**Math 281: These notes are taken from “Introductory Statistics” written by Prem S. Mann published by Wiley**

**Chapter 1:**

***Statistics*** is the science of collecting, analyzing, presenting, and interpreting data, as well as of making decisions based on such analyses.

***Descriptive Statistics*** consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures

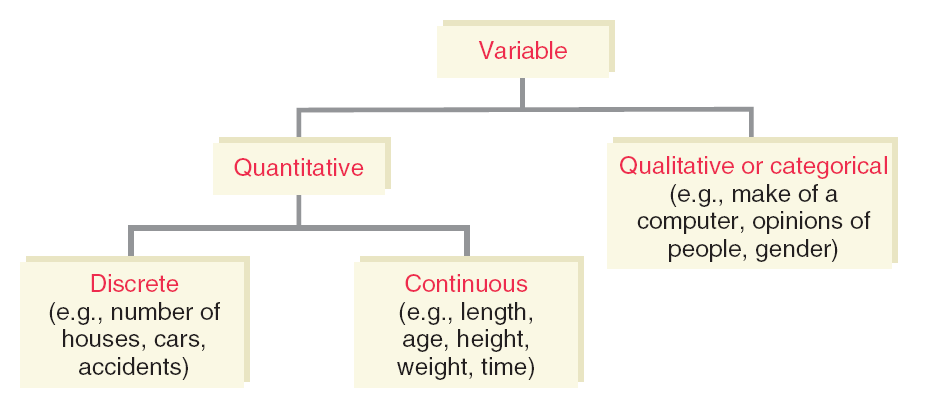
A sample that represents the characteristics of the population as closely as possible is called a ***representative sample***.

A variable that can be measured numerically is called a ***quantitative variable***. The data collected on a quantitative variable are called ***quantitative data***.

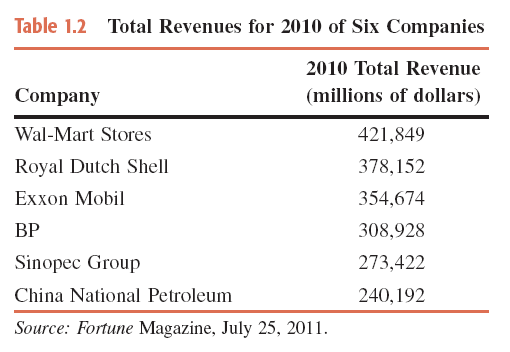
A variable whose values are countable is called a ***discrete variable***. In other words, a discrete variable can assume only certain values with no intermediate values.

A variable that can assume any numerical value over a certain interval or intervals is called a ***continuous variable***.

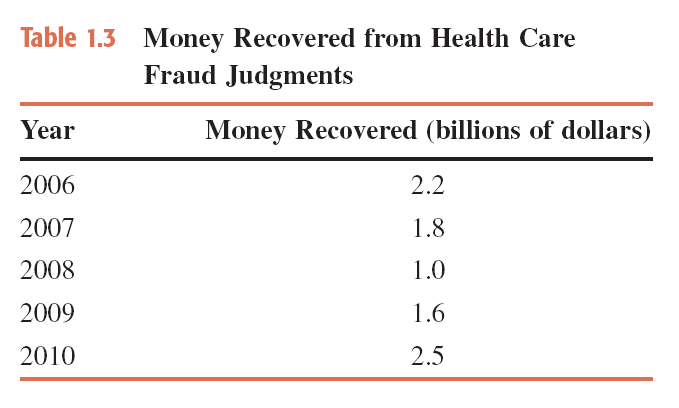
A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called a ***qualitative*** or ***categorical variable***. The data collected on such a variable are called ***qualitative data***.



Data collected on different elements at the same point in time or for the same period of time are called ***cross-section data***.



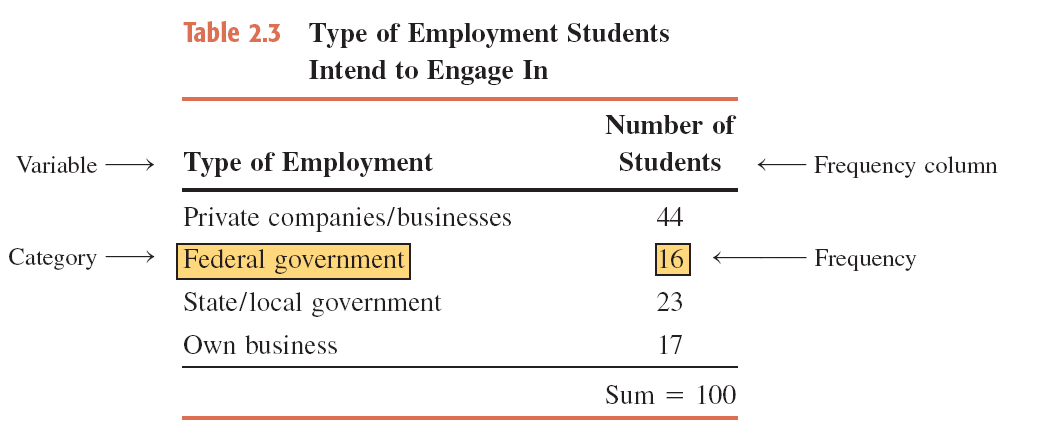
Data collected on the same element for the same variable at different points in time or for different periods of time are called ***time-series data***.

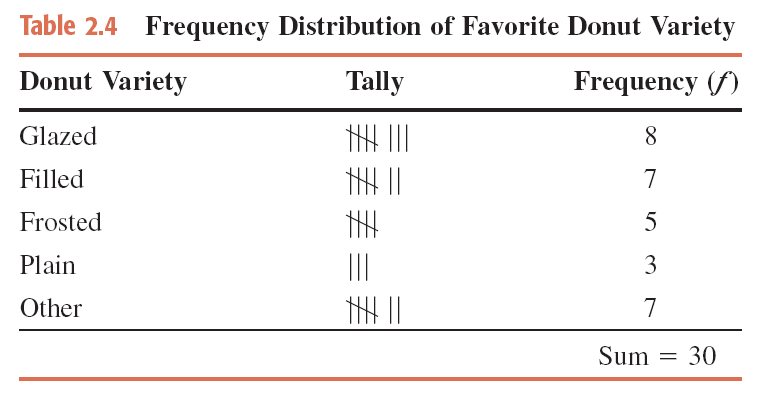


**Chapter 2: Organizing and Graphing Data**

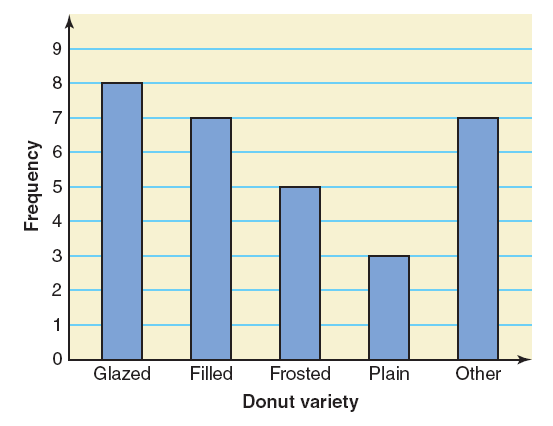
**1) Organizing and Graphing Qualitative Data**

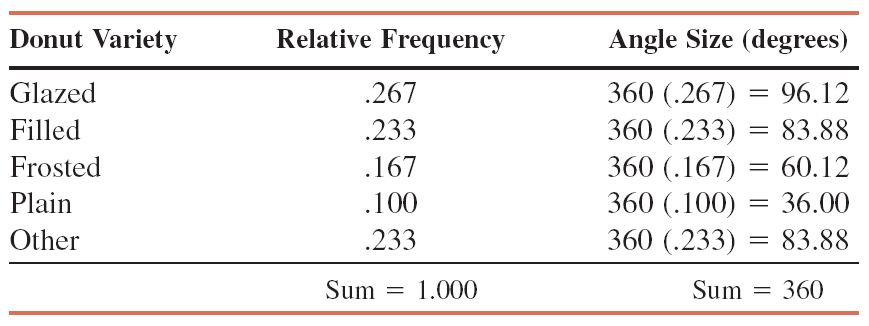
Data recorded in the sequence in which they are collected and before they are processed or ranked are called ***raw data***.



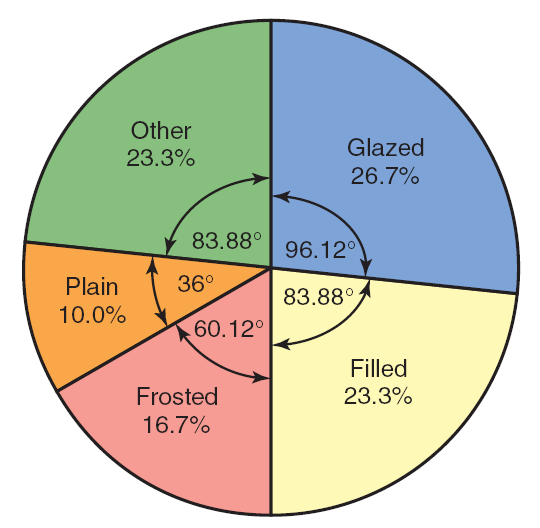


**Bar Graph:**





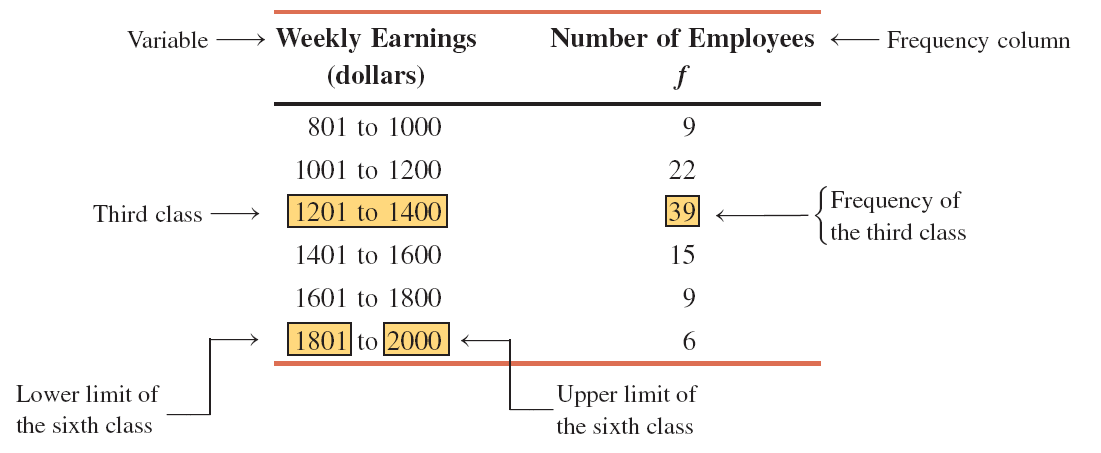
**Pie Chart**



A ***frequency*** ***distribution of a qualitative variable*** lists all categories and the number of elements that belong to each of the categories.

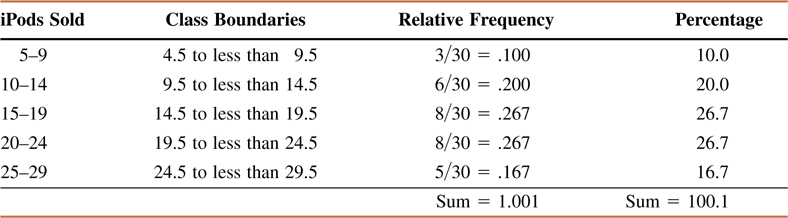
**2) Organizing and Graphing Quantitative Data**

* Frequency Distributions
* Constructing Frequency Distribution Tables
* Relative and Percentage Distributions
* Graphing Grouped Data

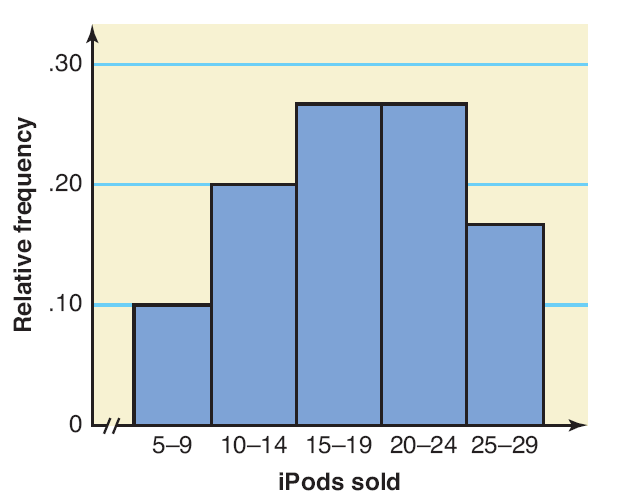


A ***frequency distribution*** ***for quantitative data*** lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called ***grouped data***.

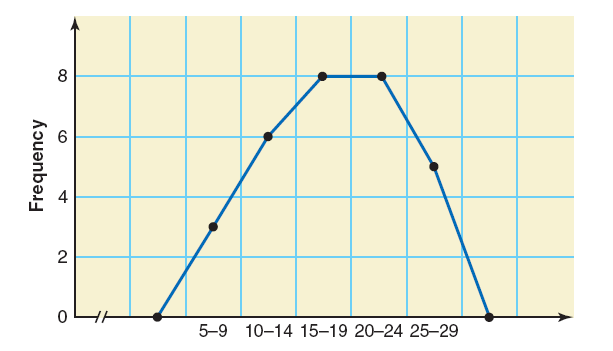




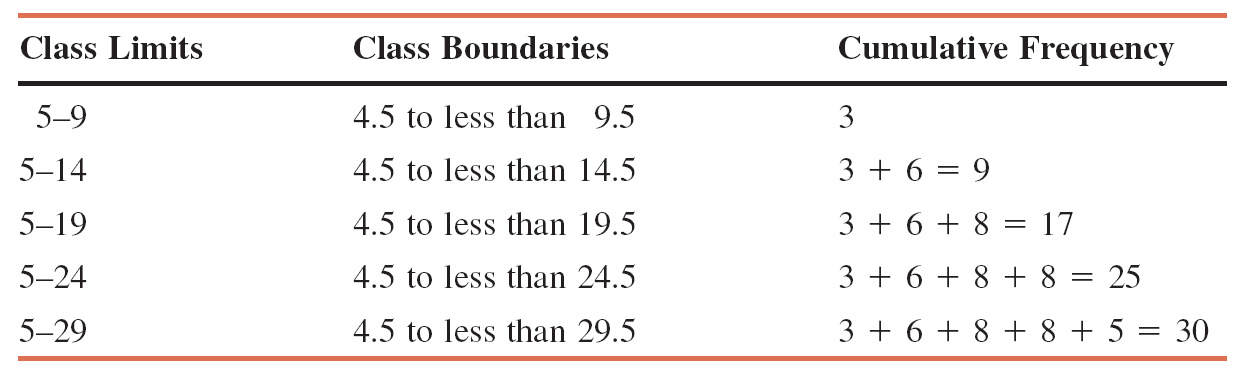
**Relative Frequency Histogram**



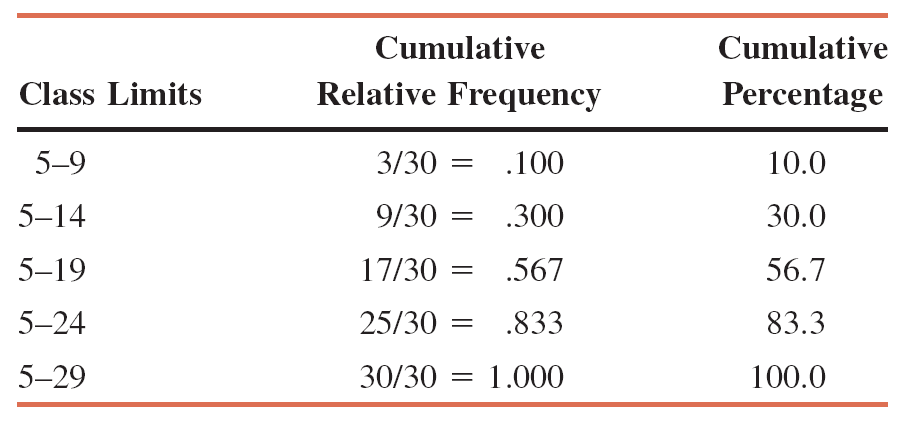
**Polygon**



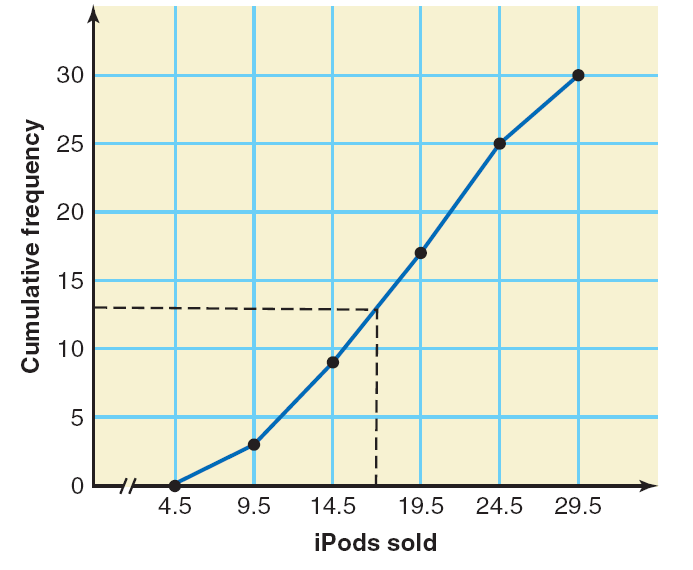
**3) Cumulative Frequency Distributions**







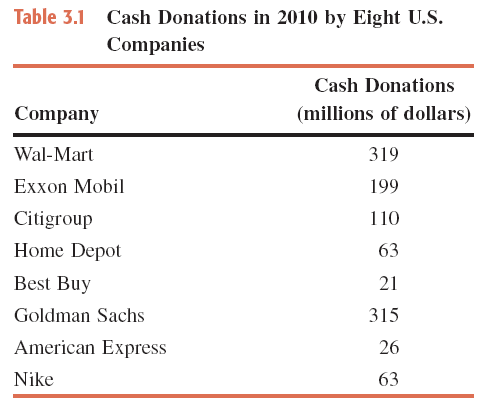
**Ogive for the Cumulative Frequency Distribution**



**Chapter 3: Numerical Descriptive Measures**

**1) Measures of Central Tendency for Ungrouped Data**



****



The population mean is 45.25 years

The ***median*** is the value of the middle term in a data set that has been ranked in increasing order.

The calculation of the median consists of the following two steps:

1) Rank the data set in increasing order.

2) Find the middle term. The value of this term is the median.

**Example:** values are listed below.

173,175 49,723 20,352 10,824 40,911 18,038 61,848

1) First, we rank the given data in increasing order as follows:



Since there are seven homes in this data set and the middle term is the fourth term,

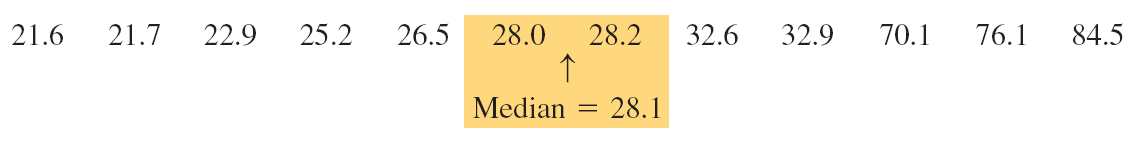
Thus, the median number of homes foreclosed in these seven states was **40,911** in 2010.

**Example:**

First we rank the given total compensations of the 12 CESs as follows:

21.6 21.7 22.9 25.2 26.5 28.0 28.2 32.6 32.9 70.1 76.1 84.5

There are 12 values in this data set. Because there are an even number of values in the data set, the median is given by the average of the two middle values.



The two middle values are the sixth and seventh in the arranged data, and these two values are 28.0 and 28.2.



Thus, the median for the 2010 compensations of these 12 CEOs is **$28.1 million**.

The ***mode*** is the value that occurs with the highest frequency in a data set.

**Example:**

In this data set, 74 occurs twice and each of the remaining values occurs only once. Because 74 occurs with the highest frequency, it is the mode. Therefore,

Mode = **74 miles per hour**

A major shortcoming of the mode is that a data set may have none or may have more than one mode, whereas it will have only one mean and only one median.

Unimodal: A data set with only one mode.

Bimodal: A data set with two modes.

Multimodal: A data set with more than two modes.

**Example:**

Last year’s incomes of five randomly selected families were $76,150, $95,750, $124,985, $87,490, and $53,740.

Because each value in this data set occurs only once, this data set contains **no mode**.

**Example:**

A small company has 12 employees. Their commuting times (rounded to the nearest minute) from home to work are 23, 36, 12, 23, 47, 32, 8, 12, 26, 31, 18, and 28, respectively.

In the given data on the commuting times of the 12employees, each of the values 12 and 23 occurs twice, and each of the remaining values occurs only once. Therefore, that data set has two modes: **12** and **23 minutes**.

**Example:**

The ages of 10 randomly selected students from a class are 21, 19, 27, 22, 29, 19, 25, 21, 22 and 30 years, respectively.

This data set has three modes: **19**, **21** and **22**. Each of these three values occurs with a (highest) frequency of 2.

**2) Measures of Dispersion for Ungrouped Data**

* The standard deviation is the most-used measure of dispersion.
* The value of the standard deviation tells how closely the values of a data set are clustered around the mean.
* In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean.
* In contrast, a larger value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively larger range around the mean.
* The standard deviation is obtained by taking the positive square root of the variance.
* The variance calculated for population data is denoted by *σ²* (read as sigma squared), and the variance calculated for sample data is denoted by *s²*.
* The standard deviation calculated for population data is denoted by *σ*, and the standard deviation calculated for sample data is denoted by *s*.
* Consequently, the standard deviation calculated for population data is denoted by *σ*, and the standard deviation calculated for sample data is denoted by *s*.

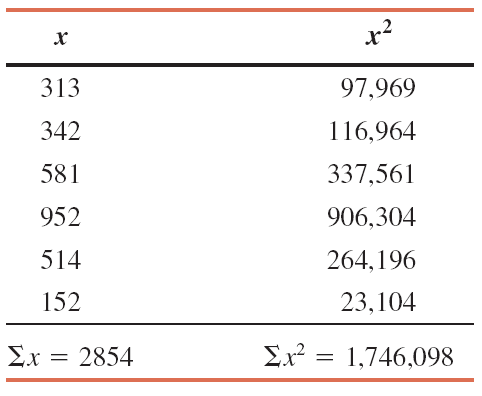
Variance and Standard Deviation



where *σ²* is the population variance, *s²* is the sample variance, *σ* is the population standard deviation, and *s* is the sample standard deviation.

**Example**

Let *x* denote the 2010 baggage fee revenue (in millions of dollars) of an airline.



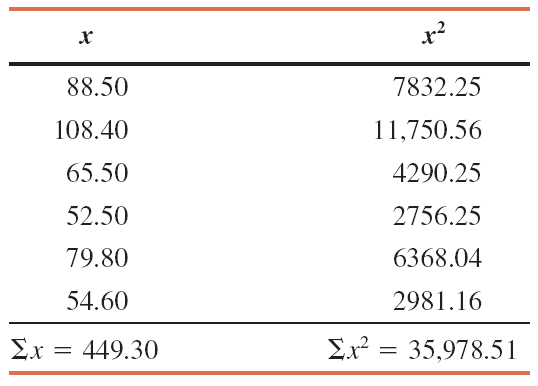


Thus, the standard deviation of the 2010 baggage fee revenues of these six airlines is **$278.76 million**.

**Example**

Following are the 2011 earnings (in thousands of dollars) before taxes for **all** six employees of a small company.

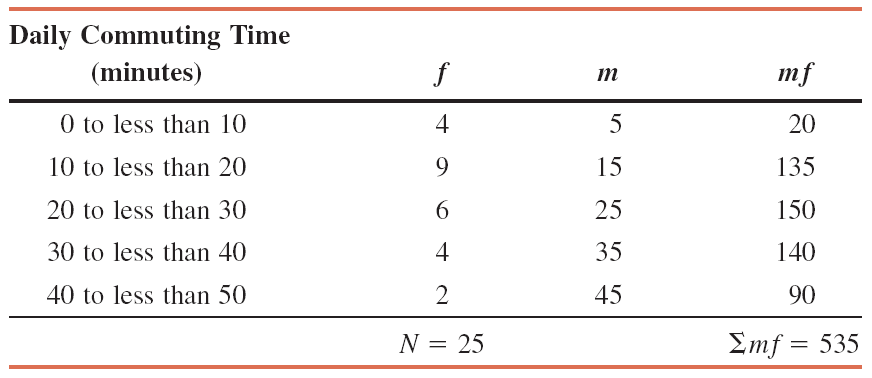
88.50 108.40 65.50 52.50 79.80 54.60





**2) Mean, Variance and Standard Deviation for Grouped Data**

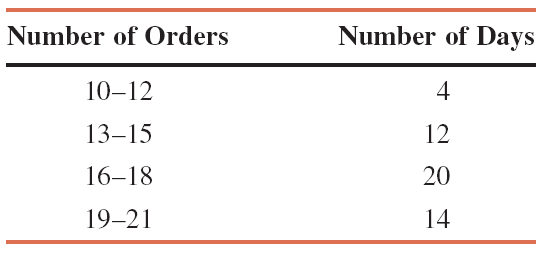






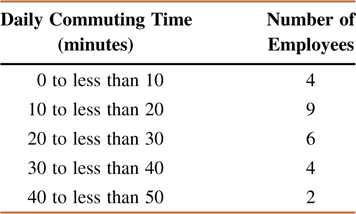
**Example:**

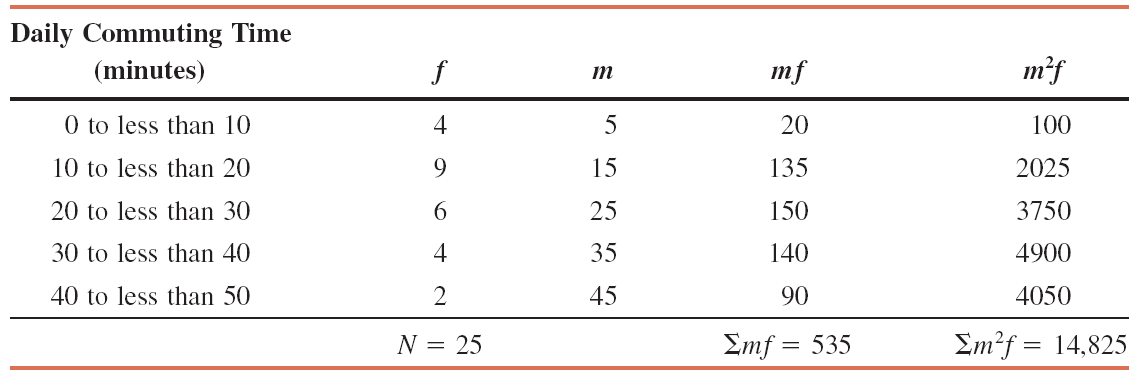
Table gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company.





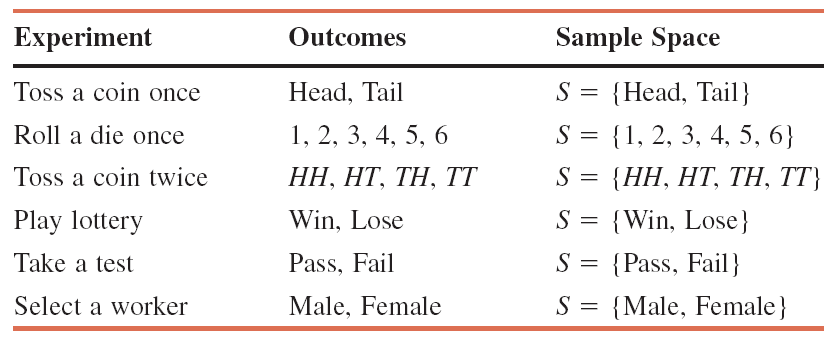
**Example:**

****

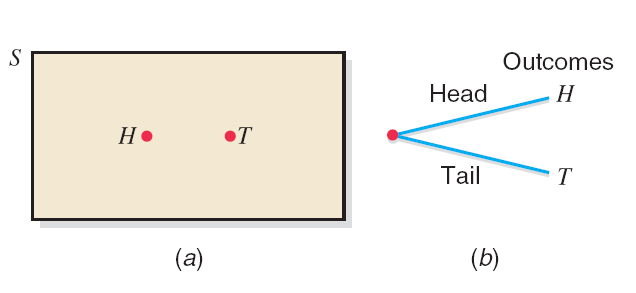




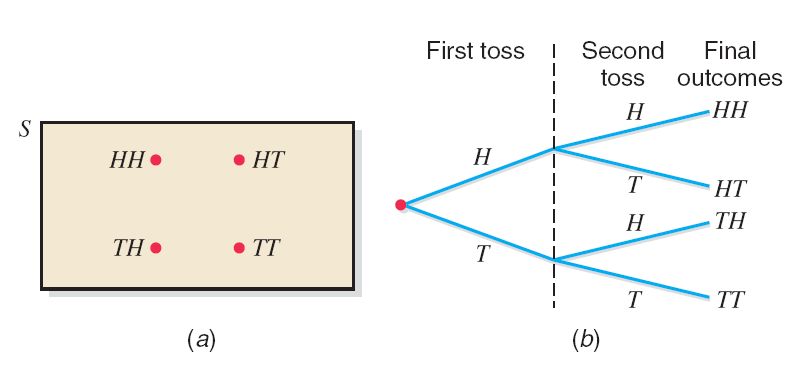
**Chapter 4: Probability**

****

**Venn diagram and Tree Diagram for One Toss of a Coin**



**Venn Diagram and Tree Diagram for two Tosses of a Coin**

****

An ***event*** is a collection of one or more of the outcomes of an experiment.

**Example:**

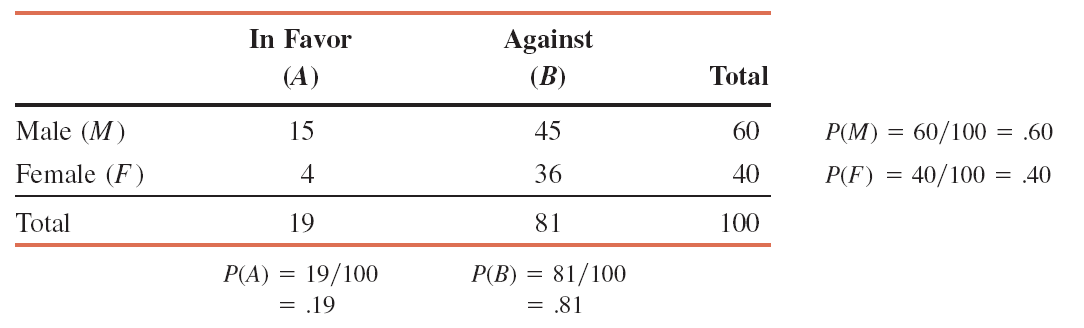
Ten of the 500 randomly selected cars manufactured at a certain auto factory are found to be lemons. Assuming that the lemons are manufactured randomly, what is the probability that the next car manufactured at this auto factory is a lemon?



**Marginal Probability, Conditional Probability and Related Probability Concepts**

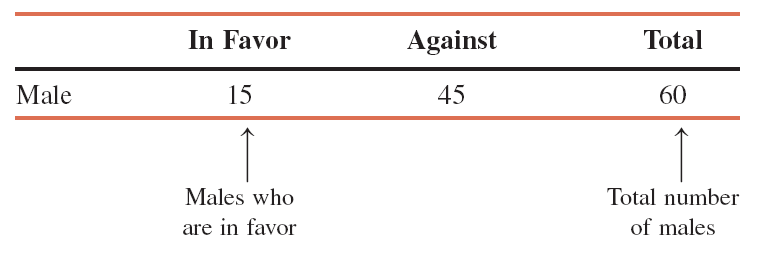
Suppose all 100 employees of a company were asked whether they are in favor of or against paying high salaries to CEOs of U.S. companies. Table gives a two way classification of the responses of these 100 employees.

**Marginal Probabilities**



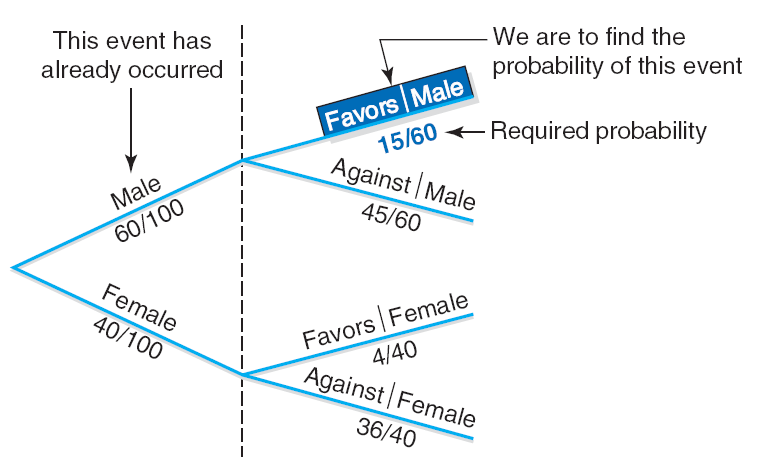
**Conditional Probabilities**

Compute the conditional probability *P* (in favor | male) for the data on 100 employees given in Table

****

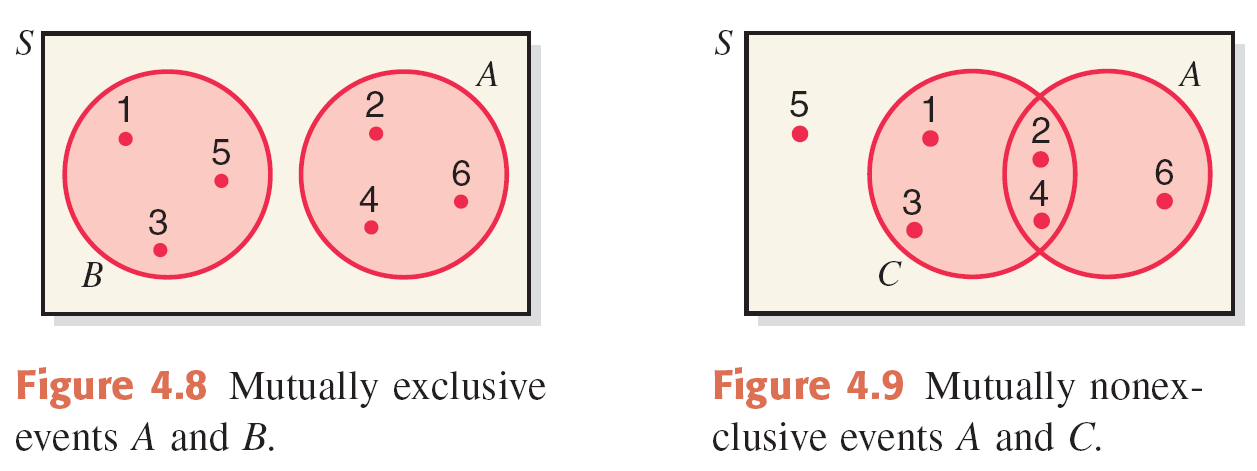


**Tree Diagram**



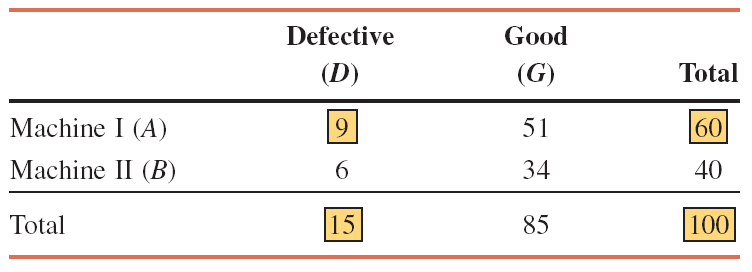
**Mutually Exclusive Events**

Events that cannot occur together are said to be *mutually exclusive events*.



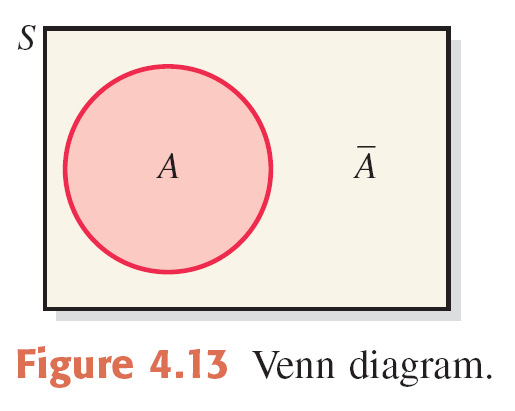
**Independent Events**

**Example:**

****

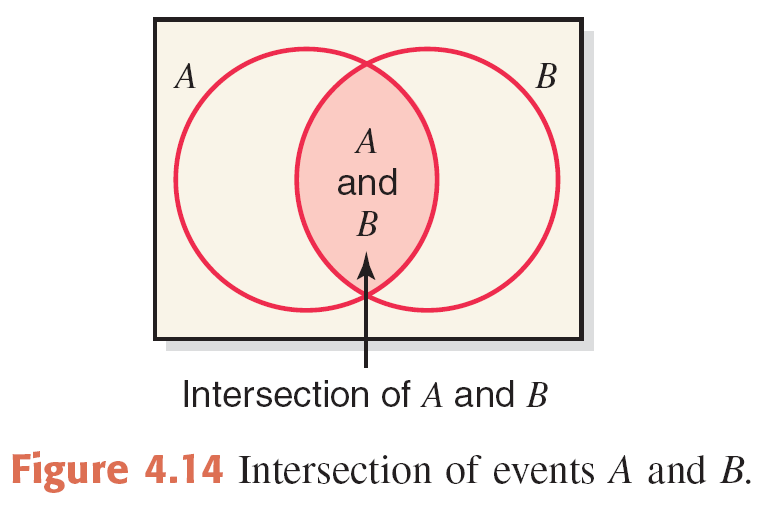
* From the given information,
* *P* (*D*) = 15/100 = .15 and   
   *P* (*D* | *A*) = 9/60 = .15
* *P* (*D*) = *P* (*D* | *A*)
* Consequently, the two events, D and A, are independent.

The ***complement of event A***, denoted by *Ā* and read as “*A* bar” or “*A* complement,” is the event that includes all the outcomes for an experiment that are not in *A*.



* Let *A* and *B* be two events defined in a sample space. The ***intersection*** of *A* and *B* represents the collection of all outcomes that are common to both *A* and *B* and is denoted by

*A* and *B*



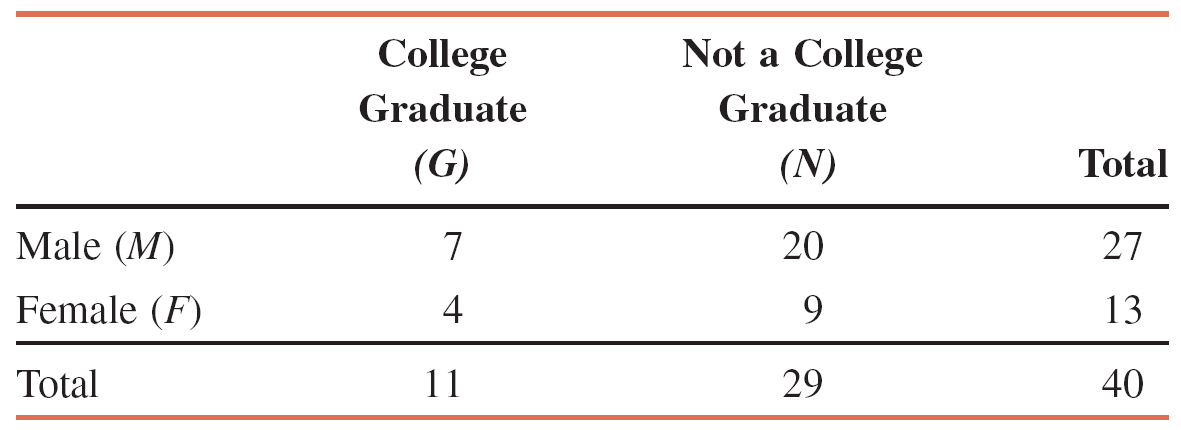
Multiplication Rule to Find Joint Probability

The probability of the intersection of two events *A* and *B* is

*P*(*A* and *B*) = *P*(*A*) *P*(*B* |*A*) = *P*(*B*) *P*(*A* |*B*)

**Example:**

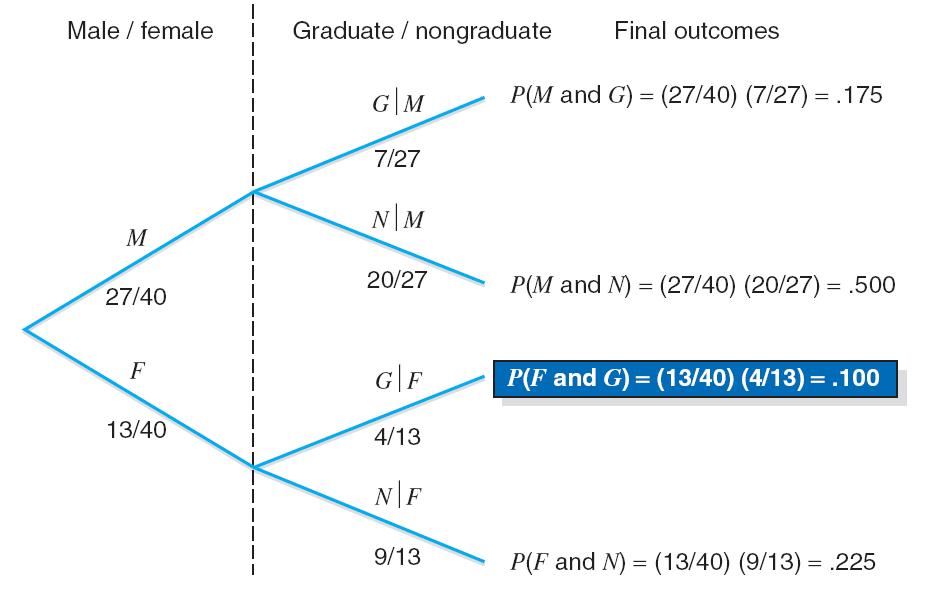
Table gives the classification of all employees of a company given by gender and college degree.



If one of these employees is selected at random for membership on the employee-management committee, what is the probability that this employee is a female and a college graduate?

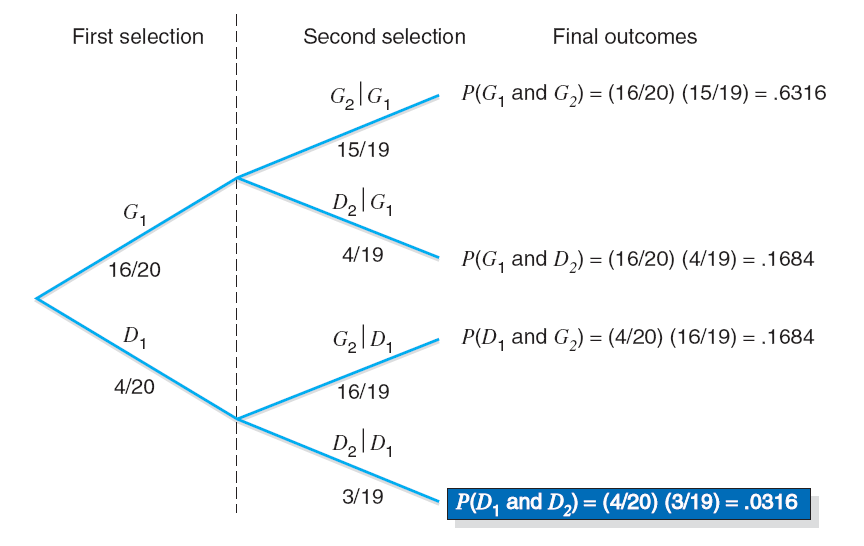
We are to calculate the probability of the intersection of the events *F and G.*

*P*(*F* and *G*) = *P*(*F*) *P*(*G* |*F*)  
 *P*(*F*) = 13/40  
 *P*(*G* |*F*) = 4/13  
 *P*(*F* and *G*) = *P*(*F*) *P*(*G* |*F*) = (13/40)(4/13) = .100



**Example:**

A box contains 20 DVDs, 4 of which are defective. If two DVDs are selected at random (without replacement) from this box, what is the probability that both are defective?



**Calculating Conditional Probability**

If *A* and *B* are two events, then,



given that *P* (*A* ) ≠ 0 and *P* (*B* ) ≠ 0.

**Union of Events**

Let *A* and *B* be two events defined in a sample space. The *union of events* *A* and *B* is the collection of all outcomes that belong to either *A* or *B* or to both *A* and *B* and is denoted by

*A* or *B*

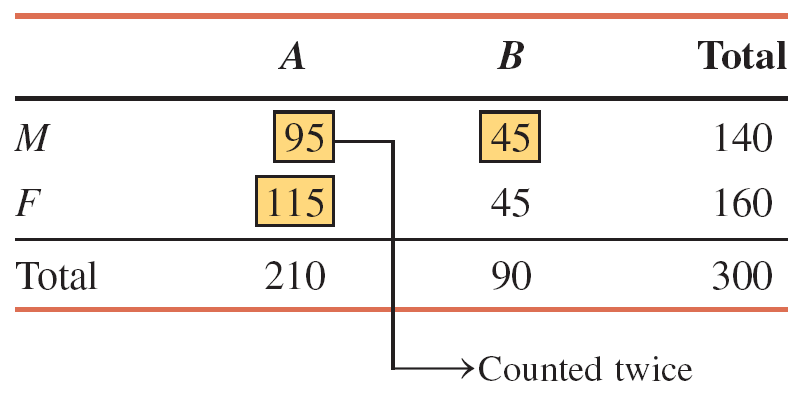
Addition Rule to Find the Probability of Union of Events

The portability of the union of two events *A* and *B* is

*P*(*A* or *B*) *= P*(*A*) *+ P*(*B*) *– P*(*A* and *B*)

**Example:**

A senior citizen center has 300 members. Of them, 140 are male, 210 take at least one medicine on a permanent basis, and 95 are male *and* take at least one medicine on a permanent basis. Describe the union of the events “male” and “take at least one medicine on a permanent basis.”



The probability of the union of two mutually exclusive events *A* and *B* is

*P*(*A* or *B*) = *P*(*A*) + *P*(*B*)

* Let us define the following events:

*M* = a senior citizen is a male

*F* = a senior citizen is a female

*A* = a senior citizen takes at least one medicine

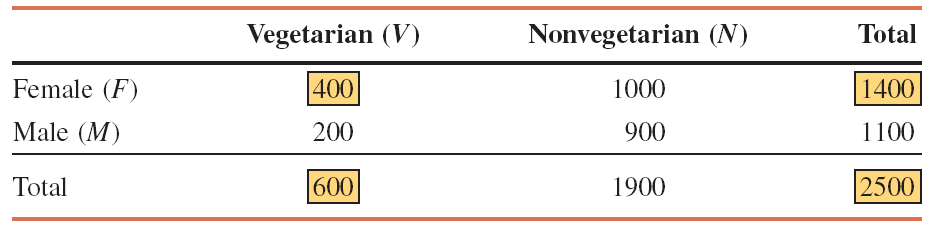
*B* = a senior citizen does not take any medicine

* The union of the events “male” and “take at least one medicine” includes those senior citizens who are either male or take at least one medicine or both. The number of such senior citizens is

140 + 210 – 95 = 255

**Example:**

In a group of 2500 persons, 1400 are female, 600 are vegetarian, and 400 are female and vegetarian. What is the probability that a randomly selected person from this group is a male or vegetarian?





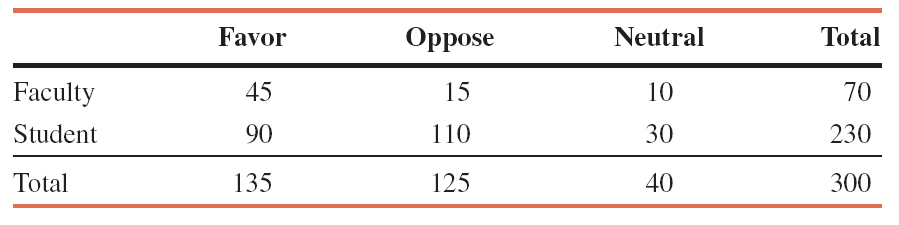
The probability of the union of two mutually exclusive events *A* and *B* is

*P*(*A* or *B*) = *P*(*A*) + *P*(*B*)

**Example:**

A university president has proposed that all students must take a course in ethics as a requirement for graduation. Three hundred faculty members and students from this university were asked about their opinion on this issue. The following table, reproduced from Table 4.9 in Example 4-30, gives a two-way classification of the responses of these faculty members and students.

What is the probability that a randomly selected person from these 300 faculty members and students is in favor of the proposal or is neutral?



Let us define the following events:

*F* = the person selected is in favor of the proposal

*N* = the person selected is neutral

From the given information,

*P*(*F*) = 135/300 = .4500

*P(N*) = 40/300 = .1333

Hence,

*P*(*F* or *N*) = *P*(*F*) + *P*(*N*) = .4500 + .1333 = **.5833**

**Factorials, Combinations and Permutations**

**Factorials**

Definition

The symbol *n!,* read as “*n* factorial,” represents the product of all the integers from *n* to 1. In other words,

*n! = n(n - 1)(n – 2)(n – 3) · · · 3 · 2 · 1*

By definition,

0! = 1

7! = 7 · 6 · 5 · 4 · 3 · 2 · 1 = **5040**

(12-4)! = 8! = 8 · 7 · 6 · 5 · 4 · 3 · 2 · 1

= **40,320**

**Combinations**

The *number of combinations* for selecting *x* from *n* distinct elements is given by the formula



where *n!*, *x!*, and *(n-x)!* are read as “*n factorial,*” “*x factorial,*” “*n minus x factorial*,” respectively.

**Example:**

An ice cream parlor has six flavors of ice cream. Kristen wants to buy two flavors of ice cream. If she randomly selects two flavors out of six, how many combinations are there?

* n = total number of ice cream flavors = 6
* x = # of ice cream flavors to be selected = 2



Thus, there are 15 ways for Kristen to select two ice cream flavors out of six.

**Permutations**

Permutations give the total selections of *x* element from *n* (different) elements in such a way that the order of selections is important. The following formula is used to find the number of permutations or arrangements of selecting *x* items out of *n* items.



which is read as “*the number of permutations of selecting x elements from n elements.*” Permutations are also called ***arrangements***.

**Example:**

A club has 20 members. They are to select three office holders – president, secretary, and treasurer – for next year. They always select these office holders by drawing 3 names randomly from the names of all members. The first person selected becomes the president, the second is the secretary, and the third one takes over as treasurer. Thus, the order in which 3 names are selected from the 20 names is important. Find the total arrangements of 3 names from these 20.



n = total members of the club = 20

x = number of names to be selected = 3

Thus, there are 6840 permutations or arrangements for selecting 3 names out of 20.

**Chapter 5: Binomial And Poison Probability Distribution**

For a binomial experiment, the probability of exactly *x* successes in *n* trials is given by the binomial formula



where

*n* = total number of trials

*p* = probability of success

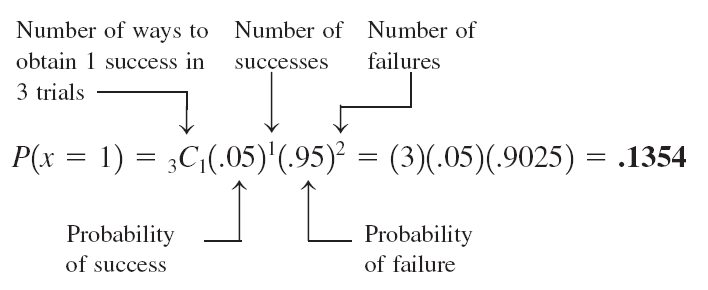
*q* = 1 – *p* = probability of failure

*x* = number of successes in *n* trials

*n* - *x* = number of failures in *n* trials

Example:

Five percent of all DVD players manufactured by a large electronics company are defective. A quality control inspector randomly selects three DVD player from the production line. What is the probability that exactly one of these three DVD players is defective?



**Example:**

At the Express House Delivery Service, providing high-quality service to customers is the top priority of the management. The company guarantees a refund of all charges if a package it is delivering does not arrive at its destination by the specified time. It is known from past data that despite all efforts, 2% of the packages mailed through this company do not arrive at their destinations within the specified time. Suppose a corporation mails 10 packages through Express House Delivery Service on a certain day.

(a) Find the probability that exactly one of these 10 packages will not arrive at its destination within the specified time.

(b) Find the probability that at most one of these 10 packages will not arrive at its destination within the specified time.

a)



Thus, there is a .1667 probability that exactly one of the 10 packages mailed will not arrive at its destination within the specified time.

b) ) At most one of the ten packages is given by the sum of the probabilities of *x* = 0 and *x* = 1



Thus, the probability that at most one of the 10 packages mailed will not arrive at its destination within the specified time is .9838.

**Poisson Probability Distribution**

Poisson Probability Distribution Formula

According to the *Poisson probability distribution*, the probability of *x* occurrences in an interval is



where *λ* (pronounced *lambda*) is the mean number of occurrences in that interval and the value of *e* is approximately 2.71828.

**Example:**

On average, a household receives 9.5 telemarketing phone calls per week. Using the Poisson distribution formula, find the probability that a randomly selected household receives exactly 6 telemarketing phone calls during a given week.



**Example:**

A washing machine in a laundromat breaks down an average of three times per month. Using the Poisson probability distribution formula, find the probability that during the next month this machine will have  
 (a) exactly two breakdowns  
 (b) at most one breakdown



**Chapter 6: The Standard Normal Distribution**

Normal Probability Distribution

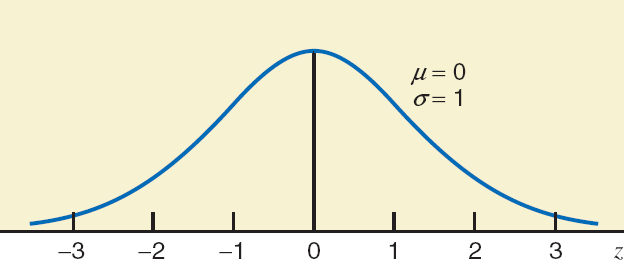
A *normal probability distribution* , when plotted, gives a bell-shaped curve such that:

1. The total area under the curve is 1.0.

2. The curve is symmetric about the mean.

3. The two tails of the curve extend indefinitely**.**

The normal distribution with μ = 0 and σ = 1 is called the *standard normal distribution*.

****

**Example:**

Find the area under the standard normal curve from *z* = -2.17 to *z* = 0.

Area from -2.17 to 0 = *P*(-2.17≤ *z* ≤ 0)

= .5000 - .0150 = **.4850**

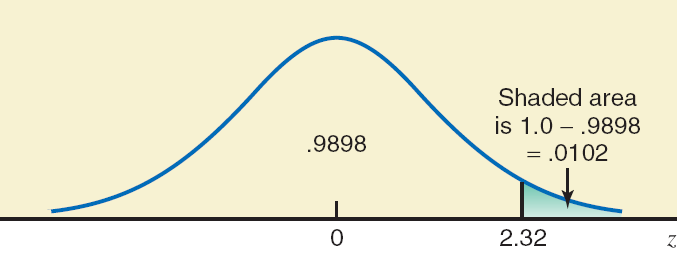
**Example:**

Find the following areas under the standard normal curve.

(a) Area to the right of z = 2.32

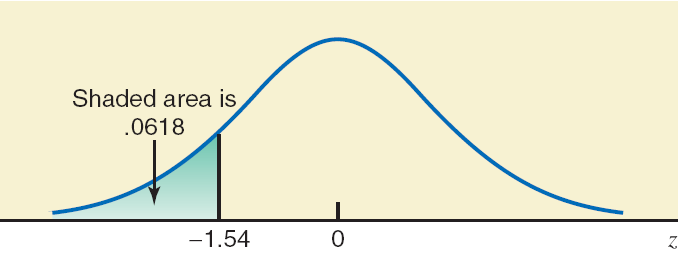
(b) Area to the left of z = -1.54

(a) To find the area to the right of z=2.32, first we find the area to the left of z=2.32. Then we subtract this area from 1.0, which is the total area under the curve. The required area is 1.0 - .9898 = .0102.



(b) To find the area under the standard normal curve to the left of z=-1.54, we find the area in Table IV that corresponds to -1.5 in the z column and .04 in the top row. This area is .0618.

Area to the left of -1.54= *P* (*z* < -1.54) = .0618



**Example:**

Find the following probabilities for the standard normal curve.

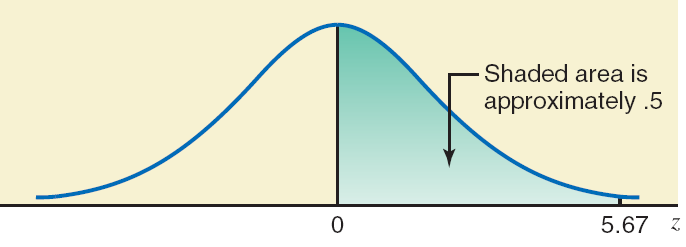
(a) *P* (0 < *z* < 5.67)

(b) *P* (*z* < -5.35)

(a) *P* (0 < *z* < 5.67) = Area between 0 and 5.67

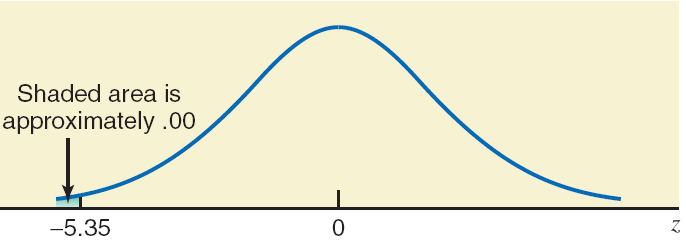
= 1.0 - .5

= **.5** approximately



(b) *P* (*z* < -5.35) = Area to the left of -5.35

= .00 approximately



**Standardizing a Normal Distribution**

Converting an *x* Value to a *z* Value

For a normal random variable *x*, a particular value of *x* can be converted to its corresponding *z* value by using the formula



where *μ* and σ are the mean and standard deviation of the normal distribution of *x*, respectively.

**Example:**

Let *x* be a continuous random variable that has a normal distribution with a mean of 50 and a standard deviation of 10. Convert the following *x* values to *z* values and find the probability to the left of these points.

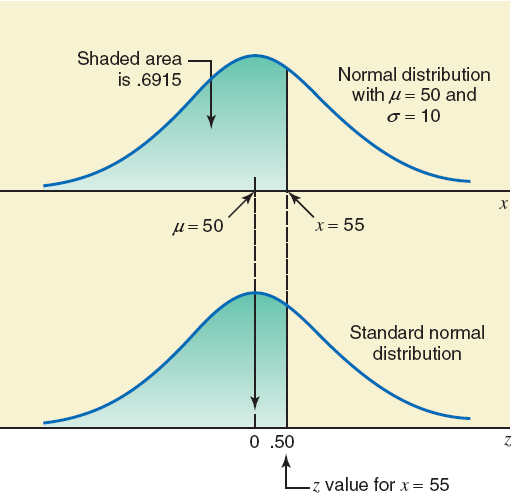
(a) x = 55

(b) x = 35

(a) x = 55



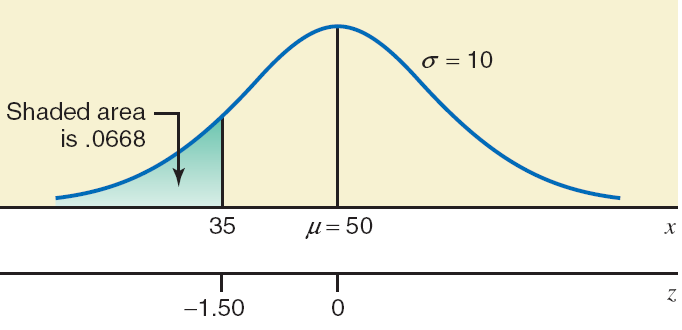
P(x < 55) = P(*z* < .50) = .6915



(b) x = 35



P(x < 35) = P(*z* < -1.50) = .0668

****

**Example:**

Let *x* be a continuous random variable that is normally distributed with a mean of 25 and a standard deviation of 4.  
Find the area

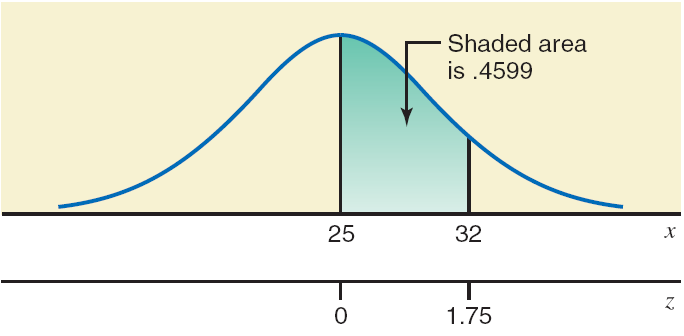
(a) between *x* = 25 and *x* = 32

(b) between *x* = 18 and *x* = 34

(a) The *z* value for *x* = 25 is 0

The *z* value for *x* = 32 is

*P* (25 < *x* < 32) = *P*(0 < *z* < 1.75)= .9599 - .5000 = .4599

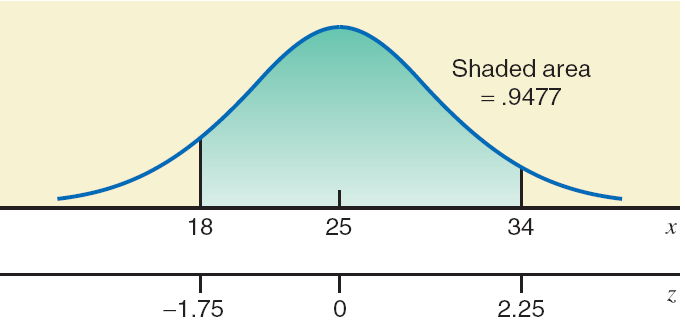


(b) For *x* = 18:



For *x* = 34:

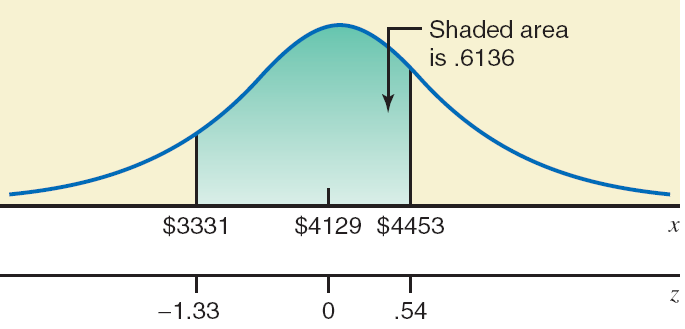
*P* (18 < *x* < 34) = *P* (-1.75 < *z* < 2.25)= .9878 - .0401 = .9477



**Applications of the Normal Distribution:**

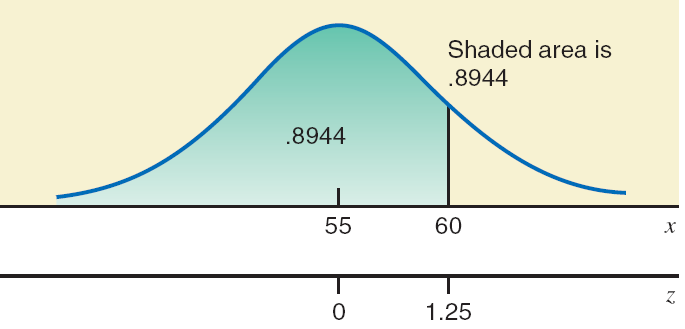
**Example:**

According to the Kaiser Family Foundation, U.S. workers who had employer-provided health insurance paid an average premium of $4129 for family coverage during 2011 (*USA TODAY*, October 10, 2011). Suppose that the premiums for family coverage paid this year by all such workers are normally distributed with a mean of $4129 and a standard deviation of $600. Find the probability that such premium paid this year by a randomly selected such worker is between $3331 and $4453.



**Example:**

A racing car is one of the many toys manufactured by Mack Corporation. The assembly times for this toy follow a normal distribution with a mean of 55 minutes and a standard deviation of 4 minutes. The company closes at 5 p.m. every day. If one worker starts to assemble a racing car at 4 p.m., what is the probability that she will finish this job before the company closes for the day?



**Example:**

Hupper Corporation produces many types of soft drinks, including Orange Cola. The filling machines are adjusted to pour 12 ounces of soda into each 12-ounce can of Orange Cola. However, the actual amount of soda poured into each can is not exactly 12 ounces; it varies from can to can. It has been observed that the net amount of soda in such a can has a normal distribution with a mean of 12 ounces and a standard deviation of .015 ounce.

(a) What is the probability that a randomly selected can of Orange Cola contains 11.97 to 11.99 ounces of soda?

(b) What percentage of the Orange Cola cans contain 12.02 to 12.07 ounces of soda?

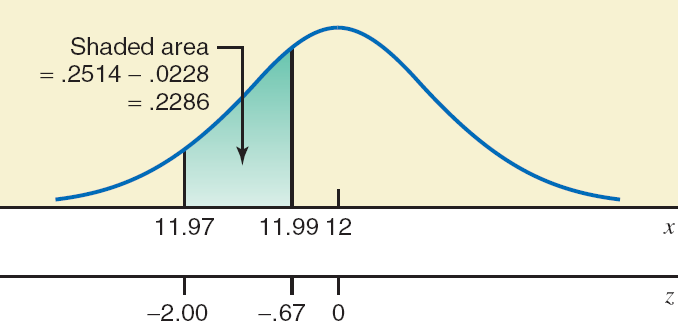
(a) For *x* = 11.97:



For *x* = 11.99:



*P* (11.97 ≤ *x* ≤ 11.99) = *P* (-2.00 ≤ *z* ≤ -.67) = .2514 - .0228 = .2286

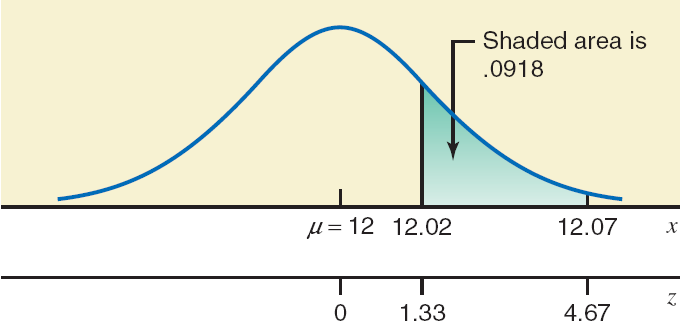


(b) For *x* = 12.02:



For *x* = 12.07:

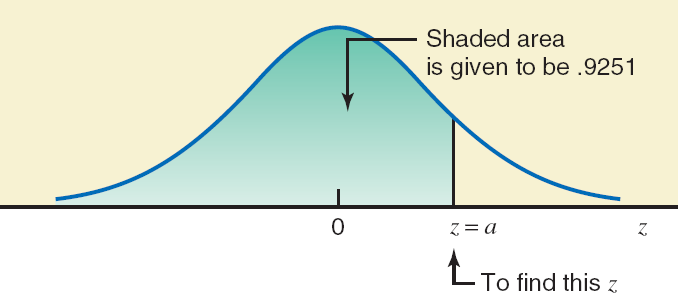
*P* (12.02 ≤ *x* ≤ 12.07) = *P* (1.33 ≤ *z* ≤ 4.67)= 1 - .9082 = .0918

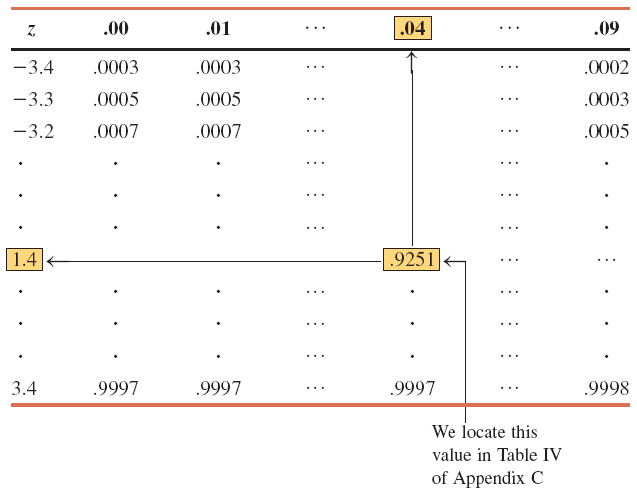


**Determining z and x values when the area under the normal curve is known.**

Example:

Find a point *z* such that the area under the standard normal curve to the left of *z* is .9251.





**Example:**

Find the value of *z* such that the area under the standard normal curve in the right tail is .0050.

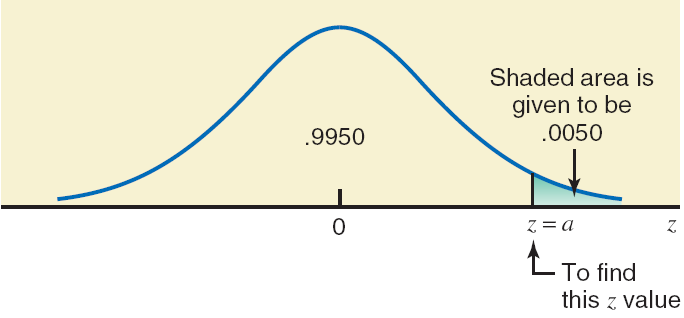
Area to the left of *z* = 1.0 - .0050 = .9950

Look for .9950 in the body of the normal distribution table. Table does not contain .9950.

Find the value closest to .9950, which is either .9949 or .9951.

If we choose .9951, the *z* = **2.58**.

If we choose .9949, the *z* = **2.57**.



**Example:**

Find the value of *z* such that the area under the standard normal curve in the left tail is .05.

Because .05 is less than .5 and it is the area in the left tail, the value of *z* is negative.

Look for .0500 in the body of the normal distribution table. The value closest to .0500 in Table is either .0505 or .0495.

If we choose .0495, the z = -1.65.

