503 STAT

Probability and Mathematical statistics

Text Book: Probability and Statistics for Engineers and Scientists.

By: R.E.Walpole and R.H.Myers

Schedule of Assessment Tasks for Students During the Semester

Assessment	Examination	Week due	Proportion of Final Assessment
1	Mid-term exam (1)	5th week	25%
2	Mid-term exam (2)	9th week	25%
3	Homework	Every week	10%
4	Final exam		40%

Definition: Statistics

Statistics is a collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting and drawing conclusions based on that data.

Types of Statistics

1- Descriptive Statistics

Descriptive Statistics are used to summarize or describe the important characteristics of a known set of data.

For example: Let us consider everyone in this room. Each one of us is a source of data. A characteristic of this data may be degree program, age, height, sex, marital status.

Types of Statistics

2- Inferential Statistics

Inferential Statistics goes beyond the description. It involves the use of sample data to make inferences about a larger set of data from which the sample was chosen.

For example: If we consider this class as a sample of KSU students and calculated the average age of the class. We could then infer that the average age of all KSU students is the same as our sample.

Population and Sample

Definition:

A **population** is the complete collection of elements (scores, people, measurements) to be studied.

A **sample** is a subcollection of elements drawn from the population.

Population and Sample

Population

(Some Unknown Parameters)

Example: KSU Students

(Height Mean)

N=Population Size

Sample = Observations

(We calculate Some Statistics)

Example: 20 Students from KSU

(Sample Mean)

n = Sample Size

Notes:

- Let $X_1, X_2, \ldots, X_N$ be the population values (in general, they are unknown)
- Let $x_1, x_2, \ldots, x_n$ be the sample values (these values are known)
- Statistics obtained from the sample are used to estimate (approximate) the parameters of the population.

Why do we study and analyze subcollections (samples) of a population

Definition (Parameter)

It is a numerical characteristics of a population that summarize the data for the entire population.

Definition (Statistic)

It is a numerical characteristics of a sample.

Definition (Variables)

A variable is a characteristic, feature or factor that varies from one individual to another in a population.

Classification of Variables

Quantitative Variables

Qualitative Variables

Discrete variables

(The number of cars in a parking lot – The number of patients in a hospital)

Continuous variables

(height - weight - time it takes to get to school)

Measures of Location (Central Tendency)

- The data (observations) often tend to be concentrated around the center of the data.
- Some measures of location are: the mean, mode, and median.
- These measures are considered as representatives (or typical values) of the data. They are designed to give some quantitative measures of where the center of the data is in the sample.

The Sample mean of the observations

Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The **sample mean**, denoted by $\bar{x}$, is

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Example:

Suppose that the following sample represents the ages (in year) of a sample of 3 men:

$$x_1 = 30, x_2 = 35, x_3 = 27$$

Then, the sample mean is:

$$\bar{x} = \frac{30+35+27}{3} = 30.67 \text{ (years)}$$

Note:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 .$$

Prove that?

Median

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the sample median is

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), & \text{if } n \text{ is even.} \end{cases}$$

Example: suppose the data set is the following: 1.7, 2.2, 3.9, 3.11, and 14.7. The sample mean and median are, respectively,

$$\bar{x} = 5.12, \quad \tilde{x} = 3.9$$

Note:

The mean is influenced by the extreme observations, whereas the median places emphasis on the true “center” of the data set.

See previous Example

Example: suppose we have the following two data sets:

data 1: 59,60,60,61

data 2: 50,60,60,70

Calculate the Mean and the Median?

What do you see?

Measures of Variability (Dispersion or Variation)

- The variation or dispersion in a set of data refers to how spread out the observations are from each other.
- The variation is small when the observations are close together. There is no variation if the observations are the same.
- Some measures of dispersion are range, variance, and standard deviation
- These measures are designed to give some quantitative measures of the variability in the data.

Range:

It is the simplest measure of variation and defined as

$$R = X_{max} - X_{\min}$$

Example:

What is the range of the following data set

42,55,47,41,57,50

Solution:

$$X_{\min} = 41$$

$$X_{\max} = 57$$

$$R = X_{\max} - X_{\min} = 57 - 41 = 16$$

The Sample Variance (S^2)

Let $x_1, x_2, \ldots, x_n$ be the observations of the sample. The sample variance is denoted by S^2 and is defined by:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} \ (\text{unit})^2$$

where $\bar{x} = \sum_{i=1}^{n} x_i / n$ is the sample mean.

Note:

(n −1) is called the degrees of freedom (df) associated with the sample variance (S^2).

The Standard Deviation (S)

The standard deviation is another measure of variation. It is the square root of the variance

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad \text{(unit)}$$

Example:

Compute the sample variance and standard deviation of the following observations (ages in year): 10, 21, 33, 53, 54.

Solution:

$n=5$

$$\bar{x}=\frac{\sum_{i=1}^{n} x_i}{n}=\frac{\sum_{i=1}^{5} x_i}{5}=\frac{10+21+33+53+54}{5}=\frac{171}{5}=34.2 \quad \text{(year)}$$

$$S^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}=\frac{\sum_{i=1}^{5}(x_i-34.2)^2}{5-1}$$

$$=\frac{(10-34.2)^2+(21-34.2)^2+(33-34.2)^2+(53-34.2)^2+(54-34.2)^2}{4}$$

$$=\frac{1506.8}{4}=376.7 \quad (\text{year})^2$$

The sample standard deviation is

$$S = \sqrt{S^2} = \sqrt{376.7} = 19.41 \qquad \text{(year)}$$

Another Formula for Calculating S^2:

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

Note:

To calculate S^2 we need:

- n = sample size
- $\sum x_i$ =The sum of the values
- $\sum x^2{}_i$=The sum of the squared values

For the above example:

x_i	10	21	33	53	54	$\sum x_i = 171$
x_i^2	100	441	1089	2809	2916	$\sum x_i^2 = 7355$

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{7355 - (5)(34.2)^2}{5-1} = \frac{1506.8}{4} = 376.7 \ (\text{unit})^2$$

Homework Exercises

Numbers 1.7 and 1.8 on page 17

Number 1.21 on page 31

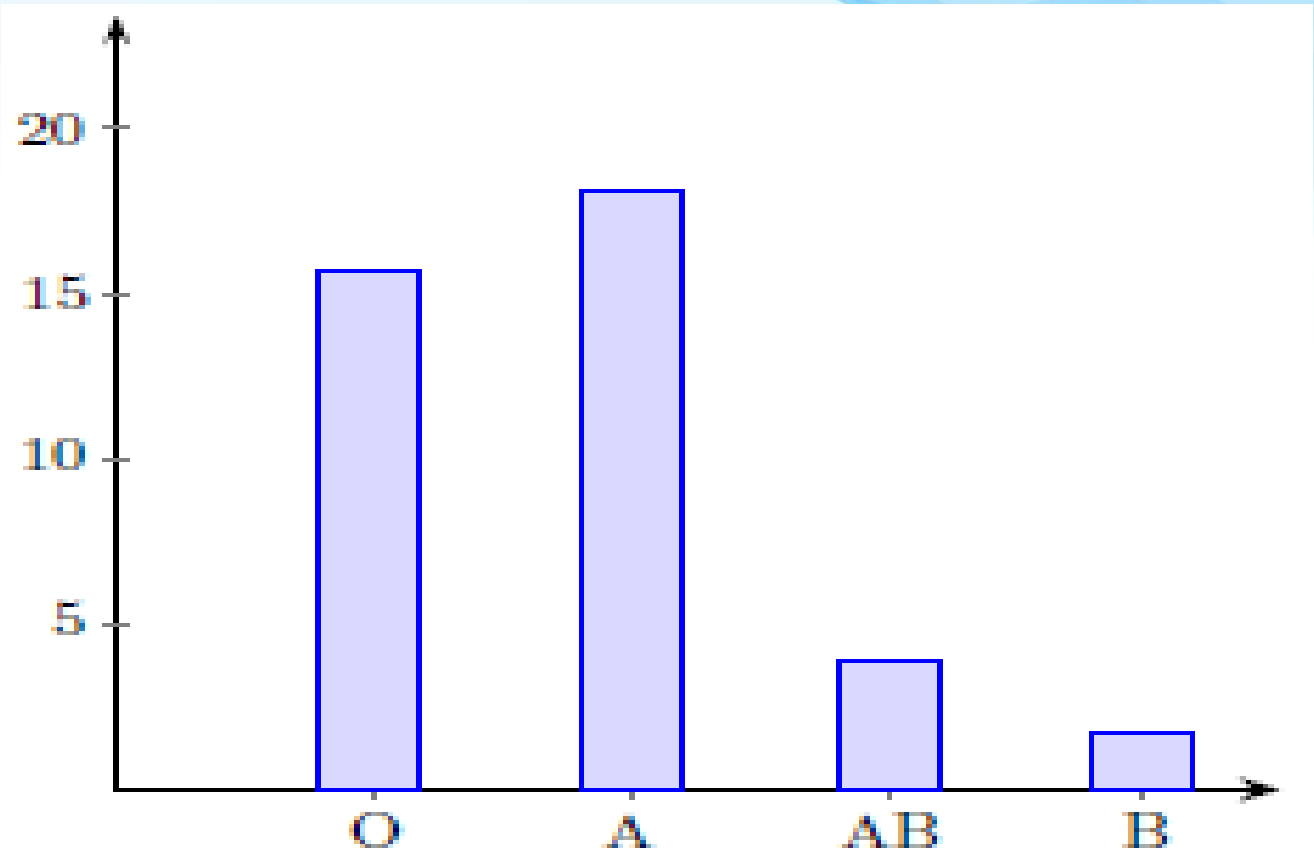
GRAPHICAL REPRESENTATIONS

(1) Bar Charts: In a bar chart, the frequency of each class is represented by a bar. The height of the bar corresponds to the frequency of the class. The width of the bar doesn't matter.

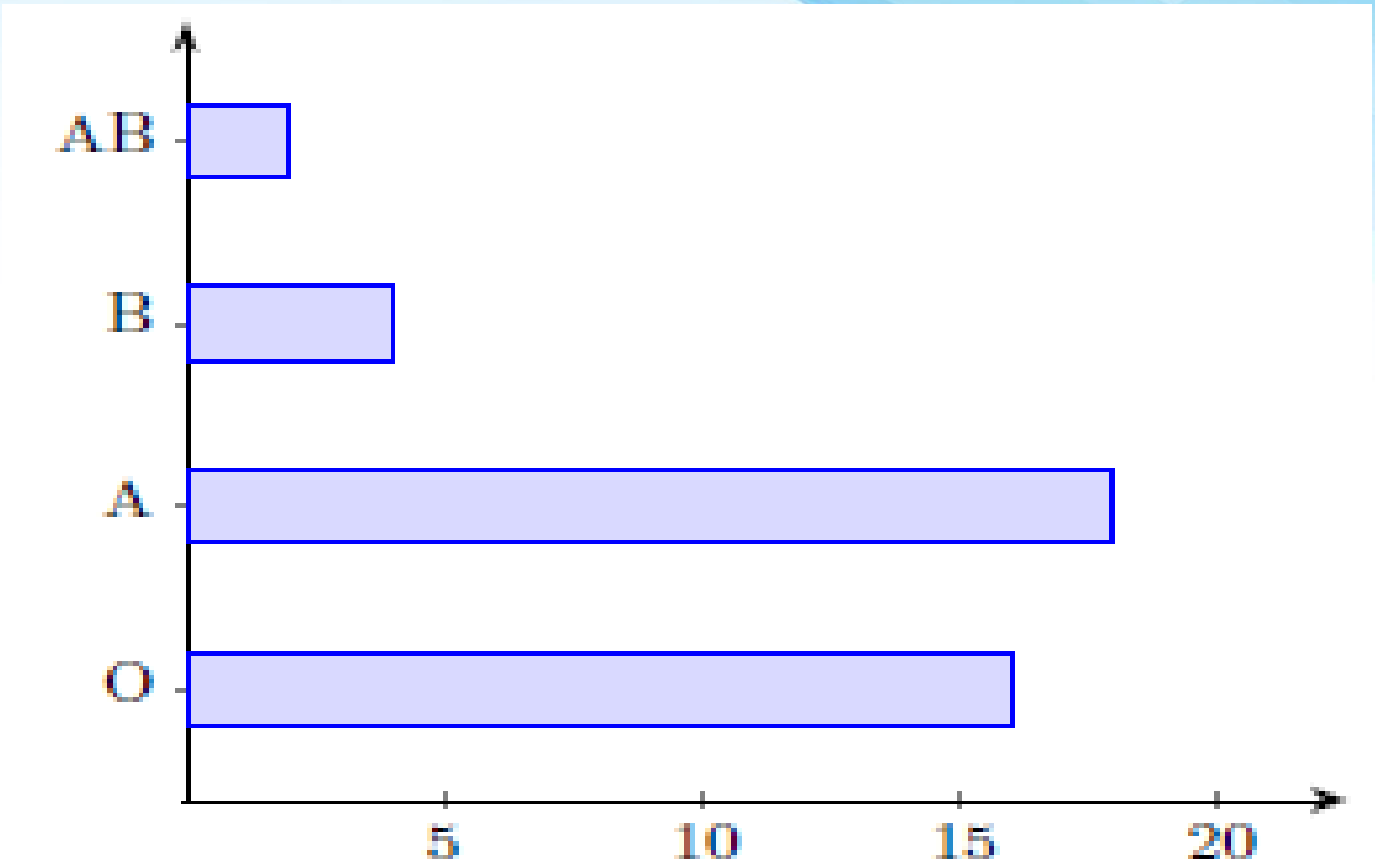
Example: Draw the bar chart for the following data

Blood group	Frequency
O	16
A	18
B	4
AB	2
Total	40

Answer:



Other presentation

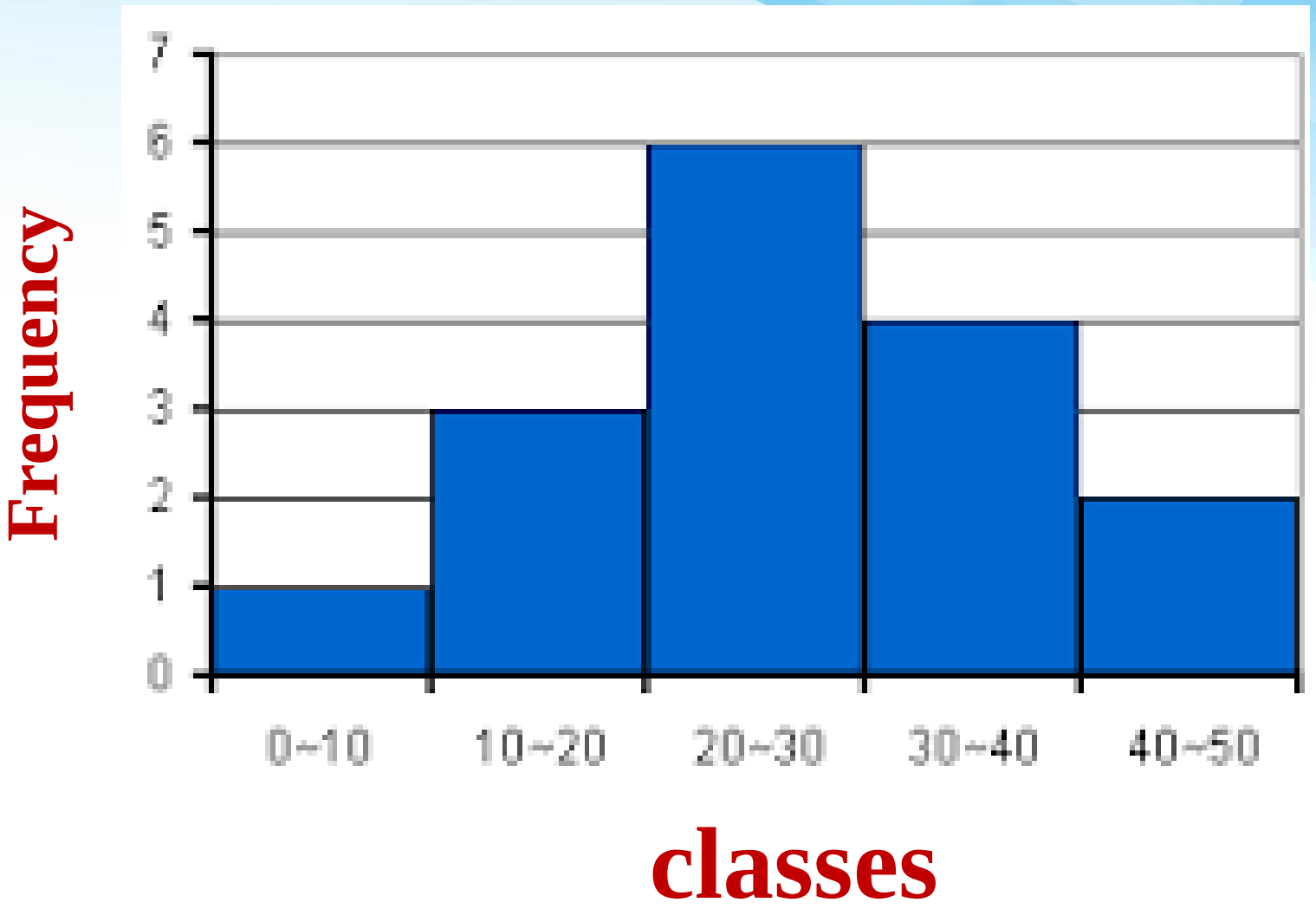


(2) Histogram: is similar to bar chart but they both have a basic difference that is in histograms, classes of the variable are adjacent to each other and the rectangular bars must touch each other. Histograms are generally used to represent quantitative data. The class intervals in a histogram are called as bins.

Example : Consider the following data

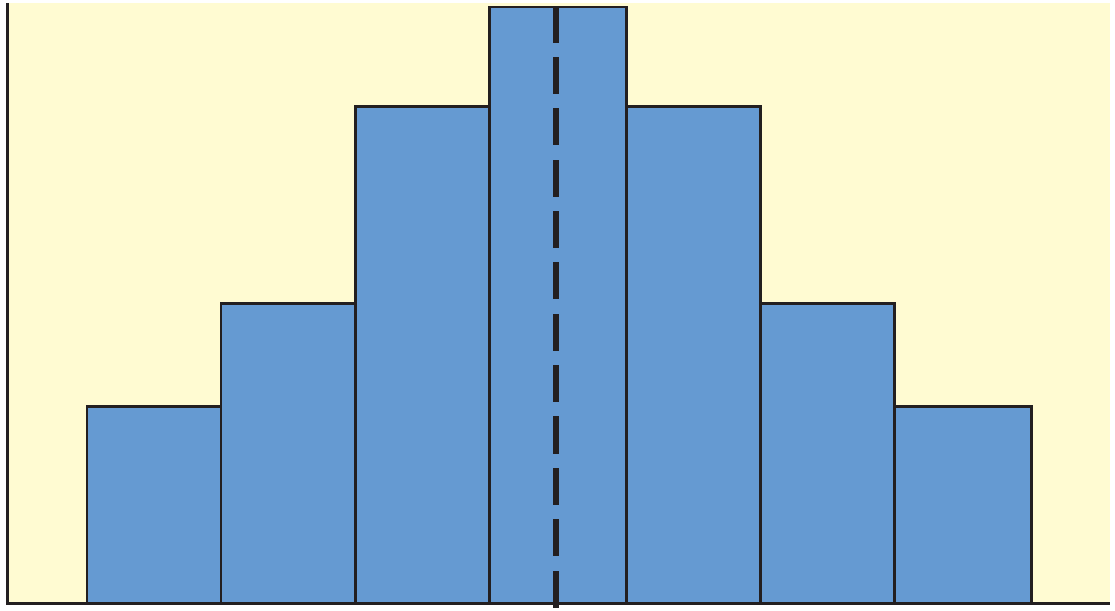
Continuous Classes	Frequency
0-10	1
10-20	3
20-30	6
30-40	4
40-50	2
Total	16

Answer:



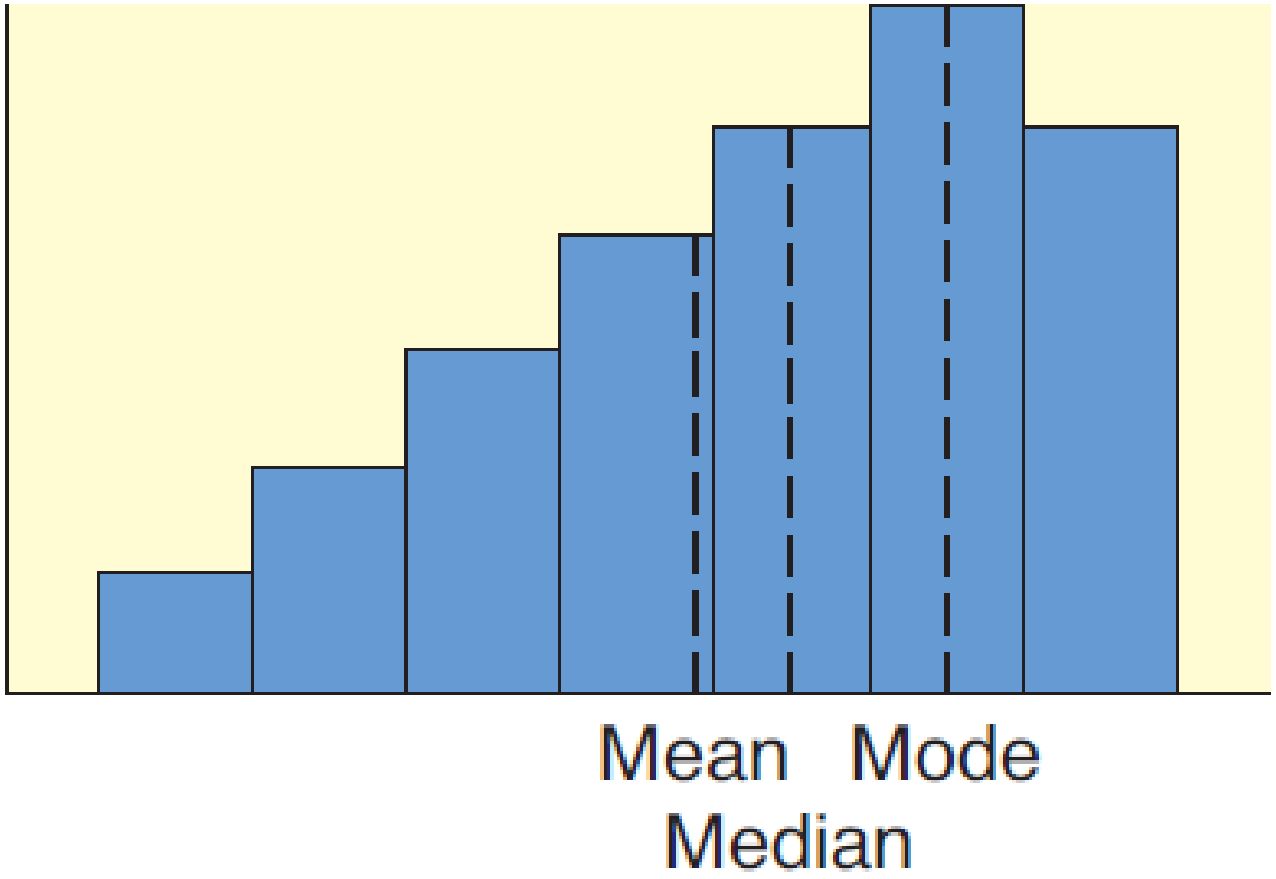
Distribution Types and Averages

(a) Mound-shaped symmetric

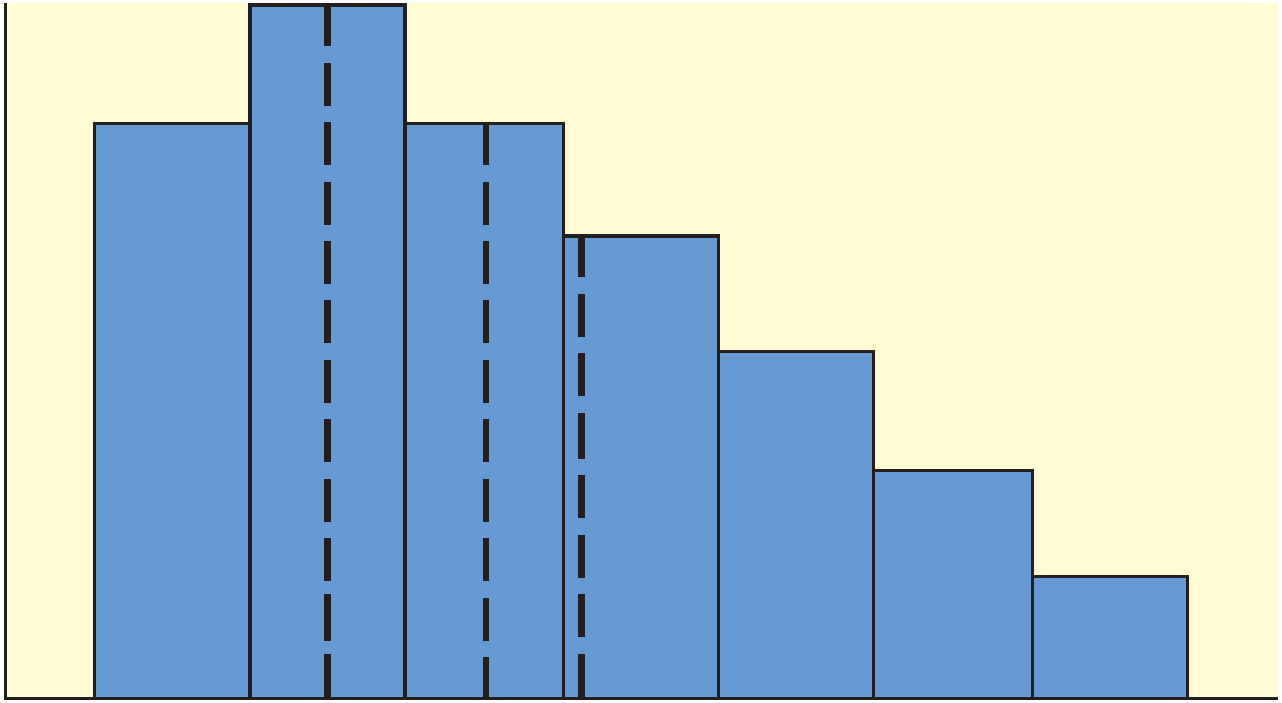


Mean
Median
Mode

(b) Skewed left



(c) Skewed right



Mode Mean
Median