a frequency distribution table with two columns: 'Class' and 'Frequency,  $f$ '. The 'Class' column contains five intervals: 1-4, 5-8, 9-12, 13-16, and 17-20. The 'Frequency,  $f$ ' column contains the corresponding counts: 4, 5, 3, 4, and 2. Blue arrows point from the text 'Upper Class Limits' to the right-hand number of each class interval. Purple arrows point from the text 'Frequencies' to the frequency value in each row.

| Class   | Frequency, $f$ |
|---------|----------------|
| 1 - 4   | 4              |
| 5 - 8   | 5              |
| 9 - 12  | 3              |
| 13 - 16 | 4              |
| 17 - 20 | 2              |

# Frequency Distributions

---

- The **class width** is the distance between lower (or upper) limits of consecutive classes.



|               | Class   | Frequency, $f$ |
|---------------|---------|----------------|
| $5 - 1 = 4$   | 1 – 4   | 4              |
| $9 - 5 = 4$   | 5 – 8   | 5              |
| $13 - 9 = 4$  | 9 – 12  | 3              |
| $17 - 13 = 4$ | 13 – 16 | 4              |
|               | 17 – 20 | 2              |

The class width is 4.

- The **range** is the difference between the maximum and minimum data entries.
-

# Constructing a Frequency Distribution

---

- **Example:**
  - The following data represents the ages of 30 students in a class. Construct a frequency distribution that has **five** classes.

| Ages of Students |    |    |    |    |    |
|------------------|----|----|----|----|----|
| 18               | 20 | 21 | 27 | 29 | 20 |
| 19               | 30 | 32 | 19 | 34 | 19 |
| 24               | 29 | 18 | 37 | 38 | 22 |
| 30               | 39 | 32 | 44 | 33 | 46 |
| 54               | 49 | 18 | 51 | 21 | 21 |

---



# Constructing a Frequency Distribution

---

- Number of classes: 5
- The minimum data entry is 18 and maximum entry is 54, so the range is 36.
- Divide the range by the number of classes to find the class width.
  - Class width =  $\frac{36}{5} = 7.2$  (Round up to 8)

|    |    |    |    |    |    |
|----|----|----|----|----|----|
| 18 | 20 | 21 | 27 | 29 | 20 |
| 19 | 30 | 32 | 19 | 34 | 19 |
| 24 | 29 | 18 | 37 | 38 | 22 |
| 30 | 39 | 32 | 44 | 33 | 46 |
| 54 | 49 | 18 | 51 | 21 | 21 |

# Constructing a Frequency Distribution

---

- Lower limit and upper limits of classes will be

- The lower class limits are 18, 26, 34, 42, and 50.
- The upper class limits are 25, 33, 41, 49, and 57.

|    |    |    |    |    |    |
|----|----|----|----|----|----|
| 18 | 20 | 21 | 27 | 29 | 20 |
| 19 | 30 | 32 | 19 | 34 | 19 |
| 24 | 29 | 18 | 37 | 38 | 22 |
| 30 | 39 | 32 | 44 | 33 | 46 |
| 54 | 49 | 18 | 51 | 21 | 21 |

| Class   | Frequency, $f$ |
|---------|----------------|
| 18 – 25 | 13             |
| 26 – 33 | 8              |
| 34 – 41 | 4              |
| 42 – 49 | 3              |
| 50 – 57 | 2              |

# Relative Frequency

---

- The relative frequency of a class is the portion or percentage of the data that falls in that class.

| Class   | Frequency, $f$ | Relative Frequency |
|---------|----------------|--------------------|
| 18 – 25 | 13             | 0.433              |
| 26 – 33 | 8              | 0.267              |
| 34 – 41 | 4              | 0.133              |
| 42 – 49 | 3              | 0.100              |
| 50 – 57 | 2              | 0.067              |
|         |                |                    |

# Cumulative Frequency

- The cumulative frequency of a class is the sum of the frequency for that class and all the previous classes.

Ages of Students

| Class   | Frequency, $f$  | Cumulative Frequency |
|---------|-----------------|----------------------|
| 18 – 25 | 13              | 13                   |
| 26 – 33 | + 8             | 21                   |
| 34 – 41 | + 4             | 25                   |
| 42 – 49 | + 3             | 28                   |
| 50 – 57 | + 2             | 30                   |
|         | $\Sigma f = 30$ |                      |

Total number of students

# Frequency Histogram

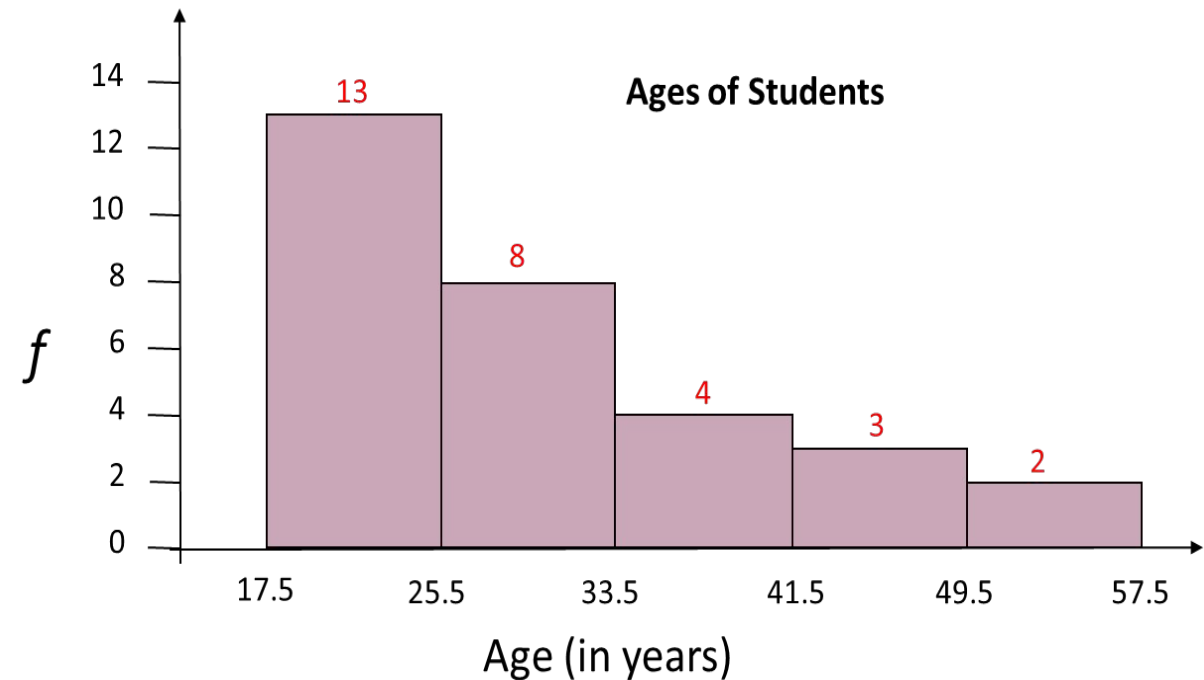
---

- A frequency histogram is a bar graph that represents the frequency distribution of a data set.
    1. The horizontal scale is quantitative and measures the data values.
    2. The vertical scale measures the frequencies of the classes.
    3. Consecutive bars must touch.
  - **Class boundaries** are the numbers that separate the classes without forming gaps between them.
  - The horizontal scale of a histogram can be marked with either the class boundaries or the midpoints.
-

# Frequency Histogram

Ages of Students

| Class   | Frequency, $f$  | Class Boundaries |
|---------|-----------------|------------------|
| 18 – 25 | 13              | 17.5 – 25.5      |
| 26 – 33 | 8               | 25.5 – 33.5      |
| 34 – 41 | 4               | 33.5 – 41.5      |
| 42 – 49 | 3               | 41.5 – 49.5      |
| 50 – 57 | 2               | 49.5 – 57.5      |
|         | $\Sigma f = 30$ |                  |

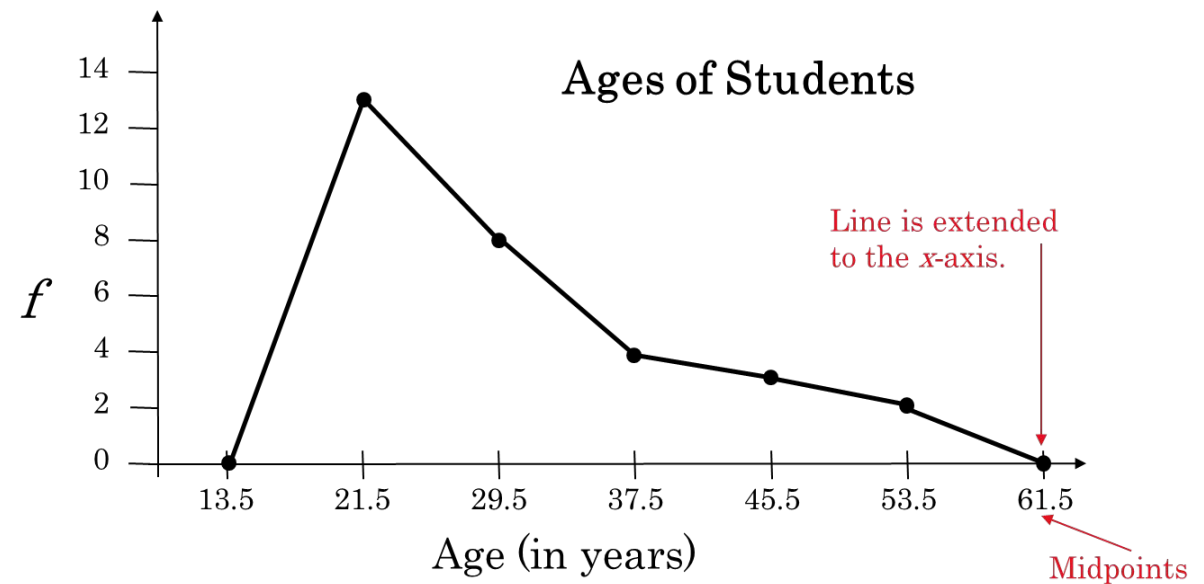


# Frequency Polygon

- A **frequency polygon** is a line graph that emphasizes the continuous change in frequencies.

Ages of Students

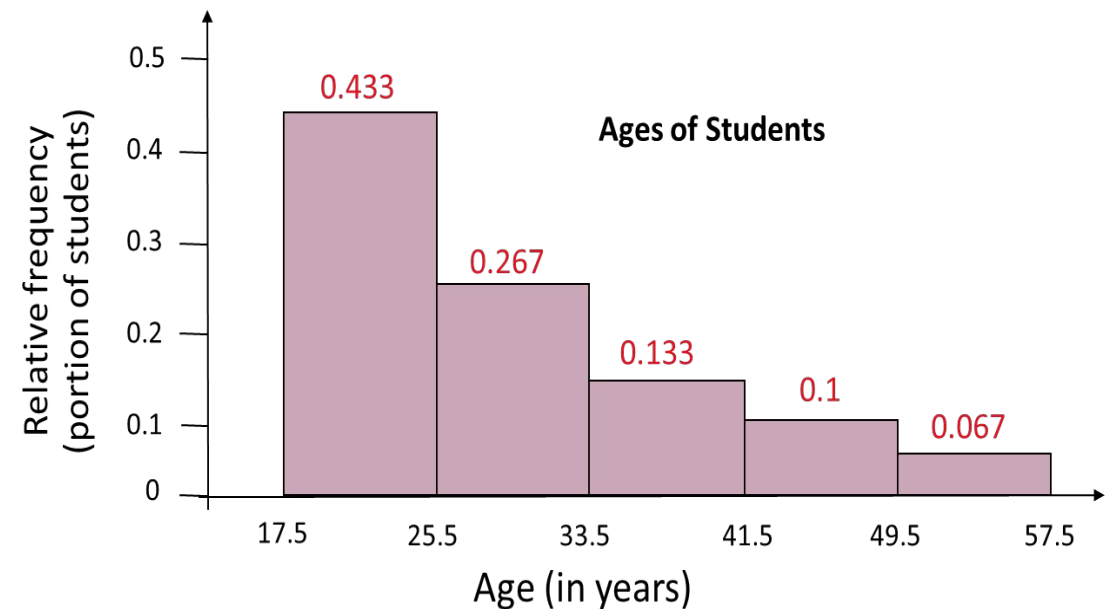
| Class   | Frequency, $f$  | Mid-Point |
|---------|-----------------|-----------|
| 18 – 25 | 13              | 21.5      |
| 26 – 33 | 8               | 29.5      |
| 34 – 41 | 4               | 37.5      |
| 42 – 49 | 3               | 45.5      |
| 50 – 57 | 2               | 53.5      |
|         | $\Sigma f = 30$ |           |



# Relative Frequency Histogram

- A **relative frequency** histogram has the same shape and the same horizontal scale as the corresponding frequency histogram.

| Class   | Frequency, $f$ | Relative Frequency |
|---------|----------------|--------------------|
| 18 – 25 | 13             | 0.433              |
| 26 – 33 | 8              | 0.267              |
| 34 – 41 | 4              | 0.133              |
| 42 – 49 | 3              | 0.100              |
| 50 – 57 | 2              | 0.067              |
|         |                |                    |



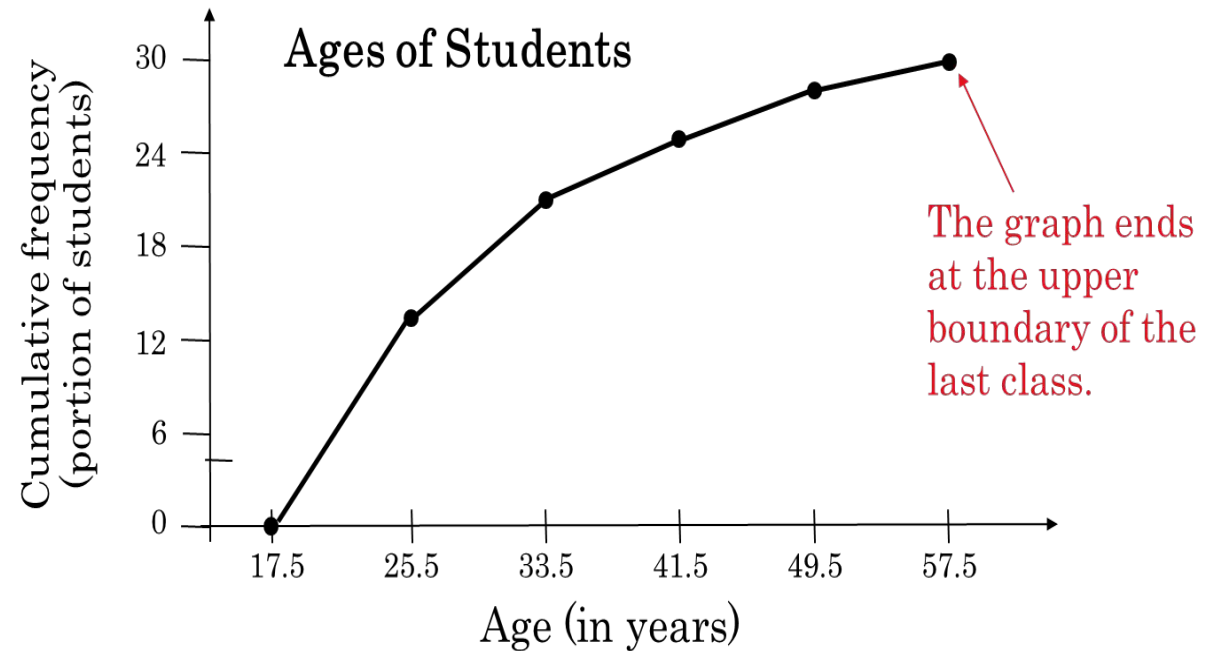


# Cumulative Frequency Graph

- A **cumulative frequency graph or ogive**, is a line graph that displays the cumulative frequency of each class at its upper class boundary.

Ages of Students

| Class   | Frequency, $f$  | Cumulative Frequency |
|---------|-----------------|----------------------|
| 18 – 25 | 13              | 13                   |
| 26 – 33 | + 8             | 21                   |
| 34 – 41 | + 4             | 25                   |
| 42 – 49 | + 3             | 28                   |
| 50 – 57 | + 2             | 30                   |
|         | $\Sigma f = 30$ |                      |



# Stem-and-Leaf Plot

---

- In a **stem-and-leaf plot**, each number is separated into a stem (usually the entry's leftmost digits) and a leaf (usually the rightmost digit).
- **Example:**
  - The following data represents the ages of 30 students in a statistics class. Display the data in a stem-and-leaf plot.

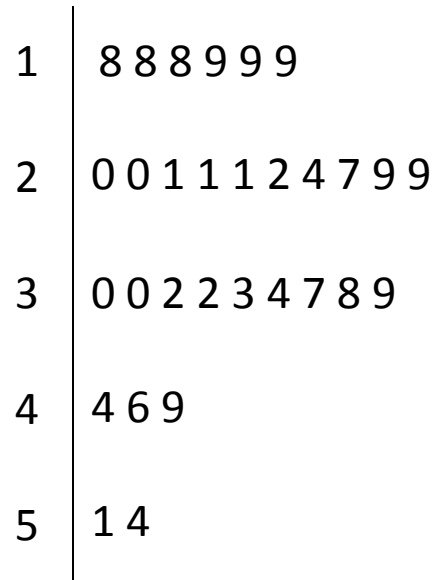
| Ages of Students |    |    |    |    |    |
|------------------|----|----|----|----|----|
| 18               | 20 | 21 | 27 | 29 | 20 |
| 19               | 30 | 32 | 19 | 34 | 19 |
| 24               | 29 | 18 | 37 | 38 | 22 |
| 30               | 39 | 32 | 44 | 33 | 46 |
| 54               | 49 | 18 | 51 | 21 | 21 |

---

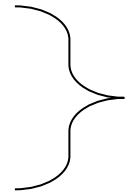
# Stem-and-Leaf Plot

---

**Ages of Students**



Key: 1|8 = 18



Most of the values lie between 20 and 39.

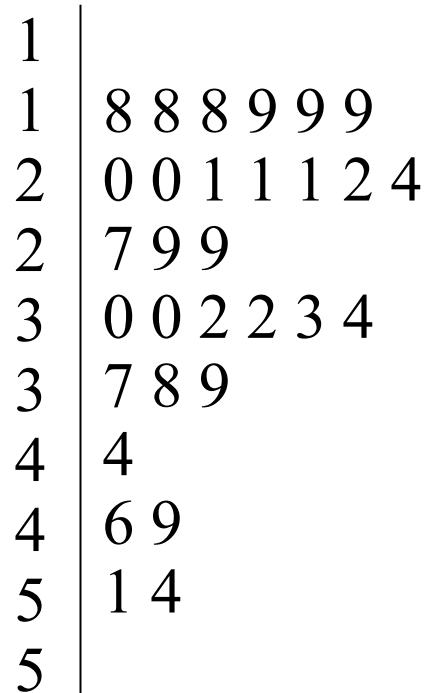
This graph allows us to see the shape of the data as well as the actual values.

# Stem-and-Leaf Plot

---

- **Example:**

- a stem-and-leaf plot that has two lines for each stem.



Key: 1|8 = 18

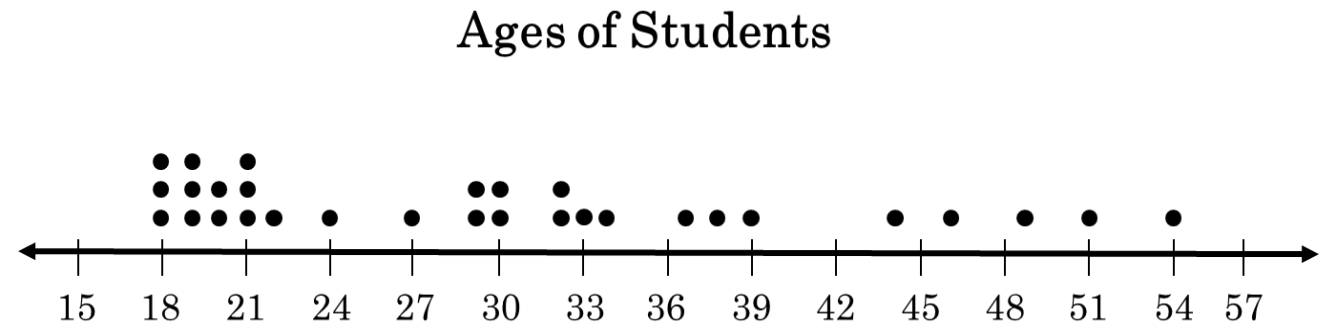
**From this graph, we can conclude that more than 50% of the data lie between 20 and 34.**

# Dot Plot

---

- In a **dot plot**, each data entry is plotted, using a point, above a horizontal axis.
- **Example:**
  - **dot plot to display the ages of the 30 students in the class.**

| Ages of Students |    |    |    |    |    |
|------------------|----|----|----|----|----|
| 18               | 20 | 21 | 27 | 29 | 20 |
| 19               | 30 | 32 | 19 | 34 | 19 |
| 24               | 29 | 18 | 37 | 38 | 22 |
| 30               | 39 | 32 | 44 | 33 | 46 |
| 54               | 49 | 18 | 51 | 21 | 21 |



From this graph, we can conclude that most of the values lie between 18 and 32.

# Pie Chart

---

- A pie chart is a circle that is divided into sectors that represent categories. The area of each sector is proportional to the frequency of each category.

Accidental Deaths in the USA in 2002

| Type                     | Frequency |
|--------------------------|-----------|
| Motor Vehicle            | 43,500    |
| Falls                    | 12,200    |
| Poison                   | 6,400     |
| Drowning                 | 4,600     |
| Fire                     | 4,200     |
| Ingestion of Food/Object | 2,900     |
| Firearms                 | 1,400     |

---

# Pie Chart

---

- To create a pie chart for the data, find the relative frequency (percent) of each category.

| Type                     | Frequency | Relative Frequency |
|--------------------------|-----------|--------------------|
| Motor Vehicle            | 43,500    | 0.578              |
| Falls                    | 12,200    | 0.162              |
| Poison                   | 6,400     | 0.085              |
| Drowning                 | 4,600     | 0.061              |
| Fire                     | 4,200     | 0.056              |
| Ingestion of Food/Object | 2,900     | 0.039              |
| Firearms                 | 1,400     | 0.019              |

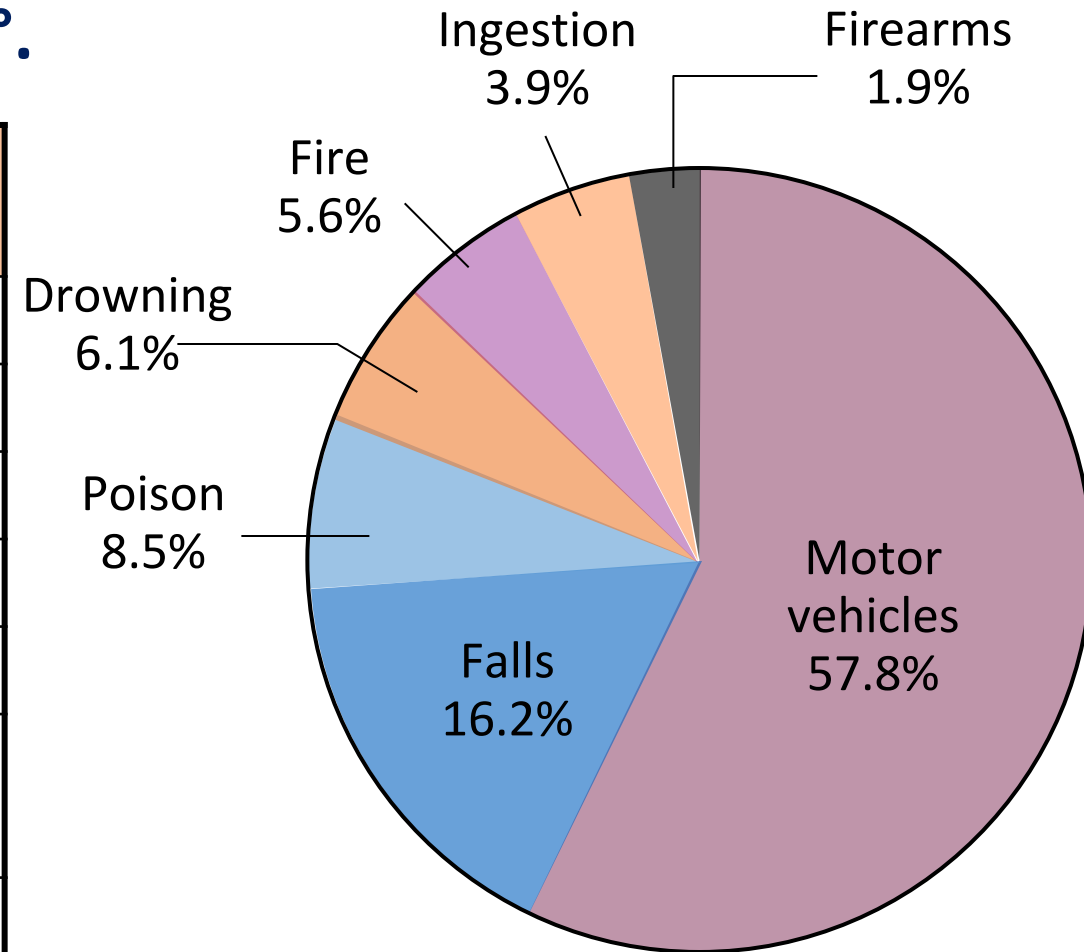
$n = 75,200$

---

# Pie Chart

- Next, find the central angle. To find the central angle, multiply the relative frequency by  $360^\circ$ .

| Type                     | Frequency | Relative Frequency | Angle         |
|--------------------------|-----------|--------------------|---------------|
| Motor Vehicle            | 43,500    | 0.578              | $208.2^\circ$ |
| Falls                    | 12,200    | 0.162              | $58.4^\circ$  |
| Poison                   | 6,400     | 0.085              | $30.6^\circ$  |
| Drowning                 | 4,600     | 0.061              | $22.0^\circ$  |
| Fire                     | 4,200     | 0.056              | $20.1^\circ$  |
| Ingestion of Food/Object | 2,900     | 0.039              | $13.9^\circ$  |
| Firearms                 | 1,400     | 0.019              | $6.7^\circ$   |



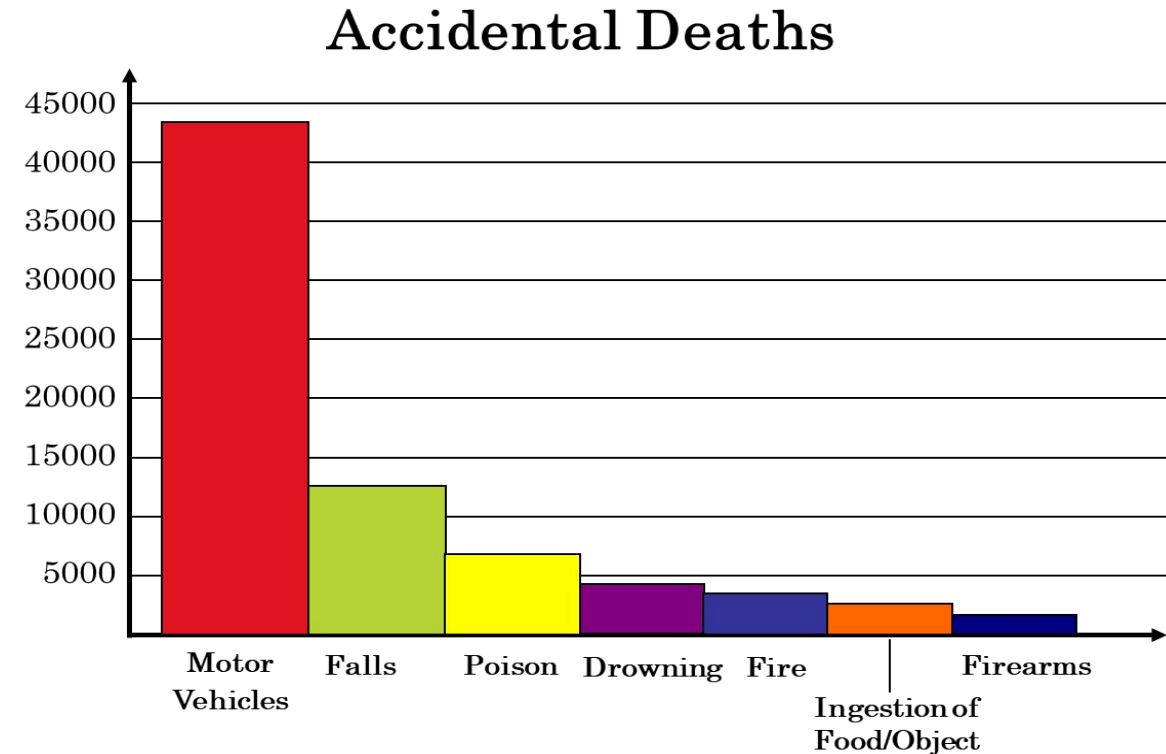


# Pareto Chart

- A Pareto chart is a vertical bar graph in which the height of each bar represents the frequency. The bars are placed in order of decreasing height, with the tallest bar to the left.

Accidental Deaths in the USA in 2002

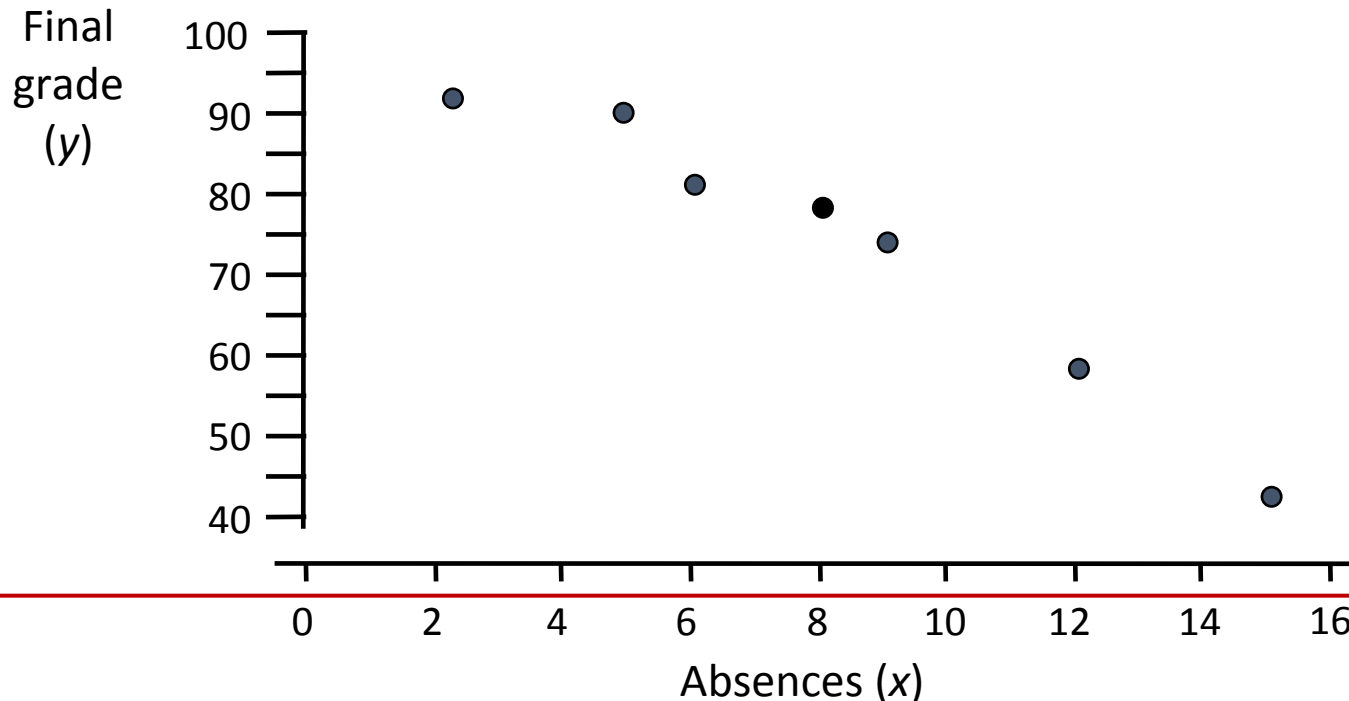
| Type                     | Frequency |
|--------------------------|-----------|
| Motor Vehicle            | 43,500    |
| Falls                    | 12,200    |
| Poison                   | 6,400     |
| Drowning                 | 4,600     |
| Fire                     | 4,200     |
| Ingestion of Food/Object | 2,900     |
| Firearms                 | 1,400     |



# Scatter Plot

- In a **scatter plot**, the ordered pairs are graphed as points in a coordinate plane. The scatter plot is used to show the relationship between two quantitative variables.
- The following scatter plot represents the relationship between the number of absences from a class during the semester and the final grade.

| Absences<br>$x$ | Grade<br>$y$ |
|-----------------|--------------|
| 8               | 78           |
| 2               | 92           |
| 5               | 90           |
| 12              | 58           |
| 15              | 43           |
| 9               | 74           |
| 6               | 81           |



From the scatter plot, you can see that as the number of absences increases, the final grade tends to decrease.

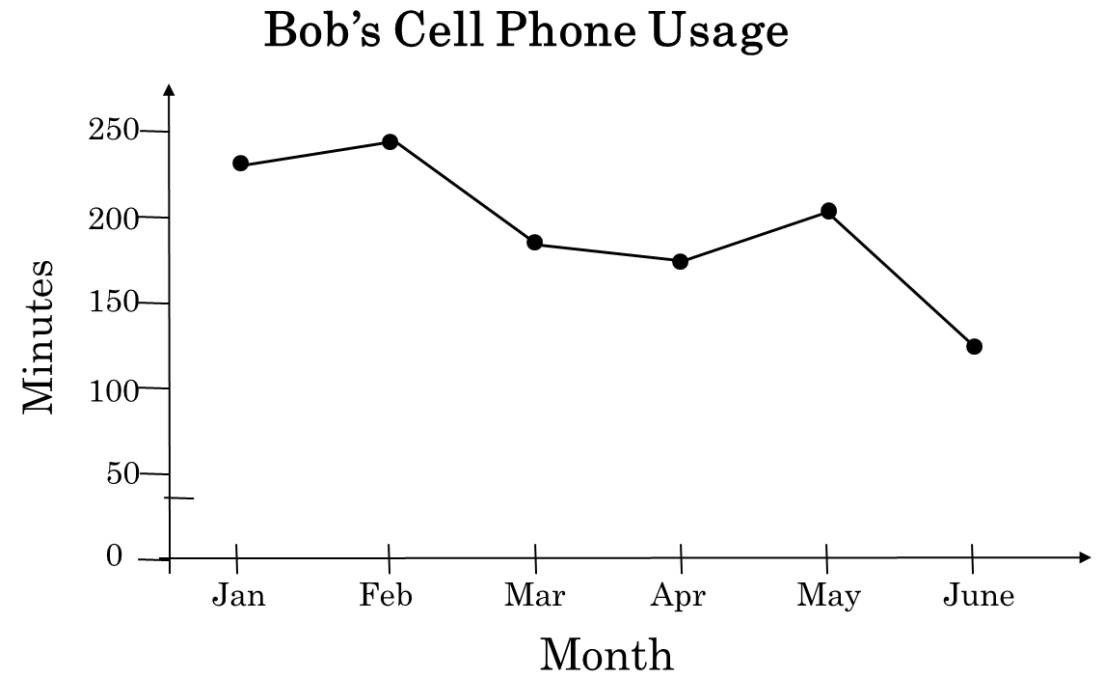
# Times Series Chart

---

- A time series chart is used to graph a time series.
- A data set that is composed of quantitative data entries taken at regular intervals over a period of time is a time series.
- **Example:**

The following table lists the number of minutes Bob used on his cell phone for the last six months.

| Month    | Minutes |
|----------|---------|
| January  | 236     |
| February | 242     |
| March    | 188     |
| April    | 175     |
| May      | 199     |
| June     | 135     |



# Measure of Central Tendency

# The Mean

---

- The mean of a data set is the sum of the data entries divided by the number of entries.
- Population mean:  $\mu = \frac{1}{N} \sum x$
- Sample mean:  $\bar{x} = \frac{1}{n} \sum x$
- Example
  - The following are the ages of all seven employees of a small company: 53, 32, 61, 57, 39, 44, 57
    - $\mu = \frac{1}{N} \sum x = \frac{342}{7} = 49 \text{ years}$

The mean age of the employees is 49 years.

# The Median

---

- The **median** of a data set is the value that lies in the middle of the data when the data set is ordered.
  - If the data set has an odd number of entries, the median is the middle data entry.
  - If the data set has an even number of entries, the median is the mean of the two middle data entries.
- **Example**
  - The following are the ages of all seven employees of a small company: 53 32 61 57 39 44 57

**SORTED DATA**

32 39 44 **53** 57 57 61

The median age of the employees is 53 years.

# The Mode

---

- The **mode** of a data set is the data entry that occurs with the greatest frequency.
  - If no entry is repeated, the data set has no mode.
  - If two entries occur with the same greatest frequency, each entry is a mode and the data set is called bimodal.
- **Example**
  - The following are the ages of all seven employees of a small company:

53   32   61   57   39   44   57

# Outliers

---

- An outlier is a data point or observation whose value is quite different from the others in the data set being analyzed.
- no absolute agreement about how to define outliers



# Comparing the Mean, Median and Mode

---

- **Example**

53 32 61 57 39 44 57

- A 29-year-old employee joins the company, and the ages of the employees are now:

53 32 61 57 39 44 57 29

- **Recalculate the mean, the median, and the mode.**

- **Mean = 46.5**

The mean takes every value into account but is affected by the outlier.

- **Median = 48.5**

- **Mode = 57**

# Weighted Mean

---

- A **weighted mean** is the mean of a data set whose entries have varying weights. A weighted mean is given by

$$\bar{x} = \frac{\sum xw}{\sum w}$$

- where **w** is the weight of each entry **x**.
- **Example**
  - **Grades in a statistics class are weighted as follows:**
    - Tests are worth 50% of the grade, homework is worth 30% of the grade and the final is worth 20% of the grade. A student receives a total of 80 points on tests, 100 points on homework, and 85 points on his final. What is his current grade?

| Source   | Score, $x$ | Weight, $w$ | $xw$ |
|----------|------------|-------------|------|
| Tests    | 80         | 0.50        | 40   |
| Homework | 100        | 0.30        | 30   |
| Final    | 85         | 0.20        | 17   |

$$\bar{x} = \frac{\sum xw}{\sum w} = \frac{87}{100} = 0.87$$

# Mean of a Frequency Distribution

---

- The mean of a frequency distribution for a sample is approximated by

$$\bar{x} = \frac{\sum(x \cdot f)}{n}$$

Note that  $n = \sum f$

- where  $x$  and  $f$  are the midpoints and frequencies of the classes.
-

# Mean of a Frequency Distribution

---

| Class   | $x$  | $f$      | $(x \cdot f)$    |
|---------|------|----------|------------------|
| 18 – 25 | 21.5 | 13       | 279.5            |
| 26 – 33 | 29.5 | 8        | 236.0            |
| 34 – 41 | 37.5 | 4        | 150.0            |
| 42 – 49 | 45.5 | 3        | 136.5            |
| 50 – 57 | 53.5 | 2        | 107.0            |
|         |      | $n = 30$ | $\Sigma = 909.0$ |

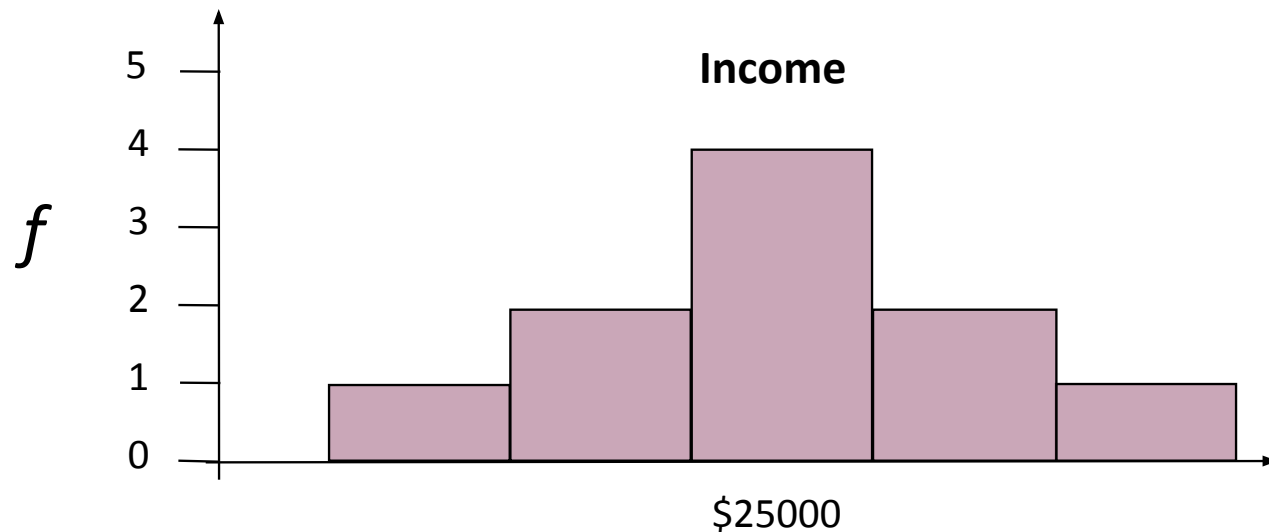
$$\bar{x} = \frac{\Sigma(x \cdot f)}{n} = \frac{909}{30} = 30.3$$

The mean age of the students is 30.3 years.

---

# Shapes of Distributions

- A frequency distribution is **symmetric** when a **vertical line** can be drawn through the **middle** of a graph of the distribution and the **resulting halves** are approximately the **mirror images**.



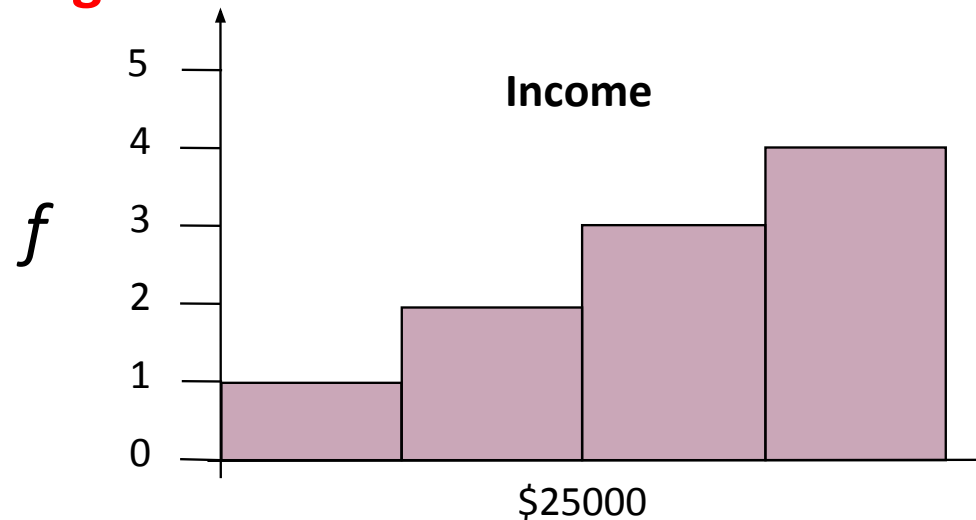
10 Annual Incomes

|        |
|--------|
| 15,000 |
| 20,000 |
| 22,000 |
| 24,000 |
| 25,000 |
| 25,000 |
| 26,000 |
| 28,000 |
| 30,000 |
| 35,000 |

mean = median = mode  
= \$25,000

# Shapes of Distributions

- A frequency distribution is skewed if the “tail” of the graph elongates more to one side than to the other.
  - A distribution is skewed left (negatively skewed) if its tail extends to the left.
  - A distribution is skewed right (positively skewed) if its tail extends to the right.



**Mean < Median**

10 Annual Incomes

|        |
|--------|
| 0      |
| 20,000 |
| 22,000 |
| 24,000 |
| 25,000 |
| 25,000 |
| 26,000 |
| 28,000 |
| 30,000 |
| 35,000 |

mean = \$23,500  
median = mode = \$25,000

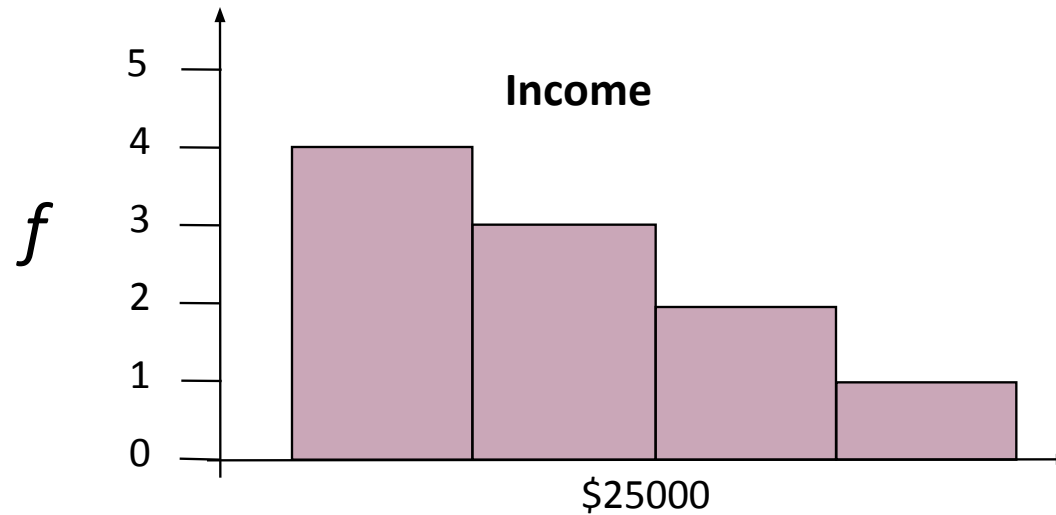
# Skewed Right Distribution

---

10 Annual Incomes

|           |
|-----------|
| 15,000    |
| 20,000    |
| 22,000    |
| 24,000    |
| 25,000    |
| 25,000    |
| 26,000    |
| 28,000    |
| 30,000    |
| 1,000,000 |

mean = \$121,500  
median = mode = \$25,000



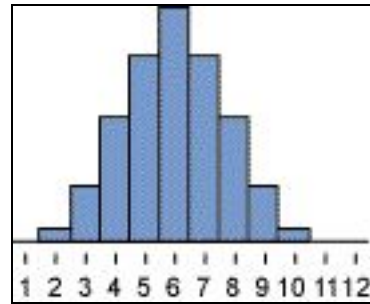
**Mean > Median**

---

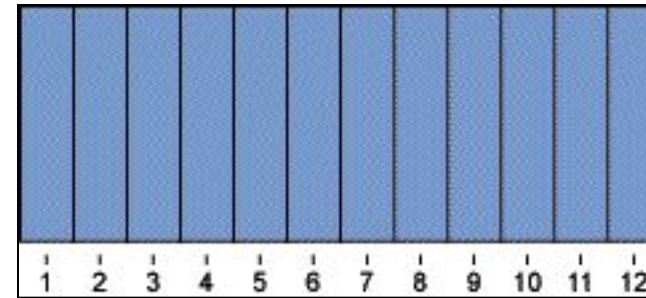
# Summary of Shapes of Distributions

---

**Symmetric**

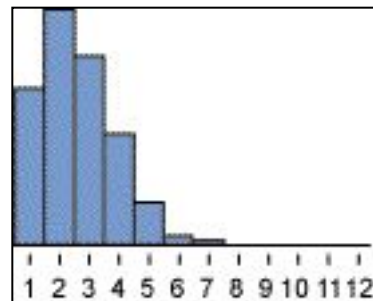


**Uniform**



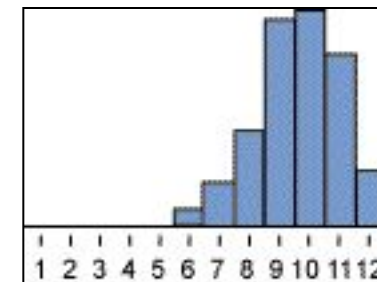
Mean = Median

**Skewed right**



Mean > Median

**Skewed left**



Mean < Median



# Measures of Dispersion

# Range

---

- The **range** of a data set is the difference between the maximum and minimum data entries in the set.
- $\text{Range} = (\text{Maximum data entry}) - (\text{Minimum data entry})$
- **Example:**
  - The following data are the closing prices for a certain stock on ten successive Fridays.

|       |    |    |    |    |    |    |    |    |    |    |
|-------|----|----|----|----|----|----|----|----|----|----|
| Stock | 56 | 56 | 57 | 58 | 61 | 63 | 63 | 67 | 67 | 67 |
|-------|----|----|----|----|----|----|----|----|----|----|

The range is  $67 - 56 = 11$ .

---

# Deviation

---

- The deviation of an entry  $x$  in a population data set is the difference between the entry and the mean  $\mu$  of the data set.
  - Deviation of  $x = x - \mu$
- Example:
  - The following data are the closing prices for a certain stock on five successive Fridays.

The mean stock price is  
 $\mu = 305/5 = 61$ .

| Stock<br>$x$     | Deviation<br>$x - \mu$ |
|------------------|------------------------|
| 56               | $56 - 61 = -5$         |
| 58               | $58 - 61 = -3$         |
| 61               | $61 - 61 = 0$          |
| 63               | $63 - 61 = 2$          |
| 67               | $67 - 61 = 6$          |
| $\Sigma x = 305$ | $\Sigma(x - \mu) = 0$  |

# Variance and Standard Deviation

---

- The population variance of a population data set of  $N$  entries is

$$\sigma^2 = \frac{\sum(\mu - )^2}{N}.$$

- The population standard deviation of a population data set of  $N$  entries is the square root of the population variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(\mu - )^2}{N}}.$$

## Finding the Population Standard Deviation

---

- **Example:**

- The following data are the closing prices for a certain stock on five successive Fridays. The population mean is 61.

| Stock<br>$x$     | Deviation<br>$x - \mu$ | Squared<br>$(x - \mu)^2$ |
|------------------|------------------------|--------------------------|
| 56               | -5                     | 25                       |
| 58               | -3                     | 9                        |
| 61               | 0                      | 0                        |
| 63               | 2                      | 4                        |
| 67               | 6                      | 36                       |
| $\Sigma x = 305$ | $\Sigma(x - \mu) = 0$  | $\Sigma(x - \mu)^2 = 74$ |

$$\sigma^2 = \frac{\Sigma(\cancel{x} - )^2}{N} = \frac{74}{5} = 14.8$$

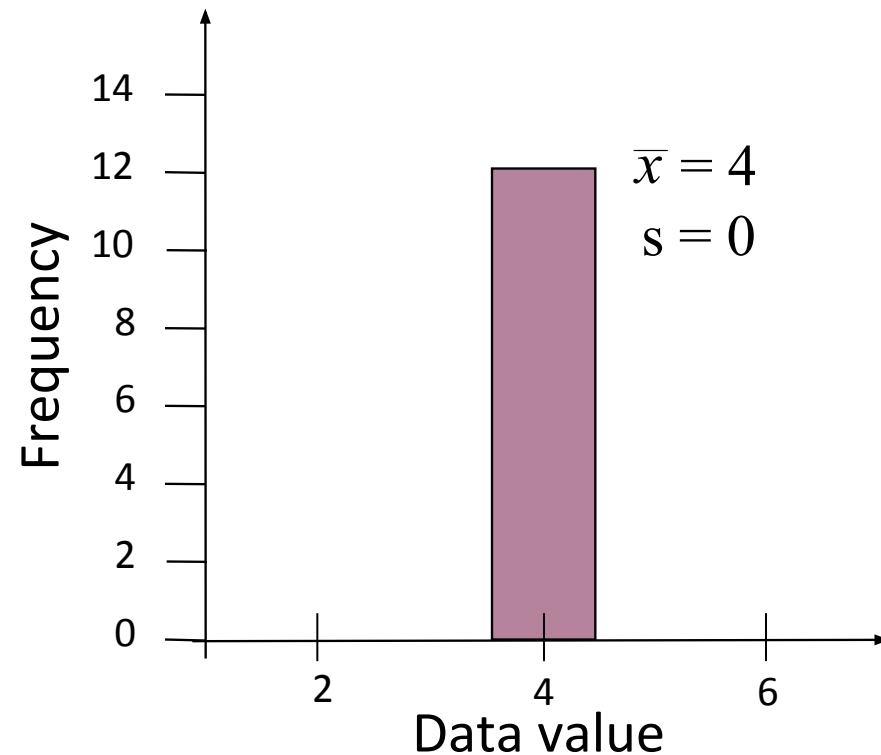
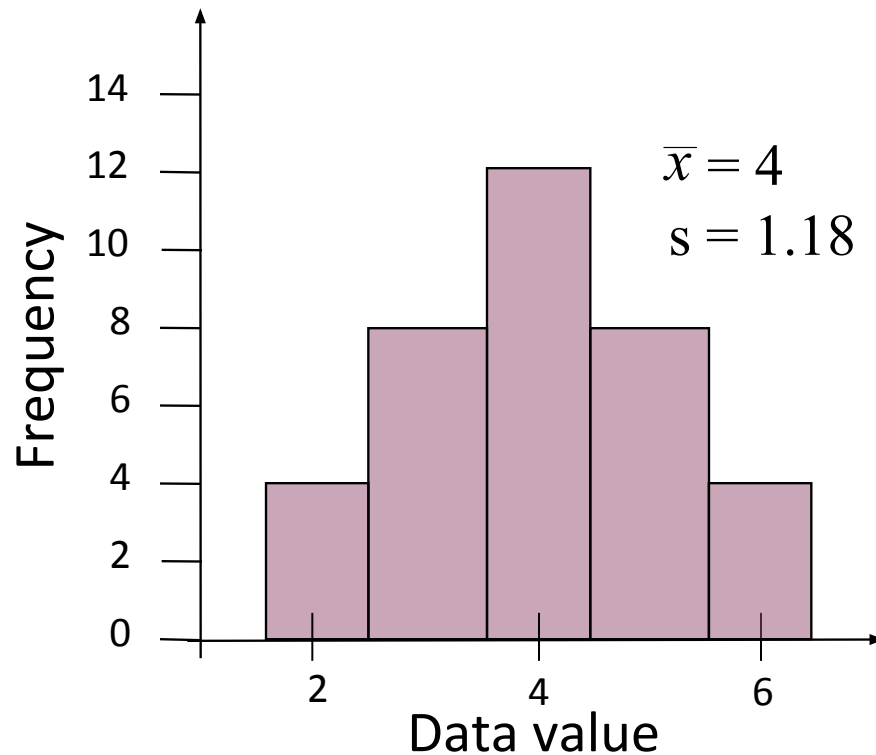
$$\sigma = \sqrt{\frac{\Sigma(\cancel{x} - )^2}{N}} = \sqrt{14.8} \approx 3.85$$

$$\sigma \approx \$3.85$$

## Interpreting Standard Deviation

---

- standard deviation is a measure of the typical amount an entry deviates from the mean.
- The more the entries are spread out, the greater the standard deviation.



# Empirical Rule (68-95-99.7%)

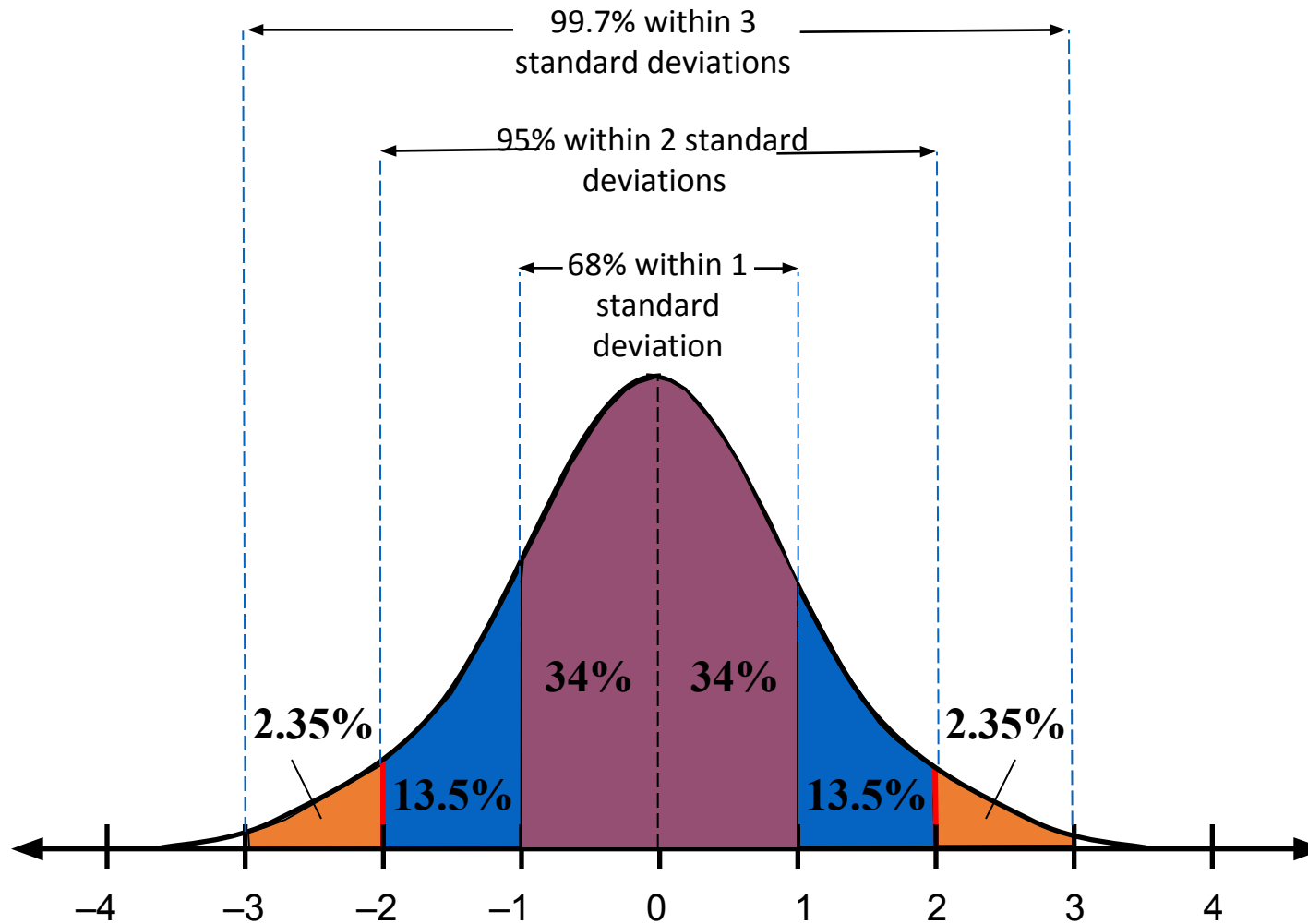
---

- Empirical Rule

- For data with a (symmetric) bell-shaped distribution, the standard deviation has the following characteristics.
  1. About 68% of the data lie within one standard deviation of the mean.
  2. About 95% of the data lie within two standard deviations of the mean.
  3. About 99.7% of the data lie within three standard deviation of the mean.

# Empirical Rule (68-95-99.7%)

---





# Standard Deviation for Grouped Data

---

- Sample standard deviation =  $s = \sqrt{\frac{\sum(x - \bar{x})^2 f}{n - 1}}$
- where  $n = \sum f$  is the number of entries in the data set, and  $x$  is the data value or the midpoint of an interval.

| Class   | $x$  | $f$      | $x - \bar{x}$ | $(x - \bar{x})^2$  | $(x - \bar{x})^2 f$ |
|---------|------|----------|---------------|--------------------|---------------------|
| 18 – 25 | 21.5 | 13       | – 8.8         | 77.44              | 1006.72             |
| 26 – 33 | 29.5 | 8        | – 0.8         | 0.64               | 5.12                |
| 34 – 41 | 37.5 | 4        | 7.2           | 51.84              | 207.36              |
| 42 – 49 | 45.5 | 3        | 15.2          | 231.04             | 693.12              |
| 50 – 57 | 53.5 | 2        | 23.2          | 538.24             | 1076.48             |
|         |      | $n = 30$ |               | $\Sigma = 2988.80$ |                     |

$$s = \sqrt{\frac{\sum(x - \bar{x})^2 f}{n - 1}} = \sqrt{\frac{2988.8}{29}} = \sqrt{103.06} = 10.2$$

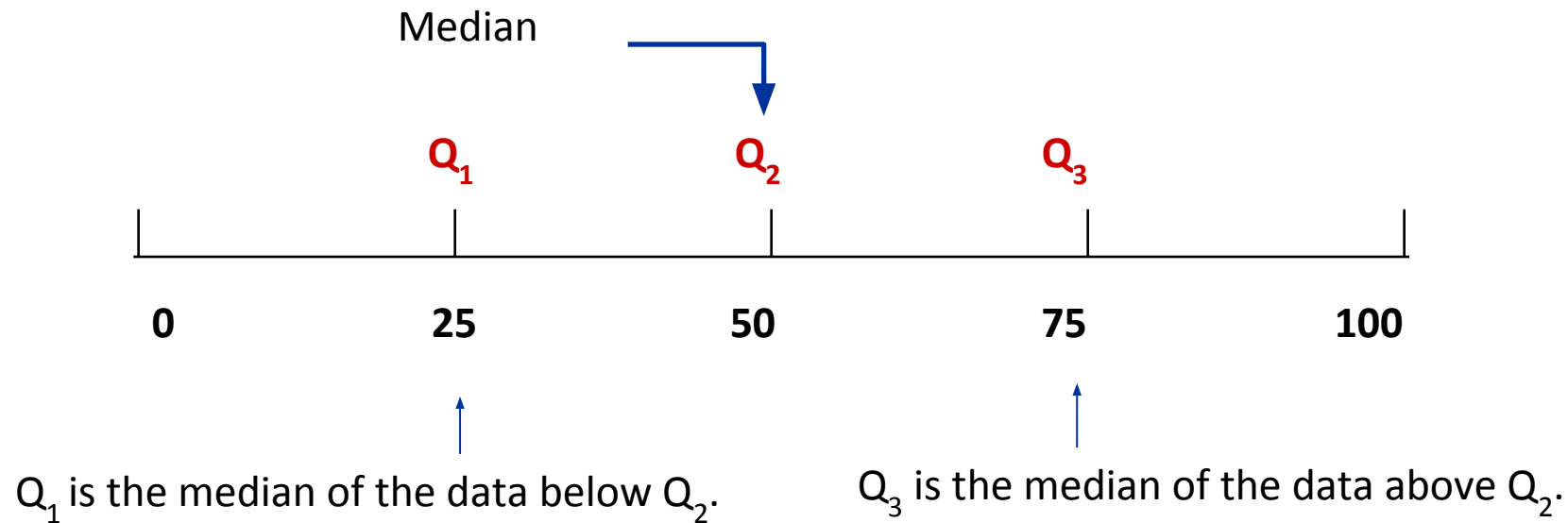
---

# Measures of Position

# Quartiles

---

- The three quartiles,  $Q_1$ ,  $Q_2$ , and  $Q_3$ , approximately divide an ordered data set into four equal parts.



# Finding Quartiles

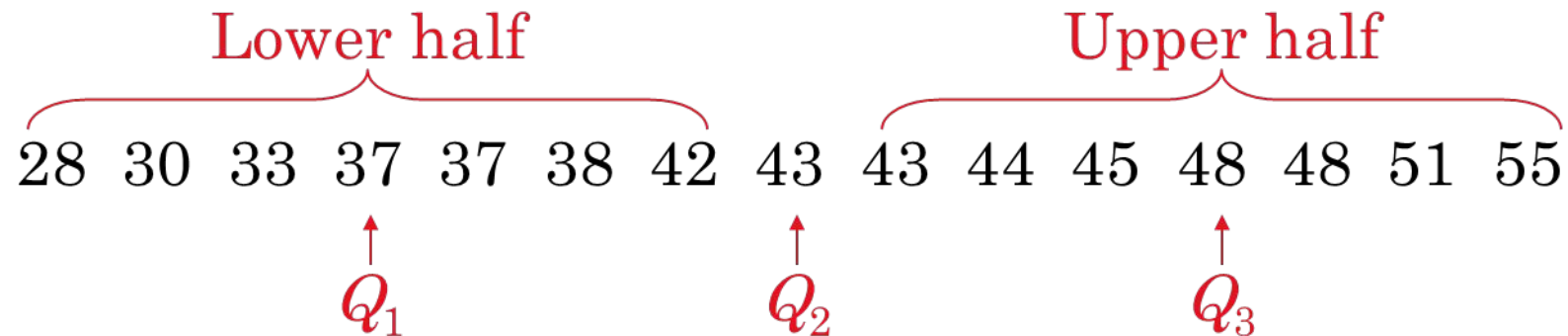
---

- **Example:**

- The quiz scores for 15 students are listed below.

28 43 48 51 43 30 55 44 48 33 45 37 37 42 38

Order the data.



- About one fourth of the students scores 37 or less; about one half score 43 or less; and about three fourths score 48 or less.
-

# Interquartile Range

---

- The interquartile range (IQR) of a data set is the difference between the third and first quartiles.
- Interquartile range (IQR) =  $Q_3 - Q_1$ .
- Example:
  - The quartiles for 15 quiz scores are listed below. Find the interquartile range.

$$Q_1 = 37$$

$$Q_2 = 43$$

$$Q_3 = 48$$

$$\begin{aligned}(\text{IQR}) &= Q_3 - Q_1 \\ &= 48 - 37 \\ &= 11\end{aligned}$$

The quiz scores in the middle portion of the data set vary by at most 11 points.

# Box and Whisker Plot

---

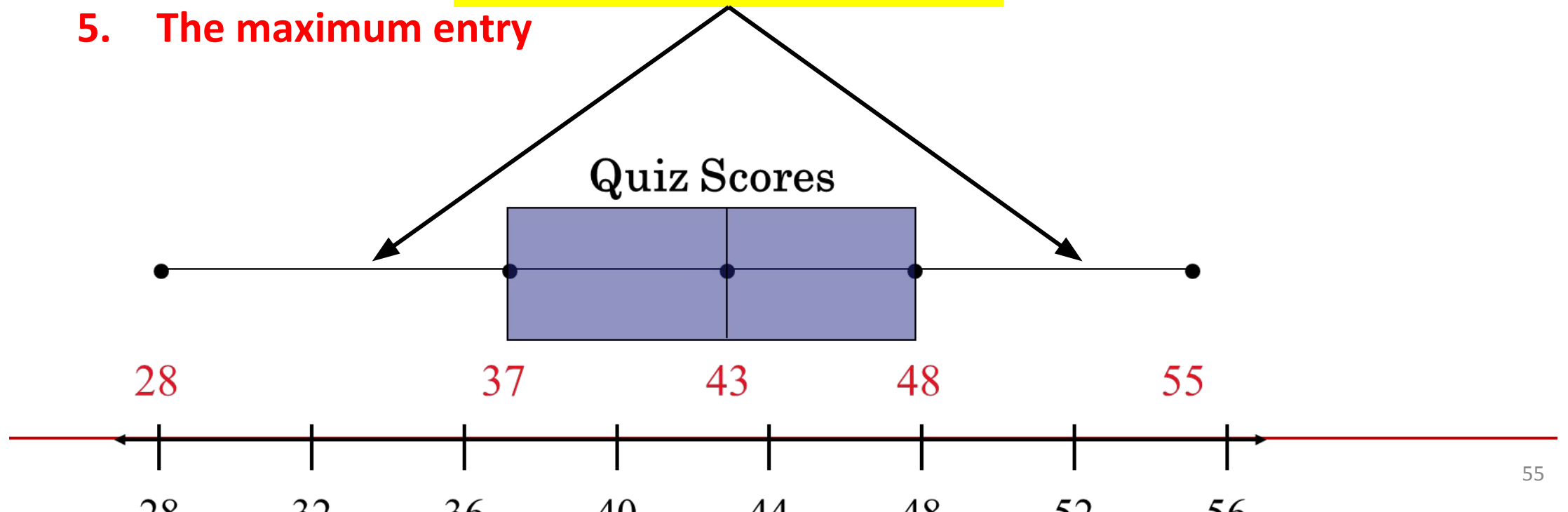
- A box-and-whisker plot is an exploratory data analysis tool that highlights the important features of a data set.
- The five-number summary is used to draw the graph.
  1. The minimum entry
  2. Q1
  3. Q2 (median)
  4. Q3
  5. The maximum entry
- Example:
  - Use the data from the 15 quiz scores to draw a box-and-whisker plot.  
28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

# Box and Whisker Plot

- Five-number summary

1. The minimum entry
2. Q1
3. Q2 (median)
4. Q3
5. The maximum entry

Whisker: Indicate variability outside the upper and lower quartiles.



# Percentiles and Deciles

---

- **Percentiles** divide an ordered data set into 100 parts. There are 99 percentiles:  $P_1, P_2, P_3 \dots P_{99}$ .
    - Percentile = (Number of Values Below “x” / Total Number of Values)  $\times 100$
    - Example:
      - The scores for student are 40, 45, 49, 53, 61, 65, 71, 79, 85, 91.
      - percentile for score 71 =  $(6/10) \times 100 = 60$
  - **Deciles** divide an ordered data set into 10 parts. There are 9 deciles:  $D_1, D_2, D_3 \dots D_9$ .
-



# Reference

---

- Probability and Statistics by Prof. Kevin M. Riordan