Chapter 1: Descriptive Statistics – PART 2

Manuella Lech Cantuaria

Victoria Blanes-Vidal

The Maersk Mc-Kinney Moller Institute

Applied AI and Data Science

Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential

1.2. Variables and Types of Data

1.3. Data organization and histograms

1.4. Measures of:

Central Tendency (Location)

Variation (Dispersion)

Position

1.5. Data representation: frequency distributions and graphs

1.6. Shapes of frequency distributions: Skewness and kurtosis

Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential
1.2. Variables and Types of Data
1.3. Data organization and histograms
1.4. Measures of:
    Central Tendency (Location)
    Variation (Dispersion)
    Position
1.5. Data representation: frequency distributions and graphs
1.6. Shapes of frequency distributions: Skewness and kurtosis

Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential

1.2. Variables and Types of Data

1.3. Data organization and histograms

1.4. Measures of:

Central Tendency (Location)

Variation (Dispersion)

Position

1.5. Data representation: frequency distributions and graphs

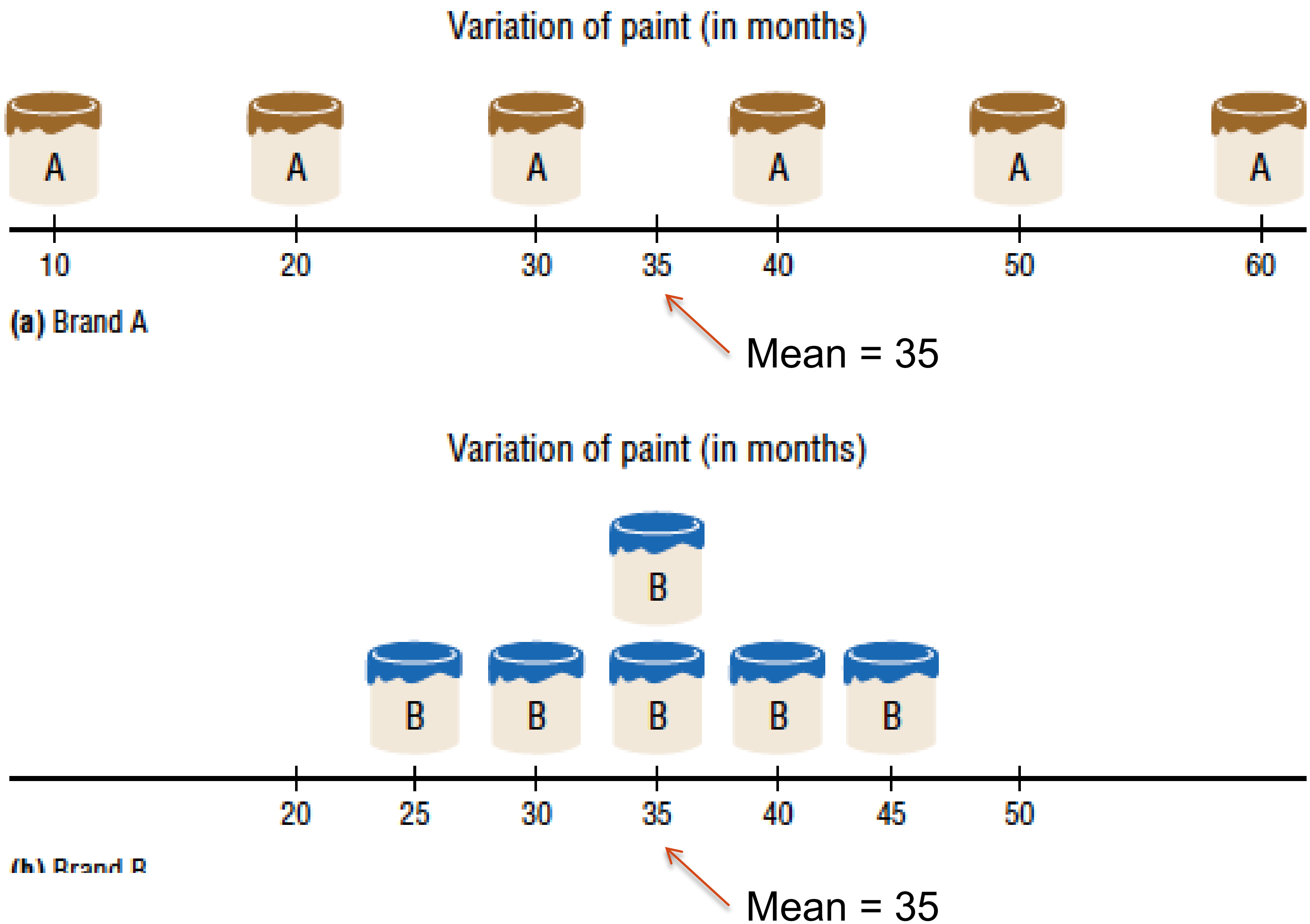
1.6. Shapes of frequency distributions: Skewness and kurtosis

Measures of Variation (Dispersion)

Example: Outdoor Paint

Two experimental brands of outdoor paint are tested to see how long each will last before fading. Six cans of each brand constitute a small sample. The results (in months) are shown. Find the mean.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25


Variation of paint (in months)
A
A
A
A
A
A
10
20
30
35
40
50
60
(a) Brand A
Mean = 35
Variation of paint (in months)
B
B
B
B
B
B
20
25
30
35
40
45
50
Mean = 35

Measures of Variation (Dispersion)

Measures of dispersion are concerned with the distribution of values around the mean in data.

How Can We Measure **Variability**?

- Range
- Variance
- Standard Deviation
- Coefficient of Variation

Range

- The **range** is the difference between the highest and lowest values in a data set.

$$R = Highest - Lowest$$

Variance & Standard Deviation

- The standard deviation and variance are measures of how spread out your data are.

- The **variance** is the average of the squares of the distance each value is from the mean.
- The **standard deviation** is the square root of the variance.

Measures of Variation: Variance & Standard Deviation (Population Theoretical Model)

- The **population variance** is

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

- The **population standard deviation** is

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

Measures of Variation: Variance & Standard Deviation (Sample Theoretical Model)

- The **sample variance** is

$$s^2 = \frac{\sum\left(X - \bar{X}\right)^2}{n-1}$$

- The **sample standard deviation** is

$$s = \sqrt{\frac{\sum\left(X - \bar{X}\right)^2}{n-1}}$$

The value of variance calculated from sample data is higher than the value that could have been found out by using population data. The logic of doing that is to compensate our lack of information about the population data.

Coefficient of Variation (relative standard deviation)

The **coefficient of variation** is the standard deviation divided by the mean, expressed as a percentage.

$$CV = \frac{s}{\bar{X}} \cdot 100\%$$

An Example Data Set

- **Daily low temperatures** recorded in a town (01/18-01/31, 2005, °F)

Jan. 18 – 11	Jan. 25 – 25
Jan. 19 – 11	Jan. 26 – 33
Jan. 20 – 25	Jan. 27 – 22
Jan. 21 – 29	Jan. 28 – 18
Jan. 22 – 27	Jan. 29 – 19
Jan. 23 – 14	Jan. 30 – 30
Jan. 24 – 11	Jan. 31 – 27

- For these 14 values, we will calculate all measures of **dispersion**

Range

- **Range** – The difference between the largest and the smallest values
- (1) **Sort** the data in ascending order

  → 11, 11, 11, 14, 18, 19, 22, 25, 25, 27, 27, 29, 30, 33
- (2) **Find** the largest value

  → max = 33
- (3) **Find** the smallest value

  → min = 11
- (4) **Calculate** the **range**

  → range = 33 – 11 = 22

Variance

- (1) **Calculate** the **mean**

  → $\bar{x} = 25.7$

- (2) **Calculate** the **deviation** for each value

  → $x_i - \bar{x}$

Jan. 18	(11 – 25.7) = -10.57	Jan. 25	(25 – 25.7) = 3.43
Jan. 19	(11 – 25.7) = -10.57	Jan. 26	(33 – 25.7) = 11.43
Jan. 20	(25 – 25.7) = 3.43	Jan. 27	(22 – 25.7) = 0.43
Jan. 21	(29 – 25.7) = 7.43	Jan. 28	(18 – 25.7) = -3.57
Jan. 22	(27 – 25.7) = 5.43	Jan. 29	(19 – 25.7) = -2.57
Jan. 23	(14 – 25.7) = -7.57	Jan. 30	(30 – 25.7) = 8.42
Jan. 24	(11 – 25.7) = -10.57	Jan. 31	(27 – 25.7) = 5.42

Variance

- (3) **Square** each of the **deviations**

$$\rightarrow \quad (x_i - \bar{x})^2$$

Jan. 18	(-10.57)^2 = 111.76	Jan. 25	(3.43)^2 = 11.76
Jan. 19	(-10.57)^2 = 111.76	Jan. 26	(11.43)^2 = 130.61
Jan. 20	(3.43)^2 = 11.76	Jan. 27	(0.43)^2 = 0.18
Jan. 21	(7.43)^2 = 55.18	Jan. 28	(-3.57)^2 = 12.76
Jan. 22	(5.43)^2 = 29.57	Jan. 29	(-2.57)^2 = 6.61
Jan. 23	(7.57)^2 = 57.33	Jan. 30	(8.43)^2 = 71.04
Jan. 24	(-10.57)^2 = 111.76	Jan. 31	(5.43)^2 = 29.57

- (4) **Sum** the **squared** deviations

$$\rightarrow \quad \sum (x_i - \bar{x})^2 = 751.43$$

Variance

- (5) **Divide** the **sum of squares** by (n-1) for a sample

  →

$$\sum (x_i - \bar{x})^2 / (n-1)$$

= 751.43 / (14-1) = 57.8

- The **variance** of the Tmin (F) data set is 57.8

Standard Deviation

- (1) – (5)

  → **s^2** = 57.8

- (6) **Take the square root** of the **variance**

  → $$\sqrt{57.8} = 7.6$$

- The **standard deviation** (*s*) of the Tmin data is 7.6 (°F)

Coefficient of Variation

- (1) **Calculate mean**

→ $$\bar{x} = 25.7$$

- (2) **Calculate standard deviation**

→ $$s = \sqrt{\sum (x_i - \bar{x})^2 / (n-1)} = 7.6$$

- (3) **Divide standard deviation** by **mean**

→ $$\mathrm{CV} = \frac{s}{\bar{x}} \times 100\% = 7.6 / 25.7 \times 100\% = 29.58$$

Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential

1.2. Variables and Types of Data

1.3. Data organization and histograms

1.4. Measures of:

Central Tendency (Location)

Variation (Dispersion)

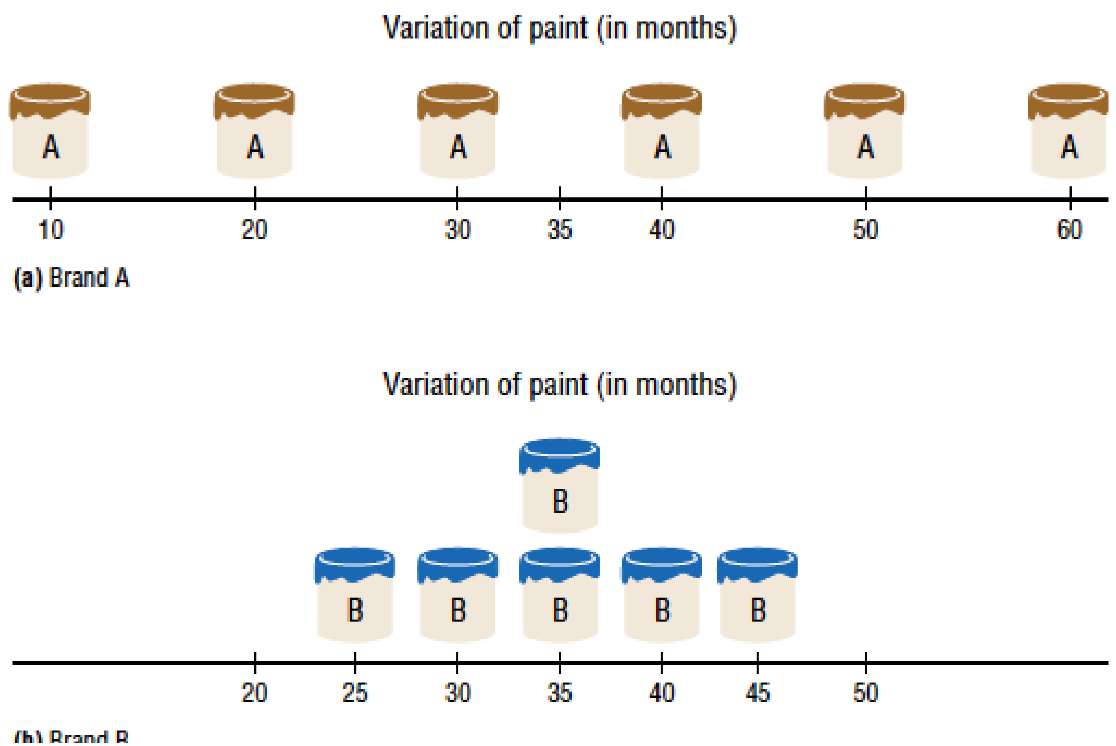
Position

1.5. Data representation: frequency distributions and graphs

1.6. Shapes of frequency distributions: Skewness and kurtosis

Measures of Position

Measures of position indicate the position of a value, relative to other values in a set of observations.



Measures of Position

Measures of position indicate the position of a value, relative to other values in a dataset.

- Z-score
- Percentile
- Decile and Quartile
- Outlier

Z-Score

Since data come from distributions with different means and different degrees of variability, it is common to **standardize** observations

One way to do this is to transform each observation into a z-score

$$z = \frac{x_i - \bar{x}}{s}$$

A **z-score** (aka, a standard score) indicates how many standard deviations an element is from the mean.

An Example Data Set

- **Daily low temperatures** recorded in a town (01/18-01/31, 2005, °F)

Jan. 18 – 11	Jan. 25 – 25
Jan. 19 – 11	Jan. 26 – 33
Jan. 20 – 25	Jan. 27 – 22
Jan. 21 – 29	Jan. 28 – 18
Jan. 22 – 27	Jan. 29 – 19
Jan. 23 – 14	Jan. 30 – 30
Jan. 24 – 11	Jan. 31 – 27

- We will calculate the **z-score** of all measures of 33°F

z-scores

- **Z-score** for maximum Tmin value (33 °F)
- (1) Calculate the **mean**

→ $$\bar{x} = 21.57$$

- (2) Calculate the **deviation**

→ $$x_i - \bar{x} = 11.43$$

- (3) Calculate the **standard deviation** (SD)

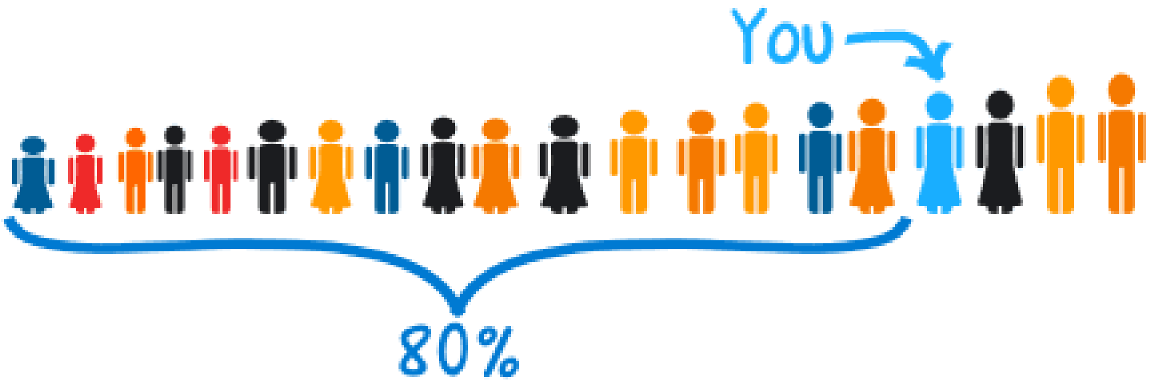
→ $$\sqrt{\sum (x_i - \bar{x})^2 / (n-1)} = 7.6$$

- (4) **Divide** the **deviation** by **standard deviation**

→ $$z = (x_i - \bar{x}) / s = 11.43 / 7.6 = 1.50$$

Measures of Position: Percentiles

- **Percentiles** separate the data set into 100 equal groups.
- A percentile rank for a datum represents the percentage of data values below the datum.



Example: You are the fourth tallest person in a group of 20 (80% of people are shorter than you):

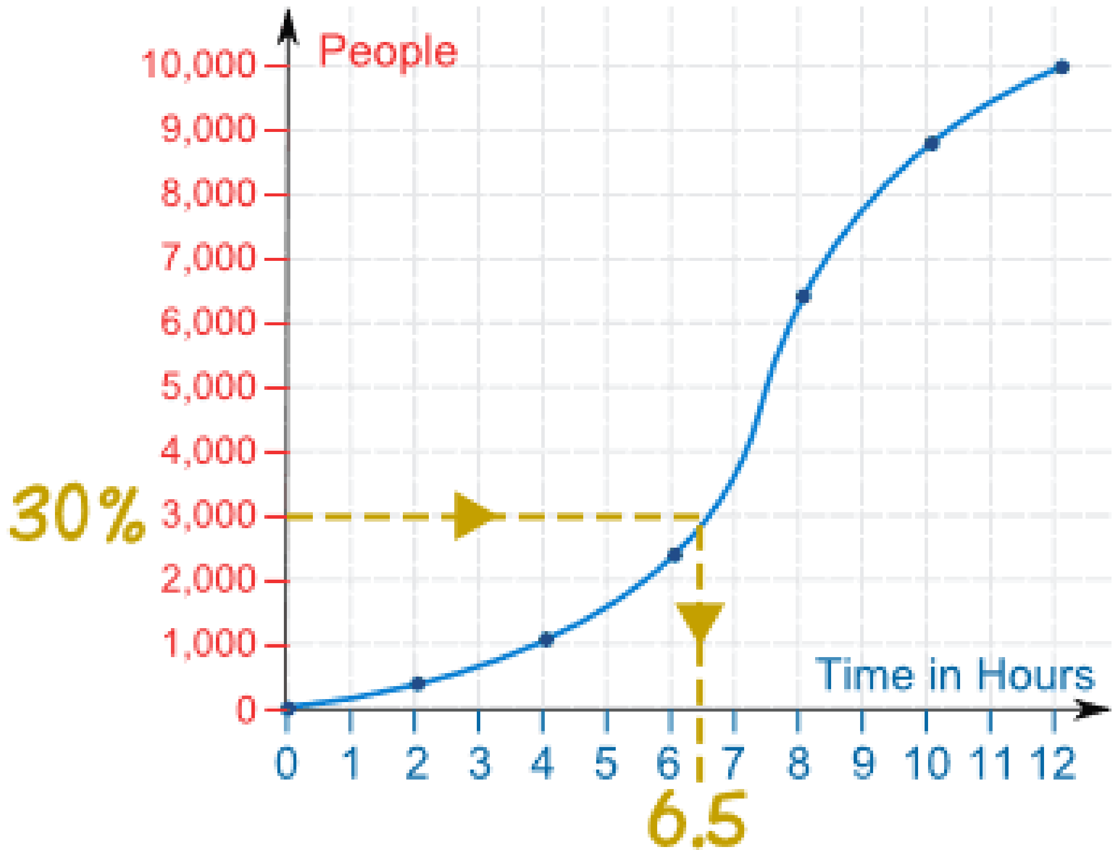
That means you are at the 80th percentile. If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

Measures of Position: Example of a Percentile Graph

A total of 10,000 people visited a shopping mall over 12 hours:

Time (hours)	People
0	0
2	350
4	1100
6	2400
8	6500
10	8850
12	10,000

Estimate the 30th percentile (when 30% of the visitors had arrived).



The 30th percentile occurs after about 6.5 hours.

Measures of Position: Quartiles and Deciles

- **Deciles** separate the data set into 10 equal groups.

  $D_1=P_{10}$, $D_4=P_{40}$
- **Quartiles** separate the data set into 4 equal groups.

  $Q_1=P_{25}$, $Q_2=MD$, $Q_3=P_{75}$

$$Q_2 = \text{median}(\text{Low},\text{High})$$
$$Q_1 = \text{median}(\text{Low},Q_2)$$
$$Q_3 = \text{median}(Q_2,\text{High})$$

- The **Interquartile Range**, $IQR = Q_3 - Q_1$.

Measures of Position: Outliers

- An **outlier** is an extremely high or low data value when compared with the rest of the data values.

- A data value:
  - less than $Q_1 - 1.5(\text{IQR})$

  Or

  - greater than $Q_3 + 1.5(\text{IQR})$

  can be considered an outlier.

Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential

1.2. Variables and Types of Data

1.3. Data organization and histograms

1.4. Measures of:

Central Tendency (Location)

Variation (Dispersion)

Position

1.5. Data representation: frequency distributions and graphs

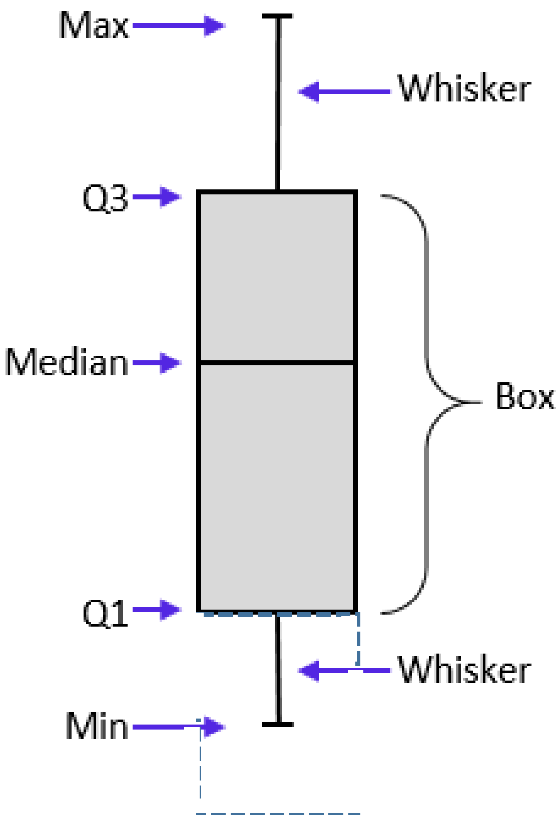
1.6. Shapes of frequency distributions: Skewness and kurtosis

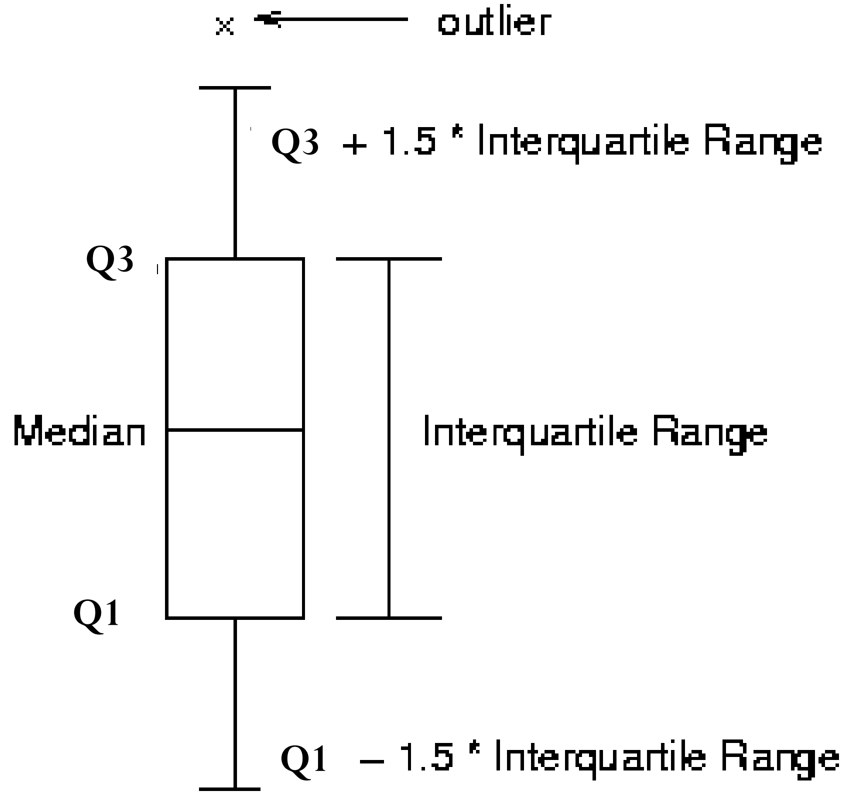
Box-plot

- The **Five-Number Summary** is composed of the following numbers: Low, Q_1, median, Q_3, High
- The Five-Number Summary can be graphically represented using a **Boxplot**.

Types of box-plots

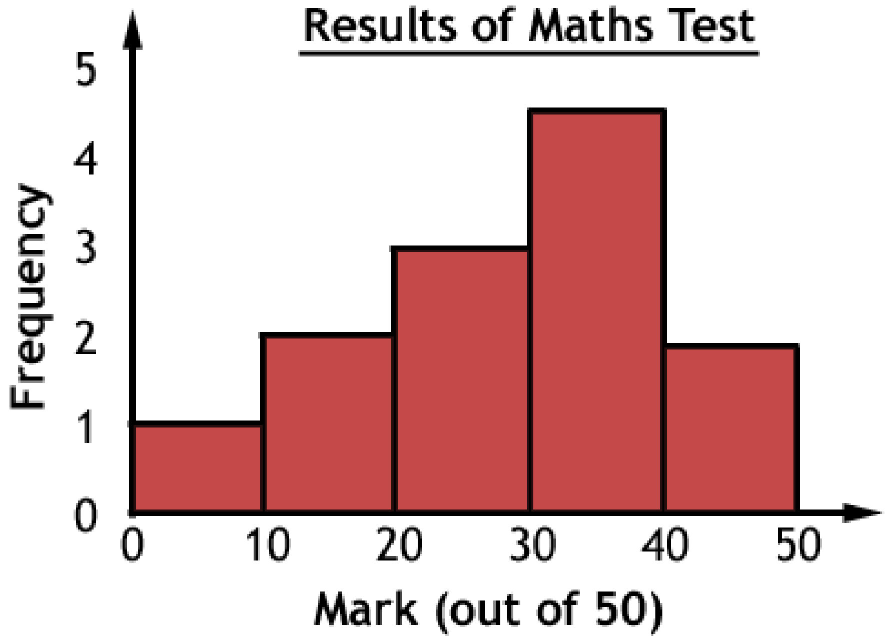
“Type 1” and “Type 2”





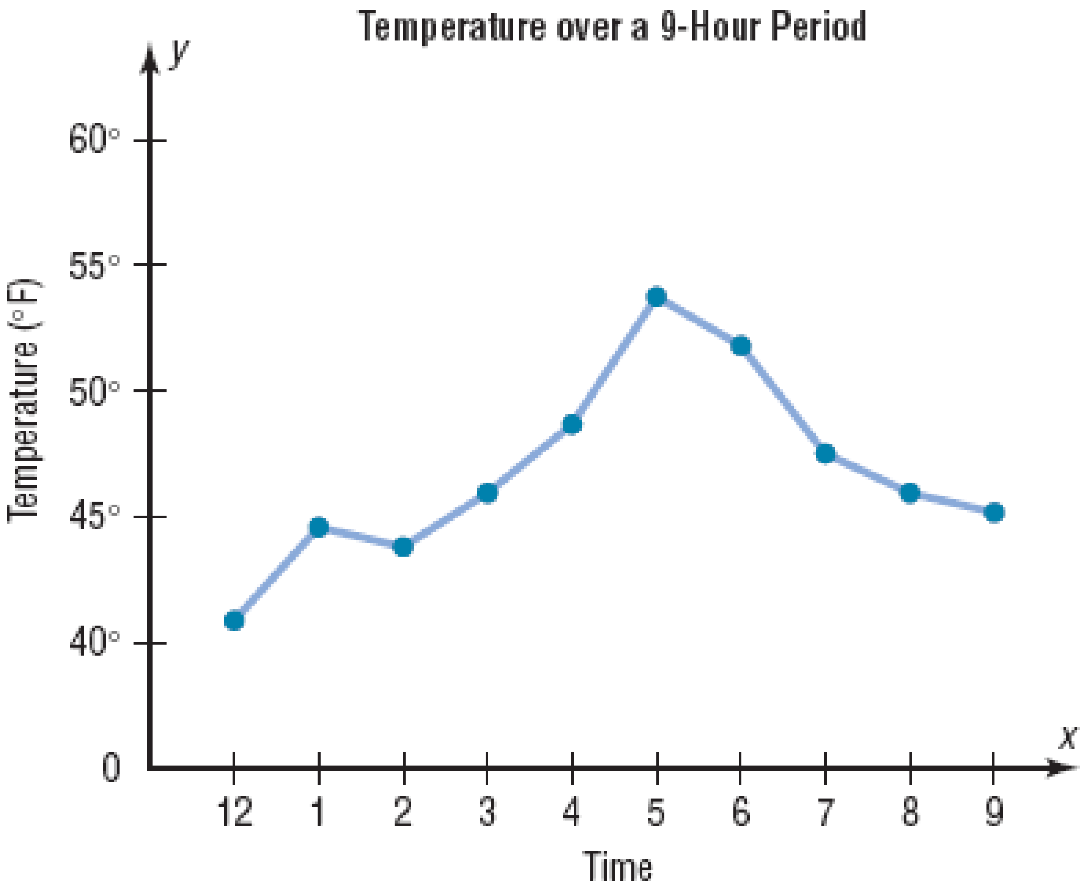
Histogram

The ***histogram*** is a graph that displays the data by using vertical bars of various heights to represent the frequencies of the classes.

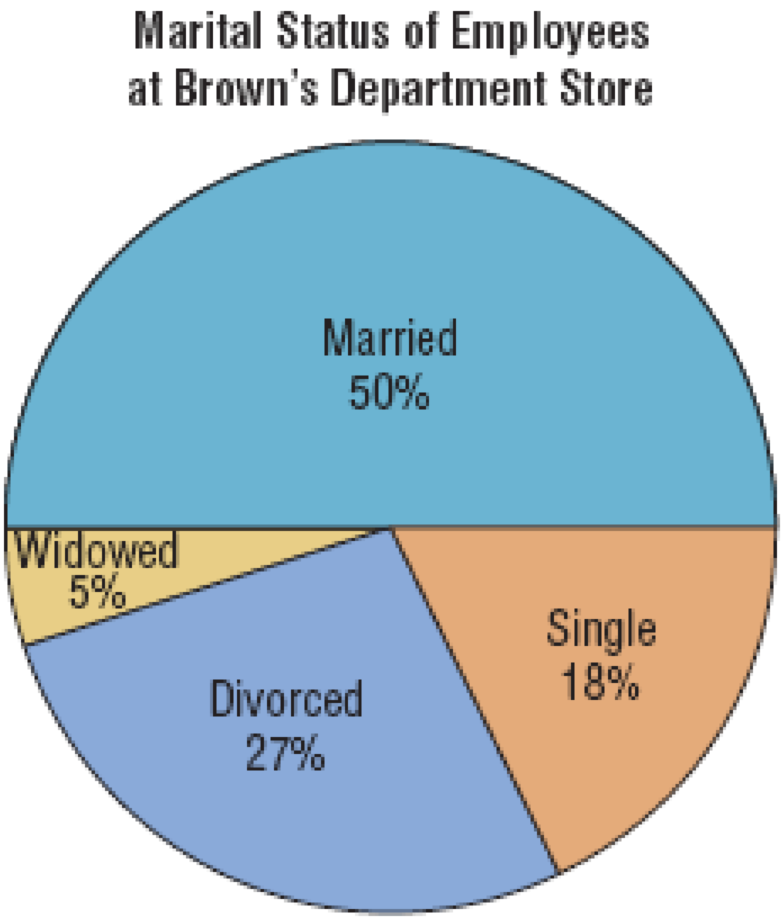


The height of each bar represents the percentage (or counts) of data values in the interval

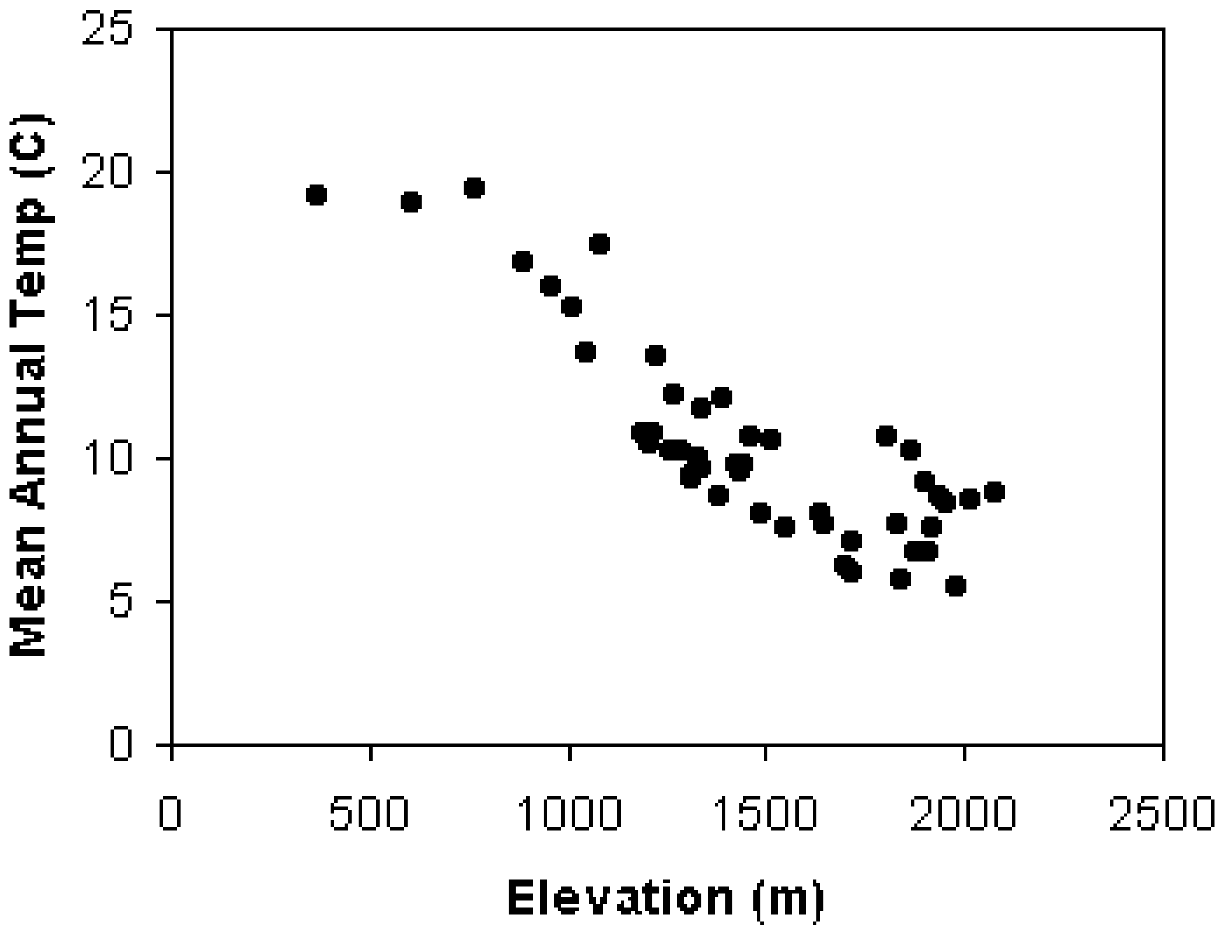
Other Types of Graphs
Time Series Graphs



Other Types of Graphs
Pie Graphs



Scatter Plot



Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential

1.2. Variables and Types of Data

1.3. Data organization and histograms

1.4. Measures of:

Central Tendency (Location)

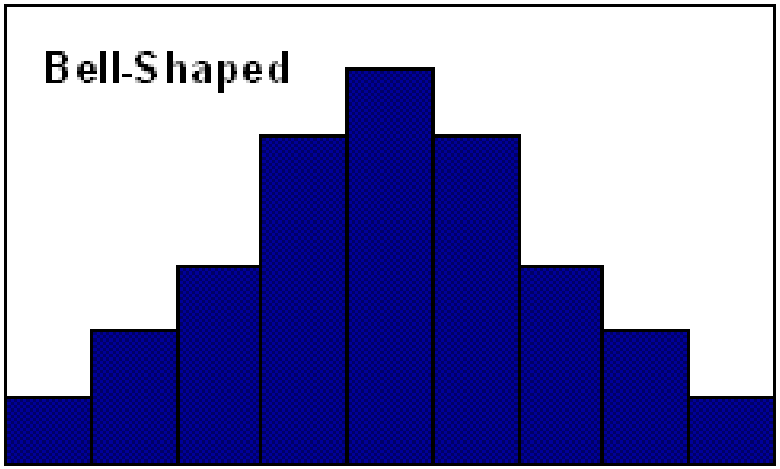
Variation (Dispersion)

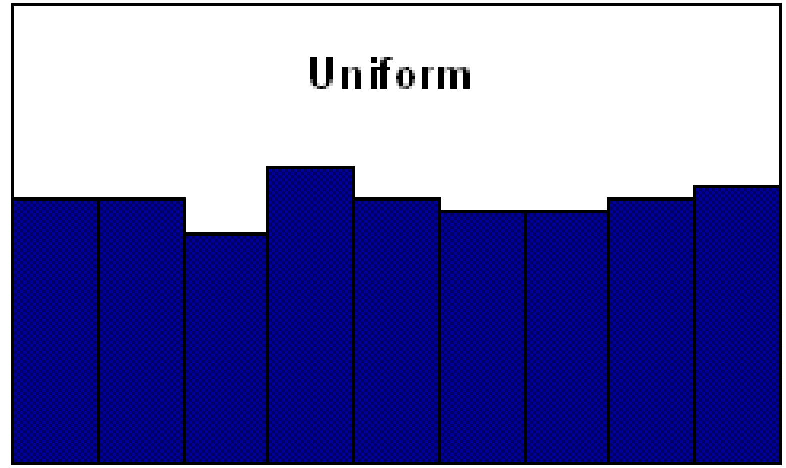
Position

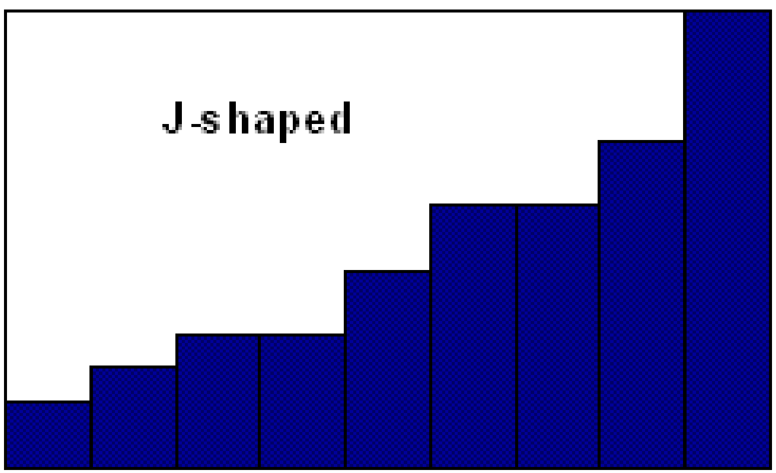
1.5. Data representation: frequency distributions and graphs

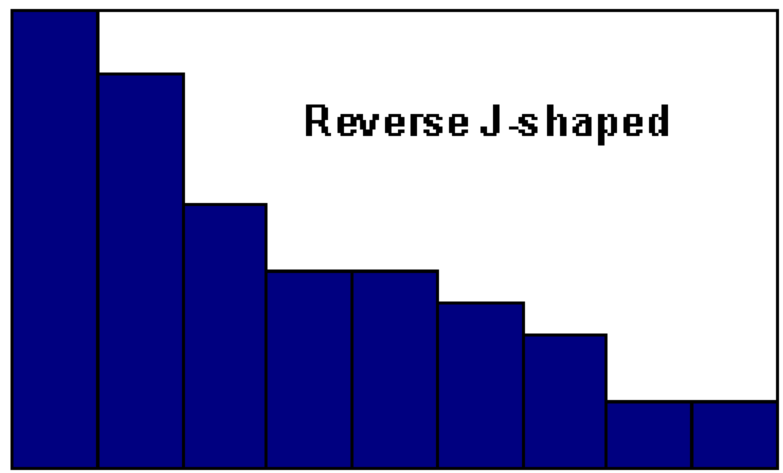
1.6. Shapes of frequency distributions: Skewness and kurtosis

Shapes of Distributions

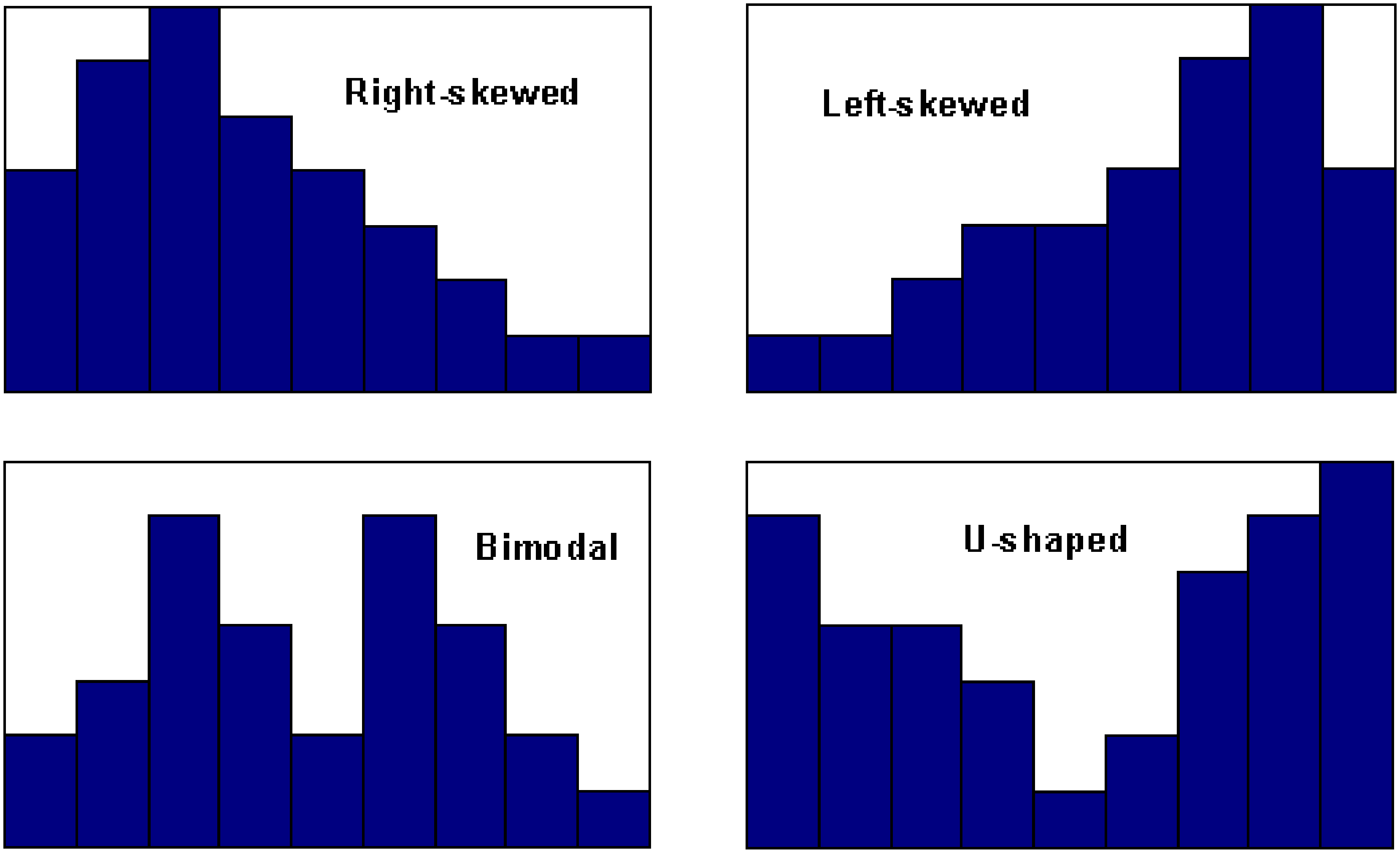








Shapes of Distributions



Skewness and kurtosis

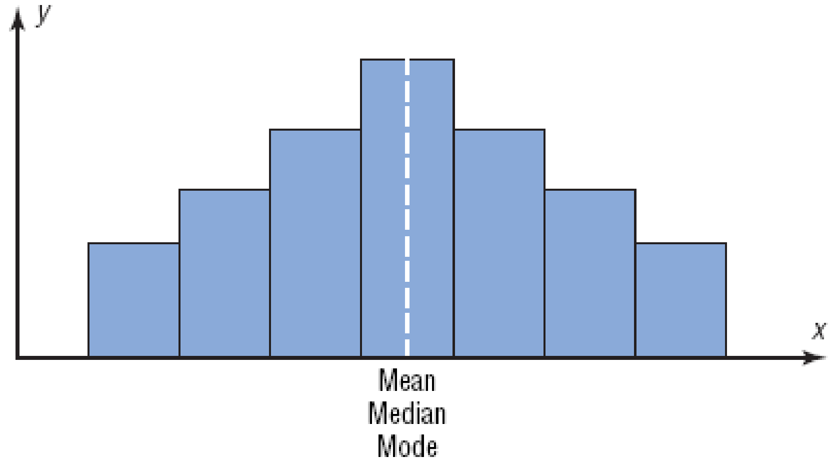
- While measures of dispersion are useful for helping us describe the width of the distribution, they tell us nothing about the **shape of the distribution**

Skewness

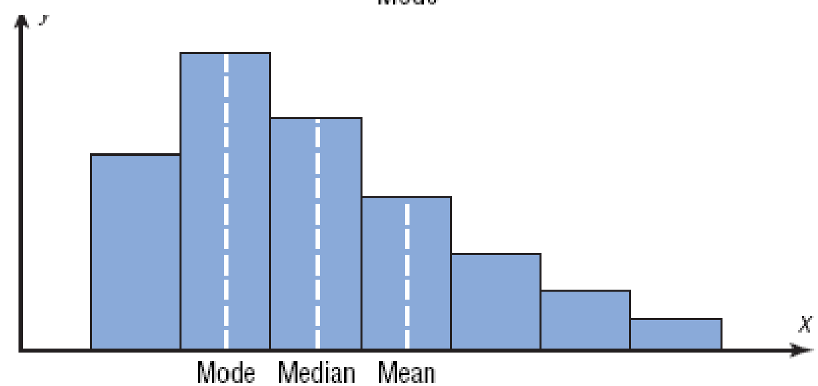
- Skewness of a distribution is a measure of symmetry, or more precisely, the lack of symmetry.
- A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Skewness

- **No-skewness**

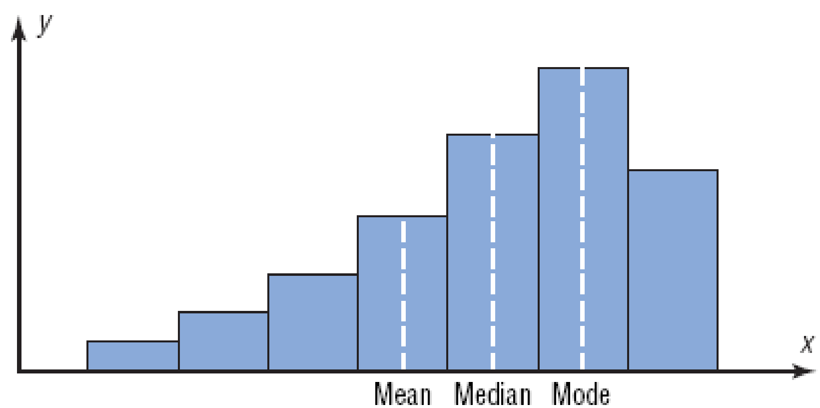


- **Positive skewness**



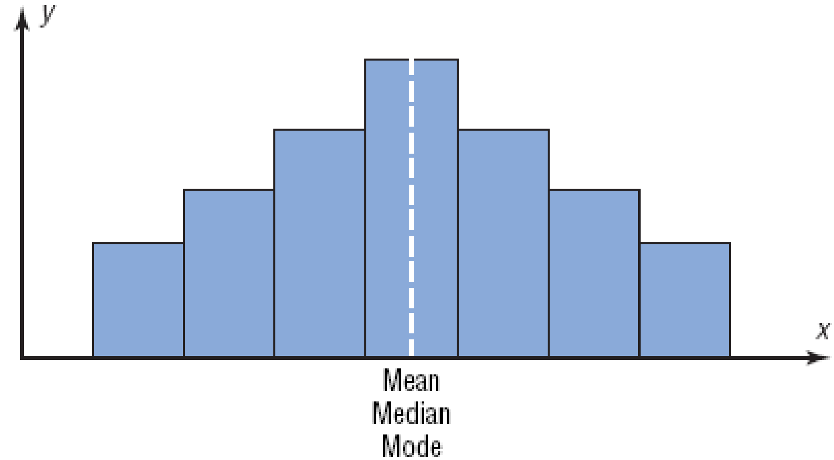
(a) Positively skewed or right-skewed

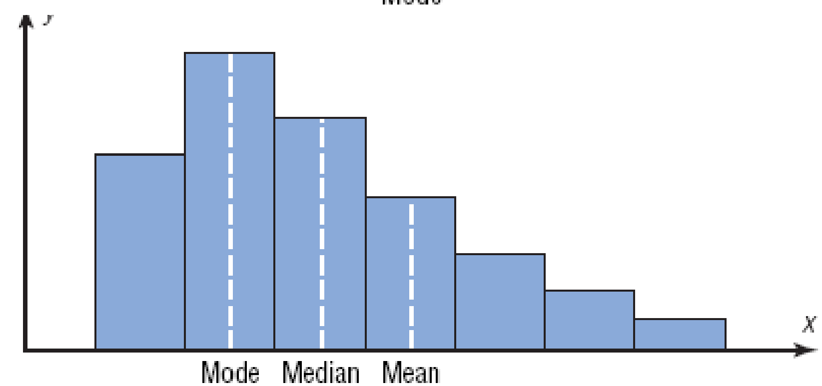
- **Negative skewness**



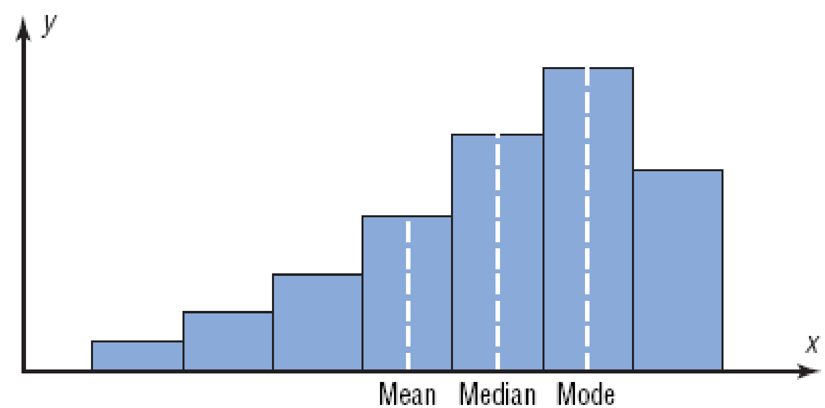
Skewness

- **No-skewness**
  - Same observations below and above the mean
  - Mean and median coincide
- **Positive skewness**
  - There are more observations below the mean than above it
  - When the mean is greater than the median
- **Negative skewness**
  - There are a small number of low observations and a large number of high ones
  - When the median is greater than the mean





(a) Positively skewed or right-skewed



- **Skewness** ("**Fisher's skewness**") measures the degree of asymmetry exhibited by the data

$$skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}$$

- If **skewness** equals zero, the histogram is **symmetric** about the mean

Fisher's Skewness > 1.00 moderate right skewness

> 2.00 severe right skewness

Fisher's Skewness < -1.00 moderate left skewness

< -2.00 severe left skewness

Kurtosis

- **Kurtosis** measures how **peaked** the histogram is:

$$kurtosis = \frac{\sum_{i}^{n}(x_i - \bar{x})^4}{ns^4} - 3$$

- The **kurtosis** of a **normal distribution** is 3
- **Kurtosis** characterizes the relative **peakedness** or **flatness** of a distribution compared to the **normal distribution**

- **Leptokurtic**– positive kurtosis indicates a relatively peaked distribution
- **Platykurtic**– negative kurtosis indicates a relatively flat distribution
- **Mesokurtic** (in between)