Engineering Statistics

Third Year

Prepared By:

Dr Taher M. Ahmed

Department of Civil Engineering

Engineering College

University of Anbar

2021-2022

Syllabus

1. Fundamentals (Introduction to Statistics)
2. Presentation of Statistical Data
3. Data Description
4. Probability and Counting Rules
5. Discrete Probability Distributions
6. Continuous Distribution
7. Confidence Intervals and Sample Size
8. Hypothesis Testing
9. Testing the Difference Between Two Means, Two Proportions, and Two Variances
10. Correlation and Regression

Chapter One:
Fundamentals (Introduction to Statistics)

1. Introduction
2. Descriptive and Inferential Statistics
3. Variables and Types of Data
4. Data Collection and Sampling Techniques
5. Observational and Experimental Studies

Chapter One
Fundamentals (Introduction to Statistics)

1. *Introduction*

❖ ***Statistics:*** Is the science the science of collecting, analyzing, presenting, and interpreting data, which often leads to the drawing of conclusions. For example :-

- ***Nearly one in seven U.S. families are struggling with bills from medical expenses even though they have health insurance. (Source: Psychology Today.)***
- ***Eating 10 grams of fiber a day reduces the risk of heart attack by 14%. (Source: Archives of Internal Medicine, Reader's Digest.)***
- ***Thirty minutes of exercise two or three times each week can raise HDLs by 10% to 15%. (Source: Prevention.)***
- ***About 15% of men in the United States are left-handed and 9% of women are left-handed. (Source: Scripps Survey Research Center.)***
- ***The median age of people who watch the Tonight Show with Jay Leno is 48.1. (Source: Nielsen Media Research.)***

❖ Populations and the Samples:

- **Populations**: is a total collection of elements, and it is often too large for us to examine each of its members due to the cost and time consuming, for this reason we deal with sampling.
- **Samples** is a subgroup of population elements; it must be representative to that population. To achieve that the sample must be chosen in a random manner.

2. Descriptive and Inferential Statistics

- **Descriptive statistics** consists of the collection, organization, summarization and presentation of data. For example the average age, income and other characteristics of the population.
- **Inferential statistics** consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

3. Variables and Types of Data

- **Variable** is a characteristic or attribute that can assume different values such as compressive strength, tensile strength, water table level, specific gravity, etc. Variables can be classified as:

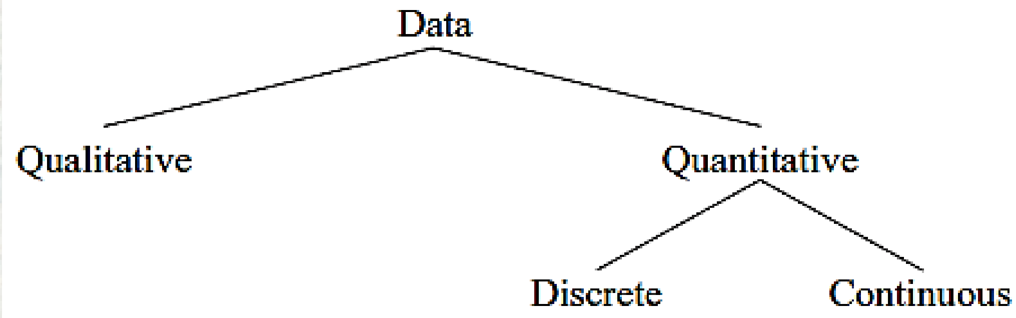
- **Qualitative variables** are variables that can be placed into distinct categories, according to some characteristic or attribute. For example, gender (male or female).

- **Quantitative variables** are numerical and can be ordered or ranked. For example, the variable *age* heights, weights, and body temperatures. Quantitative variables can be further classified into two groups:

- ✓ ***Discrete variables*** can be assigned values such as 0, 1, 2, 3 (integer values) and are said to be countable. For examples: the number of children in a family, the number of students in a classroom.

✓ **Continuous variables** can assume an infinite number of values between any two specific values. They are obtained by measuring. They often include fractions and decimals.

The classification of variables can be summarized as follows:



Statistically, variables can be divided into two types: ***independent*** and one ***dependent*** *variables.*

- The **independent variable** in an experimental study is the one that is being manipulated by the researcher.
- the **dependent variable is t**he resultant variable or the outcome variable.

$$\boldsymbol{Y = X^2 + 3} \qquad\qquad \boldsymbol{\sigma = \frac{p}{A}}$$

dependent variable

independent variable

4. Data Collection and Sampling Techniques

❖ Data Collection

Data can be collected in a variety of ways. One of the most common methods is through the use of surveys:

- ***Telephone surveys***
- ***Mailed questionnaire***
- ***Personal interview surveys***

❖ Sampling Techniques

To obtain samples that are unbiased, i.e. that give each subject in the population an equally likely chance of being selected—statisticians use four basic methods:

1. **Random** Subjects are selected by random numbers.
2. **Systematic** Subjects are selected by using every k^{th} number after the first subject is randomly selected from 1 through k.
3. **Stratified** Subjects are selected by dividing up the population into groups (strata), and subjects are randomly selected within groups.
4. **Cluster** Subjects are selected by using an intact group that is representative of the population.

5. Observational and Experimental Studies

There are several different ways to classify statistical studies for example *observational studies* and *experimental studies.*

- **Observational study** observes what is happening or what has happened in the past and tries to draw conclusions based on these observations such as accidents rate with age, …
- **Experimental study,** the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.

Thank You

Any Questions?

Chapter Two
Presentation of a Statistical Data

1. Introduction
2. Organizing Data
3. Histograms, Frequency Polygons, and Ogives
4. Other Types of Graphs

Chapter Two

Presentation of a Statistical Data

1. Introduction

- Gathering data for a particular variable under study is the primary task for presenting the data.
- The data must be organized in some meaningful way. The most convenient method of organizing data is to construct a frequency distribution.
- Then data must be presented to be understood by those who will benefit from reading the study.
- The most useful method of presenting the data is by constructing **statistical charts and grap**hs.

2. Organizing Data

- Information can be obtained from looking at raw data (Table 1), the data to be more understandable, they should be organized. One of the comment statistic methods are using so called *frequency distribution (Table 2)*.
  - A **frequency distribution** is the organization of raw data in a table form, using **classes and frequencies**.

Table 1: Raw data

49	57	38	73	81
74	59	76	65	69
54	56	69	68	78
65	85	49	69	61
48	81	68	37	43
78	82	43	64	67
52	56	81	77	79
85	40	85	59	80
60	71	57	61	69
61	83	90	87	74

Table 2: Frequency distribution table.

Class limits	Tally	Frequency
35–41	///	3
42–48	///	3
49–55	////	4
56–62	~~////~~ ~~////~~	10
63–69	~~////~~ ~~////~~	10
70–76	~~////~~	5
77–83	~~////~~ ~~////~~	10
84–90	~~////~~	5
	Total	50

Classes

Frequencies

❑ Grouped Frequency Distributions or Frequency Distributions Table

When the range of the data is large or huge, the data must be grouped into classes with the frequency of each class as shown in Table 2.

➢ Procedure for Constructing the Frequency Distribution Table

There are some concepts need to be explained as shown in the following distribution frequency table (Table 3).

- The values of the first class are called *class limits such as (24-30).* The **lower class limit (**24) represents the smallest data value that can be included in the class. The **upper class limit** (30) represents the largest data value that can be included in the class.
- The numbers in the second column are called **class boundaries.** These numbers are used to separate the classes so that there are no gaps in the frequency distribution.

Table 3: Frequency distribution table.

Class limits	Class boundaries	Tally	Frequency
24–30	23.5–30.5	///	3
31–37	30.5–37.5	/	1
38–44	37.5–44.5	~~////~~	5
45–51	44.5–51.5	~~////~~ ////	9
52–58	51.5–58.5	~~////~~ /	6
59–65	58.5–65.5	/	1
			25

Note: The *class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value and end in a 5.*

For example: the boundaries limits for the classes (31–37) & (7.8–8.8), are:

Lower limit -0.5 = 31 - 0.5 = 30.5 lower boundary

Upper limit + 0.5 = 37 + 0.5 = 37.5 upper boundary

Lower limit -0.05 = 7.8 - 0.05 = 7.75 lower boundary

Upper limit +0.05 = 8.8 + 0.05 = 8.85 upper boundary

Class limits	Class boundaries	Tally	Frequency
24–30	23.5–30.5	///	3
31–37	30.5–37.5	/	1
38–44	37.5–44.5	~~////~~	5
45–51	44.5–51.5	~~////~~ ////	9
52–58	51.5–58.5	~~////~~ /	6
59–65	58.5–65.5	/	1
			25

- ***Class width (C_w)*** *is the range between upper and lower limit of the same class.*

C_w = the lower (or upper) class limit of one class - the lower (or upper) class limit of the next class.

For example: *the class width of Table 3 is:*

31-24 = 7 OR 37-30 = 7 OR 23.5-30.5 = 7 OR 37.5-30.5 = 7

- ***Number of classes are between 5 and 20 classes.***
- **The class midpoint X_m is**

$$X_m = \frac{lower\ boundary + upper\ boundary}{2}$$

OR $$X_m = \frac{lower\ limit + upper\ limit}{2}$$

Example:

$$\frac{24 + 30}{2} = 27 \quad \text{or} \quad \frac{23.5 + 30.5}{2} = 27$$

Example 1: These data represent the record high temperatures in degrees Fahrenheit (F) for each of the 50 states. Construct a grouped Frequency distribution for the data using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Solution:

1. Find the highest value and lowest value: H = 134 and L = 100.
2. Find the range: R = highest value - lowest value = $H - L$; R = 134 – 100 = 34
3. Select the number of classes (5-20); n = 7.
4. Find the class width; $C_w = \frac{R}{n} = \frac{34}{7}$ ***= 4.9 ≈ 5 OR 4.0***
5. Select a starting point for the lowest class limit = *lowest value or less (100 or 99).*
6. *Determine the lower limits of the other class = Lower limit +* C_w *= 100 + 5 = 105, 110, 115, etc.*
7. *Determine the Upper limits of the first class =*

   lower limit (2nd class) – 1 (one unit) = ***105-1=104***
8. *Determine the upper limits of the other class = lower limit +* C_w *= 104 + 5 = 109, 114, 119, etc.*

9. Find the class boundaries: by subtracting 0.5 from each lower class limit and adding 0.5 to each upper class limit: First class : 99.5–104.5, second class: 104.5–109.5, etc.
10. Tally the data.
11. Find the numerical frequencies from the tallies.

Table 4: Frequency distribution table.

Class limits	Class boundaries	Tally	Frequency
100–104	99.5–104.5	//	2
105–109	104.5–109.5	~~////~~ ///	8
110–114	109.5–114.5	~~////~~ ~~////~~ ~~////~~ ///	18
115–119	114.5–119.5	~~////~~ ~~////~~ ///	13
120–124	119.5–124.5	~~////~~ //	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1

$n = \Sigma f = 50$

OK.

12.Further Calculations:

- The cumulative frequency distribution: It is a distribution that shows the number of data values less or higher than or equal to a specific value (usually an upper or lower boundary).

Class limits	Class boundaries	Tally	Frequency
100–104	99.5–104.5	//	2
105–109	104.5–109.5	~~////~~ ///	8
110–114	109.5–114.5	~~////~~ ~~////~~ ~~////~~ ///	18
115–119	114.5–119.5	~~////~~ ~~////~~ ///	13
120–124	119.5–124.5	~~////~~ //	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1
			$n = \Sigma f = 50$

The cumulative frequency

Ascending cumulative frequency (Less than X)

Ex: Less than 99.5 = 0

Less than 104.5 = 0 + 2 = 2

Less than 19.5 = 0+2+8+18+13 = 31

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Table 5: Ascending cumulative frequency distribution table.

Descending cumulative frequency (Greater than X)

Ex: Greater than 99.5 = 50

Greater than 104.5 = 50-2 = 48

Greater than 114.5 = 50 – 18 -8 -2 = 22

	Cumulative frequency
Greater than 99.5	50
Greater than 104.5	48
Greater than 109.5	40
Greater than 114.5	22
Greater than 119.5	9
Greater than 124.5	2
Greater than 129.5	1
Greater than 134.5	0

Table 6 Descending cumulative frequency distribution table.

Note: Cumulative frequencies are used to show how many data values are accumulated up or down to and including a specific class.

➢ Briefly

The following guides line steps can be used for constructing the frequency distribution table:

Procedure Table

Constructing a Grouped Frequency Distribution

Step 1 Determine the classes.

- Find the highest and lowest values.
- Find the range.
- Select the number of classes desired.
- Find the width by dividing the range by the number of classes and rounding up.
- Select a starting point (usually the lowest value or any convenient number less than the lowest value); add the width to get the lower limits.
- Find the upper class limits.
- Find the boundaries.

Step 2 Tally the data.

Step 3 Find the numerical frequencies from the tallies, and find the cumulative frequencies.

3. Histograms, Frequency Polygons, and Ogives

- Statistical graphs can be used to describe the data set or to analyze it.
- The purposes of using graphs are:
- ✓ to discuss an issue,
- ✓ reinforce a critical point
- ✓ summarize a data
- ✓ discover the trend or pattern in a situation over a period of time.

The three most commonly used graphs are:

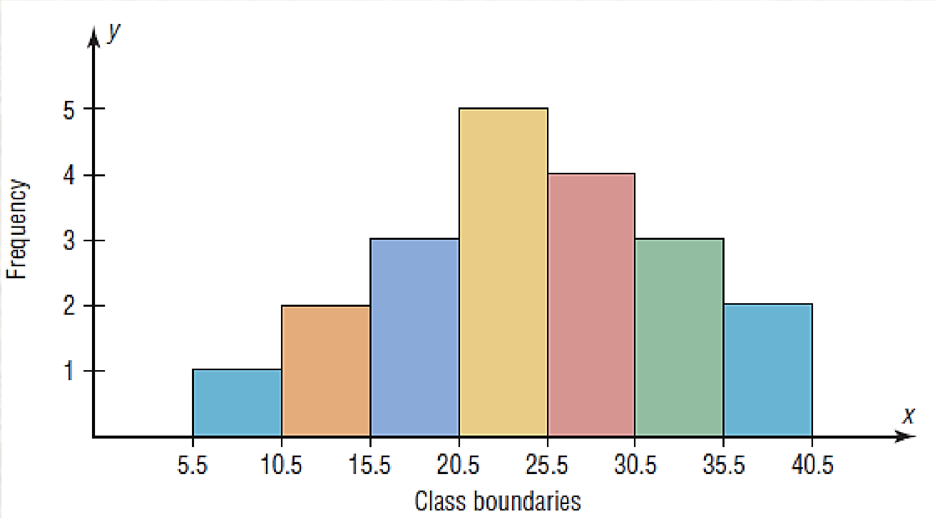
1. The histogram.
2. The frequency polygon.
3. The cumulative frequency graph, or ogive.

1. Histogram

The **histogram** is a graph that displays the data by using contiguous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes.

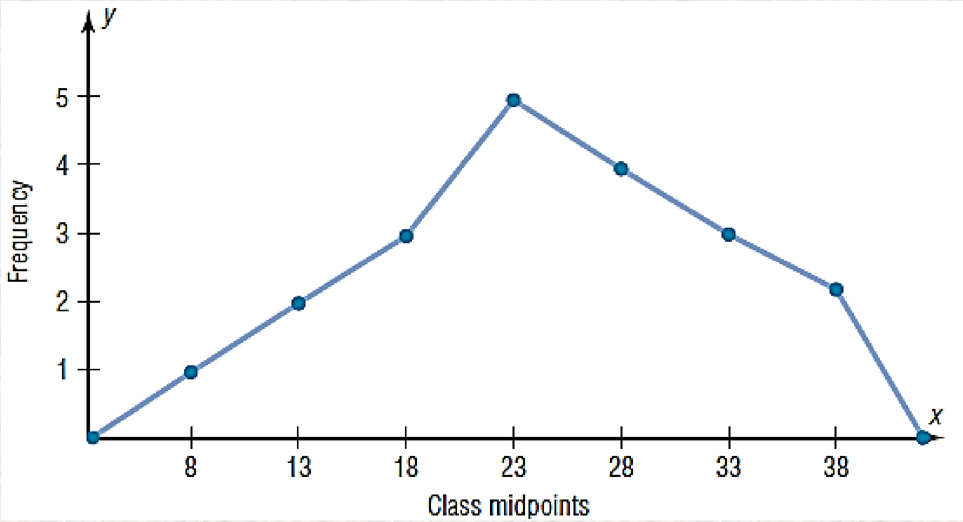
Example:

Class boundaries	Frequency
99.5–104.5	2
104.5–109.5	8
109.5–114.5	18
114.5–119.5	13
119.5–124.5	7
124.5–129.5	1
129.5–134.5	1



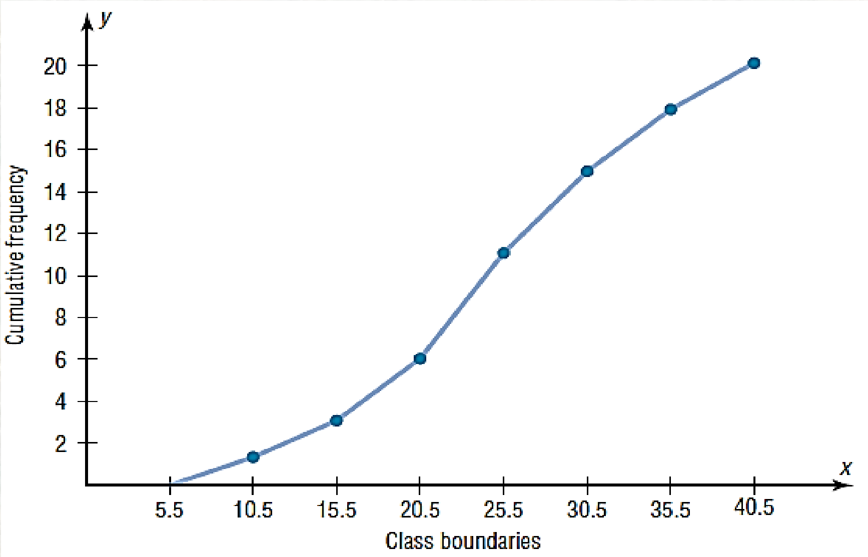
2. Frequency Polygon

The frequency **polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.



3. Ogive

This type of graph is called the **cumulative frequency graph**, or **Ogive**. The cumulative frequency is the sum of the frequencies accumulated up to the upper boundary of a class in the distribution.



Example:

Construct a histogram, polygon and Ogive to represent the data shown for the record high temperatures.

Class boundaries	Frequency
99.5-104.5	2
104.5-109.5	8
109.5-114.5	18
114.5-119.5	13
119.5-124.5	7
124.5-129.5	1
129.5-134.5	1

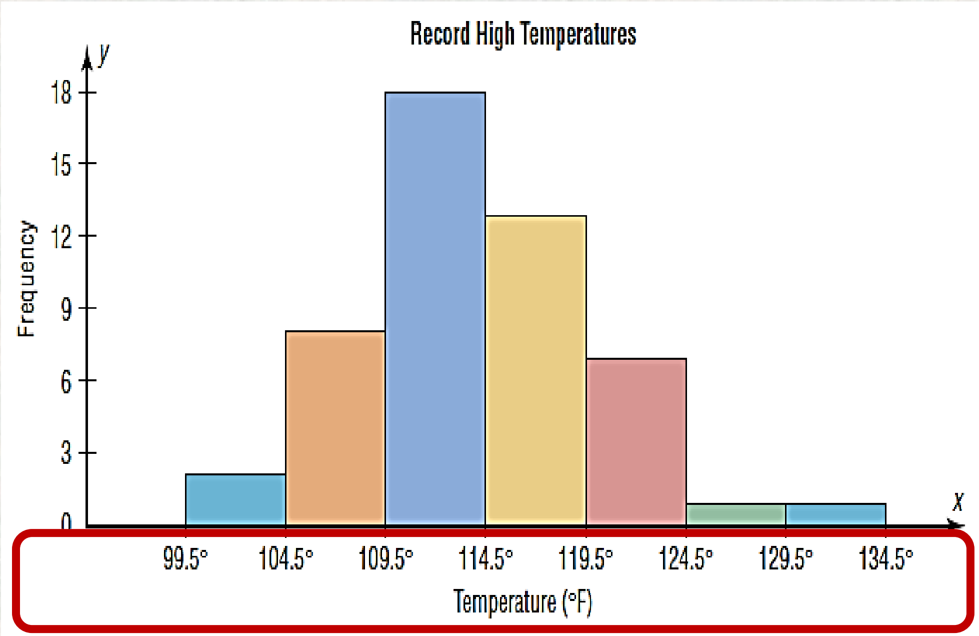
Solution

1. Find the midpoints of each class. Recall that midpoints are found by adding the upper and lower boundaries and dividing by 2.

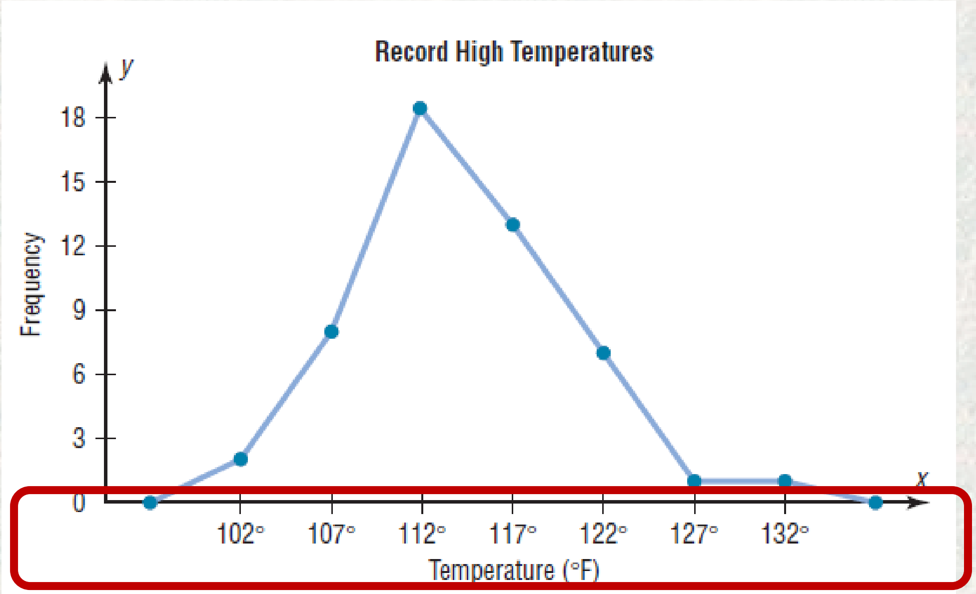
$$\frac{99.5 + 104.5}{2} = 102 \qquad \frac{104.5 + 109.5}{2} = 107$$

2. Draw and label the x and y axes. The x axis is always the horizontal axis, and the y axis is always the vertical axis.
3. Using the frequencies as the heights (Y-axes), and midpoints or boundary limits as (X-axis).

Class boundaries	Midpoints	Frequency
99.5-104.5	102	2
104.5-109.5	107	8
109.5-114.5	112	18
114.5-119.5	117	13
119.5-124.5	122	7
124.5-129.5	127	1
129.5-134.5	132	1



Boundaries limits



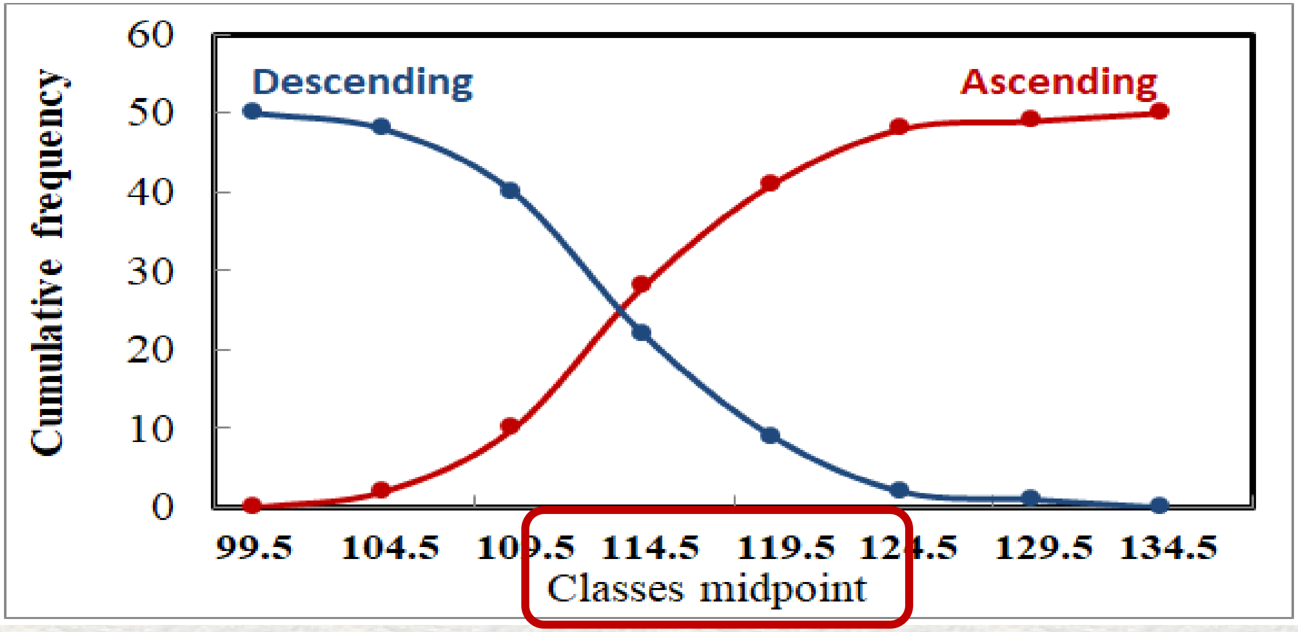
Classes midpoint

4. Find the cumulative frequency for each class.

Class boundaries	Midpoints	Frequency
99.5-104.5	102	2
104.5-109.5	107	8
109.5-114.5	112	18
114.5-119.5	117	13
119.5-124.5	122	7
124.5-129.5	127	1
129.5-134.5	132	1

Ascending cumulative frequency	
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Descending cumulative frequency	
Greater than 99.5	50
Greater than 104.5	48
Greater than 109.5	40
Greater than 114.5	22
Greater than 119.5	9
Greater than 124.5	2
Greater than 129.5	1
Greater than 134.5	0



➢ Briefly :

The following guides line steps can be used for constructing the frequency distribution table:

Procedure Table

Constructing Statistical Graphs

Step 1 Draw and label the x and y axes.

Step 2 Choose a suitable scale for the frequencies or cumulative frequencies, and label it on the y axis.

Step 3 Represent the class boundaries for the histogram or ogive, or the midpoint for the frequency polygon, on the x axis.

Step 4 Plot the points and then draw the bars or lines.

4. Relative frequency

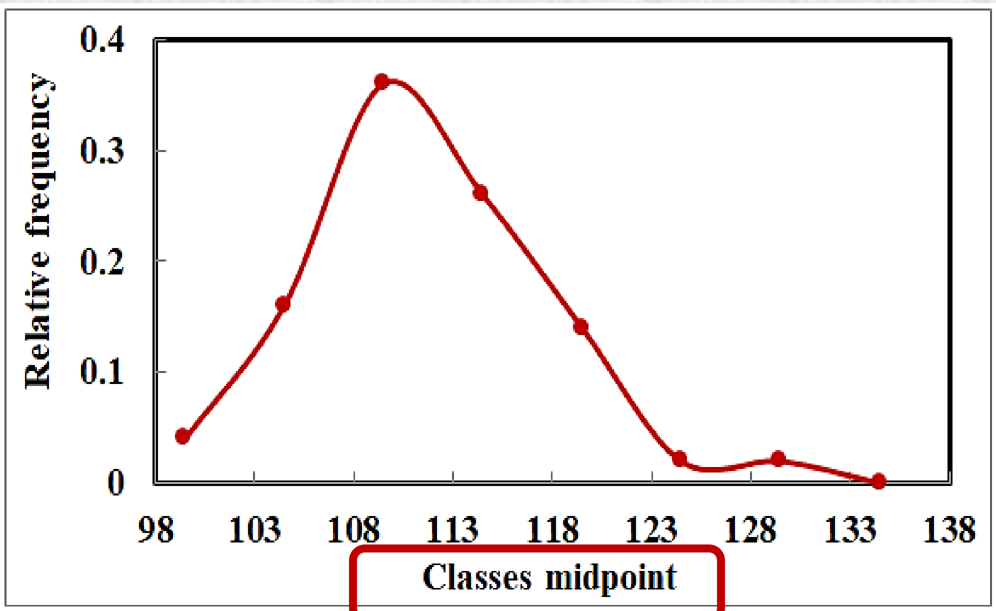
- The histogram, the frequency polygon, and the ogive shown previously were constructed by using frequencies in terms of the raw data. These distributions can be converted to distributions using *proportions* instead of raw data as frequencies. These types of graphs are called **relative frequency graphs.**

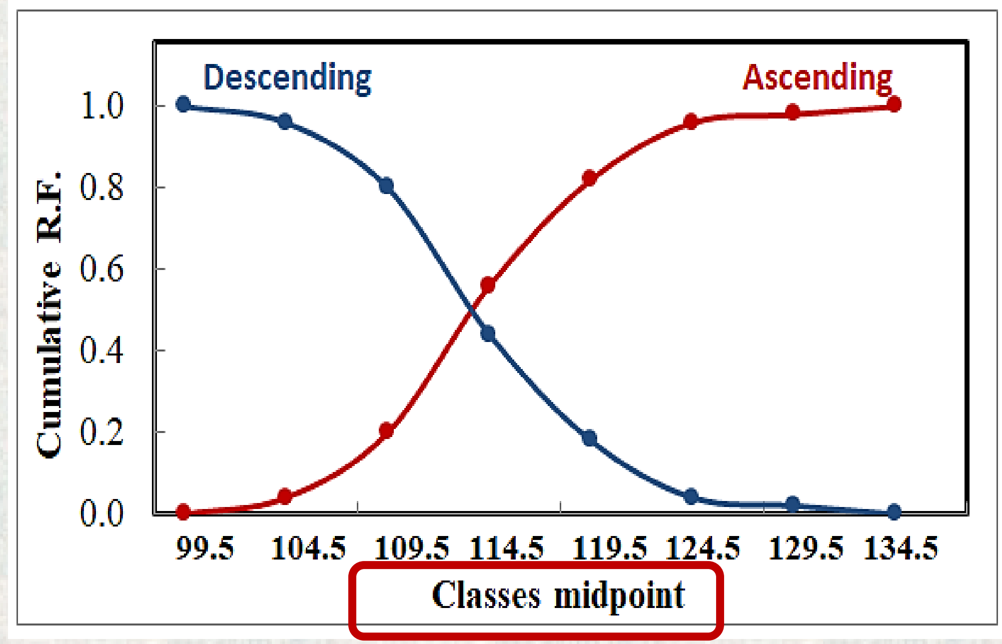
- **Relative frequency** (F_i) can be calculated by dividing the frequency for each class (f_i) by the total of the frequencies Σf_i. The sum of the relative frequencies will always be 1.

$$F_i = \frac{f_i}{\sum f_i}$$

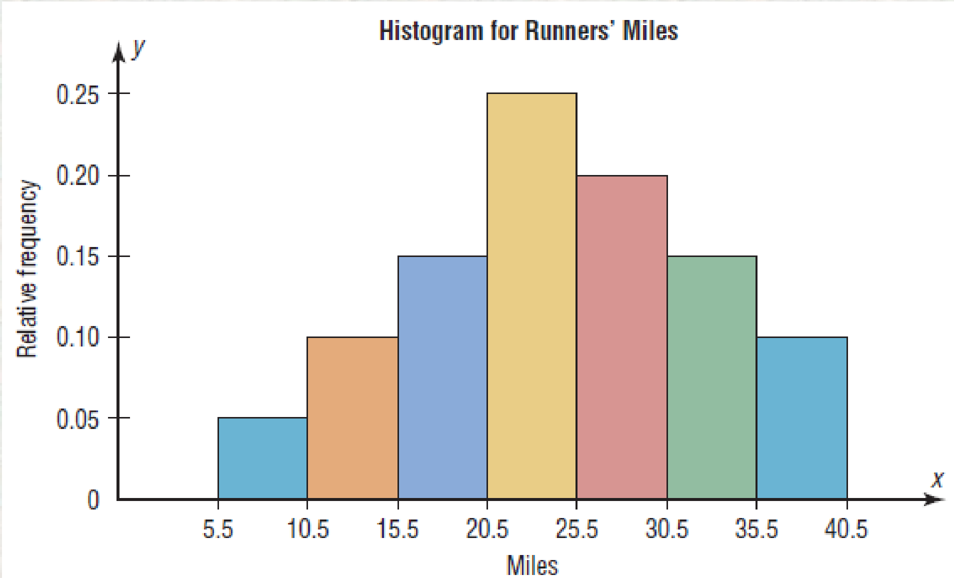
Class boundaries	Midpoints	Frequency	Relative frequency
99.5-104.5	102	2	0.04
104.5-109.5	107	8	0.16
109.5-114.5	112	18	0.36
114.5-119.5	117	13	0.26
119.5-124.5	122	7	0.14
124.5-129.5	127	1	0.02
129.5-134.5	132	1	0.02

Ascending C.R.F.	Descending C.R.F.
0.00	1.00
0.04	0.96
0.20	0.80
0.56	0.44
0.82	0.18
0.96	0.04
0.98	0.02
1.00	0.00

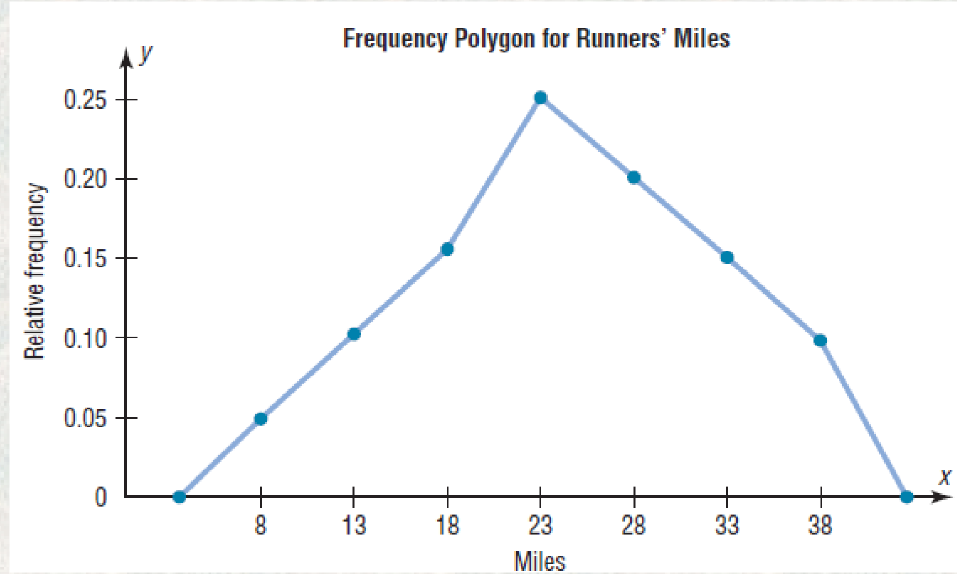




Relative frequency for histogram

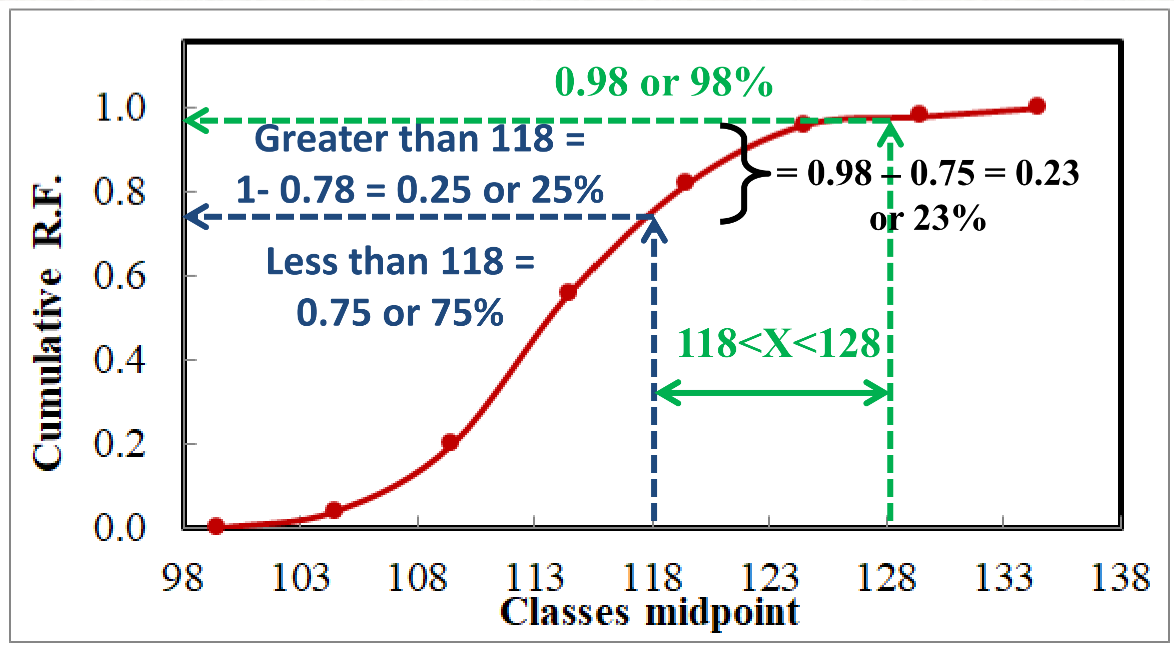


Relative Polygon for histogram



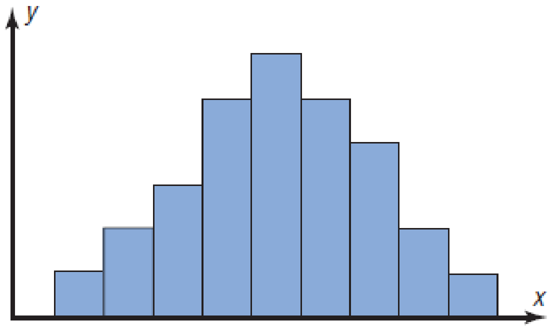
Examples for calculation

Relative frequency for histogram

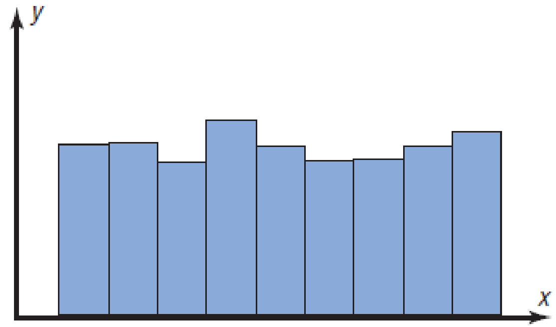


4. Distribution Shapes

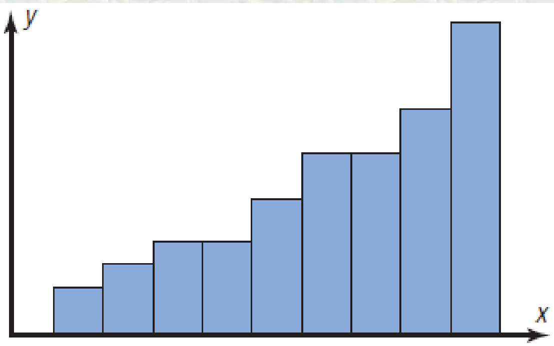
The shape of a distribution determines the appropriate statistical methods used to analyze the data.



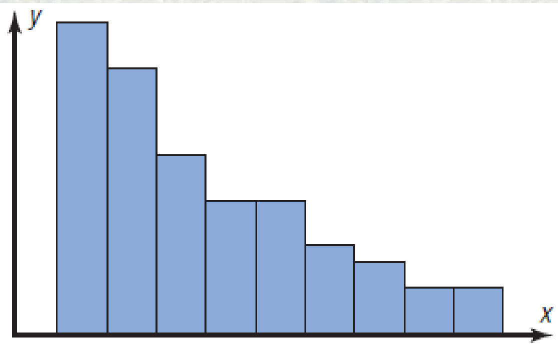
(a) Bell-shaped



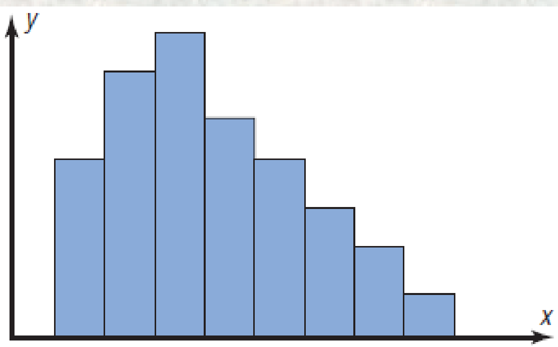
(b) Uniform



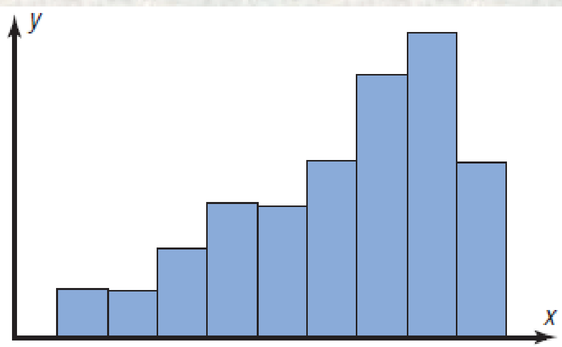
(c) J-shaped



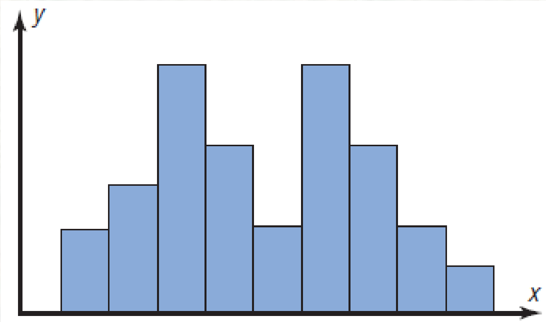
(d) Reverse J-shaped



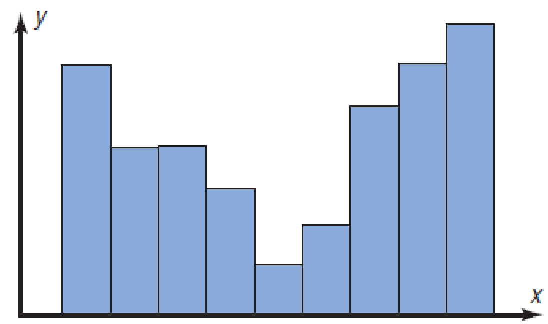
(e) Right-skewed



(f) Left-skewed

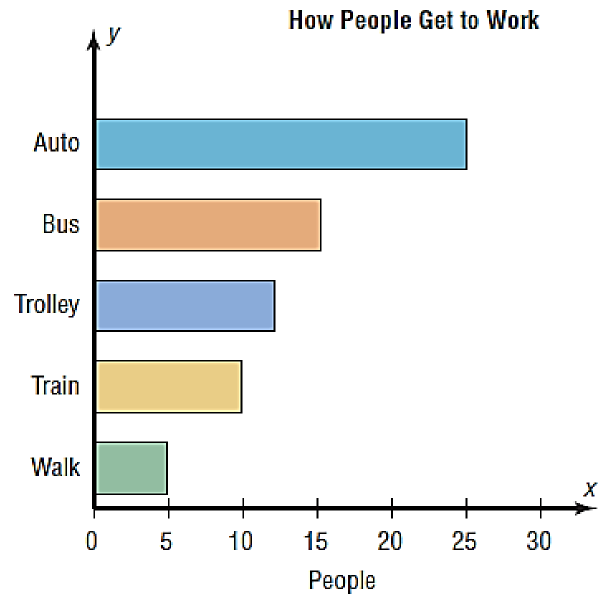


(g) Bimodal



(h) U-shaped

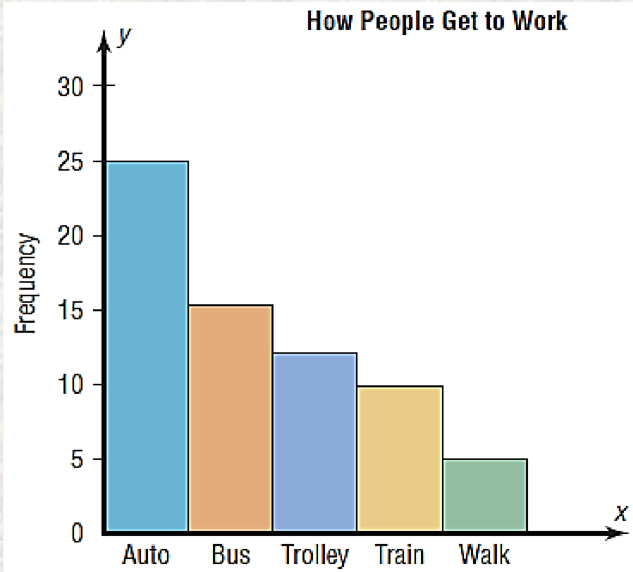
5. Other Types of Graphs



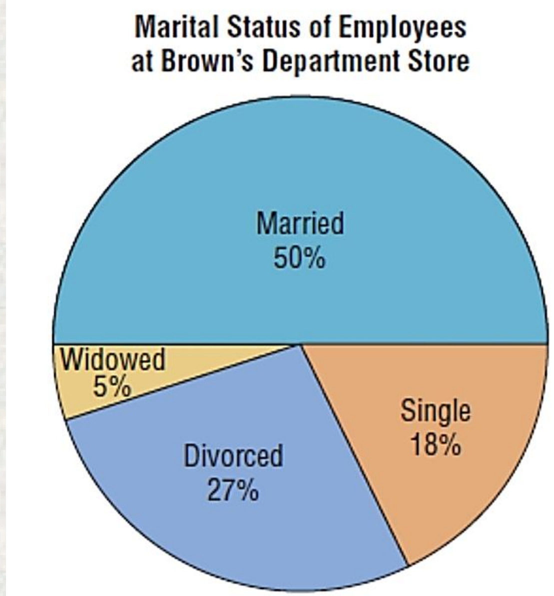
(a) Bar graph

(a) A **bar graph** represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.

(b) A **Pareto chart** is used to represent a frequency distribution for a categorical variable, and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.



(b) Pareto graph



(c) Pie graph

(c) A **pie graph** is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

Thank You

Any Questions?

Chapter Three

Data Description

1. Measures of Central Tendency
2. Measures of Variation
3. Measures of Position
4. Exploratory Data Analysis

Assi. Prof. Dr. Taher M. Ahmed
Civil Engineering Department
University of Anbar

Chapter Three
Data Description

1. Measures of Central Tendency

➢ Measures of average are called *measures of central tendency* and include several measurements such as: *mean, median, mode, midrange, etc. Two concepts must be defined:*

1. **Statistic** is a characteristic or measure obtained by using the data values from a sample.
2. **Parameter** is a characteristic or measure obtained by using all the data values from a specific population.

1.1. The Mean

The *mean,* also is known as the *arithmetic average*

The **mean** is the sum of the values, divided by the total number of values. The symbol $\bar{X}$ represents the sample mean.

$$\bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + \dots\dots\dots + X_n}{n}$$ **Raw data**

$$\bar{X} = \frac{\sum X_i f_i}{\sum f_i} = \frac{X_1 f_1 + X_2 f_2 + \dots\dots\dots + f_n X_n}{\sum fn}$$ **Tabulated data**

- where n represents the total number of values in the sample, X_i is the statistic and f_i is the frequency.
- For a population, the Greek letter (μ) is used for the mean and N the represents the total number of values in the population.

Example 1: The data show the number of patients in a sample of six hospitals who acquired an infection while hospitalized. Find the mean. 110 76 29 38 105 31

Solution:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{110+76+29+38+105+31}{6} = 64.8$$

Example 2: The data represent the number of miles run during one week for a sample of 20 runners. Find the mean.

Classes	5.5-10.5	10.5-15.5	15.5-20.5	20.5-25.5	25.5-30.5	30.5-35.5	35.5-40.5
Frequency	1	2	3	5	4	3	2

Solution

1. **Make a table as shown.**
2. **Find the midpoints of each class and enter them in column C.**

$$X_m = \frac{5.5 + 10.5}{2} = 8 \qquad \frac{10.5 + 15.5}{2} = 13$$

3. **For each class, multiply the frequency by the midpoint:** **$f_1 \times X_1 = 1\times 8 = 8$, $f_2 \times X_2 = 2\times 13 = 26$ etc.**

Classes	Frequency	Midpoint X_m	$f_i \times X_m$
5.5-10.5	1	8	8
10.5-15.5	2	13	26
15.5-20.5	3	18	54
20.5-25.5	5	23	115
25.5-30.5	4	28	112
30.5-35.5	3	33	99
35.5-40.5	2	38	76

4. **Find the sum of** $\sum X_i f_i$

$$\sum f_i = 20 \qquad \sum X_i f_i = 490$$

5. **Divide the sum by** $\sum f_i$ **to get the mean.**

$$\bar{X} = \frac{\sum X_i f_i}{\sum f_i} = \frac{490}{20} = 24.5\ mile.$$

Procedure Table

Finding the Mean for Grouped Data

Step 1 Make a table as shown.

A	B	C	D
Class	Frequency f	Midpoint X_m	$f \cdot X_m$

Step 2 Find the midpoints of each class and place them in column C.

Step 3 Multiply the frequency by the midpoint for each class, and place the product in column D.

Step 4 Find the sum of column D.

Step 5 Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

The formula for the mean is

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n}$$

[*Note:* The symbols $\Sigma f \cdot X_m$ mean to find the sum of the product of the frequency (f) and the midpoint (X_m) for each class.]

1.2. The Median

- The ***median*** is the halfway point in a data set. Before you can find this point, the data must be arranged in order. When the data set is ordered, it is called a **data array.**

- **Steps in computing the median of a data array**

(1) Arrange the data in order. (2) Select the middle point.

Note: (Raw data)

- **For odd number of values in the data set; the median was an actual data value.** $MD = \frac{n}{2}$
- **When there are an even number of values in the data set, the median will fall between two given values (average of the two values)** $MD = \frac{\frac{n}{2}+(\frac{n}{2}+1)}{2}$

Example 3: The number of tornadoes that have occurred in the United States over an 8-year period follows. Find the median. 684, 764, 656, 702, 856, 1133, 1132, 1303

Solution: (Odd values) (Arrange values)

656, 684, 702, 764, 856, 1132, 1133, 1303

n/2 ↑ (n/2)+1

Median

MD = (764 + 856)/2 = 810

Example 4: The number of children with asthma during a specific year in seven local districts is shown. Find the median. 253, 125, 328, 417, 201, 70, 90

Solution: (Even values)

70, 90, 125, 201, 253, 328, 417

Median (n/2) = 201

Tabulated Data

1. Determine ascending C.F.
2. Determine the order of median $(\frac{\sum f_i}{2} = \frac{20}{2} = 10)$.
3. Determine the class of median (between 20.5 to 30.5). (where 10 is located between 6 and 11.
4. Determine MD using the next Eq.

$$\text{MD} = L_1 + \left\{ \frac{\left[\left(\frac{\sum f_i}{2} \right) - F_i \right]}{f_M} \right\} \times W$$

Classes	Frequency	Midpoint X_m	*Ascending C.F.*
5.5-10.5	1	8	0
10.5-15.5	2	13	1
15.5-20.5	3	18	3
20.5-25.5	5	23	6
25.5-30.5	4	28	11
30.5-35.5	3	33	15
35.5-40.5	2	38	18
			20

Median class

W = Class width = 25.5-20.5 = 5

L_1 = Lower boundary limits of MD = 20.5

F_i = C.F. before the MD class = 3

f_M = median class frequency = 11-6= 5

$$\text{MD} = 20.5 + \left\{ \frac{\left[\left(\frac{20}{2} \right) - 3 \right]}{5} \right\} \times 5$$

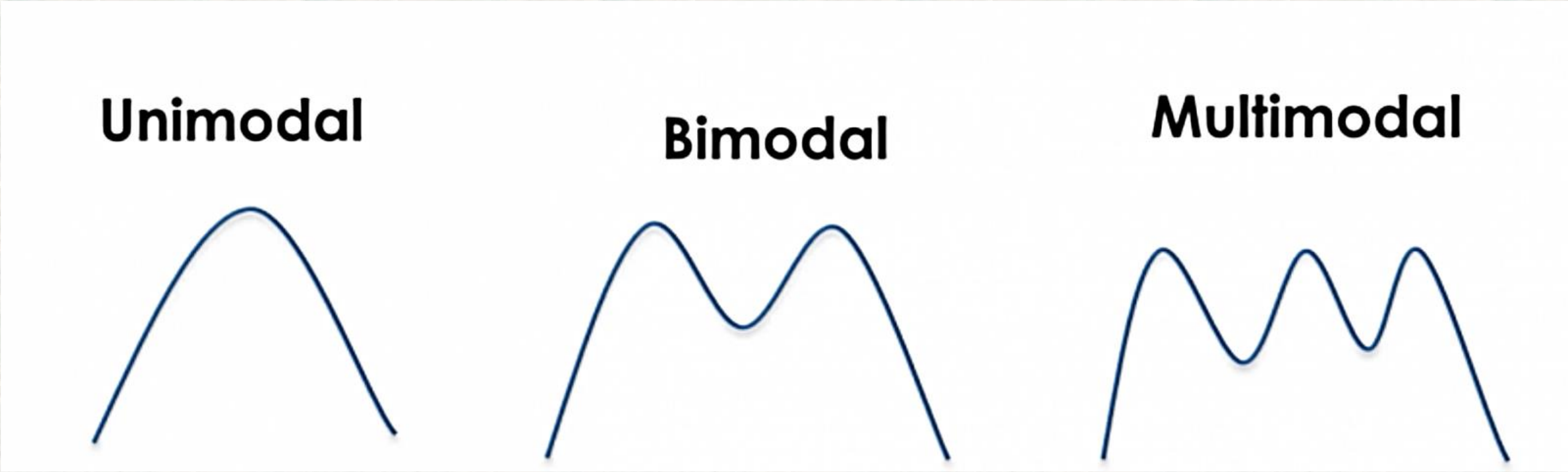
MD= 27.5

1.3. The Mode

The value that occurs most often in a data set is called the **mode.** For example, the set of data (3, 5, 8, 9, 5, 2, 5, 7, 5) **the mode is (5).**

A data set that has only :

- one value that occurs with the greatest frequency is called **unimodal.**
- two values with the same greatest frequency **bimodal.**
- more than two values with the same greatest frequency **multimodal.**



Example 4: The data show the number of licensed nuclear reactors in the United States for a recent 15-year period. Find the mode.

104 104 104 104 104
107 109 104 109 110
103 111 112 111 109

Solution

Since the values 104 occurred 6 times, the modes is 104. The data set is said to be unimodal.

Example:5 Find the mode for the number of branches that six banks have.

401, 344, 209, 201, 227, 353

Solution

Since each value occurs only once, there is no mode.

Note: Do not say that the mode is zero. That would be incorrect, because in some data, such as temperature, zero can be an actual value.

Tabulated Data

For tabulated data, mode can be calculated using the following formula:

$$M_o = L_1 + \left(\frac{d_1}{d_1 + d_2}\right) \times W$$

Where:

L_1 is the lower boundary limits For mode's class.

d_1 = the deference between mode's class and the previous class.

d_2 = the deference between mode's class and the next class.

W is the class width.

Example 6: The data represent the spot speed of a passenger cars in (km/hr) passing with a section of road. Find the mode.

Boundary limits	fi	Class midpoint
30.5-35.5	1	33
35.5-40.5	5	38
40.5-45.5	7	43
45.5-50.5	12	48
50.5-55.5	20	53
55.5-60.5	17	58
60.5-65.5	14	63
65.5-70.5	10	68
70.5-75.5	7	73
75.5-80.5	4	78
80.5-85.5	2	83
85.5-90.5	1	88

Solution:

- Determine the mode's class (50.5-550.5) where it is the highest frequency.
- $\mathbf{L_1}$ = is the lower boundary limits For mode's class (50.5).
- $\mathbf{d_1}$ = the deference between mode's class and the previous class (20-12 = 8).
- $\mathbf{d_2}$ = the deference between mode's class and the next class (20 -17 = 3).

W is the class width (5).

$$M_o = L_1 + \left(\frac{d_1}{d_1 + d_2}\right) \times W$$

$$M_o = 50.5 + \left(\frac{8}{8+3}\right) \times 5 = 54.14 \text{ km/hr}$$

2. Properties and Uses of Central Tendency

The Mean

1. The mean is found by using all the values of the data.
2. The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
3. The mean is used in computing other statistics, such as the variance.
4. The mean for the data set is unique and not necessarily one of the data values.
5. The mean is affected by extremely high or low values,

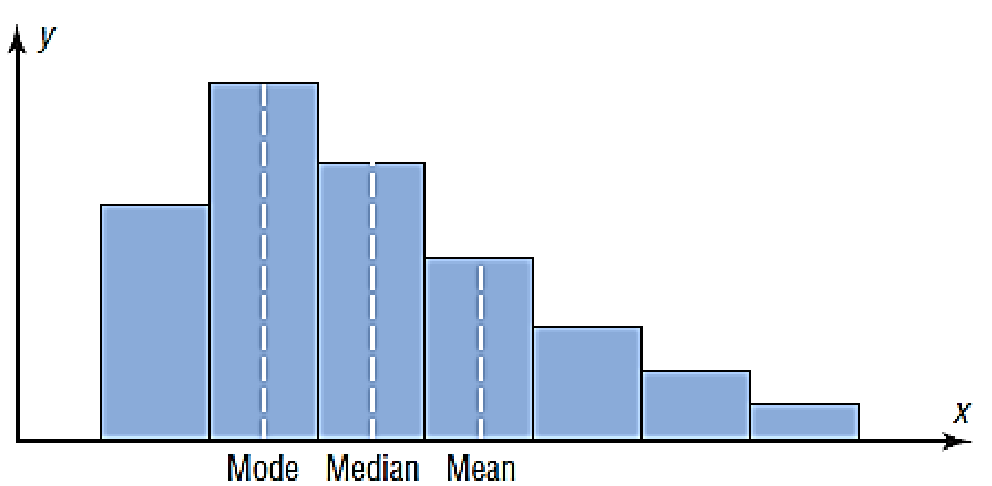
The Median

1. The median is used to find the center or middle value of a data set.
2. The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
3. The median is affected less than the mean by extremely high or extremely low values.

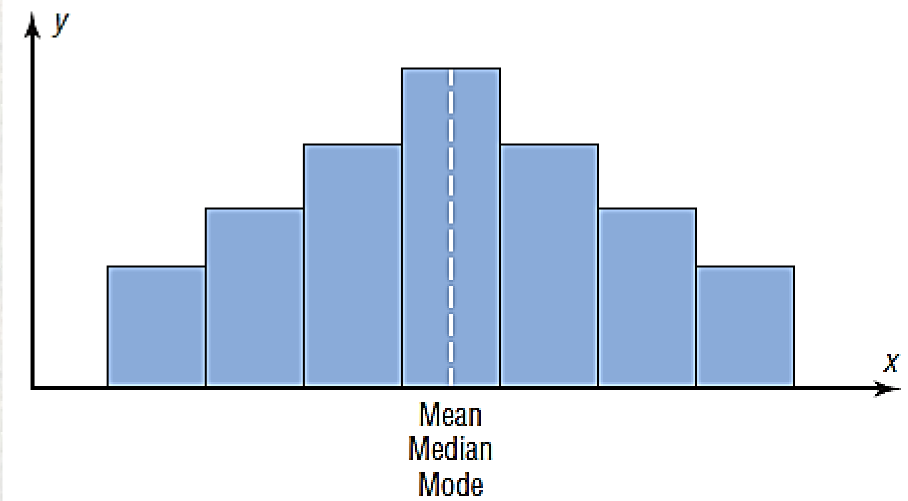
The Mode

1. The mode is used when the most typical case is desired.
2. The mode is the easiest average to compute.
3. The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

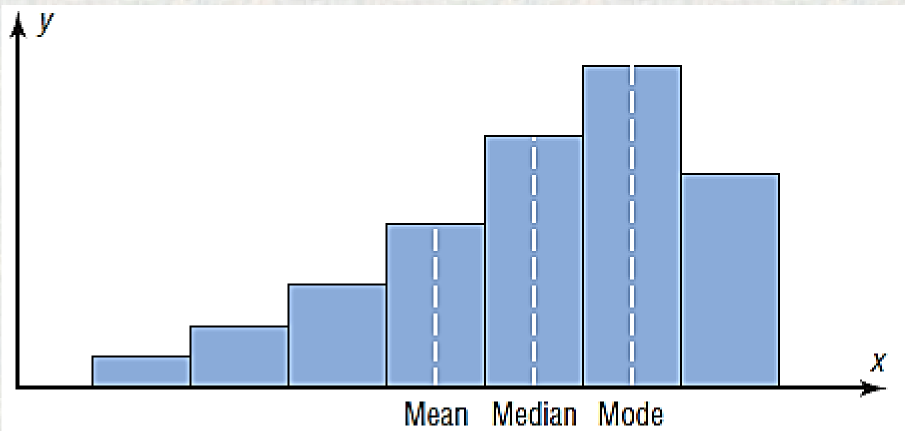
3. Distribution Shapes



(a) Positively skewed or right-skewed



(b) Symmetric



(c) Negatively skewed or left-skewed

2. Measures of Variation

2.1. Population Variance and Standard Deviation

Xi = 80, 85, 90, 98, 104, 115, 122, 130
($\bar{X}$= 103)
Yi = 101, 90, 113, 102, 103, 104, 106, 105
($\bar{Y}$= 103)

Both groups have the same mean but the variation of group X seems to be higher than group Y.

- The **variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is $\boldsymbol{\sigma^2}$ (is the Greek lowercase letter sigma). The formula for the population variance is.

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

Where: X individual value, μ population mean and N population size

- The **standard deviation** is the square root of the variance. The symbol for the population standard deviation is σ, The corresponding formula for the population standard deviation is

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

Example 6: A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading.? The data shows the results of six containers. Find the variance and standard deviation for the data set?

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

Solution: The mean for brands A and B are:

For A $\mu = \frac{\sum X}{n}$

$= \frac{10+60+50+30+40+20}{6} = 35$

For B $\mu = \frac{\sum X}{n} = 35$

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

	A	B
σ^2	291.67	41.67
σ	17.08	6.45

X_A	X_B	$(X_A-\mu)$	$(X_B-\mu)$	$(X_A-\mu)^2$	$(X_B-\mu)^2$
10	35	-25	0	625	0
60	45	25	10	625	100
50	30	15	-5	225	25
30	35	-5	0	25	0
40	40	5	5	25	25
20	25	-15	-10	225	100
$\mu =$ 35	$\mu =$ 35	$\Sigma =$ 0.0	$\Sigma =$ 0.0	$\Sigma =$ 291.67	$\Sigma =$ 41.67

????

Note: When the means are equal, the larger the variance or standard deviation is, the more variable the data are.

2.2. Sample Variance and Standard Deviation

- The formula for the sample variance, denoted by S^2, is

Where :

n = sample size

$\bar{X}$ = sample mean

X = individual value

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

- The standard deviation of a sample (denoted by S) is

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

NOTE: The expression $\boldsymbol{\sigma^2 = \frac{\sum(X-\mu)^2}{N}}$ does not give the best estimate of the population variance because when the population is large and the sample is small (usually less than 30), the variance computed by this formula usually underestimates the population variance. Therefore, instead of dividing by n, find the variance of the sample by dividing by n-1, giving a slightly larger value and an *unbiased* estimate of the population variance.

Example 7: Find the sample variance and standard deviation for the amount of European auto sales for a sample of 6 years shown. The data are in millions of dollars. **11.2, 11.9, 12.0, 12.8, 13.4, 14.3**.

Solution

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

$$S^2 = \frac{6.38}{6-1} = 1.276$$

$$S = \sqrt{S^2} = 1.13$$

X	X-$\bar{X}$	(X-$\bar{X}$)^2
11.2	**-1.4**	**1.96**
11.9	**-0.7**	**0.49**
12.0	**-0.6**	**0.36**
12.8	**0.2**	**0.04**
13.4	**0.8**	**0.64**
14.3	**1.7**	**2.89**
$\bar{X} = 35$		**Σ= 6.38**

2.3. Variance and Standard Deviation for Tabulated Data

The procedure for finding the variance and standard deviation for tabulated data is similar to that for finding the mean for tabulated data using the midpoints of each class using the following formula:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{f_i(X_m - \bar{X})^2}{\sum f_i - 1}}$$

Example 7: Find the variance and the standard deviation for the frequency distribution of the data in **Example 4.**?.

Solution

Classes	f_i	X_m	$X_m - \bar{X}$	$(X_m - \bar{X})^2$	$(X_m - \bar{X})^2 \times f_i$
5.5-10.5	1	8	-16.5	272.25	272.25
10.5-15.5	2	13	-11.5	132.25	264.5
15.5-20.5	3	18	-6.5	42.25	126.75
20.5-25.5	5	23	-1.5	2.25	11.25
25.5-30.5	4	28	3.5	12.25	49
30.5-35.5	3	33	8.5	72.25	216.75
35.5-40.5	2	38	13.5	182.25	364.5
Summation	20				1305

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{f_i(X_m - \bar{X})^2}{\sum f_i - 1}}$$

$$\bar{X} = \frac{\sum f_i \times X_m}{\sum f_i} = 24.5$$

$$\sigma = \sqrt{\frac{f_i(X_m - \bar{X})^2}{\sum f_i - 1}} = \sqrt{\frac{1305}{19}} = 8.28$$

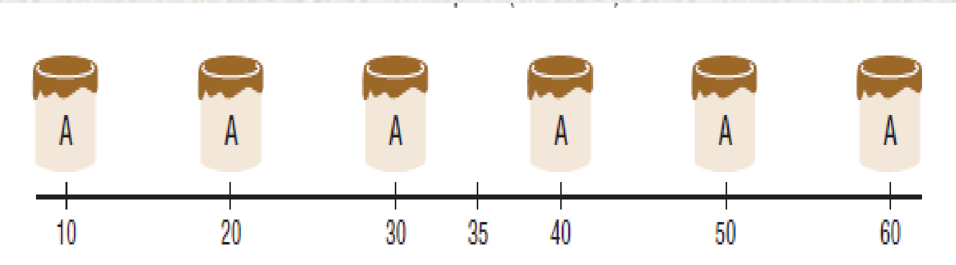
$$\sigma 2 = 68.68$$

2.4. Range

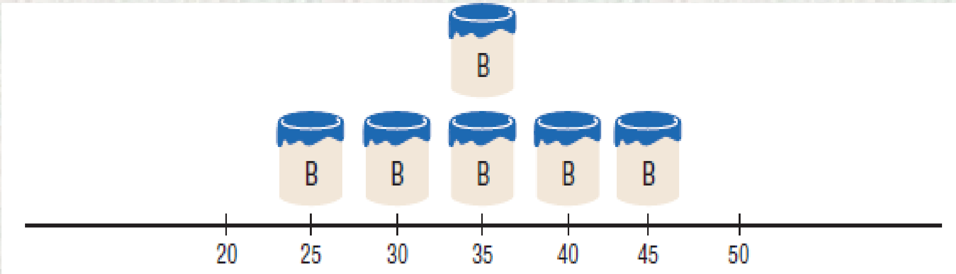
- The range is the simplest measurement of variance.
- The range is the highest value minus the lowest value. The symbol R is used for the range.

R = highest value - lowest value

Example 8: Find the ranges for the paints if last before fading in months?



(a) Brand A



(b) Brand B

Solution

- For brand A, the range is:
  R = 60 – 10 = 50 months
- For brand B, the range is:
  R = 45 – 25 = 20 months

Summary

Summary of Measures of Variation

Measure	Definition	Symbol(s)
Range	Distance between highest value and lowest value	R
Variance	Average of the squares of the distance that each value is from the mean	σ^2, s^2
Standard deviation	Square root of the variance	σ, s

Notes:

Uses of the Variance and Standard Deviation

1. As previously stated, variances and standard deviations can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed. This information is useful in comparing two (or more) data sets to determine which is more (most) variable.
2. The measures of variance and standard deviation are used to determine the consistency of a variable. For example, in the manufacture of fittings, such as nuts and bolts, the variation in the diameters must be small, or the parts will not fit together.
3. The variance and standard deviation are used to determine the number of data values that fall within a specified interval in a distribution. For example, Chebyshev's theorem (explained later) shows that, for any distribution, at least 75% of the data values will fall within 2 standard deviations of the mean.
4. Finally, the variance and standard deviation are used quite often in inferential statistics. These uses will be shown in later chapters of this textbook.

3. Coefficient of Variation

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

For samples

$$CV_{ar} = \frac{S}{\overline{X}}.100$$

For populations

$$CV_{ar} = \frac{\sigma}{\mu}.100$$

Example 9: The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is \$5225, and the standard deviation is \$773. Compare the variations of the two.

Solution

The coefficients of variation are

$$CV_{ar} = \frac{S}{\bar{X}}.100 = \frac{5}{87}.100 = 5.7\% \qquad \text{sales}$$

$$CV_{ar} = \frac{S}{\bar{X}}.100 = \frac{773}{5225}.100 = 14.8\% \qquad \text{commissions}$$

Note: Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

Example 10: The mean for the number of pages of a sample of women's fitness magazines is 132, with a variance of 23; the mean for the number of advertisements of a sample of women's fitness magazines is 182, with a variance of 62. Compare the variations.

Solution

The coefficients of variation are

$$CV_{ar} = \frac{S}{\bar{X}}.100 = \frac{\sqrt{23}}{132}.100 = 3.6\% \qquad \text{pages}$$

$$CV_{ar} = \frac{S}{\bar{X}}.100 = \frac{\sqrt{62}}{182}.100 = 4.3\% \qquad \text{advertisements}$$

Note: The number of advertisements is more variable than the number of pages since the coefficient of variation is larger for advertisements.

Chapter Four
Probability and Counting Rules

1. Sample Spaces and Probability
2. The Addition Rules for Probability
3. The Multiplication Rules and Conditional Probability
4. Counting Rules
5. Probability and Counting Rules

Assi. Prof. Dr. Taher M. Ahmed
Civil Engineering Department
University of Anbar

Chapter Four

Probability and Counting Rules

1. *Sample Spaces and Probability*

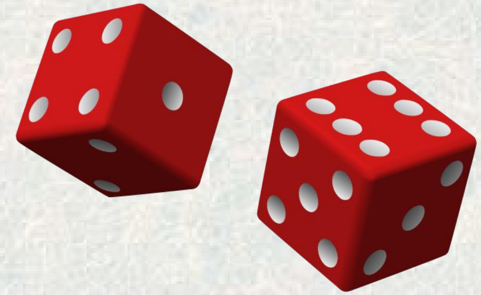
- **Probability** can be defined as the chance of an event occurring.

1.1. Basic Concepts

- **Probability experiment** is a chance process that leads to well-defined results called outcomes. such as flipping a coin, rolling a die, or drawing a card from a deck
- **Outcome** is the result of a single trial of a probability experiment. For example rolling a single die, there are six possible outcomes: 1, 2, 3, 4, 5, or 6.
- **Sample space** is the set of all possible outcomes of a probability experiment

Experiment	Sample space
Toss one coin	Head, tail
Roll a die	1, 2, 3, 4, 5, 6
Answer a true/false question	True, false
Toss two coins	Head-head, tail-tail, head-tail, tail-head

Example 1: Find the sample space for rolling two dice.



Solution

Since each die can land in six different ways, and two dice are rolled, the sample space can be presented by a rectangular array, as shown figure below:-

Die 1	Die 2					
	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Example 2: Find the sample space for the gender of the children if a family has three children. Use B for boy and G for girl.

Solution

There are two genders, male and female, and each child could be either gender. Hence, there are eight possibilities, as shown here.

BBB BBG BGB GBB GGG GGB GBG BGG

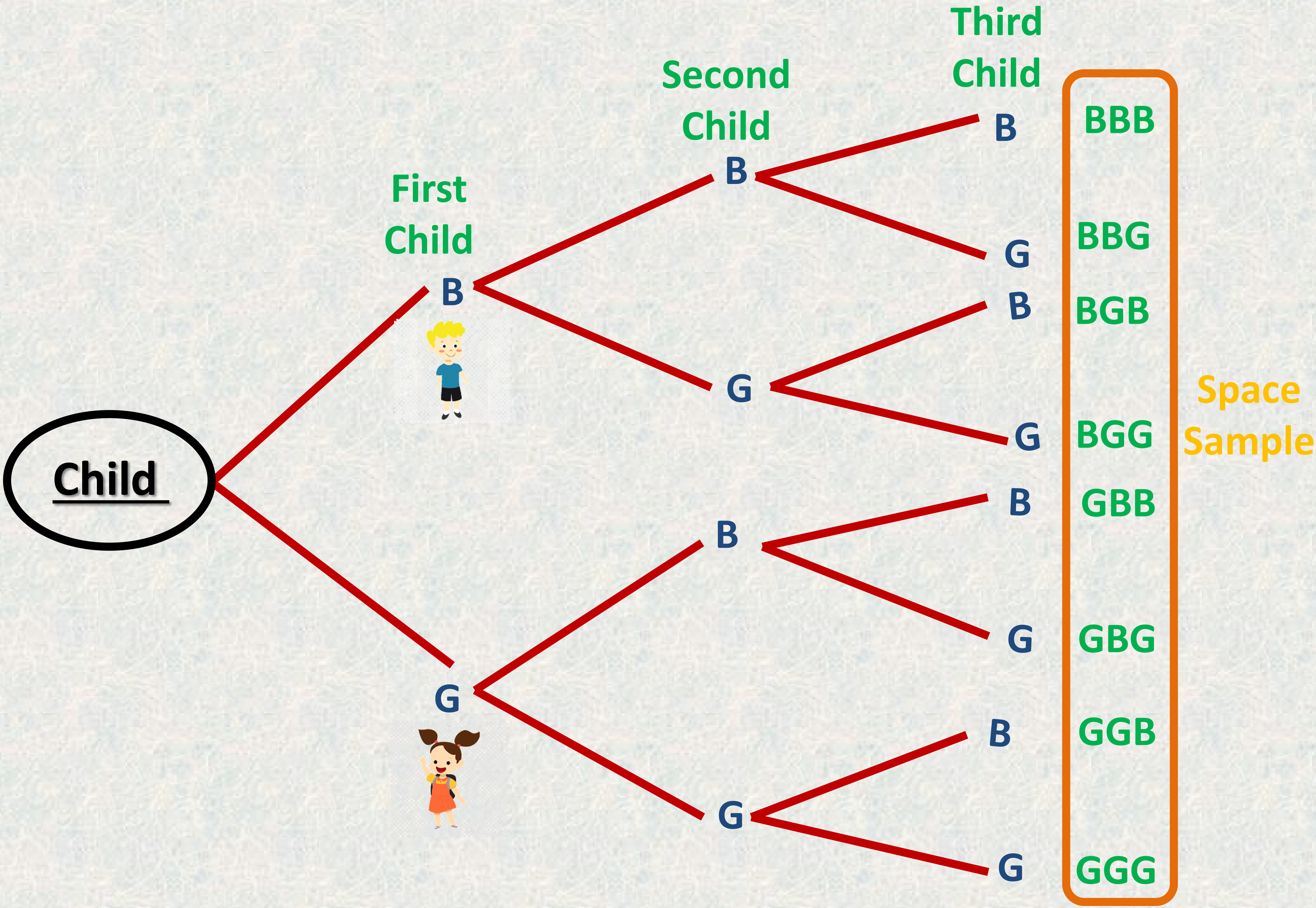
- **Tree diagram**

It is a device consisting of line segments emanating from a starting point and also from the outcome point. It is used to determine all possible outcomes of a probability experiment.

Example 3: Use a tree diagram to find the sample space for the gender of three children in a family, as in Example 2.

Solution:

Since there are two possibilities (boy or girl) for the first child, draw two branches from a starting point and label one B and the other G. Then if the first child is a boy, there are two possibilities for the second child (boy or girl), so draw two branches from B and label one B and the other G. Do the same if the first child is a girl. Follow the same procedure for the third child. The completed tree diagram is shown in the figure below. To find the outcomes for the sample space, trace through all the possible branches, beginning at the starting point for each one.


Child
First Child
B
G
Second Child
B
G
B
G
Third Child
B
G
B
G
B
G
B
G
BBB
BBG
BGB
BGG
GBB
GBG
GGB
GGG
Space
Sample

- **An event** consists of a set of outcomes of a probability experiment. Event can be one or more outcomes, for example a face from one trial dice is called **simple event**, or odd number from a single trials is called **compound event**.

Single Trial: 1 2 3 4 5 6

$\boxed{1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6}$

Simple Event

1 3 5

compound event

1.2. Classical Probability

- Classical probability assumes that all outcomes in the sample space are equally likely to occur. For example, when a single die is rolled, each outcome has the same probability of occurring which is (1/6) and for coin (1/2) and so on.
- The probability of any event *E* can be defined as:

$$\frac{\textbf{Number of outcomes in } E}{\textbf{Total number of outcomes in the sample space}}$$

OR $$P(E) = \frac{n(E)}{n(s)}$$ **OR** $$P(E) = \frac{n}{N}$$

Example 4: If a family has three children, find the probability that two of the three children are girls.

Solution:

The sample space = N = n(S) = 8: **(BBB BBG BGB GBB GGG GGB GBG BGG)**

The outcomes space = E = n(E) = n(2G) = n = 3: **(GGB GBG BGG)**

$$\boldsymbol{P(E) = \frac{n(E)}{n(s)} = \frac{n(2G)}{n(s)} = \frac{3}{8}}$$

Example 5: When a single die is rolled, find the probability of getting a 9.

Solution:

The sample space = N = n(S) = 6: **(1, 2, 3, 4, 5, 6)**

The outcomes space = E = n(E) = n(9) = n = 0: [ ]

$$\boldsymbol{P(E) = \frac{n(E)}{n(s)} = \frac{n(9)}{n(s)} = \frac{0}{6} = 0}$$

Example 6: When a single die is rolled, find the probability of getting an odd number.

Solution:

The sample space = N = n(S) = 6: **(1, 2, 3, 4, 5, 6)**

The outcomes space = E = n(Odd) = n = 3 = **(1, 3, 5)**

$$\boldsymbol{P(E) = \frac{n(E)}{n(s)} = \frac{n(odd)}{n(s)} = \frac{3}{6} = 0.5}$$

❖ Basic Probability Rules

➢ **Probability Rule 1:**

The probability of any event *E* is a number (either a fraction or decimal) between and including 0 and 1. This is denoted by: $\mathbf{0 \leq P(E) \leq 1}$

➢ **Probability Rule 2**

If an event E cannot occur (i.e., the event contains no members in the sample space), **its probability is 0**.

➢ **Probability Rule 3**

If an event E is certain, then the probability of **E is 1**.

➢ **Probability Rule 4**

The sum of the probabilities of all the outcomes in **the sample space is 1**.

Example 7: A single die is rolled, what is the probability of getting a number less than 7?

Solution

Since all outcomes (1, 2, 3, 4, 5, 6) are less than 7,

the probability is: P(X < 7) = 1 $P(X < 7) = \frac{n}{N} = \frac{6}{6} = 1$

The event of getting a number less than 7 is certain.

- **Complement of an event**

If E is the set of outcomes in the sample space that are not included in the outcomes of event E. The complement of E is denoted by $\bar{E}$ (read "E bar").

$$\boldsymbol{P(E) + P(\bar{E}) = 1 \quad \Rightarrow \quad P(E) = 1 - P(\bar{E})}$$

Example 8: If the probability that a person lives in an industrialized country of the world is , find the probability that a person does not live in an industrialized country.

Solution

P(not living in an industrialized country) = 1 - P(living in an industrialized country)= $1-\frac{1}{5}=\frac{4}{5}$ **OR** $\mathrm{P(E)}=1-\mathrm{P}(\bar{\mathrm{E}})$ **= 1-** $\frac{1}{5}=\frac{4}{5}$

1.3. Empirical Probability

Given a frequency distribution, the probability of an event being in a given class is

$$\boldsymbol{P(E) = \frac{Frequency\ for\ the\ class}{Total\ frequency\ in\ the\ distribution\ table} = \frac{f_i}{\sum f_i}}$$

This probability is called empirical probability and is based on observation.

Example 9: In the travel survey, as shown in Table below, find the probability that a person will travel by airplane over the thanksgiving holiday.

Method	Frequency
Drive	41
Fly	6
Train or bus	3
	50

Solution

$$P(E) = \frac{f_i}{\sum f_i} = \frac{6}{50} = \frac{3}{25}$$ is the probability of the person traveling by fly.

Example 10: In a sample of 50 people, 21 had type O blood, 22 had type A blood, 5 had type B blood, and 2 had type AB blood. Set up a frequency distribution and find the following probabilities. **a.** A person has type O blood. **b.** A person has type A or type B blood. **c.** A person has neither type A nor type O blood. **d.** A person does not have type AB blood.

Solution

Type	Frequency
A	22
B	5
AB	2
O	21
Total	50

$$a.\ P(O) = \frac{f_i}{\sum f_i} = \frac{21}{50}$$

b. $P(A\ or\ B) = \frac{f_i}{\sum f_i} = P(A) + P(B) = \frac{22}{50} + \frac{5}{50} = \frac{27}{50}$

c. $P(neither\ A\ nor\ O) = P(B\ and\ AB) = \left(P(AB) + P(B)\right)$

$$= \frac{2}{50} + \frac{5}{50} = \frac{7}{50}$$

d. $P(not\ AB) = 1 - P(AB) = 1 - \frac{2}{50} = \frac{48}{50} = \frac{24}{25}$

Example 11: Hospital records indicated that knee replacement patients stayed in the hospital for the number of days shown in the distribution Table. Find these probabilities:-

a. A patient stayed exactly 5 days.
b. A patient stayed less than 6 days.
c. A patient stayed at most 4 days.
d. A patient stayed at least 5 days.

Number of days stayed	Frequency
3	15
4	32
5	56
6	19
7	5
Total	127

Solution

$a. P(5) = \frac{f_i}{\sum f_i} = \frac{56}{127}$

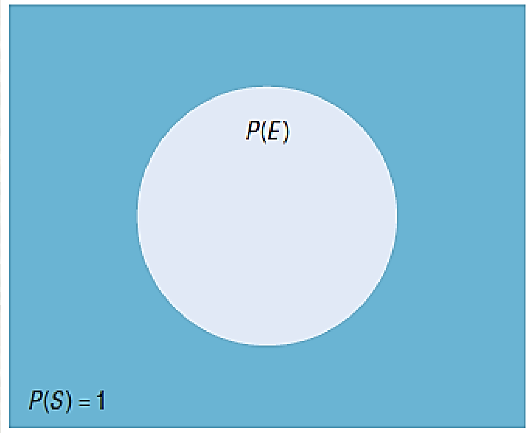
b. $P(fewer\ than\ 6\ days) = P(5) + P(4) + P(3) = \frac{56}{127} + \frac{32}{127} + \frac{15}{127} = \frac{103}{127}$

c. $P(at\ most\ 4\ days) = P(4) + P(3) = \frac{32}{127} + \frac{15}{127} = \frac{47}{127}$

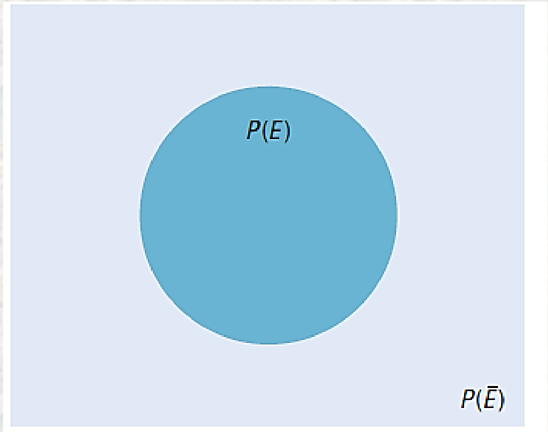
d. $P(at\ least\ 5\ days) = P(5) + P(6) + P(7) = \frac{56}{127} + \frac{19}{127} + \frac{5}{127} = \frac{80}{127}$

1.4. Venn Diagram

- It is an illustration that uses circles to show the relationships among things or finite groups of things.
- It is often useful to use a **Venn diagram** to visualize the **probabilities** of multiple events.



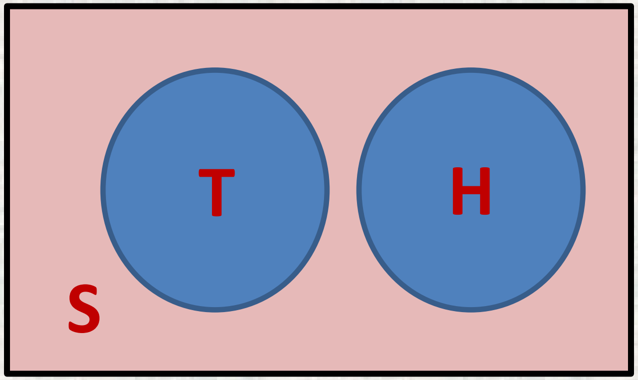
(a) Simple probability



(b) $P(\bar{E}) = 1 - P(E)$

2. The Addition Rules for Probability

Mutually exclusive events: are two or more events which cannot occur at the same time (i.e., they have no outcomes in common). For example coin experiment (H or T), on trial dice (1, or 2 or……….6)



2.1. Addition Rule 1

- **When two events *A* and *B* are mutually exclusive, the probability that *A* or *B* will occur is:**

$$P(A \text{ or } B) = P(A) + P(B)$$

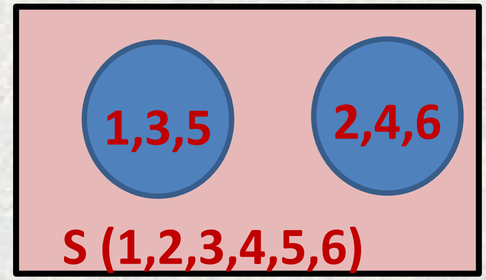
- **More than two events:**

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$

Example 12: Determine which events are mutually exclusive and which are not, when a single die is rolled: **(a)** Getting an odd number and getting an even number; **(b)** Getting a 3 and getting an odd number; **(c)** Getting an odd number and getting a number less than 4; **(d)** Getting a number greater than 4 and getting a number less than 4.

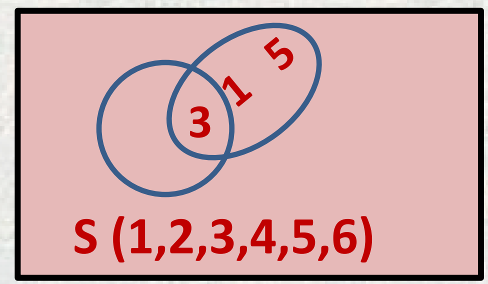
Solution

(a) The events are mutually exclusive, since the first event can be 1, 3, or 5 and the second event can be 2, 4, or 6.



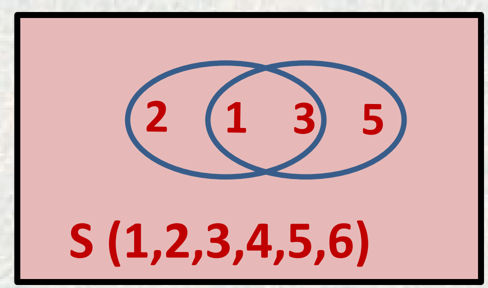
(a)

(b) The events are not mutually exclusive, since the first event is a 3 and the second can be 1, 3, or 5. Hence, 3 is contained in both events.



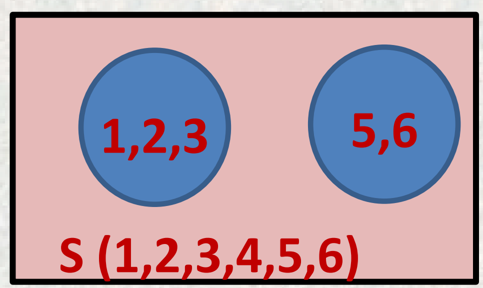
(b)

(c) The events are not mutually exclusive, since the first event can be 1, 3, or 5 and the second can be 1, 2, or 3. Hence, 1 and 3 are contained in both events.



(c)

(d) The events are mutually exclusive, since the first event can be 5 or 6 and the second event can be 1, 2, or 3.



(d)

Example 13: A city has 9 Steel factories: 3 high strength, 2 high carbon, and 4 recycled steel. If a contract selects one factory at random to buy tones of steel, find the probability that it is either high strength or recycled steel.

Solution

Since there are 3 high strength, and 4 recycled steel, and a total of 9 factories.

***P*(high strength(HS) or 4 recycled steel(RS)) = *P*(*HS*) + *P*(RS)=**$\frac{3}{9}+\frac{4}{9}=\frac{7}{9}$

The events are mutually exclusive.

Example 14: The corporate research and development centers for three local companies have the following number of employees:

U.S. Steel	**110**
Alcoa	**750**
Bayer Material Science	**250**

If a research employee is selected at random, find the probability that the employee is employed by U.S. Steel or Alcoa.

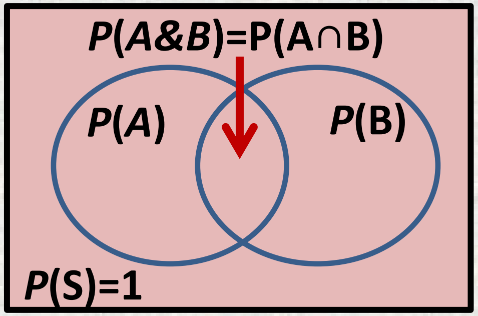
Solution

***P*(U.S. Steel or Alcoa) = P(*P*(U.S. Steel) + P(Alcoa))=**

$\frac{110}{1110}+\frac{750}{1110}=\frac{860}{1110}=\frac{86}{111}$

2.2. Addition Rule 2

This rule can also be used when the events are mutually exclusive, since *P*(*A* and *B*) will always equal 0. However, it is important to make a distinction between the two situations.

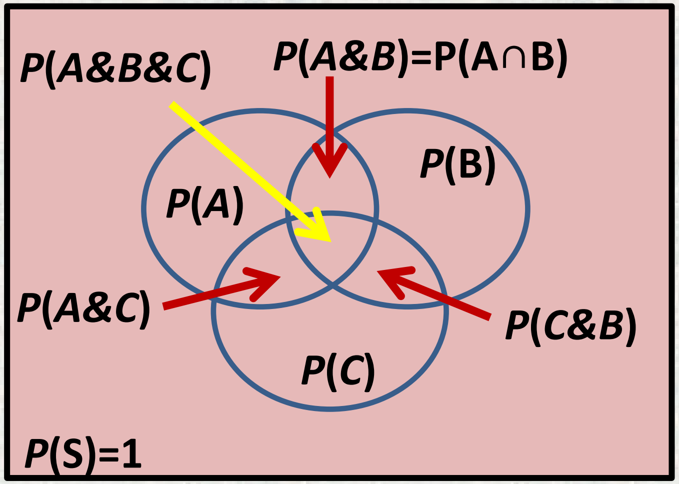


- **If *A* and *B* are *not* mutually exclusive, then:**

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

- For three events that are *not* mutually exclusive,

$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) - P(A \text{ and } B)$
$- P(A \,\&\, C) - P(B \,\&\, C) + P(A \,\&\, B \,\&\, C)$

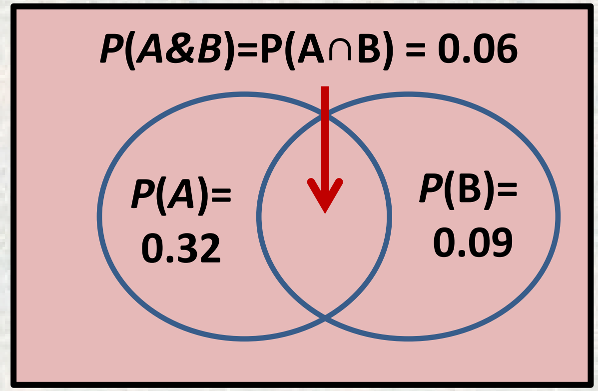


Example 14: The probability of a person driving while intoxicated is 0.32, the probability of a person having a driving accident is 0.09, and the probability of a person having a driving accident while intoxicated is 0.06. What is the probability of a person driving while intoxicated or having a driving accident?

Solution

P(intoxicated or accident) = P(intoxicated) + P(accident) - P(intoxicated and accident)

P(intoxicated or accident) = 0.32 + 0.09 - 0.06 = 0.35



2.3. The Multiplication Rules and Conditional Probability

➢ The Multiplication Rules

- The *multiplication rules* can be used to find the probability of two or more events that occur in sequence (dependent and independent events).
- Two events A and B are independent events if the fact that A occurs does not affect the probability of B occurring.
- For example: Rolling a die and getting a 6, and then rolling a second die and getting a 3.

Multiplication Rule 1:

✓ When two events are independent, the probability of both occurring is:

$$P(A \text{ and } B) = P(A)\,.\,P(B)$$

Example 14: A coin is flipped and a die is rolled. Find the probability of getting a head on the coin and a 4 on the die.

Solution

P(head and 4) = P(head) . P(4) = $\frac{1}{2}.\frac{1}{6} = \frac{1}{12}$

Sample space: coin (T,H) and Die (1,2,3,4,5,6) (both are independent): ***[T1, T2, T3, T4,T5, T6, H1, H2, H3, H4, H5, H6] = 12***

Example 15: A box contains 3 red balls, 2 blue balls, and 5 white balls. A ball is selected and its color noted. Then it is replaced. A second ball is selected and its color noted. Find the probability of each of these.

a. Selecting 2 blue balls

b. Selecting 1 blue ball and then 1 white ball

c. Selecting 1 red ball and then 1 blue ball

Solution

$P(blue) = \frac{n}{N} = \frac{2}{10}, P(red) = \frac{3}{10}$

$P(white) = \frac{5}{10}$

a. P(blue and blue) = P(blue).P(blue)

$= \frac{2}{10} . \frac{2}{10} = \frac{4}{100}$ (B1B1, B1B2, B2B1, B2B2)

b. P(blue and white) = P(blue).P(white) =

$\frac{2}{10} . \frac{5}{10} = \frac{1}{10}$

c.P(red and blue) = P(red).P(blue)

$= \frac{3}{10} . \frac{2}{10} = \frac{6}{100}$

Sample space: B1B1, B1B2, B2B1, B1R1, B1R2, B1R3, B1W1, B1W2, B1W3, B1W4, B1W5, B2B1, B2R1, B2B2, B2R3, B2W1, ……… R1B1, R1B2, R1R1, ……..W1B1, W1B2, W1W1, …..

✓ For three or more independent events by using the formula:

$$P(A \text{ and } B \text{ and } C \text{ and} \ldots \text{and } K) = P(A) \,.\, P(B) \,.\, P(C) \ldots P(K)$$

Example 16: At a signalized intersection, three cars come one by one, at the end, they have to turn left or write, determine the probability of? a) RRR, b) LRL, c) 2L1R.?

Solution

Each car will turn left or write (**independent events) ????** .

$P(R) = P(L) = \frac{1}{2}$

Sample space = RRR, RRL, RLR, LRR, LLL, $\boxed{\text{LLR, LRL, RLL}}$ = 8

a) $P(RRR) = P(R).\, P(R) \,.\, P(R) = \frac{1}{2}.\frac{1}{2}.\frac{1}{2} = \frac{1}{8}$ (in traditional probability = $\frac{n}{N} = \frac{1}{8}$)

b) $P(LRL) = P(L).P(R).P(L) = \frac{1}{2}.\frac{1}{2}.\frac{1}{2} = \frac{1}{8}$ (in traditional probability = $\frac{n}{N} = \frac{1}{8}$)

c) $\underbrace{P(2L1R) = P(LLR)+P(LRL)+P(RLL)}_{\text{Mutually exclusive event}} = \frac{1}{2}.\frac{1}{2}.\frac{1}{2} + \frac{1}{2}.\frac{1}{2}.\frac{1}{2} + \frac{1}{2}.\frac{1}{2}.\frac{1}{2} = \frac{3}{8}$

(in traditional probability = $\frac{n}{N} = \frac{3}{8}$)

Multiplication Rule 2

- When the outcome or occurrence of the first event affects the outcome or occurrence of the second event in such a way that the probability is changed, the events are said to be dependent events. For example when one ball is drawn without replacement by one.
- When two events are dependent, the probability of both occurring is

$$P(A \text{ and } B) = P(A) \, . \, P(B|A)$$

Where: the probability that event B occurs when event A has already occurred.

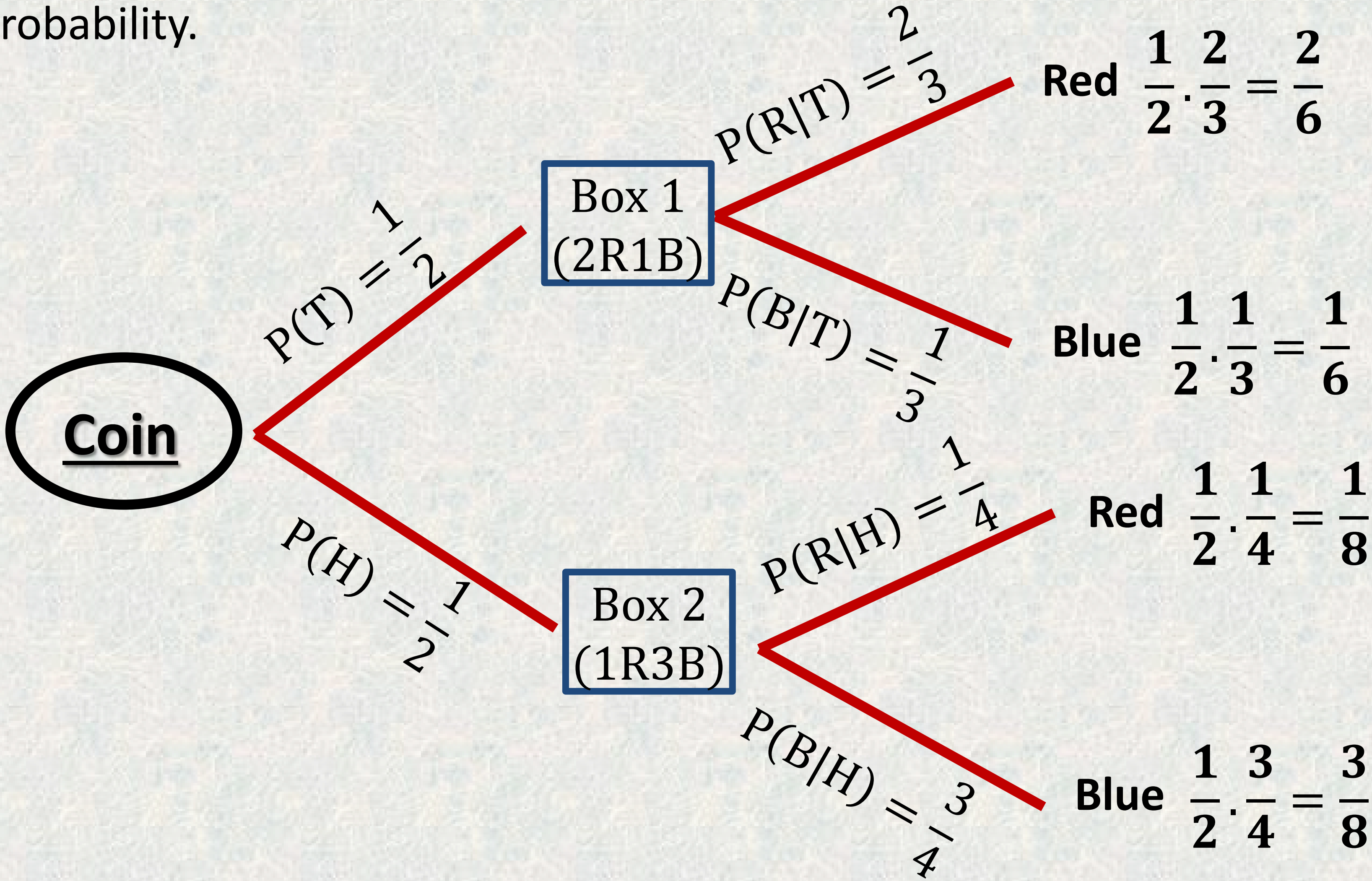
Example 17: At a university in western Pennsylvania, there were 5 burglaries reported in 2003, 16 in 2004, and 32 in 2005. If a researcher wishes to select at random two burglaries to further investigate, find the probability that both will have occurred in 2004.

Solution

In this case, the events are dependent since the researcher wishes to investigate two distinct cases. Hence the first case is selected and not replaced.

$$P(C_1 \& C_2) = P(C_1).\,P(C_2|C_1) = \frac{16}{53} \cdot \frac{15}{52} = \frac{60}{689}$$

Example 18: Box 1 contains 2 red balls and 1 blue ball. Box 2 contains 3 blue balls and 1 red ball. A coin is tossed. If it falls heads up, box 1 is selected and a ball is drawn. If it falls tails up, box 2 is selected and a ball is drawn. Find the tree probability.



Conditional Probability

The probability that the second event B occurs given that the first event A has occurred can be found by dividing the probability that both events occurred by the probability that the first event has occurred. The formula is

$$P(B/A) = \frac{P(A \text{ and } B)}{P(A)}$$

Proving:

$$P(A\ and\ B) = P(A).P(B|A)$$

$$\frac{P(A\ and\ B)}{P(A)} = \frac{\cancel{P(A)}.P(B|A)}{\cancel{P(A)}}$$

$$\Rightarrow \quad P(B/A) = \frac{P(A \text{ and } B)}{P(A)}$$

Example 19: A box has 6 red balls and 4 black balls, if two balls has been drawn one by one without replacement. Determine the probability of the second try is being red if the first is red as well?

Solution

P(R1)=6/10

P(R2)=5/9

P(R2/R1)= P(R1 . R2)/P(R1)=

(6/10×5/9)/6/10=(1/3)/(6/10)=5/9

Example 20: In a residential complex 1000 apartments, 500 residents in the northern sector and 500 others in the southern sector in each sector, 200 of the apartments contain large windows, 100 central heated, and 30% of the apartments with large windows are centrally heated. When choosing an apartment randomly, determine the probability of:

1. In the northern sector
2. In the northern sector, with a large windows
3. In the northern sector, with a large windows that are centrally heated.
4. In the southern sector and unheated centrally.

Solution

E1 = apartment in northern sector
E2 = apartment with a large windows
E3 = apartment centrally heated
N(E1)=500
N(E2)=200; P(E2) = 200/1000= 0.2 northern
N(E3)=160; P(E3) = 160/1000 = 0.16
N(E1&E2)=200; P(E1&E2) = 200/1000= 0.2
N(E1&E3)=100; P(E&E3) = 0.1
N(E2&E3)=(400*0.3)=120
N(E1&E2&E3)=200*0.3=60; (P(E1E2E3) = 0.06
N(E1&E3)=160

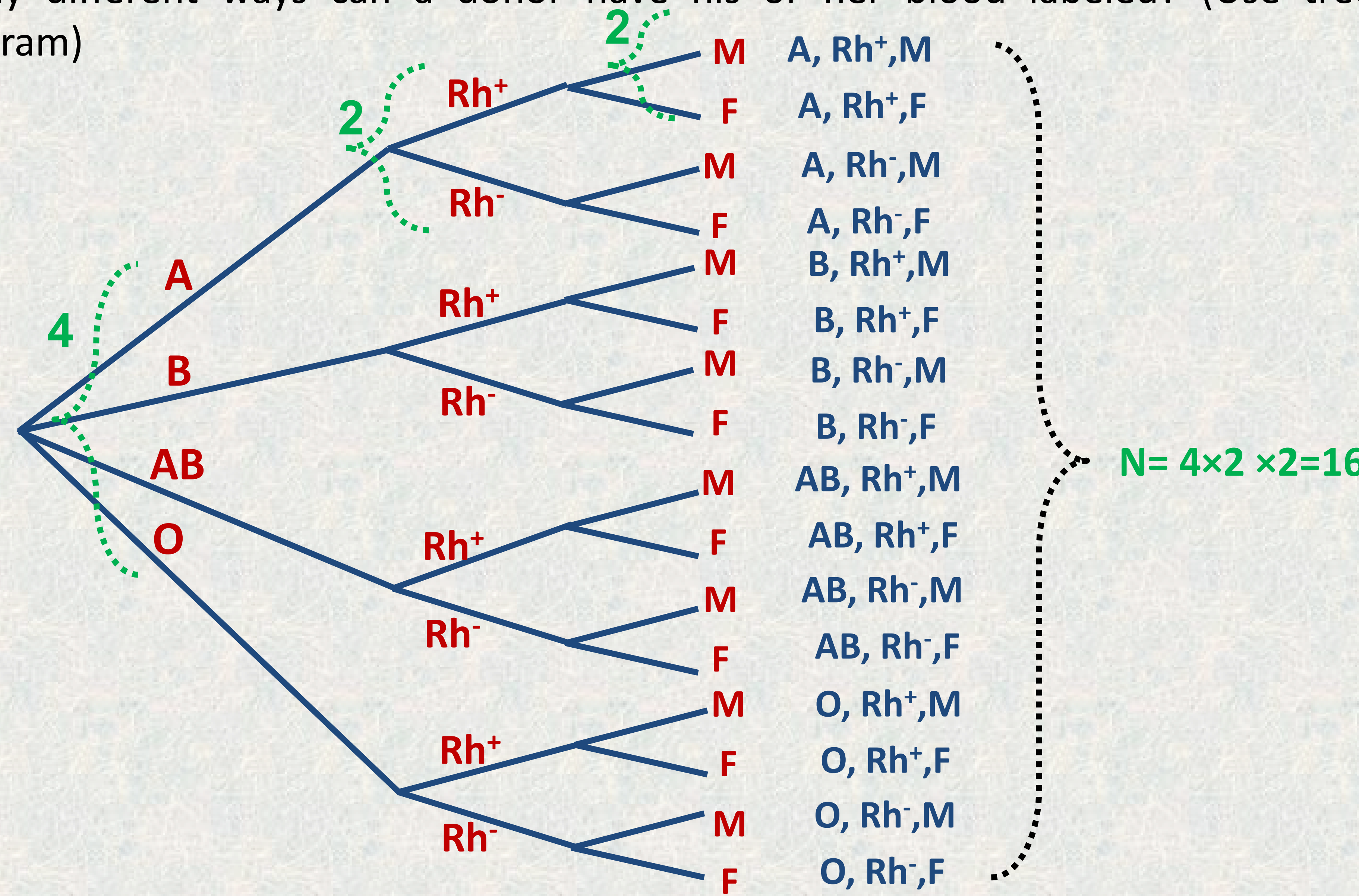
1.P(E1)= $\frac{n}{N} = \frac{500}{1000}$ = 0.5

2.P(E2|E1)= $\frac{P(E1).P(E2)}{P(E1)} = \frac{0.2\times0.5}{0.5} = 0.20$

3.P(E1E2|E3)=$\frac{P(E1E2).P(E3)}{P(E1E2)} = \frac{0.2\times0.06}{0.2}$ = 0.06

$4.P(\overline{E1E3})$= 1-P(E1E3) = 1- 160/1000=0.84

Example 21: There are four blood types, A, B, AB, and O. Blood can also be Rh and Rh. Finally, a blood donor can be classified as either male or female. How many different ways can a donor have his or her blood labeled? (Use tree diagram)



4. *Counting Rules*

4.1. Factorial Notation

For any counting *n, factorial formula is:* $n! = n(n \quad 1)(n \quad 2) \ldots\ldots 1$

$0! = 1$

Example 22: **5! = 5 . 4 . 3 . 2 . 1 = 240**

4.2. Permutations

The arrangement of n objects in a specific order using r objects at a time is called a permutation of n objects taking r objects at a time. It is written as ${}_nP_r$, and the formula is:

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

Example 23: The advertising director for a television show has 7 ads to use on the program. If she selects 1 of them for the opening of the show, 1 for the middle of the show, and 1 for the ending of the show, how many possible ways can this be accomplished?

Solution

Since order is important, the solution is:

$$ {}_n\mathbf{P}_r = \frac{\mathbf{n!}}{(\mathbf{n-r})!} \Rightarrow {}_7P_3 = \frac{7!}{(7-3)!} = \mathbf{210} $$

means there would be 210 ways to show 3 ads. $P_1P_1P_1$, $P_1P_1P_2$, $P_1P_1P_3$, $P_1P_1P_4$, $P_1P_1P_5$, $P_1P_1P_6$, ...

Example 24: A school musical director can select 2 musical plays to present next year. One will be presented in the fall, and one will be presented in the spring. If she has 9 to pick from, how many different possibilities are there?

Solution

Order is important since one play can be presented in the fall and the other play in the spring. $_9P_2 = \frac{9!}{(9-2)!}$ = 72

M1M2, M1M3,.....M2M1, M2M3, M2M4, ...

4.3. Combinations

The number of combinations of r objects selected from n objects is denoted by $_nC_r$ and is given by the formula:

$$_nC_r = \frac{n!}{(n-r)!\,r!}$$

Note: Combinations are used when the order or arrangement is not important, as in the selecting process.

Example 25: Given the letters A, B, C, and D, list the permutations and combinations for selecting two letters. Using permutation and combination.

Solution

The permutations are:

$_4P_2 = \frac{4!}{(4-2)!}$ = 12

AB	BA	CA	DA
AC	BC	CB	DB
AD	BD	CD	DC

The combination are:

$_nC_r = \frac{n!}{(n-r)!r!} = \frac{4!}{2!2!} = 6$

AB	~~BA~~	~~CA~~	~~DA~~
AC	BC	~~CB~~	~~DB~~
AD	BD	CD	~~DC~~

Example 26: In a club there are 7 women and 5 men. A committee of 3 women and 2 men is to be chosen. How many different possibilities are there?

Solution

$$\mathbf{7C3 \;.\; 5C2} = \frac{7!}{(7-3)!3!} \,.\, = \frac{5!}{(5-2)!2!} = \mathbf{350}$$

Here, you must select 3 women from 7 women, which can be done in 7C3, or 35, ways. Next, 2 men must be selected from 5 men, which can be done in 5C2, or 10, ways. Finally, by the fundamental counting rule, the total number of different ways is 35 × 10 = 350.

4.4. Probability and Counting Rules

Example 27: An exhibition has inside 8 red cars, 3 white and 9 blue. If three cars have been sold. Find the probability of: 1) 3 read, 2) 3 white, 3) 2 red 1 white, 4) at least 1 white 5) one from each color.

Solution

Counting rules depend on the concept of traditional probability ($\frac{n}{N}$) to determine n and N using combination analysis definition.

$$\mathbf{P(E)} = \frac{n}{N} = \frac{No.of\ groups\ for\ outcoms\ cases\ of\ the\ event\ E}{Number\ of\ groups\ for\ possible\ cases}$$

R1R1R1, R1R1R2, R1R1R3, ……, R1R2R1, R1R2R2,
مجاميع ثلاثية للون الاحمر فقط ....

1) $P(3R) = \frac{n}{N} = \frac{\textit{No.of groups for 3R from 8R}}{\textit{Number of groups for any 3 cars from all cars (20)}} = \frac{3C8}{3C20} = 0.49$

R1R1R1, R1R1R2, R1R1R3, ……, R1R2R1, R1R2R2, …R1R1W1,
R1R1W2, …… (مجاميع ثلاثية من جميع الالوان)

2) $P(3W) = \frac{n}{N} = \frac{\textit{No.of groups for 3W from 3W}}{\textit{Number of groups for any 3 cars from all cars (20)}} = \frac{3C3}{3C20} = 0.0008$

3) $P(2R1W) = \frac{n}{N} = \frac{(\textit{No.of groups for 2R from 8R}) \times (\textit{No.of groups for 1W from 3W})}{\textit{Number of groups for any 3 cars from all cars (20)}}$

$$= \frac{2C8 \times 1C3}{3C20} = 0.0008$$

4) At least 1W $= P(W \geq 1) = \frac{n}{N} = \frac{P(1W)+P(2W)+P(3W)}{3C20} =$

$$\frac{1C3 \times 2C17 + 2C3 \times 1C17 + 3C3 \times 0C17}{3C20} = \frac{404 + 51 + 6}{1140} = 0.403$$

<u>OR</u> $P(W \geq 1) = 1 - P(\overline{W}) = 1 - \frac{\text{3 cars are not white from 17 cars}}{3C20} = 1 - \frac{3C17}{3C20} = 0.403$

5) P(1R1W1B) $= \frac{1C8 \times \mathbf{1C3} \times \mathbf{1C9}}{\mathbf{3C20}}$ = 0.189

Example 28: A store has 6 TV Graphic magazines and 8 News-time magazines on the counter. If two customers purchased a magazine, find the probability that one of each magazine was purchased.

Solution

*P (*1 *TV Graphic* and 1 *News-time)* = $\frac{6C1 \,.8C1}{14C2} = 0.527$

End of Chapter Four

Chapter Five

Discrete Probability Distributions

1. Probability Distributions
2. Mean, Variance, Standard an Deviation
3. The Binomial Distribution
4. Other Types of Distributions

Assi. Prof. Dr. Taher M. Ahmed
Civil Engineering Department
University of Anbar

Chapter Five
Discrete Probability Distributions

1. *Probability Distributions*

This chapter explains the concepts and applications of what is called a probability distribution. In addition, special probability distributions, such as the binomial, multinomial, Poisson, and hyper-geometric distributions, are explained.

- **Random variable** is a variable whose values are determined by chance.
- **Discrete variables** which have a finite number of possible values or an infinite number of values that can be counted. The word *counted* means that they can be **enumerated** using the numbers 1, 2, 3, etc. For example, the number of family members (1, 2, 3, 4, …), number of calls, and so on..
- **Example 1:** three coins are tossed, the sample space is represented as { TTT, TTH, THT, HTT, HHT, HTH, THH, HHH}; if X is the random variable for the number of heads, then X assumes the value 0, 1, 2, or 3. (X = 0, 1, 2, 3)
 (X: 0 = no head, 1 = one head, 2 = two head, 3 = three head)

Probabilities for the values of *X* can be determined as follows:

No heads	***One heads***	***Two heads***	***Three heads***
TTT	**TTH THT HTT**	**HHT HTH THH**	**HHH**
1/8	**1/8 1/8 1/8**	**1/8 1/8 1/8**	**1/8**
1/8	**3/8**	**3/8**	**1/8**

Number of heads X	**0**	**1**	**2**	**3**
Probability P(X)	**1/8**	**3/8**	**3/8**	**1/8**

- A **discrete probability distribution** consists of the values a random variable can assume and the corresponding probabilities of the values. The probabilities are determined theoretically or by observation.

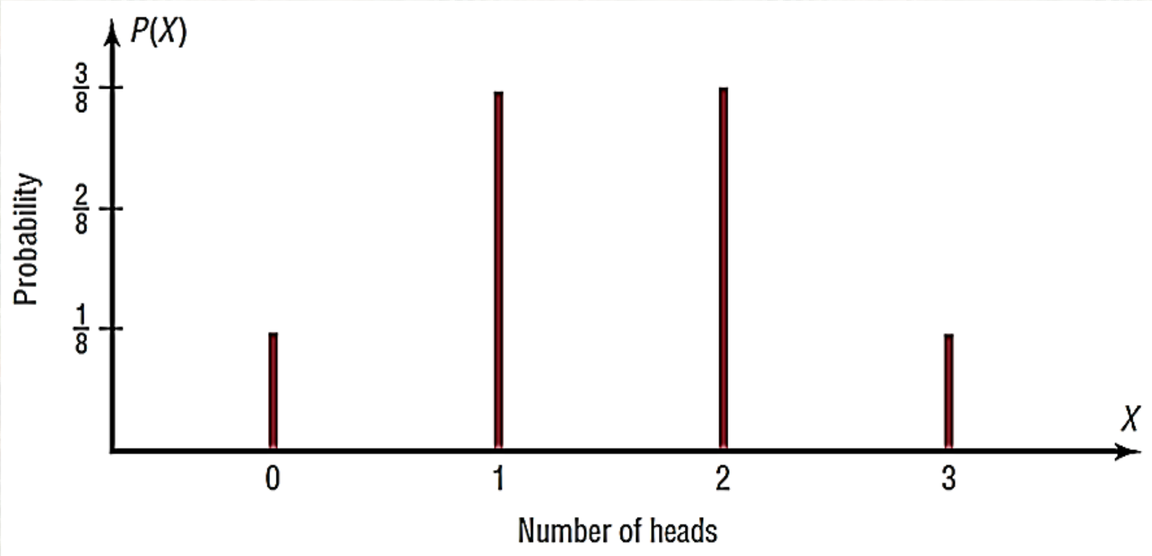
- **Discrete probability distributions** can be shown by using a graph or a table. Probability distributions can also be represented by a formula

Example 2: Represent graphically the probability distribution for Example 1.

Number of heads X	0	1	2	3
Probability P(X)	**1/8**	**3/8**	**3/8**	**1/8**

Solution:

The values that X assumes are located on the x axis, and the values for $P(X)$ are located on the y axis.



Note: Two Requirements for a Probability Distribution

1. The sum of the probabilities of all the events in the sample space must equal 1; that is, $\sum P(X) = 1$.
2. The probability of each event in the sample space must be between or equal to 0 and 1. That is, $0 \leq P(X) \leq 1$.

Example 3: Determine whether each distribution is a probability distribution.

a.

X	4	6	8	10
$P(X)$	-0.6	0.2	0.7	1.5

b.

X	1	2	3	4
$P(X)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

c.

X	8	9	12
$P(X)$	$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{6}$

d.

X	1	3	5	7	9
$P(X)$	0.3	0.1	0.2	0.4	-0.7

Solution:

a) No. It is not a probability distribution since P(X) cannot be negative or greater than 1. **b)** Yes. It is a probability distribution.

c) Yes. It is a probability distribution. **d)** No, since $P(X) \neq -0.7$.

2. *Mean, Variance, and Standard Deviation*

The mean, variance, and standard deviation for a probability distribution are computed differently from the mean, variance, and standard deviation for samples.

2.1. Mean

The mean of a random variable with a discrete probability distribution is:

$\mu = X_1 \,.\, P(X_1) + X_2 \,.\, P(X_2) + X_3 \,.\, P(X_3) + \ldots\ldots + X_n \,.\, P(X_n)$

$\quad = \Sigma X \,.\, P(X)$

Where: $X_1, X_2, X_3, \ldots, X_n$ are the outcomes and $P(X_1), P(X_2), P(X_3), \ldots, P(X_n)$ are the corresponding probabilities.

Note: $\Sigma X \,.\, P(X)$ means to sum the products.

Example 4: Find the mean of the number of spots that appear when a die is tossed.

Solution:

In the toss of a die, sample space is 1, 2, 3, 4, 5, 6; the mean can be computed thus.

Outcome X	1	2	3	4	5	6
Probability P(X)	1/6	1/6	1/6	1/6	1/6	1/6

$$\mu = \Sigma X \,.\, P(X) = 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = \frac{\mathbf{21}}{\mathbf{6}} = \mathbf{3.5}$$

Example 5: In a family with two children, find the mean of the number of children who will be girls.

Solution

The probability distribution is as follows:

Number of girls X	**0**	**1**	**2**
Probability P(X)	**1/4**	**½**	**1/4**

The mean is : $\mu = \Sigma X \cdot P(X) = 0.\frac{1}{4} + 1.\frac{1}{2} + 2 = 1$

2. Variance and Standard Deviation

- For a probability distribution, the mean of the random variable describes the measure of the so-called long-run or theoretical average, but it does not tell anything about the spread of the distribution. To measure this spread or variability, statisticians, the variance and standard deviation are used for this purpose.
- The formula for the variance of a probability distribution is: $\sigma^2 = \sum [X^2 . P(X)] - \mu^2$
- The standard deviation of a probability distribution is: $\sigma = \sqrt{\sigma^2}$ **or** $\sqrt{\sum [X^2 . P(X)] - \mu^2}$

Example 6: A box contains 5 balls. Two are numbered 3, one is numbered 4, and two are numbered 5. The balls are mixed and one is selected at random. After a ball is selected, its number is recorded. Then it is replaced. If the experiment is repeated many times, find the variance and standard deviation of the numbers on the balls.

Solution

Let X be the number on each ball. The probability distribution is

Number of ball X	**3**	**4**	**5**
Probability P(X)	**2/5**	**1/5**	**2/5**

The mean is : μ = Σ*X* . *P*(*X*) $= 3.\frac{2}{5} + 4.\frac{1}{5} + 5\,\frac{2}{5}$ **= 4**

The variance is: $\boldsymbol{\sigma^2 = \sum[X^2.P(X)] - \mu^2} \;=\; 3^2.\frac{2}{3} + 4^2.\frac{1}{5} + 5^2\,\frac{2}{5} - 4^2 = \frac{4}{5}$

The standard deviation is $\boldsymbol{\sigma = \sqrt{\sigma^2}}$

$$\boldsymbol{= \sqrt{\left(\frac{4}{5}\right)^2} = 0.894}$$

The mean, variance, and standard deviation can also be found by using vertical columns, as shown.

X	**P(X)**	**X.P(X)**	**X².P(X)**
3	**0.4**	**1.2**	**3.6**
4	**0.2**	**0.8**	**3.2**
5	**0.4**	**2.0**	**10**
ΣX.P(X) =		**4.0**	**16.8**

$\boldsymbol{\sigma = 0.894}$

Example 7: A talk radio station has four telephone lines. If the host is unable to talk (i.e., during a commercial) or is talking to a person, the other callers are placed on hold. When all lines are in use, others who are trying to call in get a busy signal. The probability that 0, 1, 2, 3, or 4 people will get through is shown in the distribution. Find the variance and standard deviation for the distribution.

X	**0**	**1**	**2**	**3**	**4**
P(X)	**0.18**	**0.34**	**0.23**	**0.21**	**0.04**

Solution

The mean is $\mu = \Sigma X \,.\, P(X) = 1.6$ The variance is $\sigma^2 = \sum \left[X^2 . P(X)\right] - \mu^2 = 1.23$

The standard deviation is: $\sigma = \sqrt{\sigma^2} = \sqrt{1.23} = 1.1$

3. The Binomial Distribution (BD)

BD is a probability experiment that satisfies the following four requirements:

1. There must be a fixed number of trials.
2. Each trial can have only two outcomes or outcomes that can be reduced to two outcomes. These outcomes can be considered as either success or failure.
3. The outcomes of each trial must be independent of one another.
4. The probability of a success must remain the same for each trial.

Binomial Probability Formula

In a binomial experiment, the probability of exactly X successes in n trials is:

$$P(X) = nCx\, P^{x} . q^{n-X} = \frac{n!}{(n-X)!\, X!} . P^{X} . q^{n-X}$$

$P(S)$ The symbol for the probability of success; $P(F)$ The symbol for the probability of failure; p The numerical probability of a success; q The numerical probability of a failure.

$P(S) = p$ and $P(F) = 1 - p = q$; n The number of trials; X The number of successes in n trials; Note that 0 X n and X 0, 1, 2, 3, . . . , n.

Example 8: A coin is tossed 3 times. Find the probability of getting exactly two heads.

Solution

This problem can be solved by looking at the sample space. There are three ways to get two heads. [HHH, HHT, HTH, THH, TTH, THT, HTT, TTT]

The answer is: 1/8+1/8+1/8 = 3/8 = 0.375

In this case, $n = 3$, $X = 2$, $p = q = 1/2$

$$P(X) = \frac{n!}{(n-X)!\, X!} . P^{X} . q^{n-X}$$

$$P(X) = \frac{3!}{(3-2)!\,2!} . \left(\frac{1}{2}\right)^{2} . \left(\frac{1}{2}\right)^{1} = 0.375$$

Example 9: A survey found that one out of five Americans say he or she has visited a doctor in any given month. If 10 people are selected at random, find the probability that exactly 3 will have visited a doctor last month.

Solution

In this case, n = 10, X = 3, $p = 1/5$, and $q = 4/5$.

$$P(3) = \frac{10!}{(10-3)!3!}\left(\frac{1}{5}\right)^3\left(\frac{4}{5}\right)^7 = 0.201$$

Example 10: A survey from Teenage Research Unlimited (Northbrook, Illinois) found that 30% of teenage consumers receive their spending money from part-time jobs. If 5 teenagers are selected at random, find the probability that at least 3 of them will have part-time jobs.

Solution

To find the probability that at least 3 have part-time jobs, it is necessary to find the individual probabilities for 3, or 4, or 5 and then add them to get the total probability.

$$P(3) = \frac{5!}{(5-3)!3!}.(0.3)^3.(0.7)^2 = 0.132$$

$$P(4) = \frac{5!}{(5-4)!4!}.(0.3)^4.(0.7)^1 = 0.028$$

$$P(5) = \frac{5!}{(5-5)!5!}.(0.3)^5.(0.7)^0 = 0.002$$

***P*(at least three teenagers have part-time jobs) = 0.132 + 0.028 + 0.002 = 0.162**

- **Mean, Variance, and Standard Deviation for the Binomial Distribution**

The mean, variance, and standard deviation of a variable that has the *binomial distribution* can be found by using the following formulas.

Mean: $\mu = n \cdot p$ Variance: $\sigma^2 = n \cdot p \cdot q$ Standard deviation: $\sigma = \sqrt{npq}$

Example 11: A die is rolled 480 times. Find the mean, variance, and standard deviation of the number of 3 that will be rolled.

Solution

This is a binomial experiment since getting a 3 is a success and not getting a 3 is considered a failure.

n = 480, p = 1/6, q= 5/6.

Mean: $\mu = n \cdot P = 480 \times 1/6 = 80$

Variance: $\sigma^2 = n \cdot p \cdot q = 480 \times 1/6 \times 5/6 = 66.67$

Standard deviation: $\sigma = \sqrt{npq}$ **= 8.16**

4. *Other Types of Distributions*

❖ The Poisson Distribution

The probability of *X* occurrences in an interval of time, volume, area, etc., for a variable where **λ** (Greek letter lambda) is the mean number of occurrences per unit (time, volume, area, etc.) is:

$$P(X;\ \lambda) = \frac{e^{-\lambda}\lambda^X}{X!}$$

where:

X=0, 1, 2, …; e = 2.718

Example 12: If there are 200 typographical errors randomly distributed in a 500-page manuscript, find the probability that a given page contains exactly 3 errors.

Solution

First, find the mean number λ of errors. Since there are 200 errors distributed over 500 pages, each page has an average of:

$$\lambda = \frac{200}{500} = 0.4 \qquad \boldsymbol{P(X;\ \lambda) = \frac{e^{-\lambda}\lambda^{X}}{X!}}$$

$$\boldsymbol{X = 3;} \quad \Rightarrow \quad \boldsymbol{P(3;\ 0.4) = \frac{e^{-0.4}(0.4)^{3}}{3!} = 0.0072}$$

Chapter Six

Continuous Probability Distributions

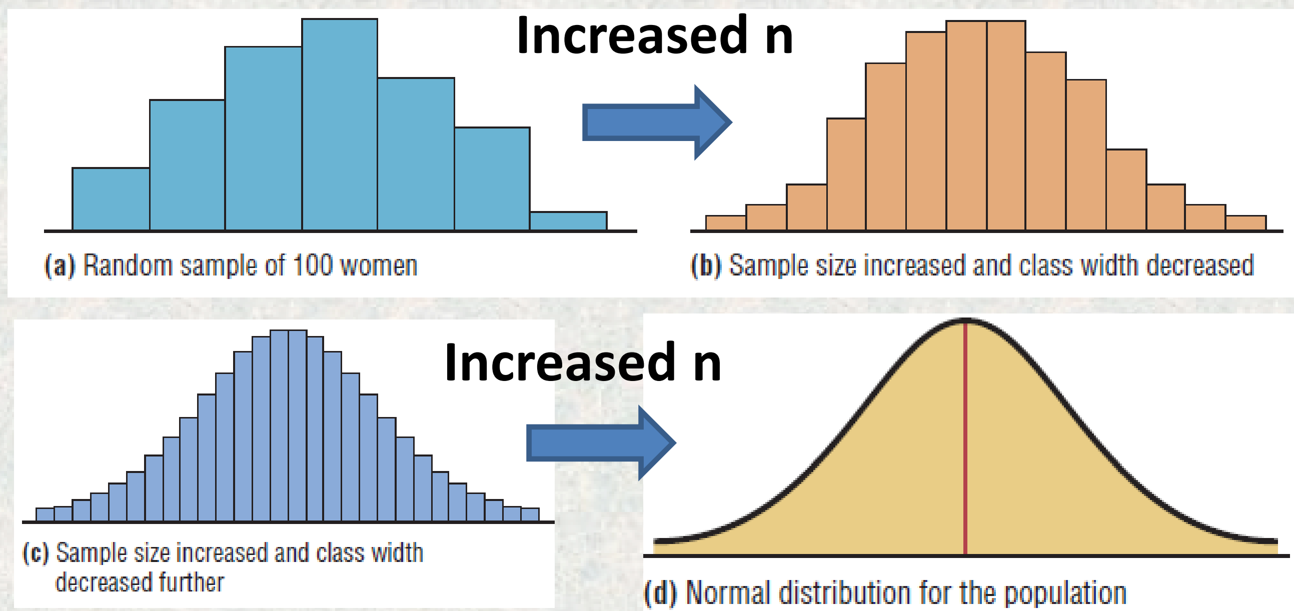
The Normal Distribution

1. Normal Distributions
2. Applications of the Normal Distribution
3. The Central Limit Theorem
4. The Normal Approximation to the Binomial Distribution

Assi. Prof. Dr. Taher M. Ahmed
Civil Engineering Department
University of Anbar

Chapter Six
Continuous Probability Distributions
Normal Distribution

- **Continuous variable** are variables that can assume to take all values between any two given values of the variables. For examples: the heights of adult men, body temperatures of rats, ground water level, and cholesterol levels of adults.
- The distributions shape takes the bell-shaped, and these are called approximately normally distributed variables. These variables approach from normal distribution as sample size increases.
- Normal distribution is known as bell curve or a Gaussian distribution, named for the German mathematician Carl Friedrich Gauss (1777–1855), who derived its equation.



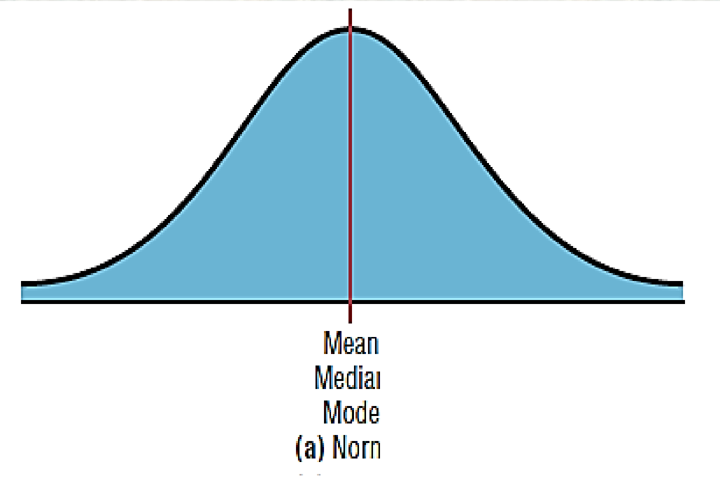
(a) Random sample of 100 women

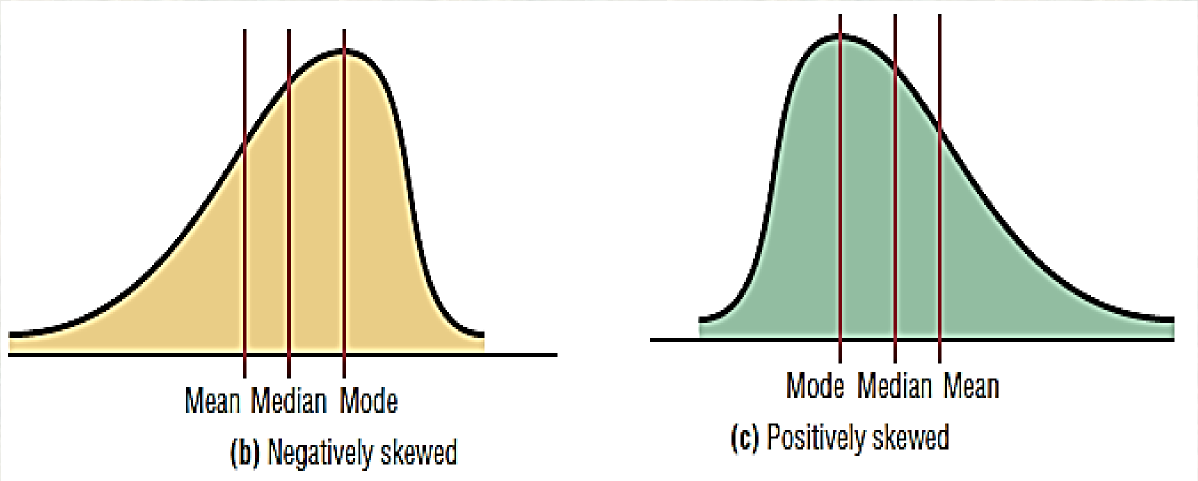
(b) Sample size increased and class width decreased

(c) Sample size increased and class width decreased further

(d) Normal distribution for the population

- When the data values are distributed about the mean, a distribution is said to be a **symmetric distribution**. While, when the majority of the data values fall to the left or right of the mean, the distribution is said to be skewed.





1. Normal Distribution

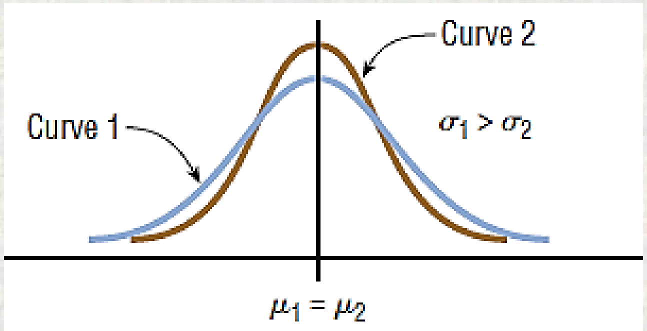
- In mathematics, curves can be represented by equations. For example, the equation of the circle, ellipse, straight motions, and so on.
- In a similar manner, the theoretical curve, called a *normal distribution curve,* can be used to study many variables that are not perfectly normally distributed but are nevertheless approximately normal.
- **A normal distribution** is a continuous, symmetric, bell-shaped distribution of a variable.

- The mathematical equation for a normal distribution is:

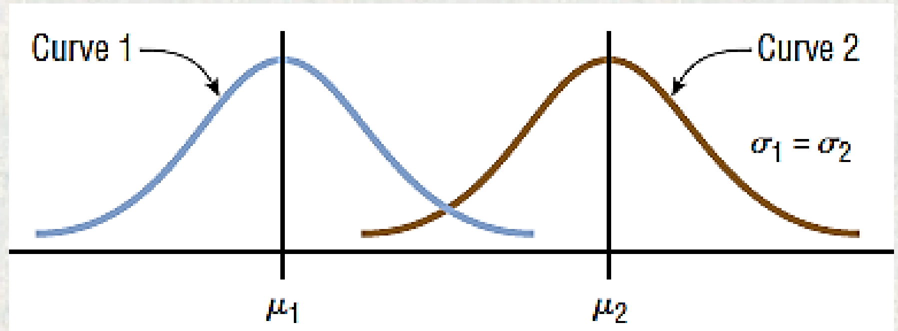
$$y = \frac{e^{-(X-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

Where: *e* ≈ 2.718 (means ≈ is approximately equal to”)
μ is population mean, σ is population standard deviation

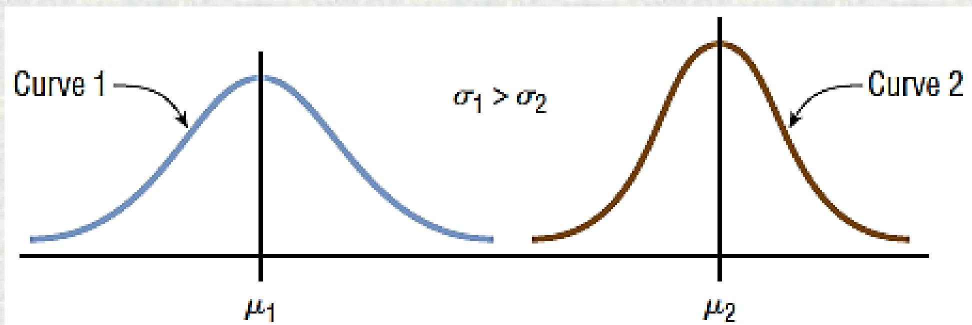
- The shape and position of a normal distribution curve depend on two parameters, the *mean* and the *standard deviation.*



(a) Same means but different standard deviations



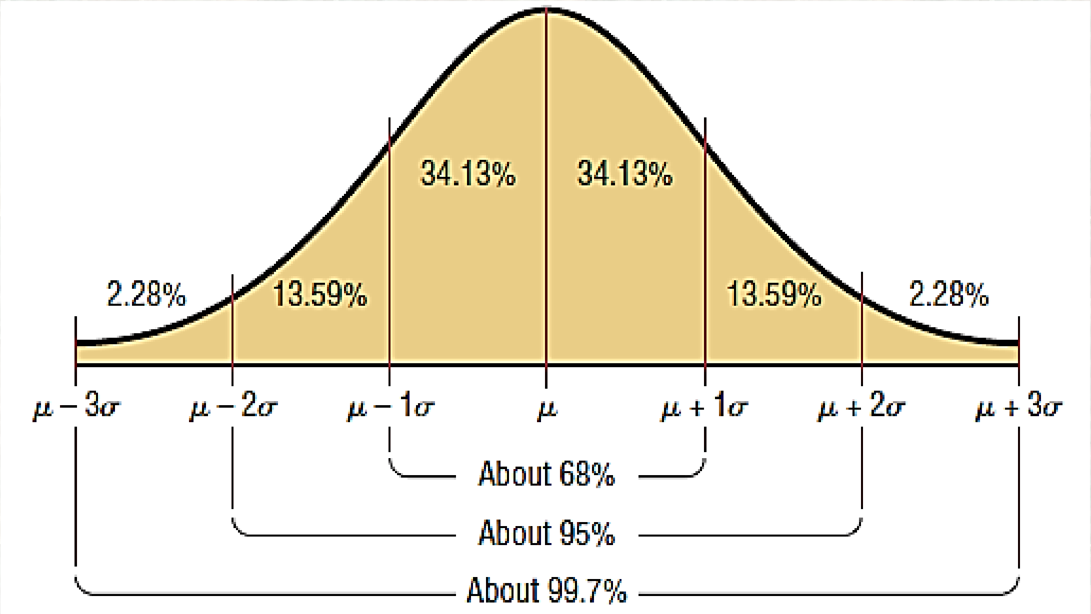
(b) Different means but same standard deviations



(c) Different means and different standard deviations

Summary of the Properties of the Theoretical Normal Distribution

1. A normal distribution curve is bell-shaped.
2. The mean, median, and mode are equal and are located at the center of the distribution.
3. The curve is symmetric about the mean, which is equivalent to saying that its shape is the same on both sides of a vertical line passing through the center.
4. The curve is continuous.
5. The curve never touches the *x* axis. The total area under a normal distribution curve is equal to 1.00, or 100%.
6. The area under the part of a normal curve that lies within 1 standard deviation of the mean is approximately 0.68, or 68%; within 2 standard deviations, about 0.95, or 95%; and within 3 standard deviations, about 0.997, or 99.7%.



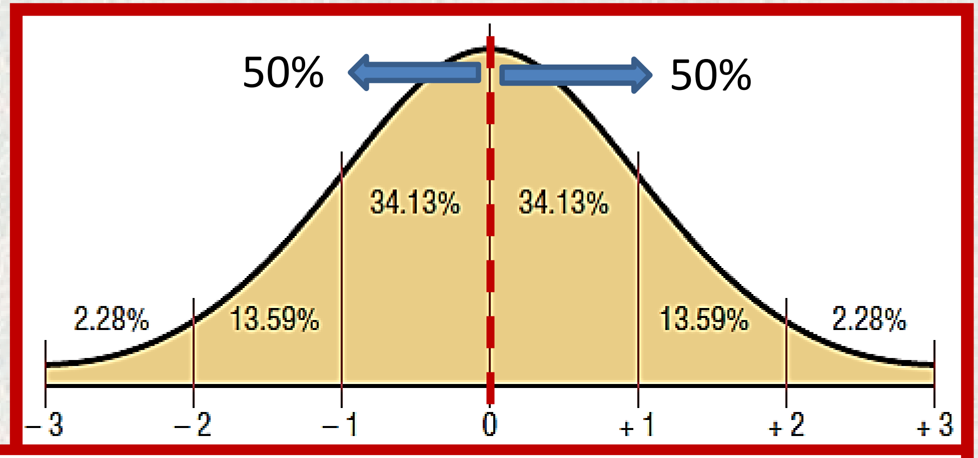
1.1. The Standard Normal Distribution

- Since each normally distributed variable has its own mean and standard deviation. So, the shape and location of these curves will vary. In practical applications, statisticians used what is called the *standard normal distribution.* Then, a table can be used to determine the area under the curve for each variable.
- The **standard normal distribution** is a normal distribution with a mean of 0 and a standard deviation of 1.
- The formula for the standard normal distribution is

$$y = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

- All normally distributed variables can be transformed into the standard normally distributed variable using the formula for the standard score:

$$z = \frac{X - \mu}{\sigma}$$



The values under the curve indicate the proportion of area in each section. For example, the area between the mean and 1 standard deviation above or below the mean is about 0.3413, or 34.13%.

1.2. Finding Areas Under the Standard Normal Distribution Curve

The area under a normal distribution curve is used to finding the probability of the continuous variables for any range would be found. A two-step process is recommended with the use of the Procedure Table shown.

Step 1: Draw the normal distribution curve and shade the area.

Step 2: Find the appropriate figure in the Procedure Table.

Example: Z = 1.39

Z	0.00	....	0.09
0.0			
⋮			
1.3			$\boxed{0.9177}$
⋮	Area = 0.9177		

Table of Z for determine the area under the curve.

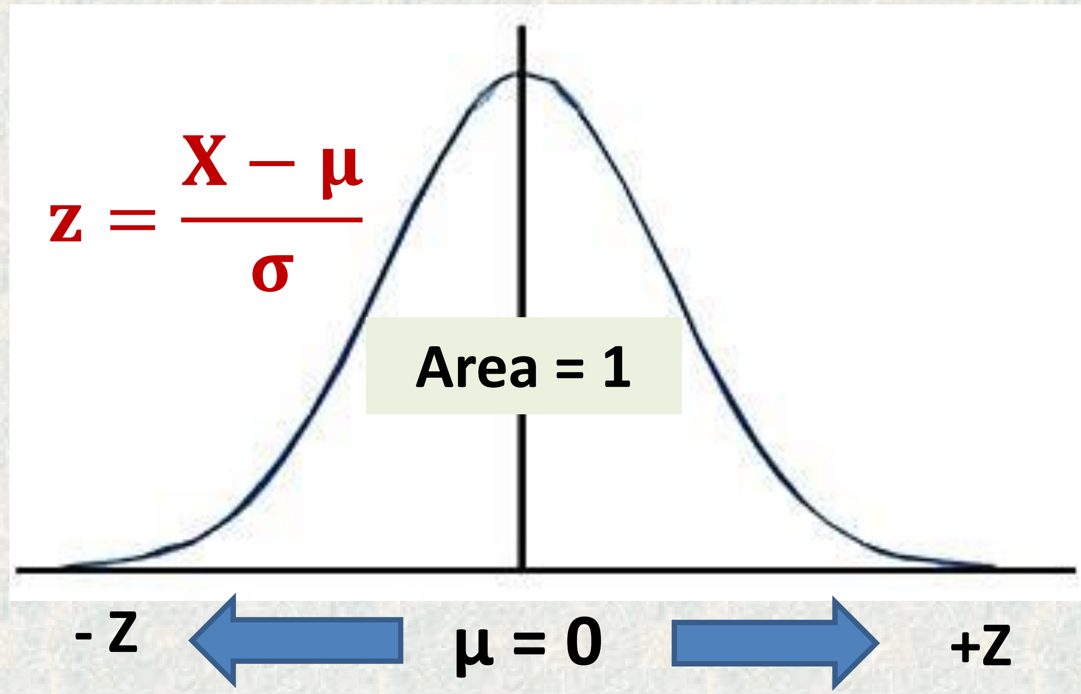


Table E The Standard Normal Distribution

Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

For z values less than −3.49, use 0.0001.

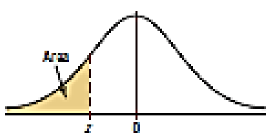
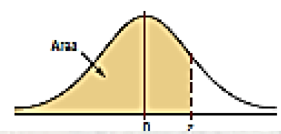


Table E (continued)

Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

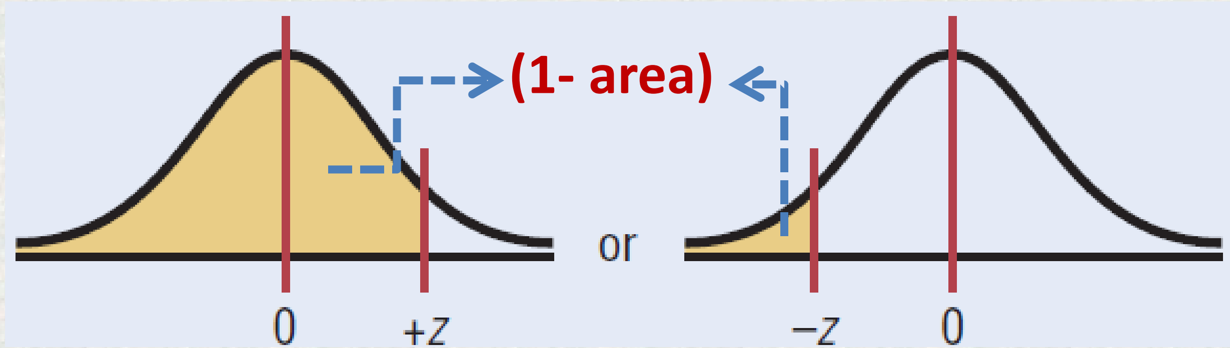
For z values greater than 3.49, use 0.9999.



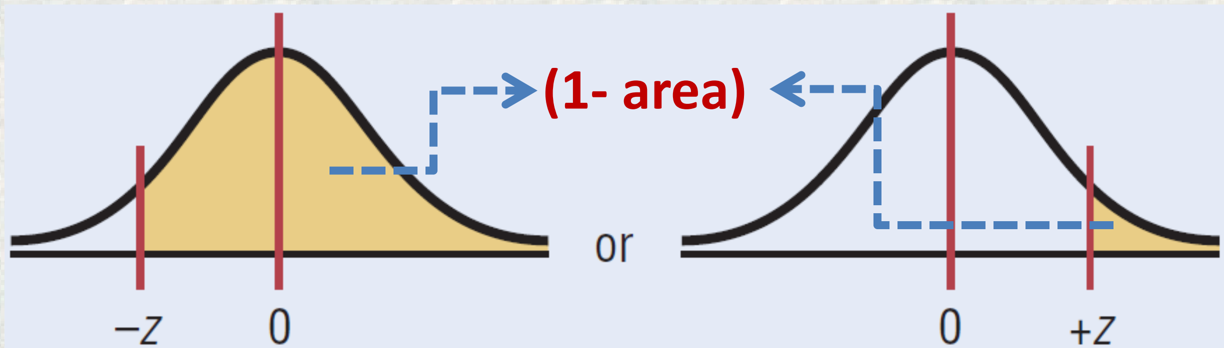
1.3. Procedure Table

Finding the Area Under the Standard Normal Distribution Curve

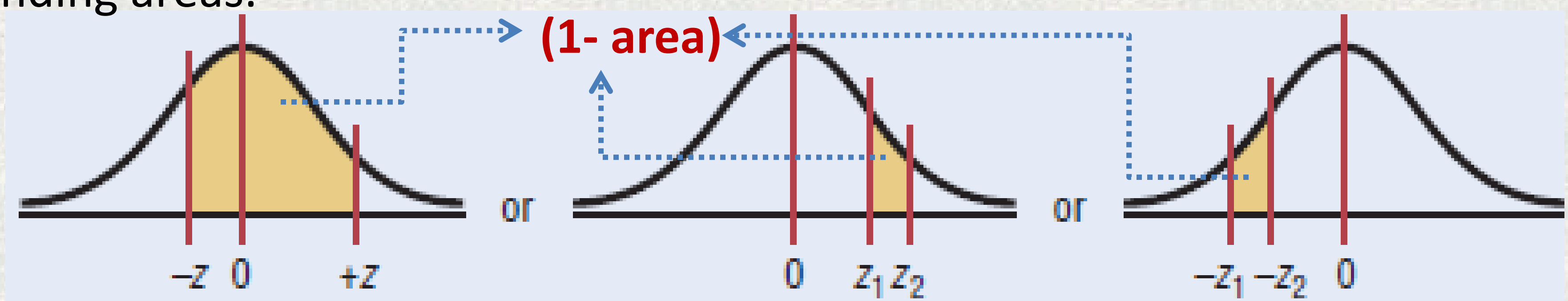
1. To the left of any z value: Look up the z value in the table and use the area given.



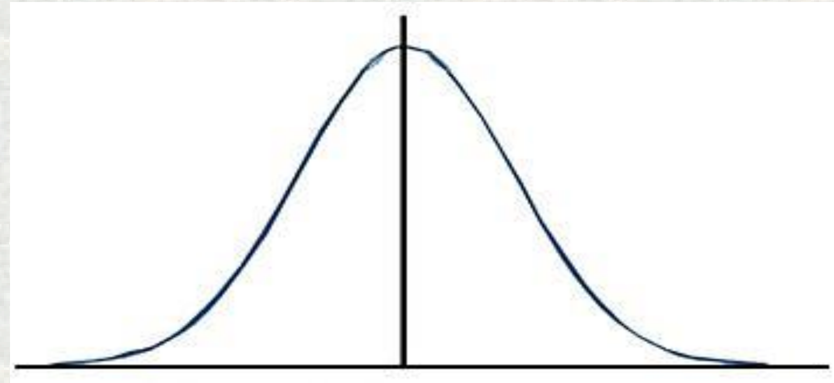
2. To the right of any z value: Look up the z value and subtract the area from 1. **(1- area)**



3. Between any two z values: Look up both z values and subtract the corresponding areas.



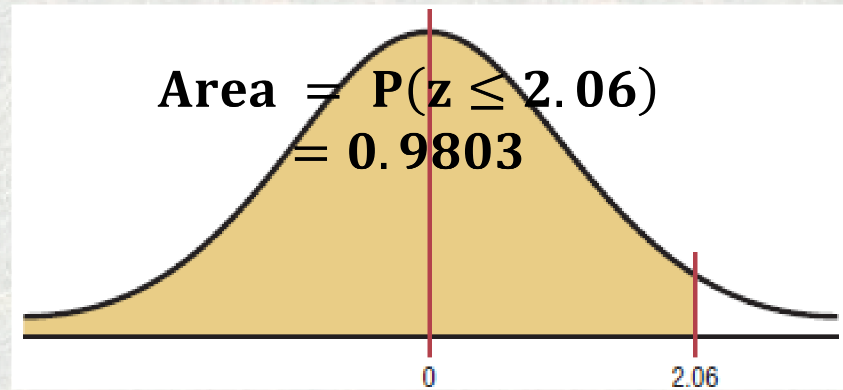
Example 1: Find the area to the left of z = 2.06.



Solution

Step 1 Draw the figure. The desired area is shown in the figure below.

Step 2 We are looking for the area under the standard normal distribution to the left of z = 2.06. Since this is an example of **the first case**, look up the area in the table. It is **0.9803.** Hence, **98.03%** of the area is less than ***z* = 2.06**.



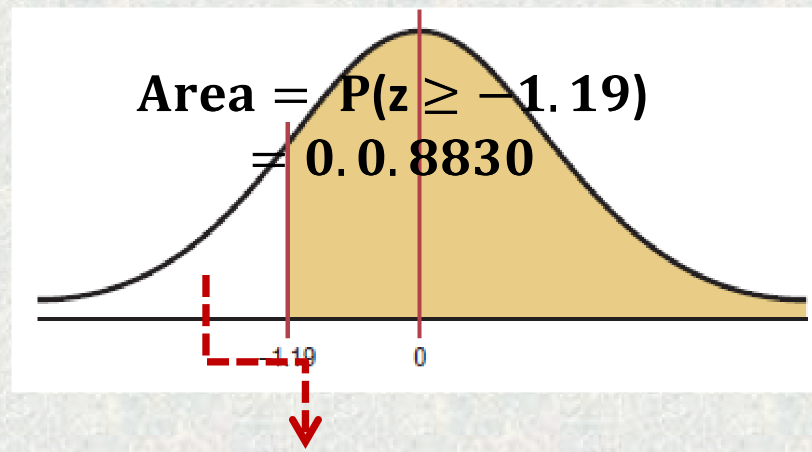
Example 1: Find the area to the right of z = -1.19.

Solution

Step 1 Draw the figure. The desired area is shown in the figure.

Step 2 This is an example of the **second case**. Look up the area for z = -1.19. It is 0.1170. Subtract it from 1.0000.

1.0000 - 0.1170 = 0.8830. (88.30%)

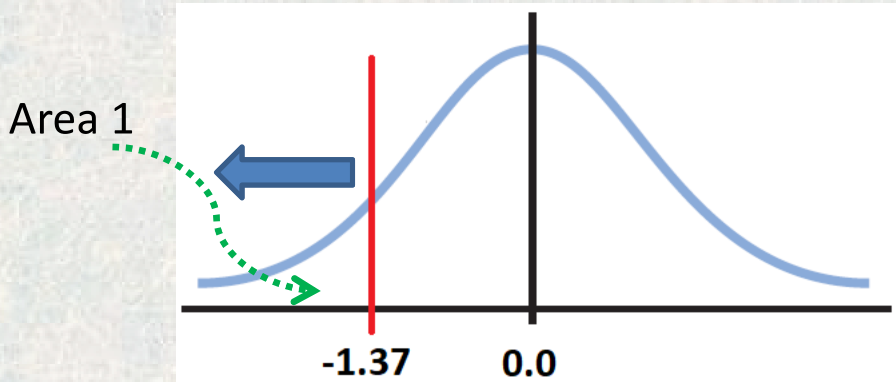


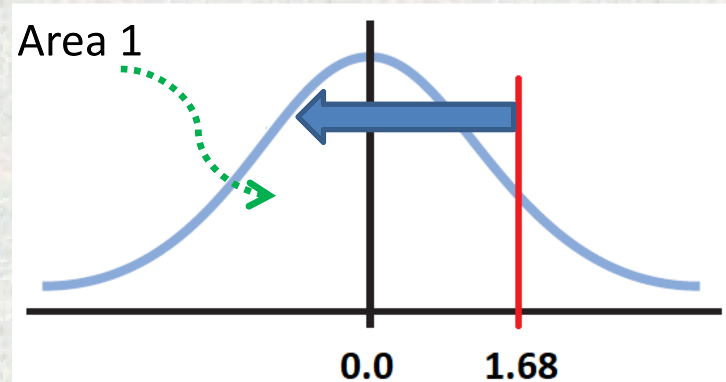
the area for ($z \leq$ -1.19) = 0.1170

Example 3: Find the area between *z* = + 1.68 and *z* = - 1.37.

Solution

This is case 3. Draw the figure as shown. The desired area is shown in the figures below.





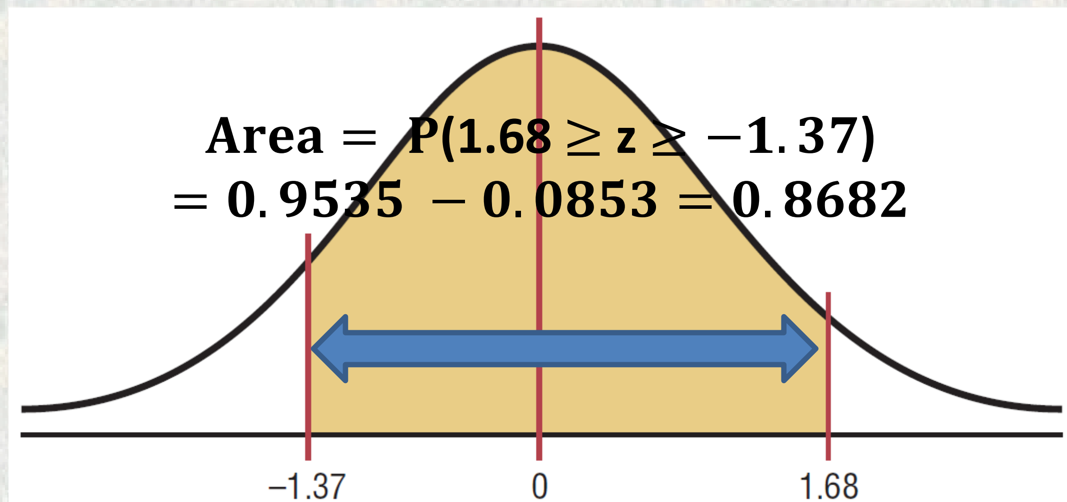
for the small area z = -1.37,
from table area = 0.0853

for the large area z = 1.68,
from table area = 0.9535

The area between the two *z* values is:

= 0.9535 - 0.0853

= 0.8682 or 86.82%.



1.4. A Normal Distribution Curve as a Probability Distribution Curve

The area under the standard normal distribution curve can be used for calculation the probability for any continuous random variable.

Example 4: Find the probability for each. a)$P(0 < z < 2.32)$; b) $P(z < 1.65)$; c) $P(z > 1.91)$

Solution

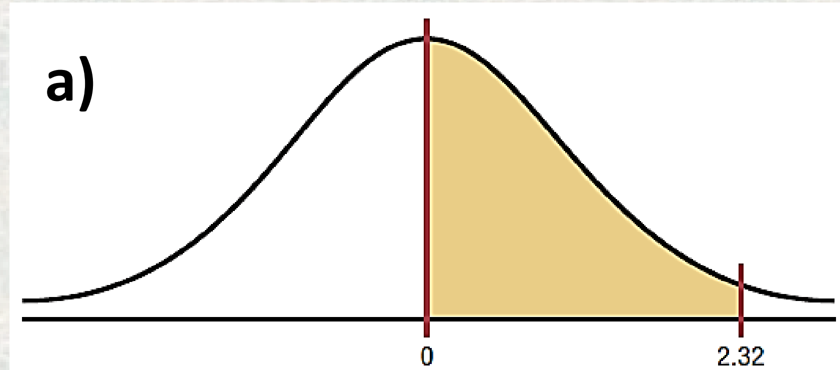
a)P(0 < z < 2.32) = P(z< 2.32) – P(z<0) OR area to the left of(z = 2.32) – area to the left of (z = 0)

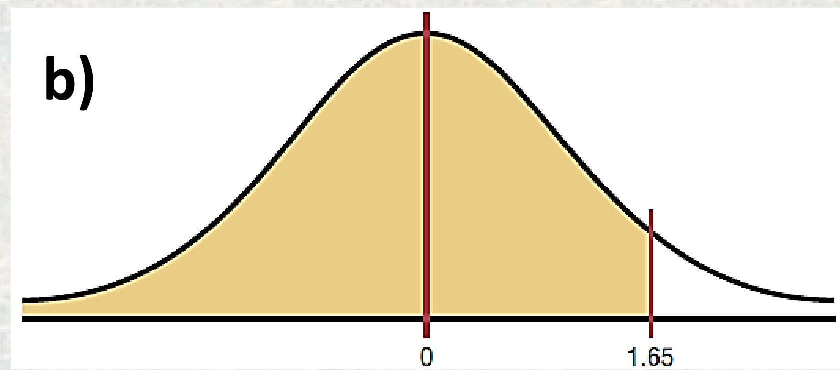
P (z < 2.32) = the area from the table = 0.9898

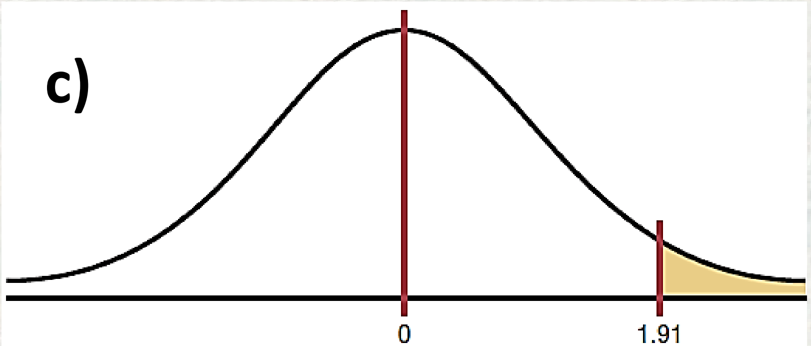
P (z< 0.0) = the area from the table = 0.5000.

❖ P(0 < z < 2.32) = 0.9898 – 0.500 = 0.4898.

b) $P(z < 1.65)$ is the area from the table to the left of Z = 1.65. $P(z < 1.6)$ = 0.9505

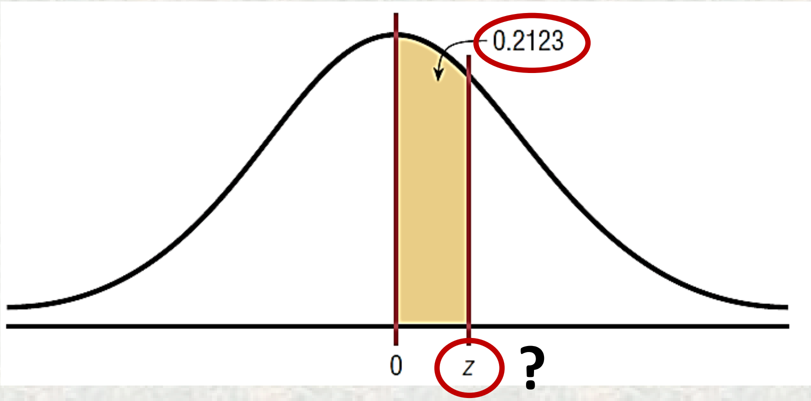






c) P(z > 1.91) = 1 – P (z < 1.91)
= 1 – the area to the left of 1.91
= 1 – 0.9719 = 0.0281, or 2.81%.

Example 5: Find the z value such that the area under the standard normal distribution curve between 0 and the z value is 0.2123.

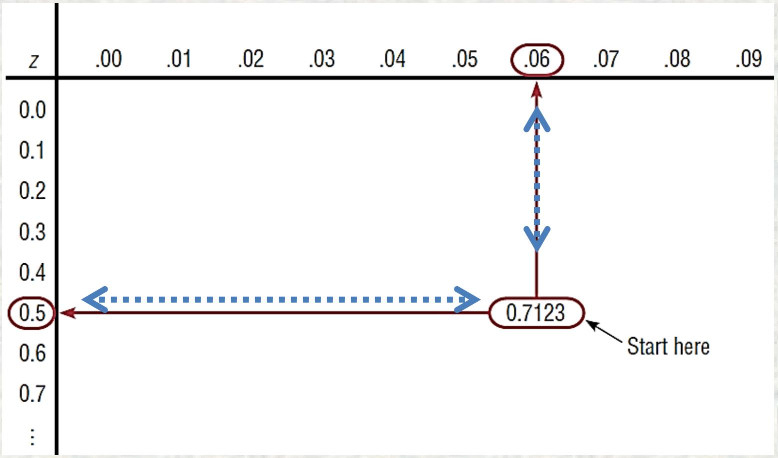


Solution

Draw the figure. The area is shown in the figure.

The total area = the area of (z = 0) + 0.2123

= 0.5000 + 0.2123 = 0.7123



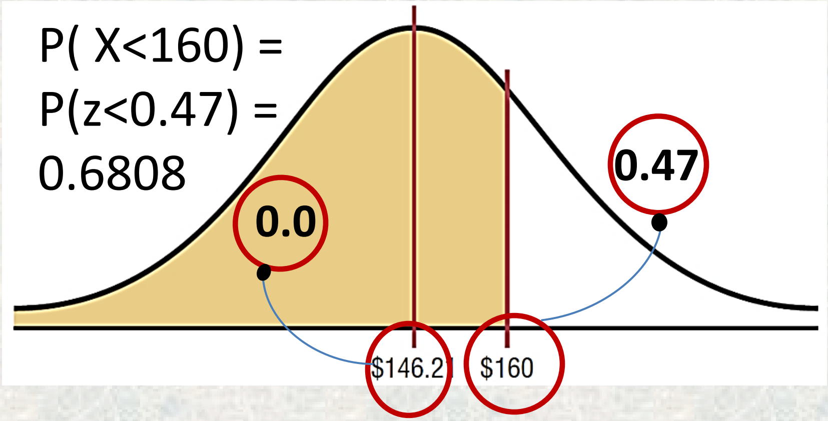
From the table. The value in the left column is 0.5, and the top value is 0.06. Add these two values to get ***z* = 0.56**.

2. Applications of the Normal Distribution

The standard normal distribution curve can be used to solve a wide variety of practical problems. To solve problems by using the standard normal distribution, transform the original variable to a standard normal distribution variable (Z) using the formula:

$$z = \frac{X - \mu}{\sigma}$$

Example 6: A survey found that women spend on average $146.21 on beauty products during the summer months. Assume the standard deviation is $29.44. Find the percentage (Probability) of women who spend less than $160.00. Assume the variable is normally distributed.



Solution

Draw the figure and represent the area as shown in the figure. Then Find the *z* value corresponding to $160.00.

$$z = \frac{X-\mu}{\sigma} = \frac{160.00-146.21}{29.44} = 0.47$$

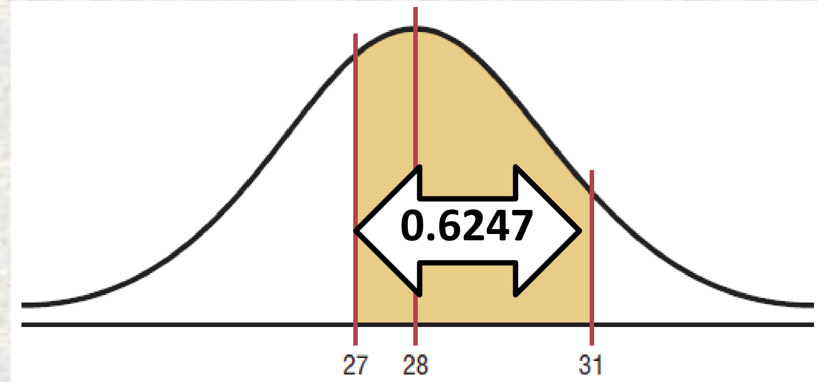
From the table: P(X<160) = P(z<0.47) = area to the left of z = 0.6808, or 68.08% (percent of the women spend less than $160.00 on beauty products).

Example 7: Each month, an American household generates an average of 28 pounds of newspaper for garbage or recycling. Assume the standard deviation is 2 pounds. If a household is selected at random, find the probability of its generating.

a) Between 27 and 31 pounds per month;
b) More than 30.2 pounds per month.

Solution

a) Draw the figure and represent the area. Then find the two *z* values.



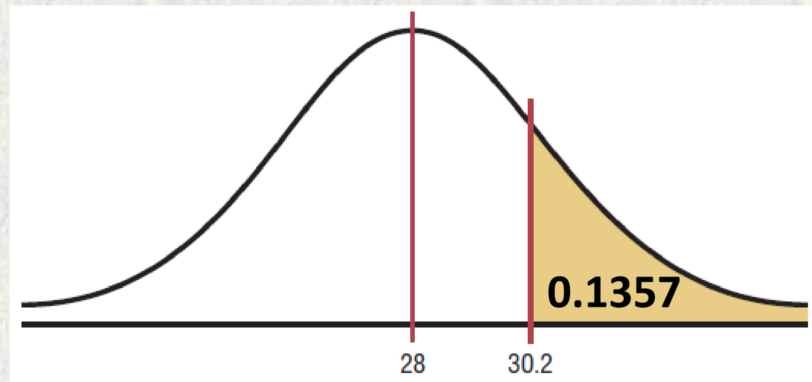
P (27<X<31) = P(-0.5<z<1.5)

$z_1 = \frac{X_1 - \mu}{\sigma} = \frac{27-28}{2}$ = - 0.5 ⇨ Area_1 = 0.3085

$z_1 = \frac{X_2 - \mu}{\sigma} = \frac{31-28}{2}$ = 1.5 ⇨ Area_2 = 0.9332

P(27<X<31) = P(-0.5<z<1.5) = 0.9332 – 0.3085 = 0.6247

b) Draw the figure and represent the area. Then find the two z values.



P(X>30.2)= P(Z>z_1)= 1–P(X<30.2) = 1 – P(Z<z_1)

$z_1 = \frac{X - \mu}{\sigma} = \frac{30.2 - 28}{2}$ = 1.1 ⇨ Area_1 = 0.8643

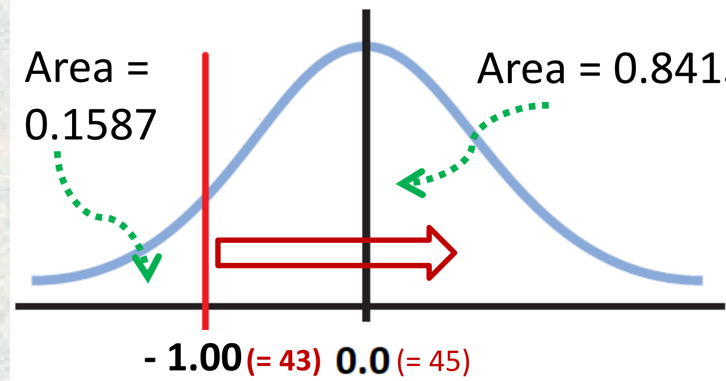
P (X > 30.2) = 1 – P(X < 30.2) = 1 – P(Z<z_1) = 1 -0.8643 = 0.1357 or 13.57%

Example 8: A steel factory produces deformed bars with average yield force 45 kN and standard deviation 2 kN, if a bare has been tested, determine the probability of ? (a) strength force ≥ 43 kN; (b) strength force ≤ 47 kN; and (c) strength force between 44 to 46 kN.

Solution

(a) $P(X \geq 43) = P\left(\left(z \geq \frac{X-\mu}{\sigma}\right) = 1 - P\left(\left(z < \frac{X-\mu}{\sigma}\right) =\right.\right.$ area to the right

$z = \frac{43-45}{2} = -1.00$ ⇨ Area to the left = 0.1587

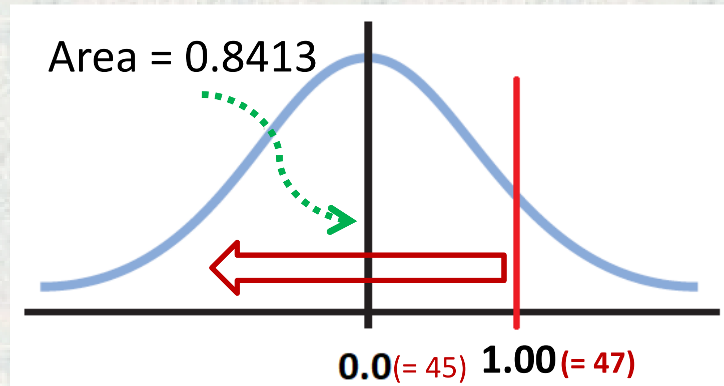


$P(X \geq 43) = P\left(\left(z \geq \frac{X-\mu}{\sigma}\right) = 1 - P\left(\left(\left(z < \frac{X-\mu}{\sigma}\right)\right) = 1 - 0.1587 = 0.8413\right.\right.$

(b) $P(X \leq 47) = P\left(\left(z \leq \frac{X-\mu}{\sigma}\right)\right.$

$z = \frac{47-45}{2} = 1.00$ ⇨ Area to the right = 0.8413

$P(X \leq 47) = P\left(\left(z \leq \frac{X-\mu}{\sigma}\right) =\right.$ 0.8413 or 84.13 %



(c) $P(44 \leq X \leq 47) = P((z_1 \leq Z \leq z_2)$

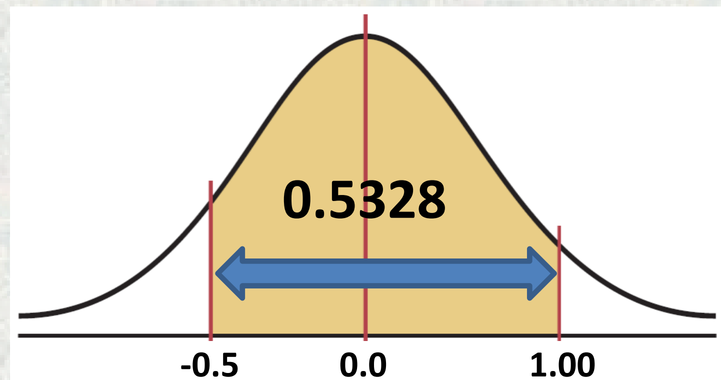
$z_1 = \frac{44-45}{2} = -0.500$ ⇨ Area to the left = 0.3085

$z_2 = \frac{47-45}{2} = 1.00$ ⇨ Area to the left = 0.8413

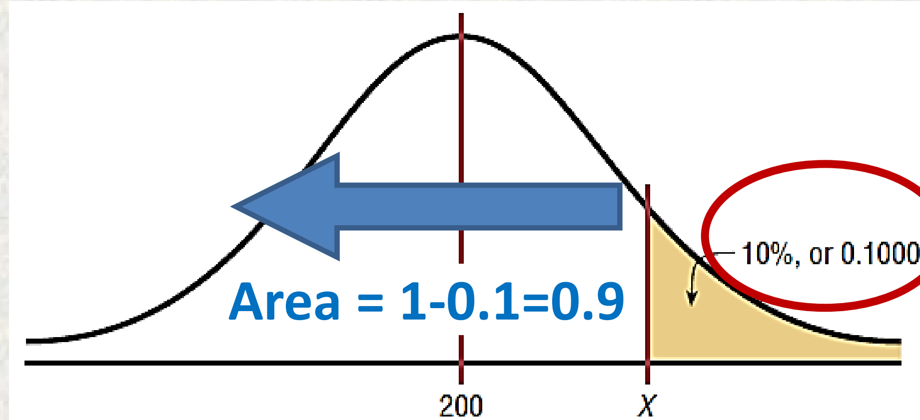
$P(44 \leq X \leq 47) = P((z_1 \leq Z \leq z_2)$

$= 0.8413 - 0.3085$

$= 0.5328$ OR 53.28 %



Example 9: To qualify of a steel factory quality, a tensile strength must score in the top 10% on a general test. The tensile mean is 200 and a standard deviation of 20. Find the lowest possible tensile strength to qualify. Assume the test scores are normally distributed.



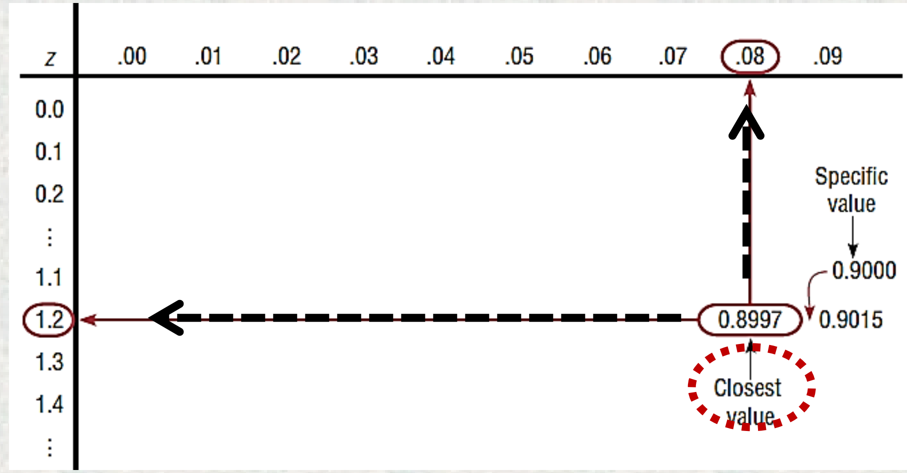
Solution

- Since the test scores are normally distributed, the area to the right test value X is 10% (0.1).
- The area to the left of X = 1 – 0.1 = 0.9
- From table the Z values that corresponding to area 0.9 ≈ 1.28

$$z = \frac{X-\mu}{\sigma} \Rightarrow 1.28 = \frac{X-200}{20} \Rightarrow X = 226$$

- **When you must find the value of *X*, you can use the following formula:**

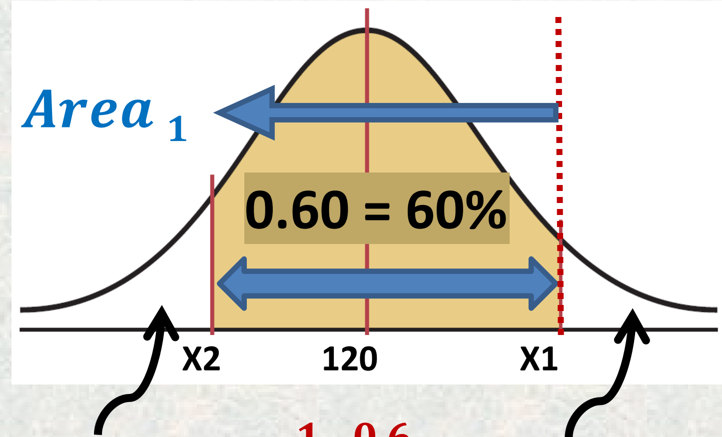
$$X = z.\sigma + \mu$$



Example 10: An engineering in PVC pipe factory wishes to select a pipe bearing pressure in the middle 60%. If the mean hydraulic pressure is 120 and the standard deviation is 8, find the upper and lower pressure that meet the requirement.

Solution

- The two values (X1 and X2) must be determined based on the area to the left side of each values



- From Table; $Area_2 = 0.2$, $z_2 = -0.84$
- From Table; $Area_1 = 0.8$, $z_1 = 0.84$

$$Area_2 = \frac{1-0.6}{2} = 0.2 = 20\%$$

$$Area_1 = 0.6 + 0.2 = 0.8$$

$$X = z.\sigma + \mu$$

$$X_1 = z_1.\sigma + \mu \Rightarrow \quad X_1 = 0.84*8 + 120 = 126.72$$

$$X_2 = z_2.\sigma + \mu \Rightarrow \quad X_2 = -0.84 \times 8 + 120 = 113.28$$

Therefore, the middle 60% will have pressure readings of: **113.28 ≤ *X* ≤ 126.72.**

3. Determining Normality

The distribution is being normally or approximately normally shaped:

- The easiest way is to draw a histogram for the data and check its shape. If the histogram is not approximately bell shaped, then the data are not normally distributed.
- Skewness coefficient (Pearson's index (PC)) $PC = \frac{3(\bar{X} - M_e)}{S}$

The Normality distribution : -1 ≤ PC ≤ +1

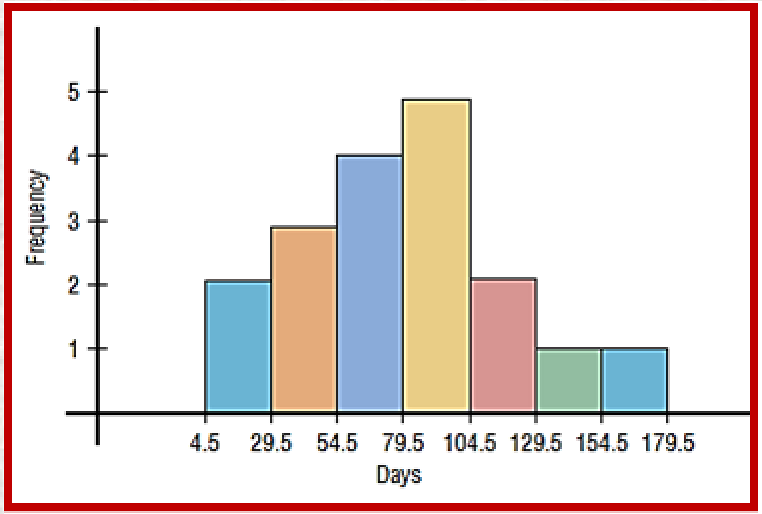
Example 11: A survey of 18 high-technology firms showed the number of days' inventory they had on hand. Determine if the data are approximately normally distributed.

5	29	34	44	45	63	68	74	74
81	88	91	97	98	113	118	151	158

Solution

- Construct a frequency distribution table and draw a histogram for the data.
- The histogram is approximately bell-shaped, we can say that the distribution is approximately normal.

Class	Frequency
5–29	2
30–54	3
55–79	4
80–104	5
105–129	2
130–154	1
155–179	1



Using PC to check the normality (average = 79.5, median = 77.5, and S = 40.5)

$PC = \frac{3(\bar{X} - M_e)}{S} = \frac{3(79.5 - 77.5)}{40.5}$ = 0.148 within -1 ≤ PC ≤ +1, it is normal distribution

End of Chapter Six

Thank you

Chapter Seven
Confidence Intervals and Sample Size

1. Preface
2. Confidence Intervals for the Mean When σ is Known
3. Confidence Intervals for the Mean When σ is Unknown
4. Confidence Intervals and Sample Size for Proportions
5. Confidence Intervals for Variances and Standard Deviations

Assi. Prof. Dr. Taher M. Ahmed
Civil Engineering Department
University of Anbar

Chapter Six

Confidence Intervals and Sample Size

1. Preface:

- A survey by the Roper Organization found that 45% of the people who were offended by a television program would change the channel, while 15% would turn off their television sets. The survey further stated that the margin of error is 3 percentage points, and 4000 adults were interviewed. Several questions arise:

1. How do these estimates compare with the true population percentages?
2. What is meant by a margin of error of 3 percentage points?
3. Is the sample of 4000 large enough to represent the population of all adults who watch television in the United States?

- Inferential statistical techniques have various assumptions that must be met before valid conclusions can be obtained.

1. The samples must be randomly selected.
2. The sample size must be greater than or equal to 30 or less.
3. The population must be normally or approximately normally distributed based on sample size.

2. Confidence Intervals for the Mean When σ is Known

2.1. A point estimate is a specific numerical value estimate of a parameter. The best point estimate of the population mean "μ" is the sample mean "$\bar{X}$".

Example: The president of university want to estimate the average age of the student **(μ)**. He could select a random sample of 100 students and find the average age **($\bar{X}$)** of these students, say, 22.3 years. From the sample mean, the president could infer that the average age of all the students is 22.3 years. So $\bar{X}$ is **estimator** for the population **(μ).**

A good estimator should satisfy the following criteria:

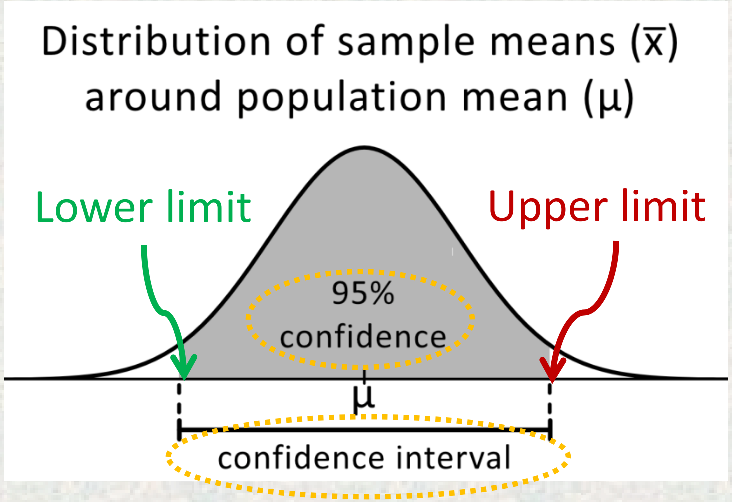
1. The estimator should be an **unbiased estimator.** That is, the expected value or the mean of the estimates obtained from samples of a given size is equal to the parameter being estimated.
2. The estimator should be consistent. For a **consistent estimator,** as sample size increases, the value of the estimator approaches the value of the parameter estimated.
3. The estimator should be a **relatively efficient estimator.** That is, of all the statistics that can be used to estimate a parameter, the relatively efficient estimator has the smallest variance

2.2. An interval estimate of a parameter is an interval or a range of values used to estimate the parameter. This estimate may or may not contain the value of the parameter being estimated.

- In an interval estimate, the parameter is specified as being between two values. For example, an interval estimate for the average age of all students might be $21.9 \le \mu \le 22.7$, or 22.3 ± 0.4 years.

2.3. Confidence Intervals

- **A confidence interval** is a specific interval estimate of a parameter determined by using data obtained from a sample and by using the specific confidence level of the estimate.



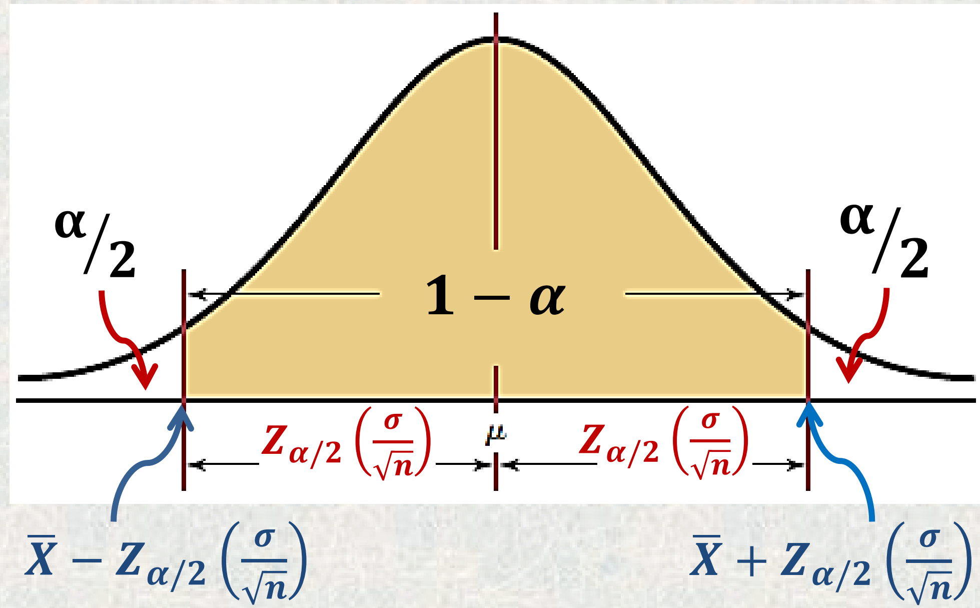
- The **confidence level** of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter, assuming that a large number of samples are selected and that the estimation process on the same parameter is repeated.
- For instance, you may wish to be 95% confident that the interval contains the true population mean.

2.3. Confidence Intervals Formula

Formula for the Confidence Interval of the Mean for a Specific α When σ is Known is:

$$\bar{X} - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

- For a 90% confidence interval, $z_{a/2}$ = 1.65; for a 95% confidence interval, $z_{a/2}$ 1.96; and for a 99% confidence interval, $z_{a/2}$ = 2.58. σ/2 is the standard error,
- The term $Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ is called the margin of error (also called the maximum error of the estimate). For a specific value, say, α = 0.05, 95% of the sample means will fall within this error value on either side of the population mean.
- The margin of error also called the maximum error of the estimate is the maximum likely difference between the point estimate of a parameter and the actual value of the parameter.



Assumptions for Finding a Confidence Interval for a Mean When σ Is Known

1. The sample is a random sample.
2. Either *n* 30 or the population is normally distributed if *n* 30.

❖ Assumptions for Finding a Confidence Interval for a Mean When σ Is Known

1. The sample is a random sample.
2. Either $n \geq 30$ or the population is normally distributed if $n < 30$.

Example 1: A researcher wishes to estimate the number of days it takes an automobile dealer to sell a Chevrolet Aveo. A sample of 50 cars had a mean time on the dealer's lot of 54 days. Assume the population standard deviation to be 6.0 days. Find the best point estimate of the population mean and the 95% confidence interval of the population mean.

Solution

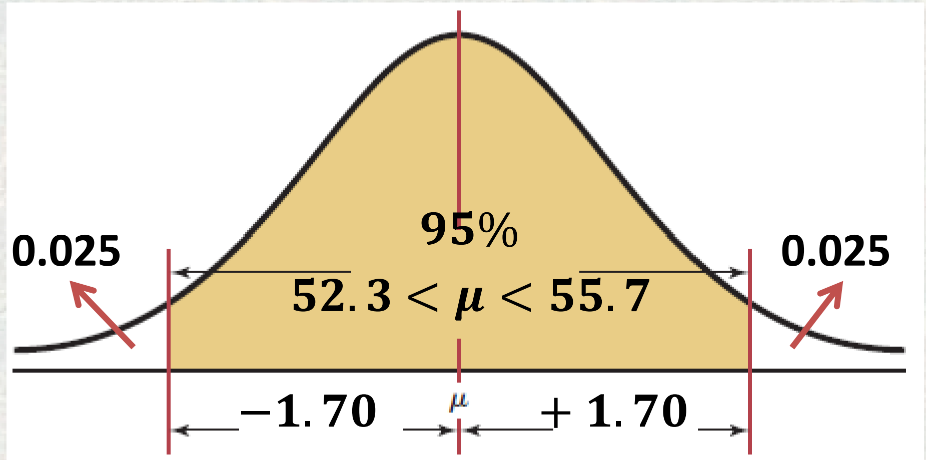
The best point estimate of the mean is 54 days. For the 95% confidence interval use $z = 1.96$ (from table E).

$$\bar{X} - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$54 - 1.96\left(\frac{6.0}{\sqrt{50}}\right) < \mu < 54 + 1.96\left(\frac{6.0}{\sqrt{50}}\right)$$

$$54 - 1.70 < \mu < 54 + 1.70$$

$$52.3 < \mu < 55.7 \quad OR \quad 54 \pm 1.70$$



One can say with 95% confidence that the interval between 52.3 and 55.7 days does contain the population mean, based on a sample of 50 automobiles.

Example 2: A survey of 30 emergency room patients found that the average waiting time for treatment was 174.3 minutes. Assuming that the population standard deviation is 46.5 minutes, find the best point estimate of the population mean and the 99% confidence of the population mean.

Solution

The best point estimate is 174.3 minutes. The 99% confidence is interval use z = 2.58 (from table E).

$$\overline{X} - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{X} + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$174.3 - 2.58\left(\frac{46.5}{\sqrt{30}}\right) < \mu < 174.3 + 2.58\left(\frac{46.5}{\sqrt{30}}\right)$$

$$174.3 - 21.9 < \mu < 174.3 + 21.9$$

$$\boxed{152.4 < \mu < 196.2}$$

One can say with 99% confidence that the mean waiting time for emergency room treatment is between 152.4 and 196.2 minutes.

Example 3: The following data represent a sample of the assets (in millions of dollars) of 30 credit unions in southwestern Pennsylvania. Find the 90% confidence interval of the mean.

12.23	16.56	4.39
2.89	1.24	2.17
13.19	9.16	1.42
73.25	1.91	14.64
11.59	6.69	1.06
8.74	3.17	18.13
7.92	4.78	16.85
40.22	2.42	21.58
5.01	1.47	12.24
2.27	12.77	2.76

Solution

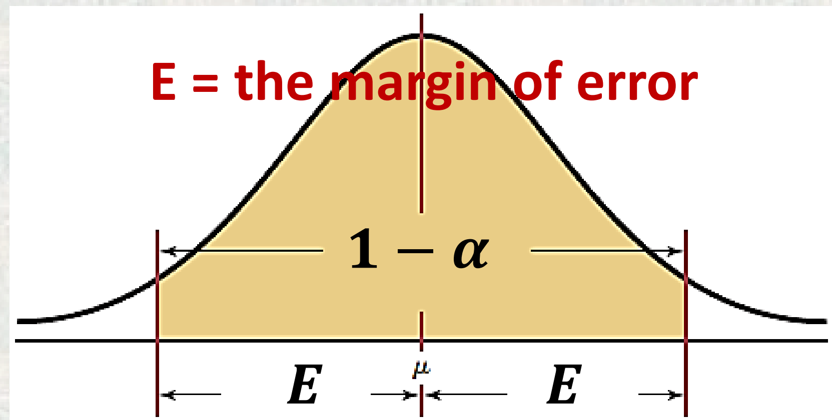
1. Find the mean and standard deviation ($\bar{X} = 11.091$, $\sigma = 14.405$.
2. Confident intervals = 0.9 ; α= 1- 0.9= 0.1 ; α/2= 0.05.
3. $Z_{\alpha/2}$ = 1.68 from table E. $\boldsymbol{\bar{X} - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}$

$$11.091 - 1.65\left(\frac{14.405}{\sqrt{30}}\right) < \mu < 11.091 + 1.65\left(\frac{14.405}{\sqrt{30}}\right)$$

$$11.091 - 4.339 < \mu < 11.091 + 4.339 \quad \Rightarrow \quad \boxed{6.752 < \mu < 15.430}$$

3. Sample Size

- Sample size depends on: the margin of error, the population standard deviation, and the degree of confidence.
- the margin of error formula is:



$$\boldsymbol{E = Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}$$

$$\mathbf{n} = \left(\frac{\mathbf{Z_{\alpha/2} \cdot \sigma}}{\mathbf{E}}\right)^{2}$$

- where *E* is the margin of error. If necessary, round the answer up to obtain a whole number. That is, if there is any fraction or decimal portion in the answer, use the next whole number for sample size *n*.

Example 4: A scientist wishes to estimate the average depth of a river. He wants to be 99% confident that the estimate is accurate within 2 feet. From a previous study, the standard deviation of the depths measured was 4.33 feet.

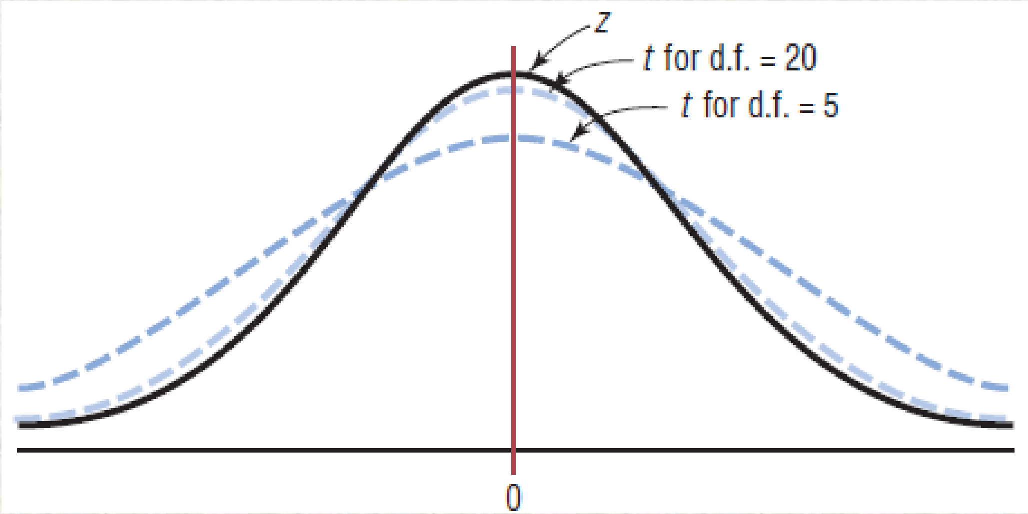
Solution $\alpha = 1 - 0.99 = 0.01$; From table E $Z_{a/2} = 2.58$ $E = 2$; $\sigma = 4.33$.

$$\mathbf{n} = \left(\frac{\mathbf{Z}_{\alpha/2}.\,\boldsymbol{\sigma}}{\mathbf{E}}\right)^2 \Rightarrow \mathbf{n} = \left(\frac{\mathbf{2.58 \times 4.33}}{\mathbf{2}}\right)^2$$

$$\mathbf{n = 31.2} \Rightarrow \mathbf{n = 32}$$

4. Confidence Intervals for the Mean When σ is Unknown

Most of the time, the value of "σ" is not known, so it must be estimated by using "S", namely, the standard deviation of the sample. When *S* is used, especially when the sample size is small, the *Student **t** distribution,* most often called the ***t*** distribution is used instead of normal distribution (Z).



$$t = \frac{\bar{X} - \mu}{S}$$

❖ The *t* distribution shares some characteristics of the normal distribution and differs from it in others.

Similarity: between t and normal distributions:

1. It is bell-shaped.
2. It is symmetric about the mean.
3. The mean, median, and mode are equal to 0 and are located at the center of the distribution.
4. The curve never touches the *x* axis.

The *t* distribution differs from the standard normal distribution in the following:

1. The variance is greater than 1.
2. The *t* distribution is actually a family of curves based on the concept of *degrees of freedom,* which is related to sample size.
3. As the sample size increases, the *t* distribution approaches the standard normal distribution.

❖ Formula for a Specific Confidence Interval for the Mean When S is Unknown

- The values for $t_{a/2}$ are found in Table F.
- Degree of freedom **d.f. = n-1**

$$\bar{X} - t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right) < \mu < \bar{X} + t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

Example 5: Find the *t*a2 value for a 95% confidence interval when the sample size is 22.

Solution

The d.f. = 22 – 1 = 21.

Find 21 in the left column and 95% in the row labeled Confidence Intervals. The intersection where the two meet gives the value for $t_{a/2}$, which is 2.080.

Table F

The *t* Distribution

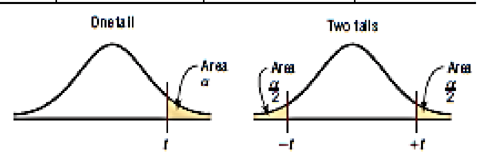
	Confidence Intervals	50%	80%	90%	95%	98%	99%
d.f.	One tail α	0.25	0.10	0.05	0.025	0.01	0.005
	Two tails α	0.50	0.20	0.10	0.05	0.02	0.01
1							
2							
3							
⋮							
21					2.080	2.518	2.831
⋮							
$(z)\infty$		0.674	1.282[a]	1.645[b]	1.960	2.326[c]	2.576[d]

Table F The *t* Distribution

d.f.	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771	2.160	2.650	3.012
14		1.345	1.761	2.145	2.624	2.977
15		1.341	1.753	2.131	2.602	2.947
16		1.337	1.746	2.120	2.583	2.921
17		1.333	1.740	2.110	2.567	2.898
18		1.330	1.734	2.101	2.552	2.878
19		1.328	1.729	2.093	2.539	2.861
20		1.325	1.725	2.086	2.528	2.845
21		1.323	1.721	2.080	2.518	2.831
22		1.321	1.717	2.074	2.508	2.819
23		1.319	1.714	2.069	2.500	2.807
24		1.318	1.711	2.064	2.492	2.797
25		1.316	1.708	2.060	2.485	2.787
26		1.315	1.706	2.056	2.479	2.779
27		1.314	1.703	2.052	2.473	2.771
28		1.313	1.701	2.048	2.467	2.763
29		1.311	1.699	2.045	2.462	2.756
30		1.310	1.697	2.042	2.457	2.750
32		1.309	1.694	2.037	2.449	2.738
34		1.307	1.691	2.032	2.441	2.728
36		1.306	1.688	2.028	2.434	2.719
38		1.304	1.686	2.024	2.429	2.712
40		1.303	1.684	2.021	2.423	2.704
45		1.301	1.679	2.014	2.412	2.690
50		1.299	1.676	2.009	2.403	2.678
55		1.297	1.673	2.004	2.396	2.668
60		1.296	1.671	2.000	2.390	2.660
65		1.295	1.669	1.997	2.385	2.654
70		1.294	1.667	1.994	2.381	2.648
75		1.293	1.665	1.992	2.377	2.643
80		1.292	1.664	1.990	2.374	2.639
90		1.291	1.662	1.987	2.368	2.632
100		1.290	1.660	1.984	2.364	2.626
500		1.283	1.648	1.965	2.334	2.586
1000		1.282	1.646	1.962	2.330	2.581
$(z)\ \infty$		1.282[a]	1.645[b]	1.960	2.326[c]	2.576[d]

[a] This value has been rounded to 1.28 in the textbook.
[b] This value has been rounded to 1.65 in the textbook.
[c] This value has been rounded to 2.33 in the textbook.
[d] This value has been rounded to 2.58 in the textbook.

Source: Adapted from W. H. Beyer, *Handbook of Tables for Probability and Statistics*, 2nd ed., CRC Press, Boca Raton, Fla., 1986. Reprinted with permission.



➢ **Assumptions for Finding a Confidence Interval for a Mean When S is Unknown**

1. The sample is a random sample.
2. Either $n \geq 30$ or the population is normally distributed if $n < 30$

Example 6: Ten randomly selected people were asked how long they slept at night. The mean time was 7.1 hours, and the standard deviation was 0.78 hour. Find the 95% confidence interval of the mean time. Assume the variable is normally distributed.

Solution

Since σ is unknown and *S* must replace it, the *t* distribution (Table F) must be used for the confidence interval.

d.f. = 10 -1 = 9 ⇒ The confidence interval = 95% ⇒ $t_{a/2} = 2.262$

Substituting in the formula. $\bar{X} - t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right) < \mu < \bar{X} + t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$

$$7.1 - 2.262\left(\frac{0.78}{\sqrt{10}}\right) < \mu < 7.1 + 2.262\left(\frac{0.78}{\sqrt{10}}\right) \Rightarrow 6.54 < \mu < 7.66$$

Therefore, 95% confident that the population mean is between 6.54 and 7.66 hr.

Example 7: The data represent a sample of the number of home fires started by candles for the past several years. (Data are from the National Fire Protection Association.) Find the 99% confidence interval for the mean number of home fires started by candles each year. **5460 5900 6090 6310 7160 8440 9930**

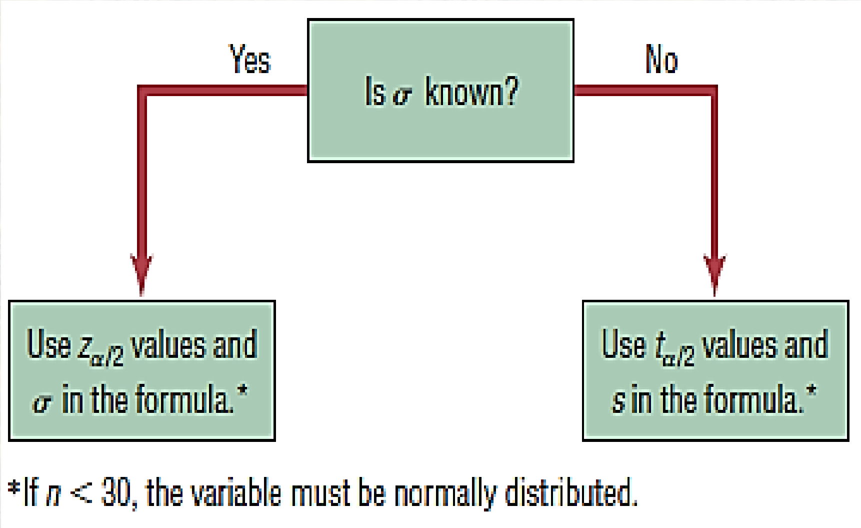
Solution

- Find the mean and standard deviation for the data. ($\bar{X} = 7041.4$ & $S = 1610.3$)
- Confidence interval = 99%; d.f. = 6. → $\boxed{\textbf{From table } t_{\alpha/2} = 3.707.}$
- Substitute in the formula and solve. $\bar{X} - t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right) < \mu < \bar{X} + t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$

$$7041.4 - 3.707\left(\frac{1610.3}{\sqrt{7}}\right) < \mu < 7041.4 + 3.707\left(\frac{1610.3}{\sqrt{7}}\right) \quad \boxed{4785.2 < \mu < 9297.6}$$

- So, at 99% confident that the population mean number of home fires started by candles each year is between 4785.2 and 9297.6.

OVERALL: As stated previously, when σ is known, $Z_{\alpha/2}$ values can be used *no matter what the sample size is,* as long as the variable is normally distributed or ***n*≥30**. When **σ** is unknown and ***n*≥30**, then ***S*** can be used in the formula and $t_{\alpha/2}$ values can be used. Finally, when σ is unknown and ***n*<30, *S*** is used in the formula and $t_{\alpha/2}$ values are used, as long as the variable is approximately normally distributed.



5. Confidence Intervals for Variances and Standard Deviations

- In statistics, the variance and standard deviation of a variable are as important as the mean. For example, when products that fit together (such as pipes) are manufactured, it is important to keep the variations of the diameters of the products as small as possible; otherwise, they will not fit together properly and will have to be scrapped. In the manufacture of medicines, the variance and standard deviation of the medication in the pills play an important role in making sure patients receive the proper dosage. For these reasons, confidence intervals for variances and standard deviations are necessary.
- To calculate these confidence intervals, a new statistical distribution is needed. It is called the chi-square distribution **(χ^2)**.
- The chi-square variable is similar to the **t** variable in that its distribution is a family of curves based on the number of degrees of freedom.
- The chi-square distribution is obtained from the values of

$$\boxed{\chi^2 = \frac{(n-1)S^2}{\sigma^2}}$$

It is normal distributed population

Sample's variance

- A chi-square variable cannot be negative, and the distributions are skewed to the right. At about 100 degrees of freedom, the chi-square distribution becomes somewhat symmetric. The area under each chi-square distribution is equal to 1.00.
- Table G gives the values for the chi-square distribution.

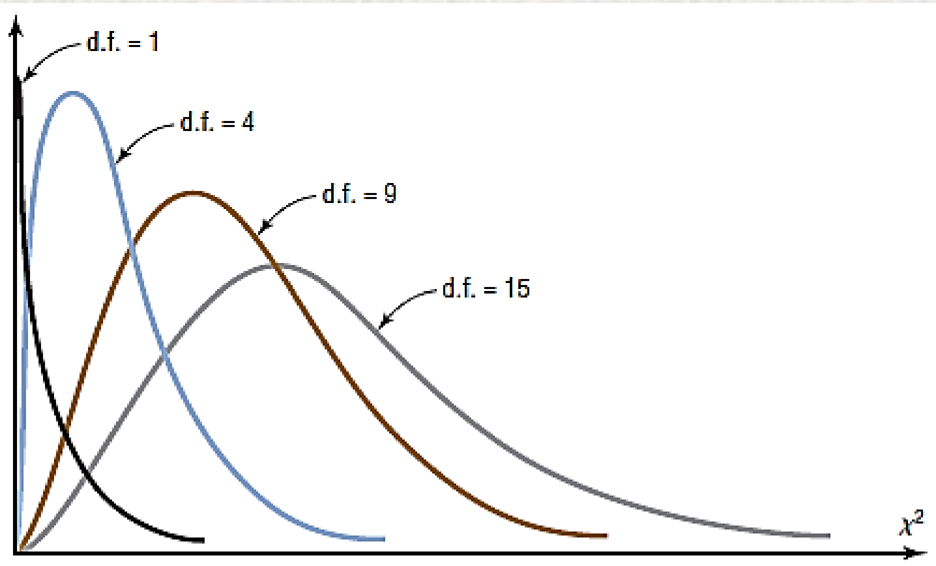
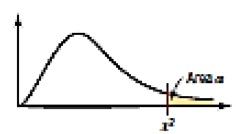


Table G The Chi-Square Distribution

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.262	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Source: Owen, *Handbook of Statistical Tables*, Table A-4 "Chi-Square Distribution Table," © 1962 by Addison-Wesley Publishing Company, Inc. Copyright renewal © 1990. Reproduced by permission of Pearson Education, Inc.



Example 8: Find the values for χ^2 right and χ^2 left for a 90% confidence interval when n 25.

Solution:

To find χ^2 right, α = 1 - 0.90 = 0.10; α/2 = 0.05.

To find χ^2 left, α = 1 - 0.95 = 0.05.

d.f. = n-1=25-1=24

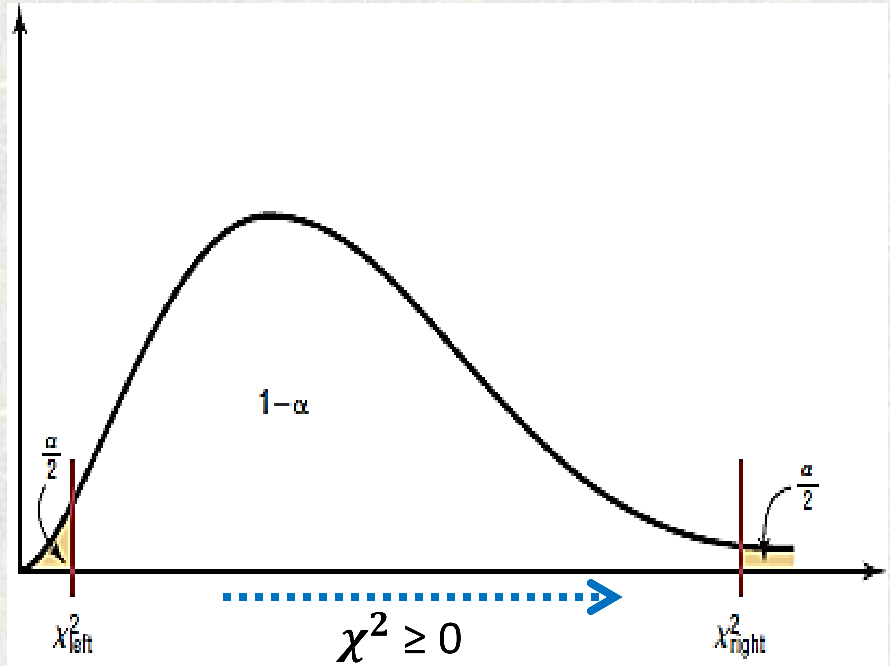
Hence, use the 0.95 and 0.05 columns and the row corresponding to d.f. = 24.

- From table: χ^2 = 36.415 at right; and χ^2 = 13.848 at left

Table G

The Chi-square Distribution

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1										
2										
⋮										
24				13.848			36.415			
				χ^2_{left}			χ^2_{right}			



5.1. Confidence Interval for a Variance

$$\frac{(n-1)S^2}{\chi^2_{right}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{left}}$$

$$d.n. = n - 1$$

5.2. Confidence Interval for a Standard Deviation

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{right}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{left}}}$$

Assumptions:

1. The sample is a random sample.
2. The population must be normally distributed

Example 9: Find the 95% confidence interval for the variance and standard deviation of the nicotine content of cigarettes manufactured if a sample of 20 cigarettes has a standard deviation of 1.6 milligrams.

Solution

$d.n. = n - 1 = 20 - 1 = 19$

$\alpha = 1 - 0.95 = 0.05$

$\alpha/2 = 0.05/2 = 0.025$ (right)

$\chi^2_{right} = \chi^2_{0.025} = 32.852$

$1 - \alpha/2 = 0.975$ (left)

$\chi^2_{left} = \chi^2_{0.975} = 8.907$

$$\frac{(n-1)S^2}{\chi^2_{right}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{left}}$$

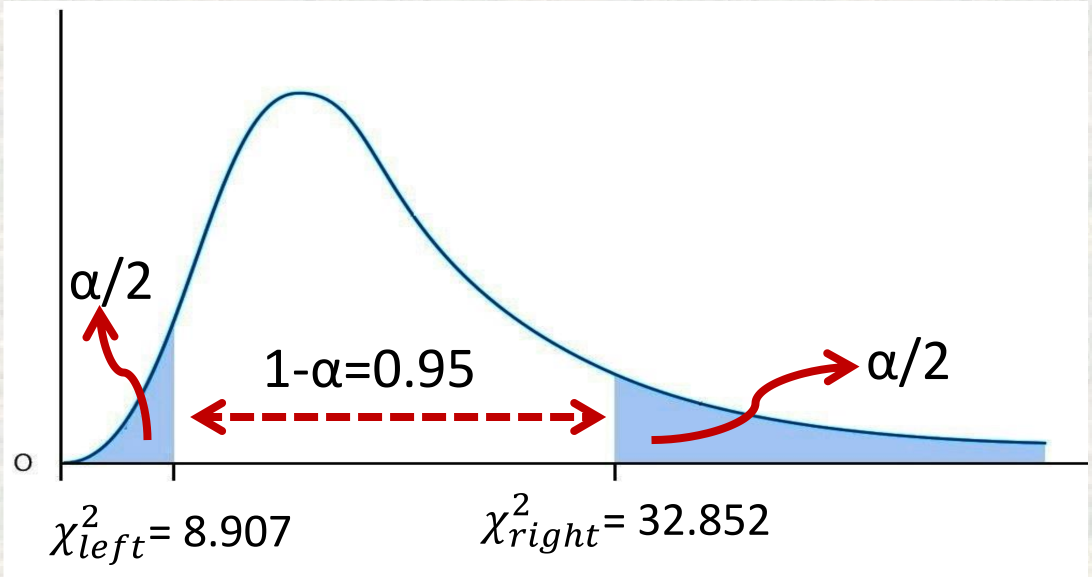
$$\frac{(20-1)(1.6)^2}{32.852} < \sigma^2 < \frac{(20-1)(1.6)^2}{8.907}$$

$$1.5 < \sigma^2 < 5.5$$

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{right}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{left}}}$$

$$\sqrt{\frac{(19)(1.6)^2}{32.852}} < \sigma < \sqrt{\frac{(19)(1.6)^2}{8.907}}$$

$$1.2 < \sigma < 2.3$$



Example 10: Find the 90% confidence interval for the variance and standard deviation for the stability test of asphalt cores in kN. The data represent a selected sample from a specific mix designed for a road. Assume the variable is normally distributed. **59 54 53 52 51 39 49 46 49 48**

Solution:

- Determine the variance for the data; S^2= 28.2.
- $1-\alpha = 1-0.9 = 0.1$; α_{left} = 0.05, α_{right} = 0.9 + 0.05= 0.95; $d.n. = n - 1 = 10 - 1 = 9$
- Find χ^2_{left} from Table = 3.325; χ^2_{right} from Table = 16.919

$$\frac{(n-1)S^2}{\chi^2_{right}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{left}} \qquad \frac{(9)(28.2)}{16.919} < \sigma^2 < \frac{(9)(28.2)}{3.325}$$

$$15 < \sigma^2 < 76.3$$

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{right}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{left}}} \Rightarrow 3.87 < \sigma < 8.73$$

Chapter Eight
Hypothesis Testing

1. Preface
2. Steps in Hypothesis Testing—Traditional Method
3. *z* Test for a Mean
4. t Test for a Mean
5. z Test for a Proportion
6. X^2 Test for a Variance or Standard Deviation

Assi. Prof. Dr. Taher M. Ahmed
Civil Engineering Department
University of Anbar

Chapter Eight

Hypothesis Testing

1. Preface:

- **Statistical Hypothesis Testing** is a decision-making process for evaluating claims about a population.
- In hypothesis testing, the researcher must define the population under study, state the particular hypotheses that will be investigated, give the significance level, select a sample from the population, collect the data, perform the calculations required for the statistical test, and reach a conclusion.

- **For Example:** a scientist might want to know whether the earth is warming up. A physician might want to know whether a new medication will lower a person's blood pressure. An educator might wish to see whether a new teaching technique is better than a traditional one. A production of concert factory is within the standard requirement. Automobile manufacturers are interested in determining whether seat belts will reduce the severity of injuries caused by accidents.

- Three methods used to test hypotheses are:

1. The traditional method
2. The P-value method
3. The confidence interval method

2. Steps in Hypothesis Testing—Traditional Method

- A **statistical hypothesis** is a conjecture about a population parameter. This conjecture may or may not be true.

- There are two types of statistical hypotheses for each situation

1. The **null hypothesis (H_0)** is a statistical hypothesis that states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters.

2. The **alternative hypothesis (H_1)**, is a statistical hypothesis that states an existence of a difference between a parameter and a specific value, or states that there is a difference between two parameters.

Example 1: A chemist invents an additive to increase the life of an automobile battery. If the mean lifetime of the automobile battery without the additive is 36 months, then her hypotheses are

$$H_0\text{: } \mu = 36 \qquad and \qquad H_1\text{: } \mu > 36$$

In this situation, the chemist is interested only in increasing the lifetime of the batteries, so her alternative hypothesis is that the mean is greater than 36 months. The null hypothesis is that the mean is equal to 36 months. This test is called right-tailed, since the interest is in an increase only.

Example 2: A contractor wishes to lower heating bills by using a special type of insulation in houses. If the average of the monthly heating bills is $78, her hypotheses about heating costs with the use of insulation are

$$H_0\text{: } \mu = \$78 \qquad and \qquad H_1\text{: } \mu < \$78$$

This test is a *left-tailed test,* since the contractor is interested only in lowering heating costs. To state hypotheses correctly, researchers must translate the *conjecture* or *claim* from words into mathematical symbols. The basic symbols used are as follows:

Equal to	$=$	*Greater than*	$>$
Not equal to	$\neq$	*Less than*	$<$

- The null and alternative hypotheses are stated together, and the null hypothesis contains the equals sign, as shown (where *k* represents a specified number).

Two-tailed test	Right-tailed test	Left-tailed test
$H_0\text{: } \mu = k$	$H_0\text{: } \mu = k$	$H_0\text{: } \mu = k$
$H_1\text{: } \mu \neq k$	$H_1\text{: } \mu > k$	$H_1\text{: } \mu < k$

Hypothesis-Testing Common Phrases	
$>$ Is greater than Is less than Is above Is below Is higher than Is lower than Is longer than Is shorter than Is bigger than Is smaller than Is increased Is decreased or reduced from	$<$ Is greater than Is less than Is above Is below Is higher than Is lower than Is longer than Is shorter than Is bigger than Is smaller than Is increased Is decreased or reduced from
$=$ Is equal to Is not equal to Is the same as Is different from Has not changed from Has changed from Is the same as Is not the same as	$\neq$ Is equal to Is not equal to Is the same as Is different from Has not changed from Has changed from Is the same as Is not the same as

Example 3:

1. A researcher thinks that if expectant mothers use vitamin pills, the birth weight of the babies will increase. The average birth weight of the population is 8.6 pounds.
2. An engineer hypothesizes that the mean number of defects can be decreased in a manufacturing process of compact disks by using robots instead of humans for certain tasks. The mean number of defective disks per 1000 is 18.
3. A psychologist feels that playing soft music during a test will change the results of the test. The psychologist is not sure whether the grades will be higher or lower. In the past, the mean of the scores was 73.

Solution

1. H_0: m = 8.6 and H_1: m > 8.6
2. H_0: m = 18 and H_1: m < 18
3. H_0: m = 73 and H_1: m $\neq$ 73

- **A statistical test** uses the data obtained from a sample to make a decision about whether the null hypothesis should be rejected. The numerical value obtained from a statistical test is called the **test value.**

- **Note**:

1. In this type of **statistical test**, the mean is computed for the data obtained from the sample and is compared with the population mean. Then a decision is made to reject or not reject the null hypothesis on the basis of the value obtained from the statistical test. If the difference is significant, the null hypothesis is rejected. If it is not, then the null hypothesis is not rejected.
2. In the hypothesis-testing situation, there are **four possible outcomes**

- A **type I error** occurs if you reject the null hypothesis when it is true.
- A **type II error** occurs if you do not reject the null hypothesis when it is false.

	H_o true	*H_o false*
Reject H_o	**Error Type I**	*Correct decision*
Do not Reject H_o	*Correct decision*	**Error Type II**

For an example: In a jury trial, there are four possible outcomes. The defendant is either guilty or innocent, and he or she will be convicted or acquitted.

	H_0 true (innocent)	H_0 false (not innocent)
Reject H_0 (convict)	1. Type I error	2. Correct decision
Do not reject H_0 (acquit)	3. Correct decision	4. Type II error

H_0: The defendant is innocent

H_1: The defendant is not innocent (i.e., guilty)

- **The level of significance**

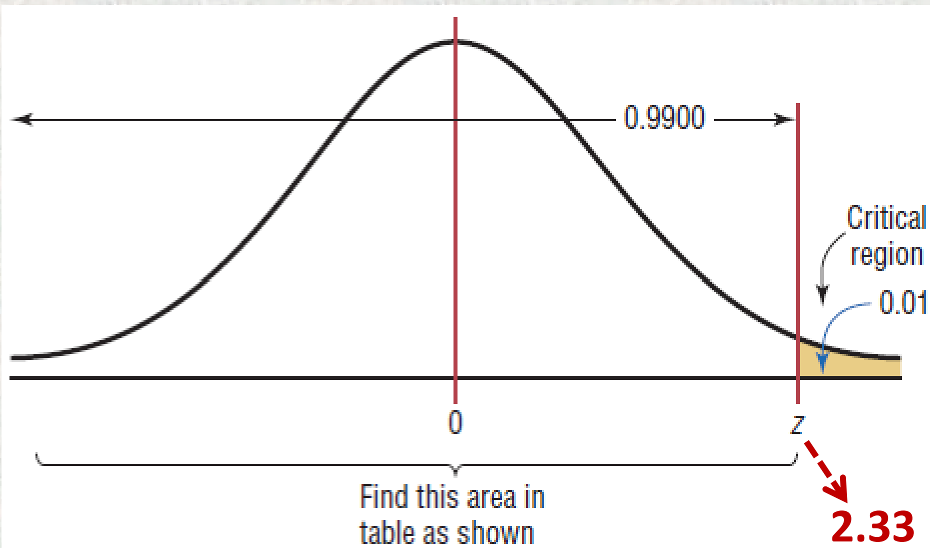
The **level of significance** is the maximum probability of committing a type I error. This probability is symbolized by **α** (Greek letter **alpha**).

- Statisticians generally agree on using three arbitrary significance levels: the 0.10, 0.05, and 0.01 levels. In other words, when **α = 0.10**, there is a 10% chance of rejecting a true null hypothesis; when **α = 0.05**, there is a 5% chance of rejecting a true null hypothesis; and when **α = 0.01**, there is a 1% chance of rejecting a true null hypothesis.
- Rejection and acceptance (not rejection) region depends on the critical value (CV) which can be determined based on the table of normal distribution (z value). The **critical value** separates the critical region from the noncritical region. CV can be either to the right side of the mean or to the left side of the mean (one or two -tailed).
- The **critical** or **rejection region** is the range of values of the test value that indicates that there is a significant difference and that the null hypothesis should be rejected.
- The **noncritical** or **non-rejection region** is the range of values of the test value that indicates that the difference was probably due to chance and that the null hypothesis should not be rejected (should be accepted).

- A **one-tailed test** indicates that the null hypothesis should be rejected when the test value is in the critical region on one side of the mean. A one-tailed test is either a **right tailed test** or **left-tailed test,** depending on the direction of the inequality of the alternative hypothesis.

Example 4: Finding the Critical Value for A 0.01 (Right-Tailed Test).

Solution: Using z table.

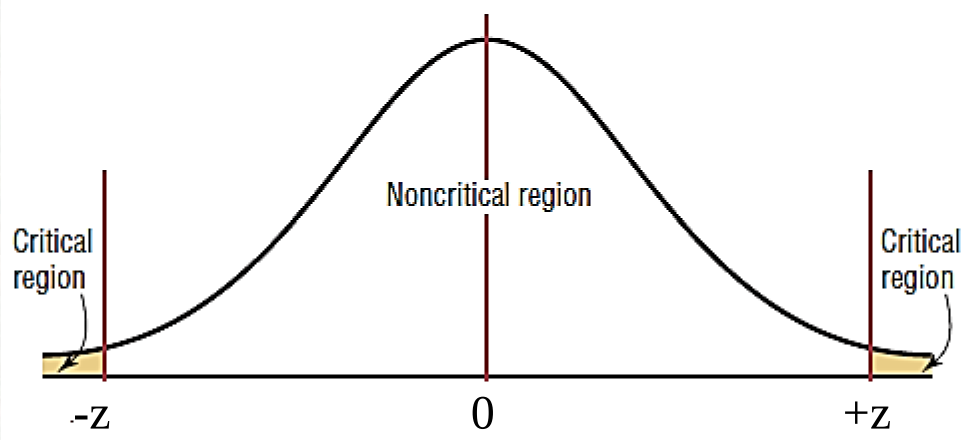


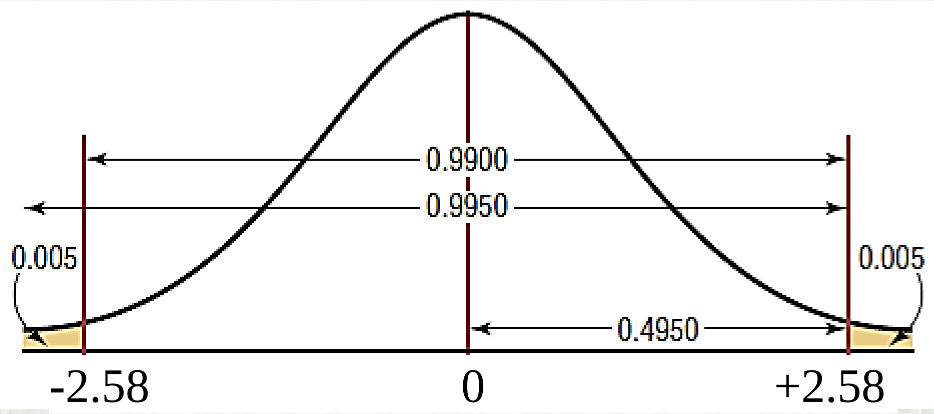
z	0.00	0.01	0.02	0.03	0.04	0.05	…
0.0							
0.1							
0.2							
0.3							
⋮							
2.1							
2.2							
2.3				0.9901			
2.4							
….							

Closest value to 0.9900

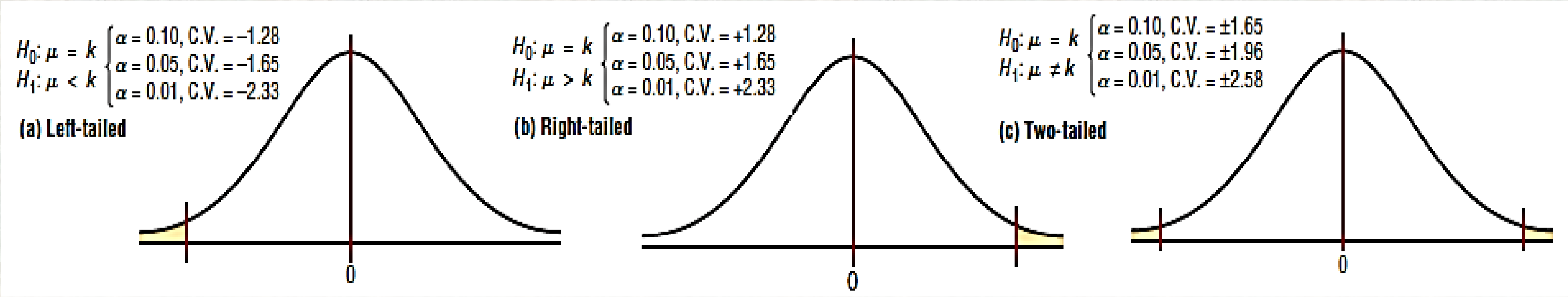
- In a **two-tailed test**, the null hypothesis should be rejected when the test value is in either of the two critical regions.

For example, a two-tailed test, the critical region must be split into two equal parts. If **α = 0.01**, then one-half of the area, or 0.005, must be to the right of the mean and one half must be to the left of the mean, as shown in the figure below.





Summery



Procedure Table

Finding the Critical Values for Specific α Values, Using Table z values.

Step 1

Draw the figure and indicate the appropriate area.

1. If the test is left-tailed, the critical region, with an area equal to **α**, will be on the left side of the mean.
2. If the test is right-tailed, the critical region, with an area equal to **α**, will be on the right side of the mean.
3. If the test is two-tailed, **α** must be divided by 2; one-half of the area will be to the right of the mean, and one-half will be to the left of the mean.

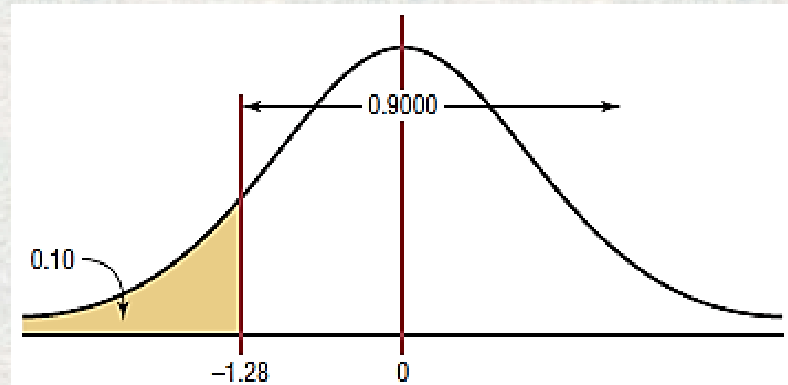
Step 2

1. For a left-tailed test, use the *z* value that corresponds to the area equivalent to **α** in Table z values.
2. For a right-tailed test, use the *z* value that corresponds to the area equivalent to (**1- α**).
3. For a two-tailed test, use the *z* value that corresponds to **α/2** for the left value. It will be negative. For the right value, use the *z* value that corresponds to the area equivalent to **1-α/2 .** It will be positive.

Example 5: Using Table E in Appendix C, find the critical value(s) for each situation and draw the appropriate figure, showing the critical region. ***a.*** A left-tailed test with a 0.10. ***b.*** A two-tailed test with a 0.02. ***c.*** A right-tailed test with a 0.005.

Solution

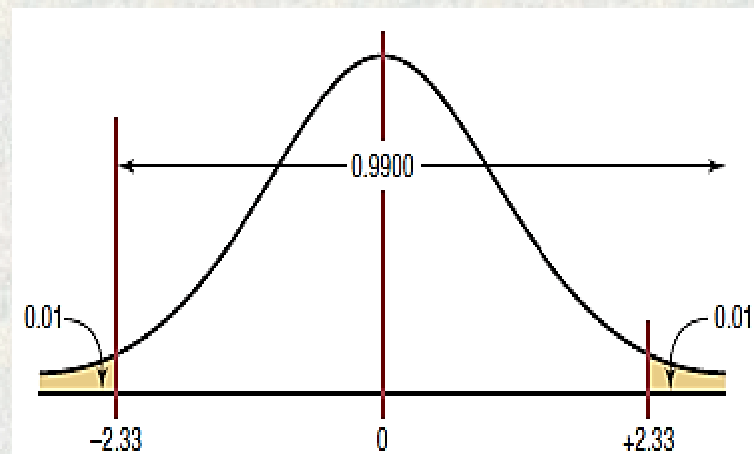
a). **Step 1:** Draw the figure and indicate the appropriate area. The area of 0.10 is located in the left tail, as shown in Figure below.



Step 2: Find the area closest to 0.1 from z's table. In this case, it is 0.1003. Find the *z* value that corresponds to the area 0.1003. It is 1.28. See Figure below.

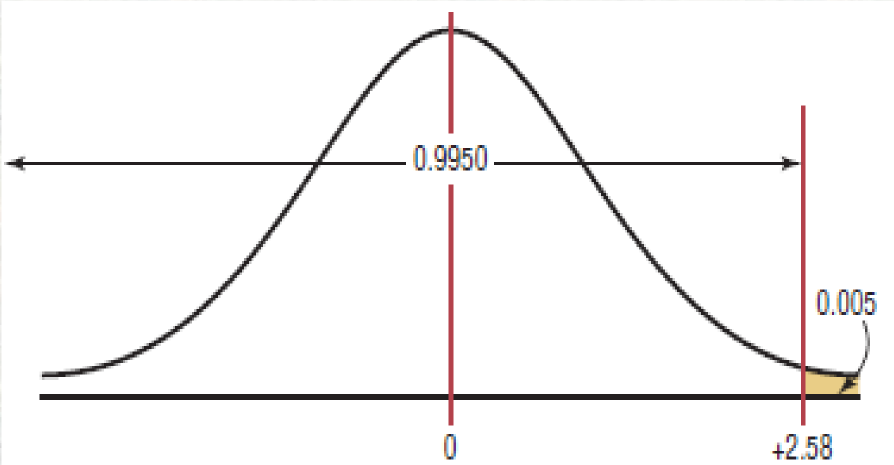
b). **Step 1:** Draw the figure and indicate the appropriate area. In this case, there are two areas equivalent to **α/2** or **0.02/2 = 0.01**.

Step 2: For the left *z* critical value, find the area closest to **α/2** , or **0.02/2 = 0.01**. In this case, it is 0.0099. For the right *z* critical value, find the area closest to **1 - α/2,** or **1 - 0.02/2 = 0.9900**. In this case, it is 0.9901. Find the *z* values for each of the areas. For 0.0099, *z*=-2.33. For the area of 0.9901, *z* = 0.9901, *z* = +2.33.



c). **Step 1:** Draw the figure and indicate the appropriate area. Since this is a right-tailed test, the area 0.005 is located in the right tail, as shown in Figure below.

Step 2: Find the area closest to **1-α** , or **1 - 0.005 = 0.9950**. In this case, it is 0.9949 or 0.9951. The two *z* values corresponding to 0.9949 and 0.9951 are 2.57 and 2.58. Since 0.9500 is halfway between these two values, find the average of the two values **(2.57 + 2.58) / 2 = 2.575**. However, 2.58 is most often used.



- **In hypothesis testing, the following steps are recommended for Traditional Method**

Step 1 State the hypotheses (null and alternative) and identify the claim.

Step 2 Find the critical value(s) from the appropriate table.

Step 3 Compute the test value.

Step 4 Make the decision to reject or not reject the null hypothesis.

Step 5 Summarize the results.

3. z Test for a Mean

- The ***z* test** is a statistical test for the mean of a population. It can be used when n 30, or when the population is normally distributed and s is known. The formula for the z test is

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

- Where: $\bar{X}$ = sample mean; μ = hypothesized population mean; σ = population standard deviation; n = sample size

- **Assumptions for the *z* Test for a Mean When σ Is Known**

  1. The sample is a random sample.
  2. Either $n \geq 30$ or the population is normally distributed if $n < 30$.

- **Procedure Steps: there are five steps for solving *hypothesis-testing* problems:**

Step 1 State the hypotheses and identify the claim.
Step 2 Find the critical value(s).
Step 3 Compute the test value.
Step 4 Make the decision to reject or not reject the null hypothesis.
Step 5 Summarize the results.

Example 6: A researcher wishes to see if the mean number of days that a basic, low-price, small automobile sits on a dealer's lot is 29. A sample of 30 automobile dealers has a mean of 30.1 days for basic, low-price, small automobiles. At $\alpha = 0.05$, test the claim that the mean time is greater than 29 days. The standard deviation of the population is 3.8 days.

Solution:

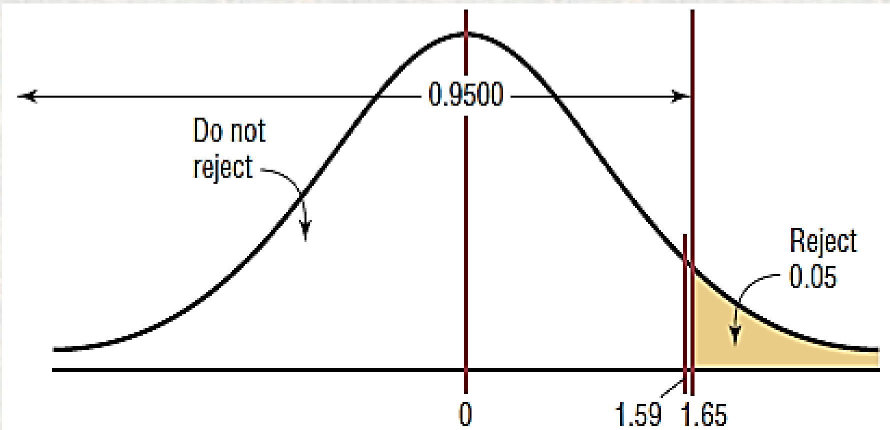
Step 1 State the hypotheses and identify the claim.

H_0: $\mu = 29$ and H_1: $\mu > 29$ (claim)

Step 2 Find the critical value. Since $\alpha = 0.05$ and the test is a right-tailed test, the critical value is $z = 1.65$ from Table.

Step 3 Compute the test value.

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} == \frac{30.1 - 29}{3.8 / \sqrt{30}} = 1.59$$



Step 4 Make the decision. Since the test value, 1.59, is less than the critical value, 1.65, and is not in the critical region, the decision is to not reject the null hypothesis. As shown in the figure

Step 5 Summarize the results. There is not enough evidence to support the claim that the mean time is greater than 29 days.

Example 7: The School Rehabilitation Foundation reports that the average cost of rehabilitation for a primary school for each 10 years is $24,672. To see if the average cost of rehabilitation is different at a particular school, a researcher selects a random sample of 35 schools to find that the average cost of their rehabilitation is $26,343. The standard deviation of the population is $3251. At α = 0.01, can it be concluded that the average cost of rehabilitation at a particular school is different from $24,672?

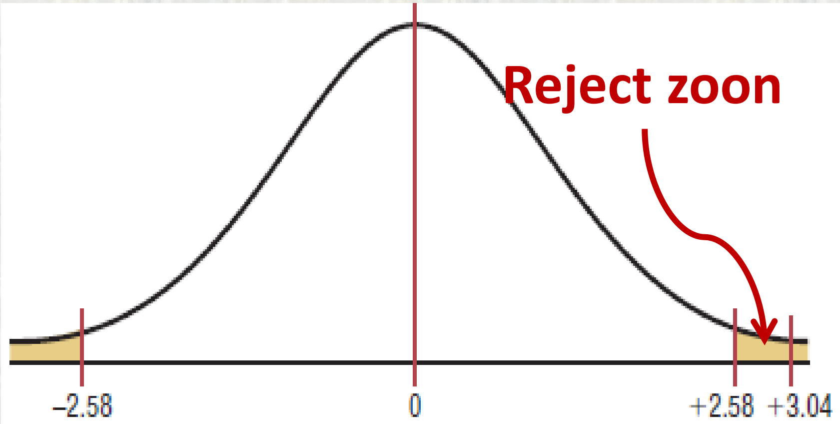
Solution

Step 1 State the hypotheses and identify the claim. H_0: **μ = $ 24,672** and H_1: **μ ≠ $24,672 (claim)**

Step 2 Find the critical value. Since α = 0.01 and the test is a two-tailed test, the critical value is *z = ± 2.58.*

Step 3 Compute the test value.

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} == \frac{26{,}343 - 24{,}672}{3251 / \sqrt{35}} = 3.04$$



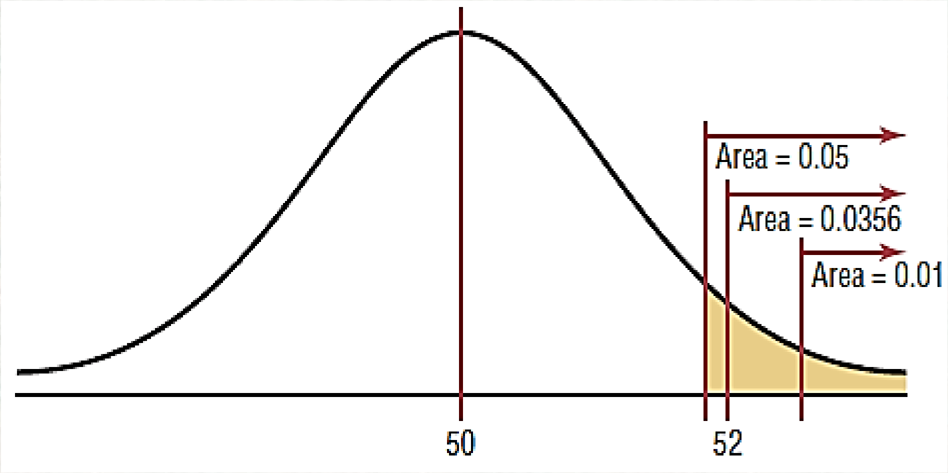
Step 4 Make the decision. Reject the null hypothesis, since the test value falls in the critical region, as shown in Figure.

Step 5 Summarize. There is enough evidence to support the claim that the average cost of rehabilitation at the particular school is different from $24,672.

4. P-Value Method for Hypothesis Testing

- The ***P*-value** (or probability value) is the probability of getting a sample statistic (such as the mean) or a more extreme sample statistic in the direction of the alternative hypothesis when the null hypothesis is true. **In other words,** the *P*-value is the actual area under the standard normal distribution curve (or other curve, depending on what statistical test is being used) representing the probability of a particular sample statistic or a more extreme sample statistic occurring if the null hypothesis is true.
- For example, suppose that an alternative hypothesis is H1: μ > 50 and the mean of a sample is $\bar{X} = 52$. If the P-value = 0.0356 for a statistical test, then the probability of getting a sample mean of 52 or greater is 0.0356 if the true population mean is 50.

The relationship between the P-value and the α value can be explained in this manner. For P = 0.0356, the null hypothesis would be rejected at α = 0.05 but not at α = 0.01



- **Procedure for Solving Hypothesis-Testing Problems (P-Value Method)**

Step 1 State the hypotheses and identify the claim.
Step 2 Compute the test value.
Step 3 Find the *P*-value.
Step 4 Make the decision.
Step 5 Summarize the results.

Example 8: A researcher claims that the average wind speed in a certain city is 8 miles per hour. A sample of 32 days has an average wind speed of 8.2 miles per hour. The standard deviation of the population is 0.6 mile per hour. At $\alpha = 0.05$, is there enough evidence to reject the claim? Use the *P*-value method.

Solution

Step 1 State the hypotheses and identify the claim. H_0: $\mu = 8$ and H_1: $\mu \neq 8$ **(claim)**

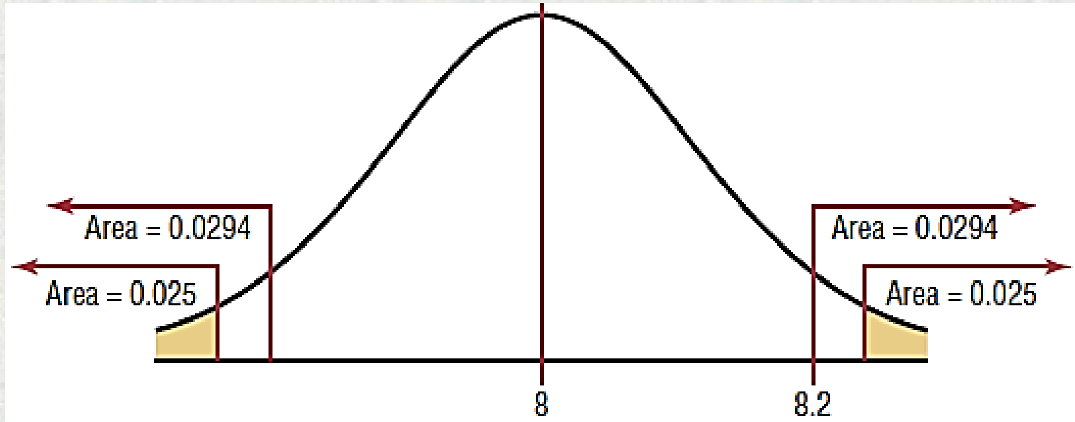
Step 2 Compute the test value.

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} == \frac{8.2 - 8.0}{0.6 / \sqrt{32}} = 1.89$$

Step 3 Find the *P*-value. Using N.D. Table, find the corresponding area for $z = 1.89$. It is 0.9706. Subtract the value from 1.0000. (1.0000 - 0.9706 = 0.0294).
Since this is a two-tailed test, the area of 0.0294 must be doubled to get the *P*-value. → (2(0.0294) = 0.0588).

Step 4 Make the decision. The decision is to not reject the null hypothesis, since the P-value is greater than 0.05. As shown in the figure.

Step 5 Summarize the results. There is not enough evidence to reject the claim that the average wind speed is 8 miles per hour.



Example 9: A producer bricks wishes to test the claim that the average of compressive strength of the product is greater than 5700 psi. He selected a random sample of 36 bricks and finds the mean to be 5950 psi. The population standard deviation is 659 psi. Is there evidence to support the claim at $\alpha = 0.05$? Use the *P*-value method.

Solution

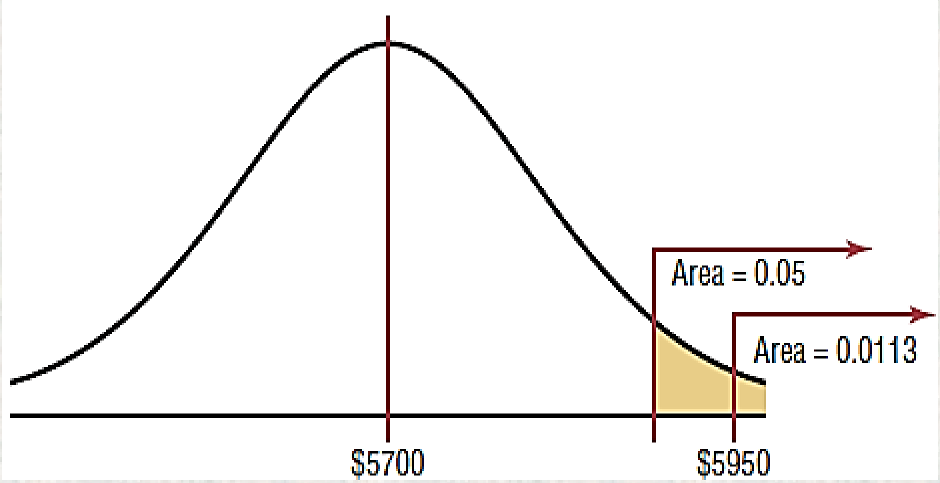
Step 1 State the hypotheses and identify the claim. H_0: **μ = 5700** and H_1: **μ > 5700 (claim)**

Step 2 Compute the test value.

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} == \frac{5950 - 5700}{659 / \sqrt{36}} = 2.28$$

Step 3 Find the *P*-value. Using N.D. Table, find the corresponding area for $z = 2.28$. It is 0.9887. Subtract the value from 1.0000. to find the area in the right tail. (1.0000 - 0.9887 = 0.0113) = *P*-value.

Step 4 Make the decision. Since the *P*-value is less than 0.05, the decision is to reject the null hypothesis. As shown in the figure.



Step 5 Summarize the results. There is enough evidence to support the claim that the compressive strength is greater than 5700 psi.

4. t - Test for a Mean

- The ***t* test** is a statistical test for the mean of a population and is used when the population is normally or approximately normally distributed, and s is unknown.
- The formula for the *t* test is.

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

The degrees of freedom are d.f. *n*= 1.

Note: The formula for the *t* test is similar to the formula for the *z* test. But since the population standard deviation σ is unknown, the sample standard deviation *s* is used instead.

Example 10: Find the critical *t* value for α = 0.05 with d.f. = 16 for a right-tailed *t* test.

Solution

Find the 0.05 column in the top row and 16 in the left-hand column. Where the row and column meet, the appropriate critical value is found; it is 1.746. as shown in the figure.

d.f.	One tail, α	0.25	0.10	0.05	0.025	0.01	0.005
	Two tails, α	0.50	0.20	0.10	0.05	0.02	0.01
1							
2							
3							
4							
5							
⋮							
14							
15							
16				1.746			
17							
18							
⋮							

t – Distribution table

Example 11: Find the critical *t* value for α = 0.01 with d.f. 22 for a left-tailed test.

Solution

Find the 0.01 column in the row labeled One tail, and find 22 in the left column. The critical value is 2.508 since the test is a one-tailed left test.

Example 12: Find the critical values for α = 0.10 with d.f. = 18 for a two-tailed *t* test.

Solution

Find the 0.10 column in the row labeled Two tails, and find 18 in the column labeled d.f. The critical values are +1.734 and -1.734.

Example 13: Find the critical value for $\alpha = 0.05$ with d.f. 28 for a right-tailed *t* test.

Solution

Find the 0.05 column in the One-tail row and 28 in the left column. The critical value is 1.701.

- **Assumptions for the *t* Test for a Mean When σ Is Unknown**

1. The sample is a random sample.
2. Either $n \geq 30$ or the population is normally distributed if $n < 30$.

- **Test hypotheses using the t test (traditional method), is the same procedure as for the z test, except use the table of t distribution.**

Step 1 State the hypotheses and identify the claim.
Step 2 Find the critical value(s) from Table F.
Step 3 Compute the test value.
Step 4 Make the decision to reject or not reject the null hypothesis.
Step 5 Summarize the results.

Example 14: An engineer of soil investigation claims that the average bearing capacity of soil is 16.3 Ton. A random sample of 10 samples had a mean bearing capacity is 17.7 ton. The sample standard deviation is 1.8 Ton. Is there enough evidence to reject the engineer's claim at $\alpha = 0.05$?

Solution

Step 1: H_0: $\mu = 16.3$ and H_1: $\mu \neq 16.3$ (claim)

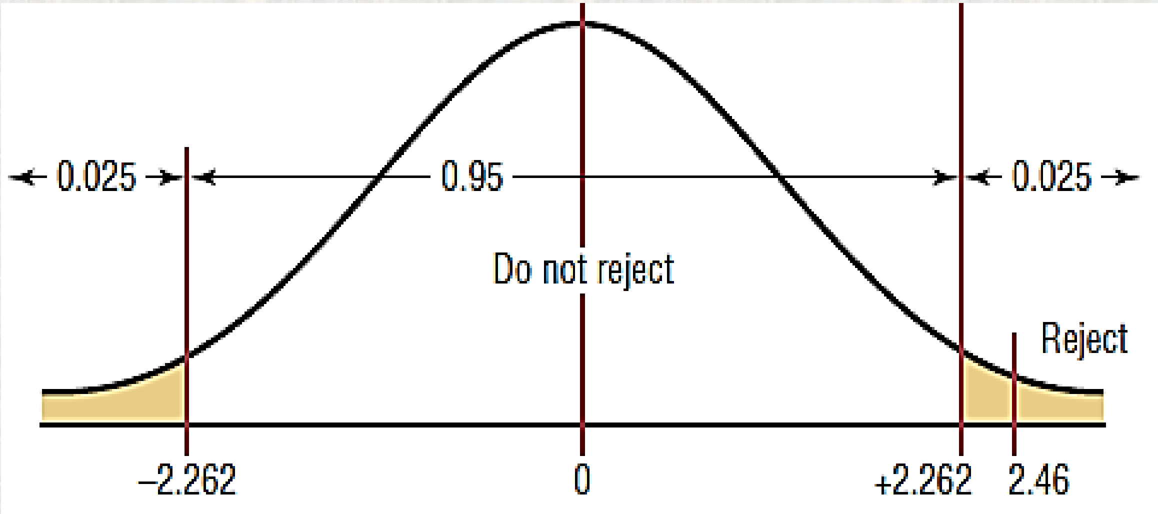
Step 2: The critical values are +2.262 and -2.262 for $\alpha = 0.05$ and d.f. = 9.

Step 3 The test value is

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} == \frac{17.7 - 16.3}{1.8/\sqrt{10}} = 2.46$$

Step 4 Reject the null hypothesis since 2.46 > 2.262. As shown in the figure.

Step 5 There is enough evidence to reject the claim that the average bearing capacity is 16.3 Ton.



Example 15: An engineer evaluated the production of a concrete factory for high strength. he claimed that the average compressive strength of the products is less than 60 MPa. A random sample of eight samples are selected as shown below. Is there enough evidence to support the engineer's claim at α = 0.10?

Compressive strength: 60, 56, 60, 55, 70, 55, 60, 55

Solution

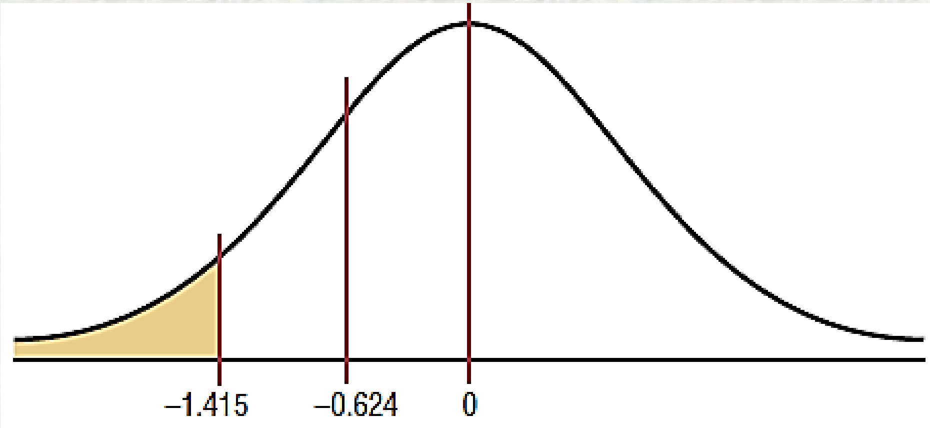
Step 1: H_0: **μ = 60** and H_1: **μ < 60 (claim)**

Step 2: At α = 0.10 and d.f. = 7. the critical value is - 1.415.

Step 3: To compute the test value, the mean and standard deviation must be found. $\bar{X} = 58.88 \; and \; S = 5.08$, find t.

$$\boldsymbol{t = \frac{\bar{X} - \mu}{S/\sqrt{n}} == \frac{58.88 - 60}{5.08/\sqrt{7}} = -0.624}$$

Step 4: Do not reject the null hypothesis since -0.624 falls in the noncritical region. As shown in the figure.



Step 5: There is not enough evidence to support the engineer's claim that the average compressive strength is less than 60 MPa.

5. z Test for a Proportion

A normal distribution can be used to approximate the binomial distribution (proportion) when **np ≥ 5** and **nq ≥ 5**, the standard normal distribution can be used to test hypotheses for proportions.

Formula for the *z* Test for Proportions

$$z = \frac{\hat{p}-p}{\sqrt{pq/n}}$$

Where: $\hat{p} = \frac{X}{n}$ (sample proportion)

p = population proportion

n = sample size

Assumptions for Testing a Proportion

1. The sample is a random sample.
2. The conditions for a binomial experiment are satisfied.
3. np ≥ 5 and nq ≥ 5.

Note: The steps for hypothesis testing are the same as those used to find critical values and *P*-values.

Example 16: A dietitian claims that 60% of people are trying to avoid trans fats in their diets. She randomly selected 200 people and found that 128 people stated that they were trying to avoid trans fats in their diets. At α = 0.05, is there enough evidence to reject the dietitian's claim?

Solution

Step 1: State the hypothesis and identify the claim.

H_0: $p = 0.60$ (claim) and H_1: $p \neq 0.60$

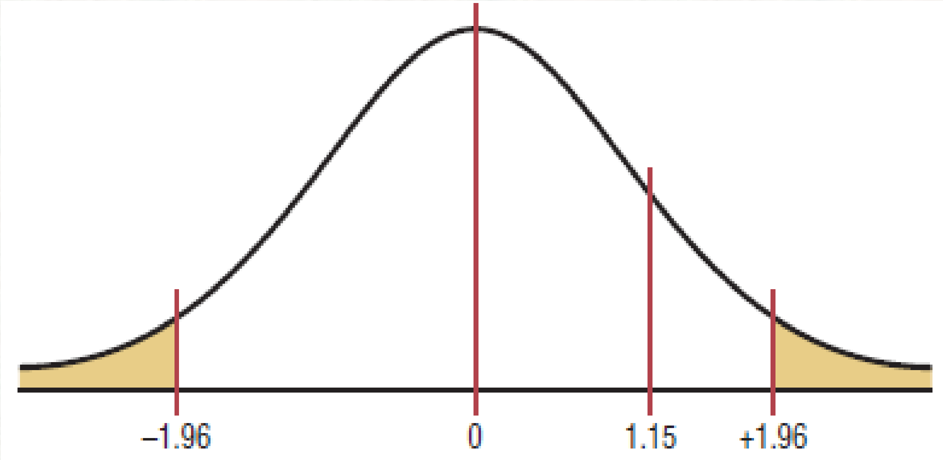
Step 2: Find the critical values. Since α = 0.05 and the test value is two-tailed, the critical values are ±1.96.

Step 3: Compute the test value. First, it is necessary to find $\hat{p}$.

$$\hat{p} = \frac{X}{n} = \frac{125}{200} = 0.64$$

$$P = 0.6 \rightarrow q = 1\text{-}0.6 = 0.4$$

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{0.64 - 0.6}{\sqrt{(0.6)(0.4)/200}} = 1.15$$



Step 4: Make the decision. Do not reject the null hypothesis since the test value falls outside the critical region, as shown in figure.

Step 5 Summarize the results. There is not enough evidence to reject the claim that 60% of people are trying to avoid trans fats in their diets.

Example 17: An engineer claims that more than 25% of all construction company advertise. A sample of 200 companies in a certain city showed that 63 had used some form of advertising. At α = 0.05, is there enough evidence to support the engineer's claim? Use the *P*-value method.

Solution

Step 1: State the hypothesis and identify the claim.

H_0: $p = 0.25$ and H_1: $p > 0.25$ (claim)

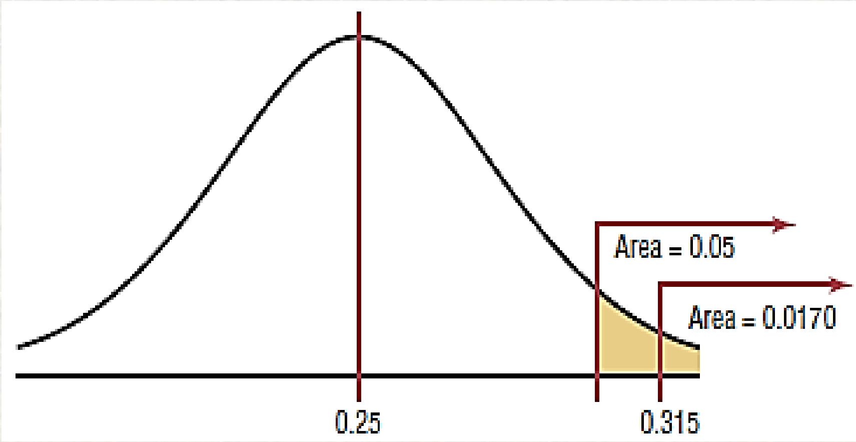
$$\hat{p} = \frac{X}{n} = \frac{63}{200} = 0.315$$

P = 0.25 → q = 1-0.25 = 0.75

Step 2: Compute the test value.

$$z = \frac{\hat{p}-p}{\sqrt{pq/n}} = \frac{0.315-0.25}{\sqrt{(0.25)(0.315)/200}} = 2.12$$

Step 3 Find the *P*-value. The area under the curve for $z =$ 2.12 is 0.9830. Subtracting the area from 1.0000, you get 1.0000 - 0.9830 = 0.0170. The *P*-value is 0.0170.



Step 4: Reject the null hypothesis, since 0.0170 0.05 (that is, *P*-value 0.05). As shown in the figure

Step 5 There is enough evidence to support the attorney's claim that more than 25% of the lawyers use some form of advertising.

6. χ^2 Test for a Variance or Standard Deviation

In Chapter 7, the chi-square distribution was used to construct a confidence interval for a single variance or standard deviation. This distribution is also used to test a claim about a single variance or standard deviation.

- **Formula for the χ^2 Test for a Single Variance**

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

with degrees of freedom equal to n=1

n = sample size

S^2 = sample variance

σ^2 = population variance

- **Assumptions for the Chi-Square Test for a Single Variance**

1. The sample must be randomly selected from the population.
2. The population must be normally distributed for the variable under study.
3. The observations must be independent of one another

- **The traditional method for hypothesis testing:**

Step 1 State the hypotheses and identify the claim.

Step 2 Find the critical value(s).

Step 3 Compute the test value.

Step 4 Make the decision.

Step 5 Summarize the results.

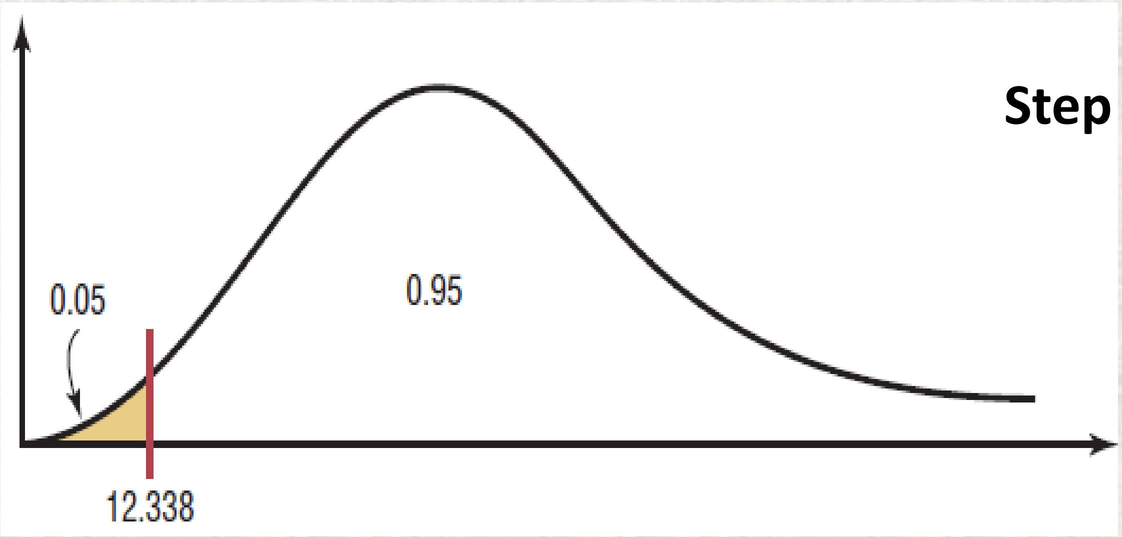
Example18: An engineer wishes to see whether the variation in the experience of construction for 23 companies is less than the variance of the population (σ^2 = 225). The variance of the companies is 198. Is there enough evidence to support the engineer's claim that the variation of the companies is less than the population variance at α = 0.05? Assume that the scores are normally distributed.

Solution

Step 1: State the hypotheses and identify the claim.

H_0: s2 225 and H_1: s2 225 (claim)

Step 2: Find the critical value. Since this test is left-tailed and α = 0.05, use the value 1 - 0.05 = 0.95. The degrees of freedom are n - 1 = 23 - 1 = 22. Hence, the critical value is 12.338. Note that the critical region is on the left, as shown in figure.

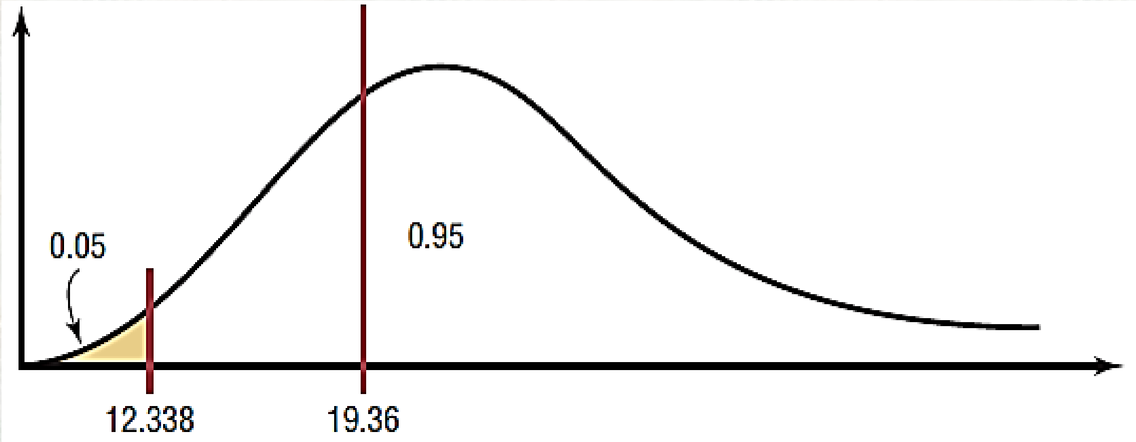


Step 3: Compute the test value.

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{(23-1)(198)^2}{(225)^2}$$

$$\chi^2 = 19.36$$

Step 4 Make the decision. Since the test value 19.36 falls in the noncritical region, as shown the figure, the decision is to not reject the null hypothesis.



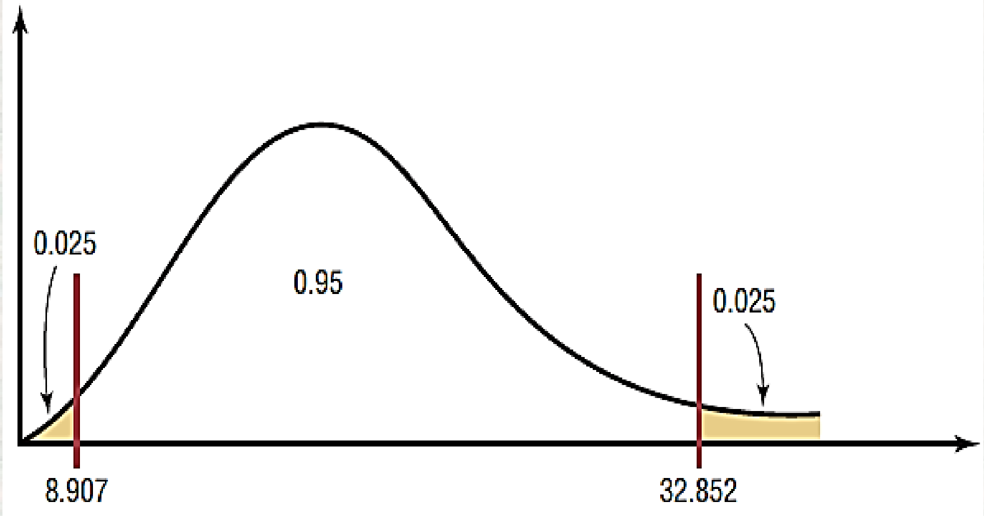
Step 5 Summarize the results. There is not enough evidence to support the claim that the variation in test scores of the engineer's claim is less than the variation in scores of the population.

Example 19: An petrol engineer wishes to test the claim that the variance of the lead content of the fuel is 0.644. Lead content is measured in milligrams, and assume that it is normally distributed. A sample of 20 bottles has a standard deviation of 1.00 milligram. At $\alpha = 0.05$, is there enough evidence to reject the manufacturer's claim?

Solution

Step 1: State the hypotheses and the claim. H_0: $\sigma^2 = 225$ and H_1: $\sigma^2 = 225$ (claim)

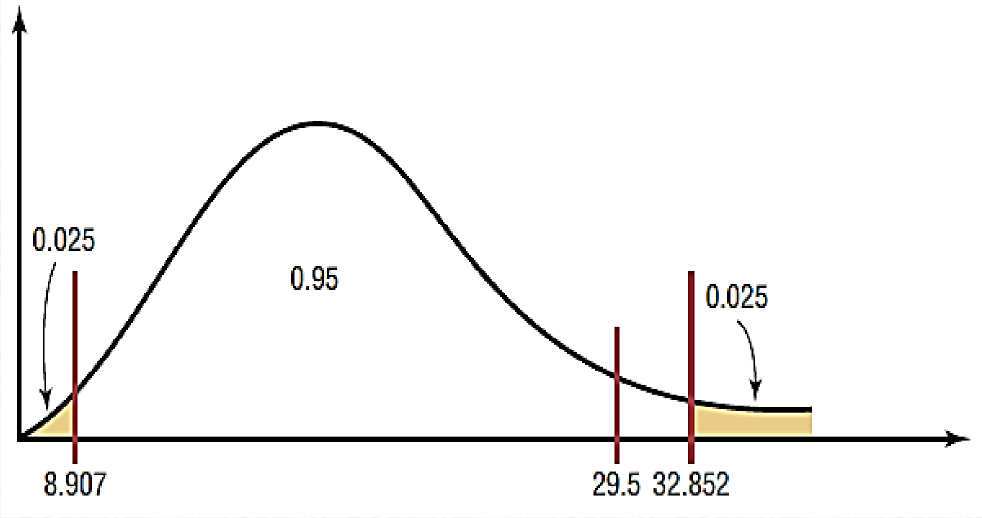
Step 2: Find the critical value. Since this test is two-tailed and α = 0.05, the critical values for 0.025 and 0.975 must be found. d.f. = 20 – 1 = 19; hence, the critical values are 32.852 and 8.907, respectively. The critical or rejection regions are shown in the figure.



Step 3: Compute the test value.

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{(20-1)(1.0)^2}{(0.644)} = 29.5$$

Step 4: Make the decision. Do not reject the null hypothesis, since the test value falls between the critical values (8.907 < 29.5 < 32.852) and in the noncritical region, as shown in the figure.



Step 5: Summarize the results. There is not enough evidence to reject the engineer's claim that the variance of the lead content of the fuel is equal to 0.644.

Thank you, the end of Ch.8