""The following content is intended solely for the educational use of the individual user.""**

These digital images may not be copied, retained, printed, shared, modified or otherwise used without the copyright holder's express written permission, except as permitted for educational uses.
Part 1

DESCRIPTIVE STATISTICS

The word statistics comes from the Latin status, meaning "state." This derivation reflects that numerical information was historically gathered and interpreted by governments.

The ability to describe data in various ways has always been valuable. Now, however, these techniques are becoming more important because of the tremendous, ever-growing number of data being collected in almost all areas of knowledge. The need to make sense out of masses of information has led to the development of formalized ways of describing data. To make intelligent decisions, as well as to understand the world around us, we must become familiar with these primary methods of portraying data.

Given a raw set of data, often we can detect no overall pattern. Perhaps some values occur more frequently, a few extreme values may stand out, and usually the range of values is apparent. The presentation of data, including summarizations and descriptions, and involving such concepts as representative or average values, measures of dispersion, positions of various values, and the shape of a distribution, falls under the broad topic of descriptive statistics, the subject of Part 1 of this book. This aspect of statistics is in contrast to analysis, the process of drawing inferences from limited data, a subject that is discussed in later chapters.

More specifically, Chapters 1–4 are concerned with (1) how to measure central tendency; (2) how to measure variability; (3) how to measure position; and (4) how to display shape.
1
CENTRAL TENDENCY

The word *average* is used in phrases common to everyday conversation. People speak of bowling and batting averages or the average life expectancy of a battery or a human being. Actually the word *average* is derived from the French *avarie*, which refers to the money shippers used to contribute to help compensate for the losses suffered by other shippers whose cargo did not arrive safely (that is, the losses were shared, with everyone contributing an *average* amount). In common usage *average* has come to mean a *representative* score or a *typical* value or the *center* of a distribution. Mathematically, there are a variety of ways to define the average of a set of data. In practice, we use whichever method is most appropriate for the particular case under consideration.

In the following paragraphs we consider the three primary ways of denoting an average:

1. The *median*, which is the middle number of a set of numbers arranged in numerical order.
2. The *mode*, which is the number that occurs most frequently.
3. The *mean*, which is found by summing items in a set and then dividing by the number of items.

**Example 1.1** Consider the following set of homerun distances (in feet) to center field in 13 ballparks: 387, 400, 400, 410, 410, 410, 414, 415, 420, 420, 421, 457, 461. What is the average?

**Answer:** The median is 414 (there are six values below 414 and six values above). The mode is 410 (occurs three times). The mean is

$$\frac{387 + 400 + 400 + 410 + \ldots + 457 + 461}{13} = 417.3 \text{ feet}$$

**MEDIAN**

The word *median* is derived from the Latin *medius* which means "middle." The values under consideration are arranged in ascending or descending order. If there is an odd number of values, the median is the middle one. If there are an even number, the median is found by adding the two middle values and
MODE

The mode, or most frequent value, is an easily understood representative score. It is clear what is meant by “The most common family size is four” or “That professor gives out more Bs than any other grade.” If exactly one value must be chosen as a basis for certain decisions, this value is often the mode. For example, if a supplier can expand production of only one single item, he may well decide upon the one most frequently sold. However, since the mode presents the typical result, its use tends to rule out further arithmetic calculations based on this result. For example, knowing that a certain player is the most popular with the fans does not lead to calculations based simply on this knowledge.

When two numbers in a set have equal frequency, and this frequency is higher than any other, we say that the set is bimodal. Even when one frequency is greater than the other, but both are above the other frequencies, the set will sometimes be termed bimodal. Occasionally, the values in a set are distributed so evenly that the idea of a modal value has very little meaning for that set.

MEAN

While the median and mode are often useful in descriptive statistics, the mean or, more accurately, the arithmetic mean, is most important for statistical inference and analysis. Also, for the layperson, the “average” is usually understood to be the “mean.”

The mean of a whole population (the complete set of items of interest) is often denoted by the Greek letter μ (mu), while the mean of a sample (a part of a population) is often denoted by x̄. For example, the mean value of the set of all houses in the United States might be μ = $56,400, while the mean value of 100 randomly chosen houses might be x̄ = $52,100 or perhaps x̄ = $63,800 or even x̄ = $124,000.

In statistics we learn how to estimate a population mean from a sample mean. Throughout this book, the word sample implies a random sample, that is, a sample selected in such a way that each element of the population has an equal chance to be included in the sample. In the real world, this process of random selection is often very difficult to achieve.

Mathematically, the mean = (Σx)/n, where Σx represents the sum of all the elements of the set under consideration, and n is the actual number of elements. Σ is the uppercase Greek letter sigma.

Unlike the mode or median, the mean is sensitive to a change in any value, as shown in Example 1.2.

Example 1.2

Suppose that the numbers of unnecessary procedures recommended by five doctors in a 1-month period are given by the set (2, 2, 8, 20, 33). Note that the median is 8, the mode is 2, and the mean is \( \frac{2 + 2 + 8 + 20 + 33}{5} = 13 \). If it is discovered that the fifth doctor recommended an additional 25 unnecessary procedures, how are the median, mode and mean affected?

Answer: The set is now (2, 2, 8, 20, 58). The median is still 8 and the mode is still 2. The mean, however, changes to \( \frac{2 + 2 + 8 + 20 + 58}{5} = 18 \).

Adding the same constant to each value will increase the mean by a like amount. Similarly, multiplying each value by the same constant will multiply the mean by a like amount. (See Study Questions 1 and 2.)

Example 1.3

Suppose the salaries of six employees are $3000, $7000, $15,000, $22,000, $23,000, and $38,000.

a. What is the mean salary?

Answer:

\[
\frac{3000 + 7000 + 15,000 + 22,000 + 23,000 + 38,000}{6} = 18,000.
\]
b. What will the new mean salary be if everyone receives a $3000 increase?

**Answer:**

\[ \frac{6000 + 10,000 + 18,000 + 25,000 + 26,000 + 41,000}{6} = 21,000. \]

Note that $18,000 + $3000 = $21,000.

c. What if everyone receives a 10% raise?

**Answer:**

\[ \frac{3300 + 7700 + 16,500 + 24,200 + 25,300 + 41,800}{6} = 19,800. \]

Note that 110% of $18,000 is $19,800.

Because many real-life applications of statistics involve comparisons of two populations, it is important for later use to note how the mean is calculated for a set of differences.

**Example 1.4** Suppose set \( X = \{2, 9, 11, 22\} \) and set \( Y = \{5, 7, 15\} \). Note that the mean of set \( X \) is \( \mu_X = \frac{2 + 9 + 11 + 22}{4} = 11 \) and the mean of set \( Y \) is \( \mu_Y = \frac{5 + 7 + 15}{3} = 9 \). Form the set \( Z \) of differences by subtracting each element of \( Y \) from each of \( X \):

\[ Z = \{2-5, 2-7, 2-15, 9-5, 9-7, 9-15, 11-5, 11-7, 11-15, 22-5, 22-7, 22-15\} = \{-3, -5, -13, 4, 2, -6, 6, 4, -4, 17, 15, 7\} \]

What is the mean of \( Z \)?

**Answer:**

\[ \mu = \frac{-3 - 5 - 13 + 4 + 2 - 6 + 6 + 4 - 4 + 17 + 15 + 7}{12} = \frac{24}{12} = 2 \]

Note that \( \mu_Z = \mu_X - \mu_Y \)

In general, the mean of a set of differences is equal to the difference of the means of the two original sets. Even more generally, if a sum is formed by picking one element from each of several sets, the mean of such sums is simply the sum of the means of the various sets.

Three secondary procedures for denoting an average are the **harmonic mean**, the **geometric mean**, and the **trimmed mean**.

### Harmonic Mean

The **harmonic mean** of a set of \( n \) numbers is defined to be

\[ H = \frac{n}{\sum \frac{1}{x}} \]

that is, \( n \) divided by the sum of the reciprocals of the numbers.

**Example 1.5** Suppose four kinds of nails cost $0.01, $0.02, $0.03, and $0.04 apiece. What is the average cost of a nail if we purchase $3.00 worth of each kind?

**Answer:** The number of $0.01 nails purchased is $3.00/$0.01 = 300. Similarly, the numbers of $0.02, $0.03, and $0.04 nails purchased are $3.00/$0.02 = 150, $3.00/$0.03 = 100, and $3.00/$0.04 = 75, respectively. The total number of nails purchased is $300 + 150 + 100 + 75 = 625$, while the total amount spent is $4($3.00$) = $12.00. Thus the average cost per nail is $12.00/625 = $0.0192.

Note that the answer could have been calculated immediately as the **harmonic mean** of the set \{0.01, 0.02, 0.03, 0.04\}:

\[ \frac{1}{0.01} + \frac{1}{0.02} + \frac{1}{0.03} + \frac{1}{0.04} = 0.0192 \]

Also note that if we bought 100 of each of the four kinds of nails (that is, same number of each kind rather than same money spent on each), then the average spent per nail would be \((0.01 + 0.02 + 0.03 + 0.04)/4 = $0.025.\)

### Geometric Mean

The **geometric mean** of a set \( \{y_1, y_2, ..., y_n\} \) is defined to be

\[ \sqrt[n]{y_1 y_2 \cdots y_n}, \]

that is, the \( n \)th root of the product of the elements. In particular, the geometric mean of two numbers is the square root of their product. The geometric mean is appropriate in situations of what is known as geometric growth.
Example 1.6 The population of the United States in 1820 was 9,638,453 and in 1840 was 17,069,453. Estimate the population in 1830 using the geometric mean.

**Answer:**

\[
\sqrt[3]{(9,638,453)(17,069,453)} = 12,826,657
\]

(The actual value was 12,866,020.)

**TRIMMED MEAN**

The trimmed mean represents an attempt to reduce the influence of extreme values (outliers) on the mean. It is calculated by arranging the terms in numerical order, deleting the first quarter and the fourth quarter of the values, and then finding the arithmetic mean of what remains. (Alternatively, instead of 25%, any percent, for example 5% or 10%, can be taken away from both the upper and lower ends, depending on the makeup of the particular set under consideration.)

Example 1.7 The World Bank has predicted that the 20 most populous countries in the year 2100 and their populations (in millions) will be as follows:

India, 1632; China, 1571; Nigeria, 509; Russia, 376; Indonesia, 356; Pakistan, 316; United States, 309; Bangladesh, 297; Brazil, 293; Mexico, 196; Ethiopia, 173; Vietnam, 168; Iran, 164; Zaire, 139; Japan, 128; Philippines, 125; Tanzania, 120; Kenya, 116; Burma, 112; and Egypt, 111. Calculate the mean and the trimmed mean for the average predicted populations of these countries.

**Answer:** The mean is \((1632 + ... + 111)/20 = 361\) million, while the trimmed mean, without the extreme influence of India, China, and Nigeria, is \((316 + ... + 128)/10 = 218\) million.

**EXERCISES**

1. In the second game of the 1989 World Series between Oakland and San Francisco (played 2 days before the northern California earthquake postponed the series), ten players went hitless, eight players had one hit apiece, and one player (Rickey Henderson) had three hits. What was the mean, median, and mode number of hits?

2. In 1993 the eight countries in which the most patents were filed were as follows: Japan, 208,347; United States, 57,890; Germany, 42,922; France, 111,187; Britain, 9333; China, 5566; Italy, 3726; and Australia, 3315 (The Economist, August 27, 1994). What are the median, mean, and trimmed mean averages for the number of 1993 patents in these eight countries?

3. On August 23, 1994, the 5-day total returns for equal investments in foreign stocks, U.S. stocks, gold, money market funds, and treasury bonds averaged 0.086% (Business Week, September 5, 1994). If the respective returns for foreign stocks, U.S. stocks, gold, and money market funds were +0.70%, −0.11%, +1.34%, and +0.05%, what was the return for treasury bonds?

4. The weights (in pounds) of the members of the 1968 Chicago Bear offensive line were as follows: Bob Pickens, 255; Jim Cadile, 240; Mike Pyle, 250; George Seals, 270; Rufus Mayes, 260. What was the average (mean) weight of these linemen? What would the average weight have been if each of the players had gained 10 pounds? If each had increased his weight by 10 percent?

5. Fecal pollution can lead to illness among swimmers in small lakes. The Environmental Protection Agency recommends that enterococcus counts at freshwater swimming areas not exceed a geometric mean of 35 per deciliter (EPA report no. 440/5-84-002). If three samples of lake water taken at different sections of a swimming area show pathogen counts of 25, 30, and 40 organisms per deciliter, determine whether the geometric mean exceeds the EPA limit.

6. The results of a recent University of Chicago study reported in The New York Times are tabulated below:

<table>
<thead>
<tr>
<th>SEX PARTNERS SINCE AGE 18</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<tr>
<td>Men</td>
</tr>
<tr>
<td>3%</td>
</tr>
<tr>
<td>Women</td>
</tr>
<tr>
<td>3%</td>
</tr>
</tbody>
</table>

In which category does the median number of partners fall for men? For women? Explain.

7. In your daily newspaper, find examples of two sets of numbers, in one of which the mean is greater than the median, and in the other the mean is less than the median. Can you find an example of data where the mean and median are equal?

**Study Questions**

1. Show that the addition of the same constant to each value in a set will increase the mean by a like amount. (Hint: Suppose that the
set has the \( n \) elements: \( x_1, x_2, x_3, \ldots, x_n \). Then the mean is:
\[
\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}
\]
Show that the mean of the set \( \{x_1 + c, x_2 + c, x_3 + c, \ldots, x_n + c\} \) is \( \mu + c \).

2. Prove that multiplying each value in a set by the same constant will multiply the mean by a like amount.

3. Questions 1 and 2 form the background for a technique called \textit{coding} which in certain cases provides a shortcut for computing the mean. Using these addition and multiplication principles, how would you mentally calculate the mean of the set \( \{25.002, 25.015, 25.013, 25.008, 25.009, 25.001\} \)?

4. Suppose \( X = \{x_1, x_2, \ldots, x_m\} \) and \( Y = \{y_1, y_2, \ldots, y_n\} \). Form the set of differences \( Z = \{x_i - y_j\}, i = 1, 2, \ldots, m \) and \( j = 1, 2, \ldots, n \). Show that \( \mu_Z = \mu_X - \mu_Y \).

[\textit{Hint:} \( \sum z = n \sum x - m \sum y \). \textit{Why? Then} \( \frac{\sum z}{mn} = ? \)]

2

\section*{VARIABILITY}

In describing a set of numbers, not only is it useful to designate an average value, but also it is important to be able to indicate the \textit{variability} or the \textit{dispersion} of the measurements. A producer of time bombs aims for small variability—it would not be good if his 30 minute fuses actually had a range of 10 to 50 minutes before detonation. On the other hand, a teacher interested in distinguishing the better from the poorer students aims to design exams with large variability in results—it would not be helpful if all her students scored exactly the same. The players on two basketball teams may have the same average height, but this fact doesn't tell the whole story. If the dispersions are quite different, one team may have a 7-foot player, whereas the other has no one over 6 feet tall. Two Mediterranean holiday cruises may advertise the same average age for their passengers. One, however, may have only passengers between 20 and 25 years old, while the other has only middle-aged parents in their 40s together with their children under age 10.

There are four primary ways of describing variability or dispersion:

1. The \textit{range}, which is the difference between the largest and smallest values.

2. The \textit{average deviation}, which is found by averaging the absolute differences of all the values from the mean.

3. The \textit{variance}, which is determined by averaging the squared differences of all the values from the mean.

4. The \textit{standard deviation}, which is the square root of the variance.

\textbf{Example 2.1} The monthly rainfall in Monrovia, Liberia, where May through October is the so-called rainy season and November through April the dry season, is as follows:

<table>
<thead>
<tr>
<th>Month</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rain (in.)</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>18</td>
<td>37</td>
<td>31</td>
<td>16</td>
<td>28</td>
<td>24</td>
<td>9</td>
<td>4</td>
</tr>
</tbody>
</table>

The mean is
\[
\frac{1 + 2 + 4 + 6 + 18 + 37 + 31 + 16 + 28 + 24 + 9 + 4}{12} = 15 \text{ inches}
\]
What are the measures of variability?

**Answer:** Range: The maximum is 37 inches (June), and the minimum is 1 inch (January). Thus the range is $37 - 1 = 36$ inches of rain.

Average deviation:

$$\frac{|1 - 15| + |2 - 15| + |4 - 15| + |6 - 15| + |18 - 15| + \ldots + |4 - 15|}{12}$$

$$= \frac{14 + 13 + 11 + 9 + 3 + 22 + 16 + 1 + 13 + 9 + 6 + 11}{12} = 10.7 \text{ inches}$$

Variance:

$$\frac{14^2 + 13^2 + 11^2 + 9^2 + 3^2 + 22^2 + 16^2 + 1^2 + 13^2 + 9^2 + 6^2 + 11^2}{12} = 143.7$$

Standard deviation: $\sqrt{143.7} = 12.0 \text{ inches}$

**RANGE**

The simplest, most easily calculated, measure of variability is the range. The difference between the largest and smallest values can be noted quickly, and it does give some impression of the dispersion. However, it is entirely dependent on the two extreme values and is insensitive to the ones in the middle.

One use of the range is to evaluate samples with very few items. For example, some quality control techniques involve taking periodic small samples and basing further action upon the range found in several such samples.

**AVERAGE DEVIATION**

Unlike the range, most other measures of variability are concerned with the dispersion around some center or representative value. The average deviation is the sum of the deviations from the mean (without regard to sign) divided by the number of elements:

$$\text{Average deviation} = \frac{\sum|x - \mu|}{n}$$

where $\mu$ is the mean and $n$ is the population size.

The average deviation gives information in an easily understood form. If a town has a mean yearly snowfall of 90 inches with an average deviation of 10 inches, the reader has an intuitive sense that a yearly snowfall between 80 and 100 inches would probably not be unusual. If two companies advertise mean starting salaries of $25,000, but in one the average deviation is $500 while in the other it is $8,000, applicants will have some feeling for what this difference means. The average deviation is also useful with regard to discrepancies in populations where all the values are planned rather than related to chance occurrence. In such cases it is worthwhile to discuss how far from the mean each value is chosen to be, and what the average is of these differences.

**VARIANCE**

In many situations, dispersion is not planned, but rather is the result of various chance happenings. In such cases, the average deviation is not the proper tool for measuring variability. For example, consider the motion of microscopic particles suspended in a liquid. The unpredictable motion of any particle is the result of many small movements in various directions caused by random bumps from other particles. If we average the total displacements of all the particles from their starting points (this corresponds to the average deviation), the result does not increase in direct proportion to time. If, however, we average the squares of the total displacements of all the particles, this result does increase in direct proportion to time.

The same holds true for the movement of paramecia. Their seemingly random motions as noted under a microscope can be described by the fact that the average of the squares of the displacements from their starting points is directly proportional to time. Also, consider ping-pong balls dropped straight down from a high tower and subjected to chance buffeting in the air. We can measure the deviations from a center spot on the ground to the spots where the balls actually strike. As the height of the tower is increased, the average of the squared deviations increases proportionately.

In a wide variety of cases we are in effect trying to measure dispersion from the mean due to a multitude of chance effects. The proper tool in these cases is the average of the squared deviations from the mean. This is called the variance and is denoted by $\sigma^2$ ($\sigma$ is the lower case Greek letter sigma):

$$\sigma^2 = \frac{\sum(x - \mu)^2}{n}$$

For circumstances specified in Chapter 11, the variance of a sample, denoted by $s^2$, is calculated as

$$s^2 = \frac{\sum(x - \overline{x})^2}{n - 1}$$

**Example 2.2** During the years 1929 through 1939 of the Great Depression, the weekly average hours worked in manufacturing jobs were 45, 43, 41, 39, 39, 35, 37, 40, 39, 36, and 37. What is the variance?
Answer:
\[ \mu = \frac{45 + 43 + 41 + 39 + 39 + 35 + 37 + 40 + 39 + 36 + 37}{11} = 39.2 \text{ hours} \]
\[ \sigma^2 = \frac{(45 - 39.2)^2 + (43 - 39.2)^2 + \ldots + (37 - 39.2)^2 + (36 - 39.2)^2}{11} = 81 \]

Important for later use is a procedure to calculate the variance for a set of differences.

Example 2.3
Consider the following from Chapter 1:

\[ X = \{2, 9, 11, 22\}, Y = \{5, 7, 15\}, \text{ and } Z = \text{ the set of differences} \{ -3, -5, -13, 4, 2, -6, 6, 4, -4, 17, 15, 7 \}. \text{ What is the variance of } Z? \]

Answer:
\[ \mu_X = 11, \sigma^2_X = \frac{(2 - 11)^2 + (9 - 11)^2 + (11 - 11)^2 + (22 - 11)^2}{4} = \frac{206}{4} = 51.5 \]
\[ \mu_Y = 9, \sigma^2_Y = \frac{(5 - 9)^2 + (7 - 9)^2 + (15 - 9)^2}{3} = \frac{56}{3} = 18.67 \]
\[ \mu_Z = 2, \sigma^2_Z = \frac{(-3 - 2)^2 + (-5 - 2)^2 + (-13 - 2)^2 + \ldots + (7 - 2)^2}{12} = \frac{842}{12} = 70.17 \]

How are \( \sigma^2_X, \sigma^2_Y, \text{ and } \sigma^2_Z \) related? Note that in the above example, \( \sigma^2_X = \sigma^2_Y = \sigma^2_Z \)? This is true for the variance of any set of differences. More generally, if a total is formed by a procedure that adds or subtracts one element from each of several sets, the variance of the resulting totals is simply the sum of the variances of the several sets.

Not only can we sum variances as shown above to calculate total variance, but we can also reverse the process and determine how the total variance is split up among its various sources. For example, we can find what portion of the variance in numbers of sales by a company's sales representatives is due to the individual salesperson, what portion is due to the territory, what portion is due to the particular products sold by each salesperson, and so on. (The average deviation cannot be partitioned in this way.)

An arithmetic tool for calculating the variance is:
\[ \sigma^2 = \frac{\sum x^2}{n} - \mu^2 \]

In words, the variance can be found by subtracting the square of the mean from the average of the squared scores.

Example 2.4
Let \( X = \{3, 7, 15, 23\} \). What is the variance?

Answer:
\[ \Sigma x = 3 + 7 + 15 + 23 = 48 \]
\[ \Sigma x^2 = 3^2 + 7^2 + 15^2 + 23^2 = 812 \]
\[ \mu = \frac{\Sigma x}{n} = \frac{48}{4} = 12 \]

The variance can be calculated from its definition:
\[ \sigma^2 = \frac{\Sigma (x - \mu)^2}{n} \]
\[ = \frac{(3 - 12)^2 + (7 - 12)^2 + (15 - 12)^2 + (23 - 12)^2}{4} = 59 \]

Or it can be calculated as follows:
\[ \sigma^2 = \frac{\Sigma x^2 - \mu^2}{n} = \frac{812}{4} - 12^2 = 203 - 144 = 59 \]

Similarly, there is an arithmetical tool for calculating the variance of a sample:
\[ s^2 = \frac{\sum x^2 - (\Sigma x)^2}{n - 1} \]

STANDARD DEVIATION

Suppose we wish to pick a representative value for the variability of some population. The preceding discussions indicate that a natural choice is the value whose square is the average of the squared deviations from the mean. Thus we are led to consider the square root of the variance. This is called the standard deviation, denoted by \( \sigma \), and is calculated as follows:
\[ \sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \mu^2} \]
Similarly, the standard deviation of a sample (see Chapter 11) is denoted by and is calculated as follows:

\[ \sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - (\sum x)^2}{n-1}} \]

**Example 2.5** The approximate numbers of bank suspensions during the years 1928 through 1933 were 500, 650, 1350, 2300, 1450, and 4000, respectively. What is the standard deviation?

**Answer:**

\[
\mu = \frac{500 + 650 + 1350 + 2300 + 1450 + 4000}{6} = 1700
\]

\[
\sigma^2 = \frac{1200^2 + 1050^2 + 375^2 + 75^2 + 250^2 + 2300^2}{6} = 1,394,375
\]

\[
\sigma = \sqrt{1,394,375} \approx 1181
\]

Two secondary procedures for measuring variability or dispersion are the **interquartile range** and the **relative variability**.

**INTERQUARTILE RANGE**

This is one method of removing the influence of extreme values on the range. The **interquartile range** is calculated by arranging the data in numerical order, removing the upper and lower one-quarter of the values, and then stating the range of the remaining values.

**Example 2.6** Suppose that farm sizes in a small community have the following characteristics: the smallest value is 16.6 acres, 10% of the values are below 23.5 acres, 25% are below 41.1 acres, the median is 57.6 acres, 60% are below 87.2 acres, 75% are below 101.9 acres, 90% are below 124.0 acres, and the top value is 201.7 acres.

a. What is the range?

**Answer:** The range is 201.7 - 16.6 = 185.1 acres.

b. What is the interquartile range?

**Answer:** The interquartile range, with the highest and lowest one-quarter of the values removed, is 101.9 - 41.1 = 60.8 acres. Thus, while the largest farm is 185 acres more than the smallest, the middle 50% of the farm sizes range over a 61 acre interval.

**RELATIVE VARIABILITY**

A comparison of two variances is more meaningful if the means of the populations are also taken into consideration. **Relative variability** is defined to be the quotient obtained by dividing the standard deviation by the mean. Usually the result is then expressed as a percentage.

**Example 2.7** Suppose that the mean salary for police officers is $45,000 with a standard deviation of $9000, while the mean salary for fire fighters is $35,000 with a standard deviation of $7700. What are the respective relative variabilities?

**Answer:** The relative variabilities are 9000/45,000 and 7700/35,000, or 20% and 22%, respectively. Thus, while the police officers’ salaries vary more in absolute terms, the fire fighters’ salaries vary more in relative terms when the difference in means is taken into consideration.

**EXERCISES**

1. The number of criminals executed in the United States during the 1930s and 1940s were as follows: 1930—155; 1931—153; 1932—140; 1933—160; 1934—168; 1935—199; 1936—195; 1937—147; 1938—190; 1939—185; 1940—124; 1941—123; 1942—147; 1943—131; 1944—120; 1945—117; 1946—131; 1947—153; 1948—119; 1949—119. What are the mean, the range, and the standard deviation?
2. According to a 1988 *New York Times* article, the ten car models with the highest theft rates were as follows:

<table>
<thead>
<tr>
<th>Vehicle Model</th>
<th>Thefts per 1000 Cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pontiac Firebird</td>
<td>30.14</td>
</tr>
<tr>
<td>Chevrolet Camaro</td>
<td>26.02</td>
</tr>
<tr>
<td>Chevrolet Monte Carlo</td>
<td>20.28</td>
</tr>
<tr>
<td>Toyota MR2</td>
<td>19.25</td>
</tr>
<tr>
<td>Buick Regal</td>
<td>14.70</td>
</tr>
<tr>
<td>Mitsubishi Starion</td>
<td>14.70</td>
</tr>
<tr>
<td>Ferrari Mondial</td>
<td>13.60</td>
</tr>
<tr>
<td>Mitsubishi Mirage</td>
<td>12.80</td>
</tr>
<tr>
<td>Pontiac Fiero</td>
<td>12.68</td>
</tr>
<tr>
<td>Oldsmobile Cutlass</td>
<td>11.73</td>
</tr>
</tbody>
</table>

a. What are the mean, range, and standard deviation of these theft rates?
b. How will each of these values change if each theft rate increases by 1.5% by 15%?
3. Because Brazilian doctors do not use generic prescriptions, black markets exist in Brazil for many drugs (The Lancet, October 15, 1994). For example, the black market price per tablet of acetaminophen for seven brands are as follows: Acetofen, $0.23; Dorico, $0.10; Pacemol, $0.17; Parador, $0.14; Tylenol, $0.18; Paracetamol, $0.08; and Nevralgina, $0.27. What are the mean, range, and standard deviation of these prices?

Study Questions
1. Suppose that the same constant is added to every value in a population. How does this affect the range? The average deviation? The standard deviation?
2. Suppose that each value in a set is multiplied by the same constant. How does this affect the range? The average deviation? The variance? The standard deviation?
3. Does it make sense to talk about the average of the deviations from the mean (with regard to sign)? In other words, what can be said about \( \frac{\sum (x - \mu)}{n} \)?
4. If the median was the most significant measure of central tendency, the appropriate measure of variability might well be \( \frac{\sum |x - M|}{n} \). Show that \( \sum |x - K| \) is smallest when \( K = M \) (the median). [Hint: Think about the position of \( M \) among the \( x \). As we move away from \( M \) by a small amount \( h \), what happens to each \( |x - M| \) in changing to \( |x - K| \) where \( K = M + h \)?]
5. If you've had an introductory calculus course, use calculus to show that \( \sum (x - t)^2 \) is smallest when \( t = \mu \). [Hint: \( \frac{d}{dt} \sum (x - t)^2 = ? \) Set equal to 0 and solve for \( t \).]
Part 2

PROBABILITY

In the world around us, unlikely events sometimes take place. At other times, events that seem inevitable do not occur. Because of the myriad and minute origins of various happenings, it is often impracticable, or simply impossible, to predict exact outcomes. However, while we may not be able to foretell a specific result, we can sometimes assign what is called a probability to indicate the likelihood that a particular event will occur.

For our study of statistics, we need an understanding of the probability that a given elementary event will happen many times, each time under the same circumstances. We want to be able to deduce the chance or prospect of occurrence of such events, and then use the result to draw inferences about more complex circumstances for which complete information is unavailable. For example, we might analyze the past movements of various stock prices given various economic conditions, calculate specific probabilities, and then ask what can be said, with what degree of confidence, about future movements.

In Chapters 5–7 we concentrate on the development of the specific techniques necessary to appreciate the basic principles of statistical analysis which will be considered later. In particular, we need to understand (1) how to use counting techniques; (2) how to calculate elementary probabilities, including binomial probabilities, and (3) how to calculate expected values.
Counting techniques are important in solving probability problems. In this chapter we develop the concept of combinations, a counting procedure critical in later applications. Along the way we learn the multiplicative rule and the way to count permutations.

Examples 5.1 and 5.2 illustrate the use of the multiplicative rule.

**Example 5.1** A company must choose one of eight applicants for a secretarial position, and one of six applicants for a janitorial position. In how many ways can this be done?

**Answer:** We can pair up each of the eight secretarial applicants with each of the six janitorial applicants for a total of $8 \times 6 = 48$ ways.

<table>
<thead>
<tr>
<th>8</th>
<th>6</th>
<th>$8 \times 6 = 48$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ways of picking secretaries</td>
<td>ways of picking janitors</td>
<td></td>
</tr>
</tbody>
</table>

**Example 5.2** A manufacturer must choose one of two warehouses, one of five assembly plants and one of three distribution centers. In how many ways can this be done?

**Answer:** There are $2 \times 5 = 10$ ways of picking a warehouse and an assembly plant, and for each of these there are 3 ways to pick a distribution center, for a total of $10 \times 3 = 30$ ways.

<table>
<thead>
<tr>
<th>2</th>
<th>5</th>
<th>3</th>
<th>$2 \times 5 \times 3 = 30$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ways of picking warehouses</td>
<td>ways of picking assembly plants</td>
<td>ways of picking distribution centers</td>
<td></td>
</tr>
</tbody>
</table>

The multiplicative rule can be stated in more general terms.
MULTIPLICATIVE RULE

If one event can occur in \( m \) ways, and, for each of these ways, a second event can occur in \( n \) ways, then the two events can occur together in \( mn \) ways. The same principle can be extended to three or more events, as in Example 5.2.

The multiplicative rule is the underlying principle behind various counting techniques.

**Example 5.3**

A president and vice president are to be picked from six candidates. In how many ways can this be done?

**Answer:** There are 6 choices for president and then 5 remaining choices for vice president for a total of \( 6 \times 5 = 30 \) ways.

<table>
<thead>
<tr>
<th>6</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ways of choosing president</td>
<td>ways of choosing vice-president</td>
</tr>
<tr>
<td>( 6 \times 5 = 30 )</td>
<td></td>
</tr>
</tbody>
</table>

**Example 5.4**

Three bonuses, one for $10,000, one for $5000 and one for $2000, are to be given to three employees chosen from a group of ten. In how many ways can this be done?

**Answer:** There are 10 choices for the $10,000 bonus, 9 remaining choices for the $5000 bonus, and then 8 remaining choices for the $2000 bonus for a total of \( 10 \times 9 \times 8 = 720 \) ways.

<table>
<thead>
<tr>
<th>10</th>
<th>9</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>choices for $10,000 bonus</td>
<td>choices for $5000 bonus</td>
<td>choices for $2000 bonus</td>
</tr>
<tr>
<td>( 10 \times 9 \times 8 = 720 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The answers to Examples 5.3 and 5.4 can be expressed in terms of factorials (\( 3! = 3 \times 2 \times 1 = 6, 5! = 5 \times 4 \times 3 \times 2 \times 1 = 120 \), etc.) as follows:

\[
6 \times 5 = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1} = \frac{6!}{4!}
\]

\[
10 \times 9 \times 8 = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{10!}{7!}
\]

Note that \( 6 \times 5 \) has 2 factors, \( 6 - 2 = 4 \), and \( 6 \times 5 = 6!/(6 - 2)! \). Also \( 10 \times 9 \times 8 \) has 3 factors, \( 10 - 3 = 7 \), and \( 10 \times 9 \times 8 = 10!/(10 - 3)! \).

Similarly, if we have \( r \) factors beginning with \( n \), we get \( n!/(n - r)! \). This is the basis of the permutation rule.

**PERMUTATION RULE**

The number of ways of choosing \( r \) distinct objects from among \( n \) objects, where "order is important" is \( n!/(n - r)! \). Each of the different ways of making this choice is called a permutation, and the total number of such permutations is denoted by \( P(n,r) \). Thus, \( P(n,r) = n!/(n - r)! \). We can see that "order is important" in Examples 5.3 and 5.4; for example, choosing Tom to be president and Mary to be vice president is different from choosing Mary to be president and Tom to be vice president.

**Example 5.5**

For a special advertising campaign, a company plans to pick one of its nine products for TV, one for radio, one for magazines, and one for newspapers. In how many ways can this be done?

\[
P(9,4) = \frac{9!}{(9-4)!} = \frac{9!}{5!} = 3024
\]

In situations where order is not important, we must modify the counting procedure.

**Example 5.6**

In how many ways can a two person committee be chosen from six people?

**Answer:** Compare this problem with Example 5.3. In that case, the answer was \( 6 \times 5 = 30 \), but that answer is too large for this example. While choosing Tom as president and Mary as vice president is different from choosing Mary as president and Tom as vice president, there is only one two-person committee consisting of Tom and Mary. Thus 30 is two times too large, and the correct answer is \( 30/2 = 15 \).

**Example 5.7**

In how many ways can three employees be chosen from a group of ten to receive identical $4000 bonuses?

**Answer:** Compare this problem with Example 5.4. In that case, the answer was \( 10 \times 9 \times 8 = 720 \), but again, that answer is too large for this example. There is only one way to give Jane, David, and Ann the same $4000 raise, but how many ways can we distribute three different raises among three people? The answer is \( 3 \times 2 \times 1 \) because there are three choices for the big raise, two remaining choices for the middle raise, and only one remaining choice for whom to give the last raise. Thus 720 is \( 3 \times 2 \times 1 = \)
$3! = 6$ times too large for this example, and the correct answer is $720/6 = 120$.

The answers to Examples 5.6 and 5.7 can be expressed in terms of factorials as follows: $6 \times 5$ has 2 factors, and we had to divide $6 \times 5 = 6 \frac{1}{4}$! by the number 2 or 2!. $10 \times 9 \times 8$ has 3 factors, and we had to divide $10 \times 9 \times 8 = 10! / 7!$ by the number 6 or 3!. Also, $6! / 4!$ divided by 2! is $\frac{6!}{4!} \times 2!$, and $10! / 7!$ divided by 3! is $\frac{10!}{7!}$. Similarly, if we divide $n! / (n - r)!$ by the number $r!$; the result is $\frac{n!}{(n - r)!} / r!$.

**COMBINATION RULE**

The number of ways of choosing $r$ distinct objects from $n$ objects, where “order is not important,” is $n! / (n - r)! r!$. Each of the different ways of making this choice is called a combination, and the total number of such combinations is denoted by $C(n,r)$ or $\binom{n}{r}$. Thus

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

**Example 5.8** In how many ways can a company select four of its nine products for an advertising campaign? (Note the difference from Example 5.5.)

**Answer:**

$$C(9,4) = \binom{9}{4} = \frac{9!}{(9 - 4)! 4!} = \frac{9!}{54!} = 126$$

For ease in future calculations, we note how to simplify factorial divisions. For instance,

$$\frac{10!}{6!} = \frac{10 \times 9 \times 8 \times 7 \times 6!}{6!} = 10 \times 9 \times 8 \times 7 = 5040$$

$$\frac{8!}{3!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3!}{3!} = 8 \times 7 \times 6 \times 5 \times 4 = 6720$$

When evaluating combinations, it is simplest to cancel the larger factorial in the denominator against part of the factorial in the numerator.

**Example 5.9**

$$C(7,4) = \frac{7!}{34!} = \frac{7 \times 6 \times 5}{3 \times 2} = 35$$

We canceled the 4! in the denominator against part of the 7! in the numerator, and we were left with $3! = 3 \times 2$ in the denominator. Similarly:

$$C(12,10) = \frac{12!}{210!} = \frac{12 \times 11}{2} = 66$$

$$C(20,4) = \frac{20!}{164!} = \frac{20 \times 19 \times 18 \times 17}{4 \times 3 \times 2} = 4845$$

We define 0! to be 1, and so, for example, $C(5,5) = \frac{5!}{0! 5!} = 1$. This result is expected since there is only one way of choosing five items from a group of five when order is not important. Similarly, $\binom{4}{2} = 1$, $\binom{2}{1} = 1$, $C(10,10) = 1$, etc. Another combination in which 0! arises has the form $C(6,0) = \frac{6!}{6! 0!} = 1$. (There is exactly one way to choose no item from a group of six.)

Examples of other readily computed combinations are $C(57,1) = \frac{57!}{56! 1!} = 57$ (there are 57 ways to pick one item from a group of 57) and $C(32,31) = 32$ (there are 32 ways to pick all but one item from a group of 32). Similarly, $C(85,1) = 85$ and $C(23,22) = 23$.

The results of other combinations are not as obvious and require some work. However, patterns can be noticed. For example, consider the following:

$$\begin{matrix}
(0) & 1 \\
(1) & 1 & 1 \\
(2) & 1 & 2 & 2 & 1 \\
(3) & 1 & 3 & 3 & 3 & 1 \\
(4) & 1 & 4 & 4 & 4 & 4 & 1 \\
(5) & 1 & 5 & 5 & 5 & 5 & 5 & 1 \\
\end{matrix}$$

These numbers form a pattern known as Pascal’s triangle:

$$\begin{array}{ccccccccccc}
1 \\
1 & 1 \\
1 & 2 & 1 \\
1 & 3 & 3 & 1 \\
1 & 4 & 6 & 4 & 1 \\
1 & 5 & 10 & 10 & 5 & 1 \\
\end{array}$$

If one line of Pascal’s triangle is known, how can the next line be found?
Answer: The next line will start and end with 1. Each other number can be found by adding the two numbers immediately above and slightly to the right and left. For example, the next line after the last one above is

\[ 1 \hspace{1cm} 1+5=6 \hspace{1cm} 5+10=15 \hspace{1cm} 10+10=20 \hspace{1cm} 10+5=15 \hspace{1cm} 5+1=6 \hspace{1cm} 1 \]

Thus,

\[ \binom{6}{0} = 1 \hspace{1cm} \binom{6}{1} = 6 \hspace{1cm} \binom{6}{2} = 15 \hspace{1cm} \binom{6}{3} = 20 \hspace{1cm} \binom{6}{4} = 15 \hspace{1cm} \binom{6}{5} = 6 \hspace{1cm} \binom{6}{6} = 1 \]

Finally, for later use, we comment on how to find \( \binom{n}{r} \) if we know \( \binom{n}{r-1} \).

\[ \frac{n-r+1}{r} \binom{n}{r-1} = \binom{n}{r} \]  
(See Study Question 1)

Example 5.10  
Given that \( \binom{8}{3} = 56 \), what is \( \binom{8}{4} \)?

Answer:

\[ \frac{8-4+1}{4} = \frac{5}{4} \text{ and } \frac{5}{4} \times 56 = 70 \]

Given that \( \binom{15}{9} = 5005 \), what is \( \binom{15}{10} \)?

Answer:

\[ \frac{15-10+1}{10} = \frac{6}{10} \text{ and } \frac{6}{10} \times 5005 = 3003 \]

EXERCISES

1. In his novel *Alaska*, James Michener talks about the different factors that must combine to produce an ice age. Then, mathematically, he explains that "if four different factors in an intricate problem operate in cycles of 13, 17, 23, and 37 years respectively, and if all have to coincide to produce the desired result, you might have to wait 188,071 years ..." Explain how Michener calculated 188,071.

2. As reported in the Allentown Morning Call, Boston Chicken advertised that their customers had a choice of 3360 combinations of three side dishes from among their 16 side dishes offered. A high school teacher pointed out the mistake; can you?

3. Johann Gregor Mendel characterized garden peas by the following seven attributes: the stems are either long or short; the flowers are either red or white, and are either terminal or axial; the pods are either green or yellow, and are either constricted or inflated; and the seeds are either green or yellow, and are either smooth or wrinkled. How many different varieties of garden peas are possible?

Study Questions

1. Show that \( \frac{n-r+1}{r} \binom{n}{r-1} = \binom{n}{r} \). [Hint: Write out the expressions for \( \binom{n}{r-1} \) and for \( \binom{n}{r} \), and note that \( r(r-1)! = r! \) and \( (n-r+1)(n-r)! = (n-r+1)! \].

2. Prove the formula \( \binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r-1} \) which we used in Example 5.10 to determine an entry in Pascal's triangle from the two above entries. [Hint: Write out the algebraic expressions for each of the two combinations on the right, and then convert to the common denominator: \( r!(n-r)! \].

3. Suppose we have a set with \( n \) elements. Then the number of subsets that contain exactly \( r \) elements is \( C(n,r) \). What proportion of these subsets contain some designated element of the original set? [Hint: Reason that the answer is \( C(n-1,r-1)/C(n,r) \), and show algebraically that this is equal to \( r/n \)].

4. a. What is the total number of subsets of a set with two elements? Of a set with three elements?

b. Do you think that a set could have exactly ten subsets? Sixteen subsets?
6
PROBABILITY CONCEPTS

The probability of a particular outcome of an experiment is a mathematical statement about the likelihood of that event occurring. Probabilities are always between 0 and 1 with a probability close to 0 meaning that an event is unlikely to occur, and a probability close to 1 meaning that the event is likely to occur. The sum of the probabilities of all the separate outcomes of an experiment is always 1. In this chapter we list the basic probability concepts that are needed for our discussion of probability distributions.

ELEMENTARY PROBABILITY

Complementary events
The probability that an event will not occur, that is, the probability of its complement, is equal to 1 minus the probability that the event will occur:

\[ P(X') = 1 - P(X) \]

Example 6.1 If the probability that a company will win a contract is .3, what is the probability that it will not win the contract?

Answer: 1 − .3 = .7.

Addition principle
If two events are mutually exclusive, that is, they cannot occur simultaneously, then the probability that at least one event will occur is equal to the sum of the respective probabilities of the two events. That is:

\[ P(X \cup Y) = P(X) + P(Y) \]

where \( X \cap Y \), read “\( X \) intersect \( Y \),” means that both \( X \) and \( Y \) occur, while \( X \cup Y \), read “\( X \) union \( Y \),” means that either \( X \) or \( Y \) or both occur.

Example 6.2 If the probabilities that Jane, Tom, and Mary will be chosen chairperson of the board are .5, .3, and .2, respectively, then the probability that the chairperson will be either Jane or Mary is .5 + .2 = .7.
When two events are not mutually exclusive, then the sum of their probabilities counts their shared occurrence twice. This leads to the

**General addition formula**

For any pair of events \( X \) and \( Y \),

\[
P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)
\]

**Example 6.3** Suppose that the probability that a construction company will be awarded a certain contract is \( .25 \), the probability that it will be awarded a second contract is \( .21 \), and the probability that it will get both contracts is \( .13 \). What is the probability that the company will win at least one of the two contracts?

**Answer:** \( .25 + .21 - .13 = .33 \)

**Independence principle**

If the chance that one event will happen is not influenced by whether or not a second event happens, the probability that both events will occur is the product of their separate probabilities.

**Example 6.4** The probability that a student will receive a state grant is \( 1/3 \), while the probability that she will be awarded a federal grant is \( 1/2 \). If whether or not she receives one grant is not influenced by whether or not she receives the other, what is the probability of her receiving both grants?

**Answer:** \( 1/3 \times 1/2 = 1/6 \)

The independence principle can be extended to more than two events; that is, given a sequence of independent events, the probability that all will happen is equal to the product of their individual probabilities.

**Example 6.5** Suppose a reputed psychic in an ESP experiment has called heads or tails correctly on ten successive tosses of a coin. What is the probability that guessing would yield this perfect score?

**Answer:** \( (1/2)(1/2)...(1/2) = (1/2)^{10} = 1/1024 \)

In many applications, such as coin tossing, there are only two possible outcomes. For example, on each toss either the psychic guesses correctly or she guesses incorrectly, either a radio is defective or it is not defective, either the workers will go on strike or they will not walk out, either the manager's salary is above \( $50,000 \) or it does not exceed \( $50,000 \). In some applications such two-outcome situations are repeated many times. For example, we may ask about the chances that at most one of four tires is defective, or that at least three of five unions will vote to go on strike, or that exactly two of three executives have salaries above \( $50,000 \). For applications in which a two-outcome situation is repeated some number of times, and the probability of each of the two outcomes remains the same for each repetition, the resulting calculations involve what are known as binomial probabilities.

**Example 6.6** Suppose the probability that a light bulb is defective is \( .1 \). What is the probability that four light bulbs are all defective?

**Answer:** \( (.1)(.1)(.1)(.1) = (.1)^4 = .0001 \)

**Example 6.7** Again suppose the probability that a light bulb is defective is \( .1 \). What is the probability that exactly two of three light bulbs are defective.

**Answer:** We subdivide the problem as follows. The probability that the first two bulbs are defective, and the third is good, is \( (.1)(.1)(.9) = .009 \). (Note that, if the probability of being defective is \( .1 \), the probability of being good is \( .9 \).) The probability that the first bulb is good, and the other two are defective, is \( (.9)(.1)(.1) = .009 \). Finally, the probability that the second bulb is good, and the other two are defective is \( (.1)(.9)(.1) = .009 \). Summing, we find that the probability that exactly two out of three bulbs are defective is \( .009 + .009 + .009 = .027 \).

**Example 6.8** If the probability of a defective light bulb is \( .1 \), what is the probability that exactly three out of eight light bulbs are defective?

**Answer:** We can subdivide the problem again. For example, the probability that the first, third, and seventh bulbs are defective, and the rest are good, is

\[
(.1)(.9)(.1)(.9)(.1)(.9)(.1)(.9) = (.1)^3(.9)^5 = .00059049
\]

The probability that the second, third, and fifth bulbs are defective, and the rest are good, is

\[
(.9)(.1)(.1)(.9)(.1)(.9)(.9)(.9) = (.1)^3(.9)^5 = .00059049
\]

As can be seen, the probability of any particular arrangement of three defective and five good bulbs is \( (.1)^3(.9)^5 = .00059049 \). How many such arrangements are there? In other words, in how many ways can we pick three of eight positions for the defective bulbs...
(the remaining five positions will be for good bulbs). From Chapter 5 the answer is given by combinations:

\[ C(8,3) = \frac{8!}{5!3!} = \frac{8 \times 7 \times 6}{3 \times 2} = 56 \]

Each of these 56 arrangements has a probability of .00059049. Thus, the probability that exactly three out of eight light bulbs are defective is \[ 56 \times .00059049 = .03306744 \]

\[ C(8,3)(.1)^3(.9)^5 = \frac{8!}{3!5!} (.1)^3(.9)^5 = .03306744 \]

**Example 6.9** Suppose 30% of the employees in a large factory are smokers. What is the probability that there will be exactly two smokers in a randomly chosen five-person work group?

**Answer:** We reason as follows. The probability that a person smokes is \(0.3 = 0.3\), so the probability that he or she does not smoke is \(1 - 0.3 = 0.7\). The probability of a particular arrangement of two smokers and three nonsmokers is \((0.3)^2(0.7)^3 = 0.03087\). The number of such arrangements is \(C(5,2) = 5!2!3! = 10\). Each such arrangement has probability \(0.03087\), so the final answer is \(10 \times 0.03087 = 0.3087\).

\[ C(5,2)(.3)^2(.7)^3 = \frac{5!}{2!3!} (.3)^2(.7)^3 = .3087 \]

We can state the general principle as the binomial formula.

**BINOMIAL FORMULA**

Suppose an experiment has two possible outcomes, called *success* and *failure*, with the probability of success equal to \(p\) and the probability of failure equal to \(q\) (of course, \(p + q = 1\)). Suppose further that the experiment is repeated \(n\) times, and the outcome at any particular time has no influence over the outcome at any other time. Then the probability of exactly \(x\) successes (and thus \(n-x\) failures) is

\[ C(n,x) p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} \]

**Example 6.10** A manager notes that there is a .125 probability that any employee will arrive late for work. What is the probability that exactly one person in a six-person department will arrive late?

**Answer:** If the probability of being late is .125, then the probability of being on time is \(1 - .125 = .875\). If one person of six is late, then \(6 - 1 = 5\) will be on time. Thus the desired probability is

\[ C(6,1)(.125)^1(.875)^5 = 6(.125)(.875)^5 = .385 \]

Many, perhaps most, applications of probability involve such phrases as *at least*, *at most*, *less than*, and *more than*. In these cases, solutions involve summing two or more cases.

**Example 6.11** A manufacturer has the following quality control check at the end of a production line: If at least eight of ten randomly picked articles meet all specifications, the whole shipment is approved. If, in reality, 85% of a particular shipment meet all specifications, what is the probability that the shipment will make it through the control check?

**Answer:** The probability of meeting specifications is \(0.85\), so the probability of not meeting specifications must be \(0.15\). We want the probability that at least eight of ten articles meet specifications, that is, the probability that exactly eight or exactly nine or exactly ten articles meet specifications. We sum the three binomial probabilities:

\[ C(10,8)(.85)^8(.15)^2 + C(10,9)(.85)^9(.15)^1 + C(10,10)(.85)^{10}(.15)^0 \]

\[ = \frac{10!}{8!2!} (.85)^8(.15)^2 + \frac{10!}{9!1!} (.85)^9(.15)^1 + (.85)^{10} = .820 \]

**Example 6.12** For the problem in Example 6.11, what is the probability that a shipment in which only 70% of the articles meet specifications will make it through the control check?

**Answer:**

\[ C(10,8)(.7)^8(.3)^2 + C(10,9)(.7)^9(.3)^1 + C(10,10)(.7)^{10}(.3)^0 \]

\[ = 45(.7)^8(.3)^2 + 10(.7)^9(.3) + (.7)^{10} = .383 \]

In some situations it is easier to calculate the probability of the complementary event and subtract this value from 1.

**Example 6.13** Joe DiMaggio had a career batting average of .325. What was the probability that he would get at least one hit in five official times at bat?
Answer: We could sum the probabilities of exactly one hit, two hits, three hits, four hits, and five hits. However, the complement of "at least one hit" is "zero hits." The probability of no hit is

\[ C(5,0)(.325)^0(.675)^5 = (.675)^5 = .140, \]

and thus the probability of at least one hit in five times at bat is

\[ 1 - .140 = .860. \]

Example 6.14

A grocery store manager notes that 35\% of the customers who buy a particular product make use of a store coupon to receive a discount. If seven people purchase the product, what is the probability that fewer than four will use a coupon?

Answer: In this situation, "fewer than four" means zero or one or two or three.

\[
\begin{align*}
C(7,0)(.35)^0(.65)^7 + C(7,1)(.35)^1(.65)^6 + C(7,2)(.35)^2(.65)^5 + C(7,3)(.35)^3(.65)^4 \\
= (.65)^7 + 7(.35)(.65)^6 + 21(.35)^2(.65)^5 + 35(.35)^3(.65)^4 = .800
\end{align*}
\]

Sometimes we are asked to calculate the probability of each of the possible outcomes (the results should sum to 1).

Example 6.15

If the probability of a male birth is .51, what is the probability that a five-child family will have all boys? Exactly four boys? Exactly three boys? Exactly two boys? Exactly one boy? All girls?

Answer:

\[
\begin{align*}
P(5 \text{ boys}) &= C(5,5)(.51)^5(.49)^0 = (.51)^5 = .0345 \\
P(4 \text{ boys}) &= C(5,4)(.51)^4(.49)^1 = 5(.51)^4(.49) = .1657 \\
P(3 \text{ boys}) &= C(5,3)(.51)^3(.49)^2 = 10(.51)^3(.49)^2 = .3185 \\
P(2 \text{ boys}) &= C(5,2)(.51)^2(.49)^3 = 10(.51)^2(.49)^3 = .3060 \\
P(1 \text{ boy}) &= C(5,1)(.51)^1(.49)^4 = 5(.51)(.49)^4 = .1470 \\
P(0 \text{ boys}) &= C(5,0)(.51)^0(.49)^5 = (.49)^5 = .0283
\end{align*}
\]

A table such as the one in Example 6.15 shows the entire probability distribution, which in this case refers to a listing of the probabilities of all outcomes.

For later use we need to know how to calculate the probability of \( x \) successes if we know the probability of \( x - 1 \) successes:

\[
P(x \text{ successes}) = C(n, x)p^x q^{n-x}
\]

\[
P(x - 1 \text{ successes}) = C(n, x - 1)p^{x-1}q^{n-x+1}
\]

\[
[n - (x - 1) = n - x + 1]
\]

Note that \( P(x \text{ successes}) \) has one more \( p \) and one less \( q \) than does \( P(x - 1 \text{ successes}) \), so we should expect a factor \( p/q \). Also, from Chapter 5 we have \( C(n, x) = [(n - x + 1)/x]C(n, x - 1) \). Combining these results gives

\[
P(x \text{ successes}) = \frac{n - x + 1}{x} \frac{p}{q} P(x - 1 \text{ successes})
\]

Since \( q = 1 - p \), this is often written

\[
P(x \text{ successes}) = \frac{n - x + 1}{x} \frac{p}{1 - p} P(x - 1 \text{ successes})
\]

Example 6.16

A sharpshooter can hit a bullseye target 95\% of the time. Given that the probability of exactly eight bullseyes in ten shots is .0746, what is the probability of exactly nine bullseyes in ten shots?

Answer:

\[
P(10 \text{ bullseyes}) = \frac{10 - 9 + 1}{9} \frac{.95}{1 -.95} (.0746) = .315
\]

EXERCISES

1. As reported in The New York Times (February 19, 1995), the Russian Health Ministry announced that one-quarter of the country's hospitals had no sewage system and one-seventh had no running water. What is the probability that a Russian hospital will have at least one of these problems if:
   a. the two problems are independent?
   b. hospitals with the running water problem are a subset of those with the sewage problem?

2. In the November 27, 1994, issue of Parade magazine, the "Ask Marilyn" section contained this question:
   "Suppose a person was having two surgeries performed at the same time. If the chances of success for surgery A are 85\%, and the chances of success for surgery B are 90\%, what are the chances that both would fail?"

   What do you think of Marilyn's solution? (.15)(.10) = .015 or 1.5\%?

3. In November 1994, Intel announced that a "subtle flaw" in its Pentium chip would affect one in 9,000,000,000 division problems. Suppose a computer performed 20,000,000 divisions (a not unreasonable number) in the course
of a particular program. What is the probability of no error? Of at least one error?

4. According to a CBS/New York Times poll taken in 1992, 15% of the public has responded to a telephone call-in poll. In a random group of five people, what is the probability that exactly two have responded to a call-in poll? That at least two have responded? That at most two have responded?

5. In a 1974 "Dear Abby" letter a woman lamented that she had just given birth to her eighth child, and all were girls! Her doctor had assured her that the chance of the eighth child being a girl was only one in 100.
   a. What was the real probability that the eighth child would be a girl?
   b. Before the birth of the first child, what was the probability that the woman would give birth to eight girls in a row?

6. The yearly mortality rate for American men from prostate cancer has been constant for decades at about 25 of every 100,000 men. (This rate has not changed in spite of new diagnostic techniques and new treatments.) In a group of 100 American men, what is the probability that at least one will die from prostate cancer in a given year?

Study Questions

1. (Geometric distribution) Suppose the probability that someone will make a major mistake on an income tax return is .08. One day, an IRS agent plans to audit as many returns as necessary until she finds one with a major mistake. What is the probability that a major mistake will be found on the first return? The second? The third? The fourth? The fifth? Note that, while the probability of a major mistake on any given return is always .08, the probability of which return the first major mistake is found on steadily decreases. The associated cumulative probability distribution is called a geometric distribution. Reason why, in this question, if the first three returns have no major mistake, the probability that more than five additional returns will have to be audited is the same as the probability at the beginning of the day that more than five returns will have to be audited.

2. (Negative binomial distribution) During the 1990 World Series, the Oakland Athletics were heavily favored over the Cincinnati Reds. Suppose the probability that Cincinnati would win any given game was 4/9. What was the probability that Cincinnati would win the series, that is, would win four games before Oakland won four games? [Hint: The probabilities of separate cases must be summed; for each case, consider how many wins and losses occurred before the final game.] The above gives rise to what is called a negative binomial distribution. For a negative binomial distribution, as well as for the geometric distribution described in Study Question 1, the number of "successes" is fixed, while the total number of trials is allowed to vary. However, for the original binomial distribution, the number of "successes" is allowed to vary, while the number of trials is fixed.

7

EXPECTED VALUES

A bettor placing a chip on a number in roulette has a small chance of winning a lot and a large chance of losing a little. To determine whether the bet is "fair," we must be able to calculate the "expected value" of the game to the bettor. An insurance company in any given year pays out a large sum of money to each of a relatively small number of people, while collecting a small sum of money from each of many other individuals. To determine policy premiums an actuary must calculate the "expected values" relating to deaths in various age groups of policy holders. In this chapter we introduce the concept of a random variable, and then show how to determine expected values or means and variances relating to this concept. In particular, we consider how to calculate means and standard deviations in the case of binomial probabilities.

RANDOM VARIABLES

In Chapter 6 we determined the probability for each outcome in various examples. Often each outcome has not only an associated probability, but also an associated real number. For example, the probability may be 1/2 that there are 5 defective batteries; the probability may be .01 that a company receives 7 contracts; the probability may be .95 that 3 people recover from a disease. If X represents the different numbers associated with the potential outcomes of some chance situation, we call X a random variable.

Example 7.1 A prison official knows that 1/2 of the inmates he admits stay only 1 day, 1/4 stay 2 days, 1/5 stay 3 days, and 1/20 stay 4 days before they are either released or are sent on to the county jail. If X represents the number of days, then X is a random variable that takes the value 1 with probability 1/2, the value 2 with probability 1/4, the value 3 with probability 1/5, and the value 4 with probability 1/20.

The random variable in Example 7.1 is called discrete because it can assume only a countable number of values. The random variable in Example 7.2, however, is said to be continuous because it can assume values associated with a whole line interval.
Example 7.2
Let $X$ be a random variable whose values correspond to the speeds at which a jet plane can fly. The jet may be traveling at 623.478 ... miles per hour or at any other speed in some whole interval. We might ask what the probability is that the plane is flying at between 300 and 400 miles per hour.

A probability distribution for a discrete variable is a listing or formula giving probability for each value of the random variable.

Example 7.3
Concessionaires know that attendance at a football stadium will be 60,000 on a clear day, 45,000 if there is light snow, and 15,000 if there is heavy snow. Furthermore, the probability of clear skies, light snow, or heavy snow on any particular day is 1/2, 1/3, and 1/6 respectively. (Here we have a random variable $X$ that takes the values 60,000, 45,000, and 15,000.)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>clear skies</th>
<th>light snow</th>
<th>heavy snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>1/2</td>
<td>1/3</td>
<td>1/6</td>
</tr>
<tr>
<td>Random variable</td>
<td>60,000</td>
<td>45,000</td>
<td>15,000</td>
</tr>
</tbody>
</table>

What average attendance should be expected for the season?

**Answer:** We reason as follows. Suppose on, say, 12 game days we had clear skies $1/2$ of the time (6 days), a light snow $1/3$ of the time (4 days), and a heavy snow $1/6$ of the time (2 days). Then the average attendance would be

$$\frac{(6 \times 60,000) + (4 \times 45,000) + (2 \times 15,000)}{12} = 47,500$$

Note that we could have divided the 12 into each of the three terms in the numerator to obtain

$$\left( \frac{6}{12} \times 60,000 \right) + \left( \frac{4}{12} \times 45,000 \right) + \left( \frac{2}{12} \times 15,000 \right)$$

or, equivalently,

$$\left( \frac{1}{2} \times 60,000 \right) + \left( \frac{1}{3} \times 45,000 \right) + \left( \frac{1}{6} \times 15,000 \right) = 47,500$$

Actually, there was no need to consider 12 days. We could have simply multiplied probabilities times corresponding attendances and summed the resulting products.

**EXPECTED VALUE (MEAN) OF A RANDOM VARIABLE**

The final calculation in Example 7.3 motivates our definition of expected value. The expected value (or average or mean) of a random variable $X$ is the sum of the products obtained by multiplying each value $x$ by the corresponding probability $P(x)$:

$$E(X) = \sum x \cdot P(x)$$

**Example 7.4**
In a lottery, 10,000 tickets are sold at $1 each with a prize of $7500 for one winner. What is the average result for each bettor?

**Answer:** The actual winning payoff is $7499 because the winner paid $1 for a ticket, so we have:

<table>
<thead>
<tr>
<th>Outcome</th>
<th>win</th>
<th>lose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>1/10,000</td>
<td>9,999</td>
</tr>
<tr>
<td>Random variable</td>
<td>7499</td>
<td>-1</td>
</tr>
</tbody>
</table>

Expected value $= 7499 \left( \frac{1}{10,000} \right) + (-1) \left( \frac{9,999}{10,000} \right) = -0.25.$

Thus the average result for each person betting the lottery is a 25¢ loss.

**Example 7.5**
A manager must choose among three options. Option $A$ has a 10% chance of resulting in a $250,000 gain, but otherwise will result in a $10,000 loss. Option $B$ has a 50% chance of gaining $40,000 and a 50% chance of losing $1000. Finally, option $C$ has a 5% chance of gaining $800,000, but otherwise will result in a loss of $20,000. Which option should the manager choose?

---

1The development here is limited to random variables with a finite number of values, but this definition can be extended somewhat using infinite sums.
**Answer:** Calculate the expected values of the three options:

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Option A</th>
<th>Option B</th>
<th>Option C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability:</td>
<td>gain</td>
<td>loss</td>
<td>gain</td>
</tr>
<tr>
<td>Probability:</td>
<td>.10</td>
<td>.90</td>
<td>.50</td>
</tr>
<tr>
<td>Random variable:</td>
<td>250,000</td>
<td>-10,000</td>
<td>40,000</td>
</tr>
</tbody>
</table>

\[
E(A) = .10(250,000) + .90(-10,000) = 16,000 \\
E(B) = .50(40,000) + .50(-2000) = 19,000 \\
E(C) = .05(800,000) + .95(-20,000) = 21,000
\]

The manager should choose option \( C \). However, although Option \( C \) has the greatest mean (expected value), the manager may well wish to consider the relative riskiness of the various options. It, for example, a \$5000\ loss would be disastrous for the company, the manager might well decide to choose option \( B \) with its maximum possible loss of \$2000. (It should be intuitively clear that some concept of variance would be helpful here in measuring the risk; later in this chapter the variance of a random variable will be defined.)

Suppose we have a binomial random variable, that is, a random variable whose values are the numbers of “successes” in some binomial probability distribution.

**Example 7.6**

Of the automobiles produced in a particular plant, 40% had a certain defect. Suppose a company purchases five of these cars. What is the expected value for the number of cars with defects?

**Answer:** We might guess that the average or mean or expected value is 40% of 5 = .4 × 5 = 2, but let’s calculate from the definition. Letting \( X \) represent the number of cars with the defect, we have:

\[
P(0) = C(5,0)(.4)^0(.6)^5 = (.6)^5 = .07776 \\
P(1) = C(5,1)(.4)(.6)^4 = 5(.4)(.6)^4 = .25920 \\
P(2) = C(5,2)(.4)^2(.6)^3 = 10(.4)^2(.6)^3 = .34560 \\
P(3) = C(5,3)(.4)^3(.6)^2 = 10(.4)^3(.6)^2 = .23040 \\
P(4) = C(5,4)(.4)^4(.6)^1 = 5(.4)^4(.6) = .07680 \\
P(5) = C(5,5)(.4)^5(.6)^0 = (.4)^5 = .01024
\]

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>0 cars</th>
<th>1 car</th>
<th>2 cars</th>
<th>3 cars</th>
<th>4 cars</th>
<th>5 cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability:</td>
<td>.07776</td>
<td>.25920</td>
<td>.34560</td>
<td>.23040</td>
<td>.07680</td>
<td>.01024</td>
</tr>
</tbody>
</table>

\[
E(X) = 0(.07776) + 1(.25920) + 2(.34560) + 3(.23040) + 4(.07680) + 5(.01024) = 2
\]

Thus, the answer turns out to be the same as would be obtained by simply multiplying the probability of “success” times the number of cases.

The following is true: If we have a binomial probability situation with the probability of success equal to \( p \) and the number of trials equal to \( n \), the expected value or mean number of successes for the \( n \) trials is \( np \).

One proof of this statement involves algebraic manipulation with the binomial probability formula to show that \( \Sigma xP(x) = np \). Another insight is as follows:

If \( n = 1 \) we have:

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>success</th>
<th>failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability:</td>
<td>( p )</td>
<td>( 1-p )</td>
</tr>
<tr>
<td>Random Variable (successes):</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

with expected value \( = 1(p) + 0(1 - p) = p \). For larger \( n \), consider each trial independently with a resulting expected value of \( p \) for each trial. The number of successes in \( n \) trials is the sum of the successes in all \( n \) trials, and so the expected value or mean of the \( n \) trials is the sum of the expected values of all \( n \) trials. We calculate

\[
p + p + p + ... + p = np
\]

One \( p \) for each of \( n \) trials

**Example 7.7**

An insurance salesperson is able to sell policies to 15% of the people she contacts. Suppose that she contacts 120 people during a 2-week period. What is the expected value for the number of policies she sells?

**Answer:** We have a binomial probability with the probability of success \(.15\) and the number of trials \(120\), so the mean or expected value for the number of successes is \(120 \times .15 = 18\).

**VARIANCE OF A RANDOM VARIABLE**

We have seen that the mean of a random variable is \( \Sigma xP(x) \). However, not only is the mean important, but also we would like to measure the variability for the values taken on by a random variable. Since we are dealing with chance events, the proper tool is variance. Variance was defined in Chapter 2 to be the mean of the squared deviations \((x - \mu)^2\). If we regard the \((x - \mu)^2\) terms as the values of some random variable (whose probability is the same as the probability of \( x \)), then the mean of this new random variable is
\[ \sum (x - \mu)^2 \text{ } P(x), \text{ which is precisely how we define the variance } \sigma^2 \text{ of a random variable:} \]

\[ \sigma^2 = \sum (x - \mu)^2 \text{ } P(x) \]

As before, the standard deviation \( \sigma \) is the square root of the variance.

**Example 7.8**

A highway engineer knows that his crew can lay 5 miles of highway on a clear day, 2 miles on a rainy day, and only 1 mile on a snowy day. Suppose the probabilities are as follows:

<table>
<thead>
<tr>
<th>Outcome</th>
<th>clear</th>
<th>rain</th>
<th>snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>.6</td>
<td>.3</td>
<td>.1</td>
</tr>
</tbody>
</table>

| Random Variable (miles of highway) | 5    | 2    | 1    |

What are the mean (expected value) and the variance?

**Answer:**

\[ \mu = \sum x \text{ } P(x) = 5(.6) + 2(.3) + 1(.1) = 3.7 \]

\[ \sigma^2 = \sum (x - \mu)^2 \text{ } P(x) \]

\[ = (5 - 3.7)^2(.6) + (2 - 3.7)^2(.3) + (1 - 3.7)^2(.1) = 2.61 \]

**Example 7.9**

Look again at Example 7.6. We calculated the mean to be 2. What is the variance?

**Answer:**

\[ \sigma^2 = (0-2)^2(.07776) + (1-2)^2(.2592) + (2-2)^2(.3456) \]

\[ + (3-2)^2(.2304) + (4-2)^2(.0768) + (5-2)^2(.01024) = 1.2 \]

Could we have calculated the above result more easily? In this case, we have a binomial random variable, so we will use the same type of argument that we used in showing that \( \mu = np \). We first consider the binomial distribution for \( n = 1 \):

<table>
<thead>
<tr>
<th>Outcome</th>
<th>success</th>
<th>failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>( p )</td>
<td>( q )</td>
</tr>
</tbody>
</table>

\( \mu = p(1) + q(0) = p, \) so

\[ \sigma^2 = (1 - \mu)^2 p + (0 - \mu)^2 q \]

\[ = \sigma^2(p + \mu) \]

\[ = pq(p + q) \]

\[ = pq \text{ [factoring]} \]

\[ = pq \text{ [} p + q = 1 \text{]} \]

For larger \( n \), we reason as follows. The number of successes is the sum of the successes in all trials, so the variance is the sum of the variances from all trials (Chapter 2):

\[ pq + pq + pq + \ldots + pq = npq \]

One \( pq \) for each of \( n \) trials

More formally, this result can be obtained through algebraic manipulations (see Study Question 3).

**Example 7.10**

How can we use this method to more simply calculate the variance in Example 7.9?

**Answer:**

\[ npq = 5(.4)(.6) = 1.2. \]

Thus, for a random variable \( X, \)

\[ \mu = \sum x \text{ } P(x) \]

\[ \sigma^2 = \sum (x - \mu)^2 \text{ } P(x) \]

**Example 7.11**

Sixty percent of all new car buyers choose automatic transmissions. For a group of five new car buyers, calculate the mean and standard deviation for the number of buyers choosing automatics.

**Answer:**

\[ \mu = np = 5(.6) = 3.0 \]

\[ \sigma = \sqrt{np(1 - p)} = \sqrt{5(1 - .6)} = 1.1 \]

Note that this could have been calculated through the more involved

\[ \mu = \sum x \text{ } P(x) = 0(.4)^5 + 15(.6)^4(.4) + 210(.6)^3(.4)^2 \]

\[ + 35(10)(.6)^2(.4)^3 + 4(5)(.6)^3(.4)^4 + 5(.6)^5 = 3.0 \]

\[ \sigma = \sqrt{\sum (x - \mu)^2 \text{ } P(x)} \]

\[ = \sqrt{(9)(.01024) + 4(.0768) + 1(.2304) + 0(.3456) + 1(.2592) + 4(.07776) \]

\[ = 1.1 \]
EXERCISES

1. Alan Dershowitz, one of O.J. Simpson’s lawyers, has stated that only one out of every 1000 abusive relationships ends in murder each year. If he is correct, and if there are approximately 1,500,000 abusive relationships in the United States, what is the expected value for the number of people who are killed each year by their abusive partners?

2. If 3% of the population is allergic to the malaria fighting drug chloroquine, what is the expected value for the number of allergic people in a town of 1200 people?

3. A television game show has three payoffs with the following probabilities:

<table>
<thead>
<tr>
<th>Payoff ($)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.6</td>
</tr>
<tr>
<td>1000</td>
<td>.3</td>
</tr>
<tr>
<td>10,000</td>
<td>.1</td>
</tr>
</tbody>
</table>

What are the mean and standard deviation for the payoff variable?

4. Companies proved to have violated pollution laws are being fined various amounts with the following probabilities:

<table>
<thead>
<tr>
<th>Fine ($)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>.4</td>
</tr>
<tr>
<td>10,000</td>
<td>.3</td>
</tr>
<tr>
<td>50,000</td>
<td>.2</td>
</tr>
<tr>
<td>100,000</td>
<td>.1</td>
</tr>
</tbody>
</table>

What are the mean and standard deviation for the fine variable?

5. The New York Times (September 21, 1994) reported that an American woman diagnosed with ovarian cancer has a 37.5% chance of survival, and that approximately 20,000 American women are diagnosed with ovarian cancer each year.
   a. What is the expected number of deaths annually from this disease?
   b. What are the mean, variance, and standard deviation for a binomial with \( n = 20,000 \) and \( p = .375 \)?

6. Of the coral reef species in the Great Barrier Reef off Australia, 73% are poisonous. If a tourist boat bringing divers to different points off the reef encounters an average of 25 coral reef species, what are the mean and standard deviation for the expected number of poisonous species seen?

Study Questions

1. A famous problem asks you to choose between two envelopes, one of which has twice as much money as the other. You arbitrarily pick one, open it, and find $100. You are then given the chance to switch envelopes. You reason that the other envelope has either $50 or $200, each with a .5 probability. Applying your understanding of expected value, you calculate .5($50) + .5($200) = $125 and conclude that you should switch envelopes. Comment on this reasoning.
Part 3

PROBABILITY DISTRIBUTIONS

With an understanding of binomial probabilities and expected values, we can now develop and explore several probability distributions of general interest in statistics. Such knowledge not only is necessary for our future study, but also is immediately useful as a decision making tool for certain classes of problems.

For example, knowing the probability of finding oil given certain geological conditions, we can use the binomial distribution to calculate the probabilities of various numbers of positive strikes for a given number of test sites. Knowing the average number of Supreme Court vacancies during previous presidential terms, we can use the Poisson distribution to calculate the probabilities of various numbers of vacancies arising during the next 4-year term. Knowing the mean and variance of heights of U.S. Marines, we can use the normal distribution to calculate the probability that any Marine has a height greater than a specified value.

Our discussion in Chapter 8 of the properties of the binomial distribution arises from the concepts and calculations of Chapters 6 and 7. The Poisson distribution in Chapter 9 can be viewed as a limiting case of the binomial when \( n \) is large and \( p \) is small, and the normal distribution in Chapter 10 as a limiting case of the binomial when \( p \) is constant but \( n \) increases without bound. With this approach, it will be clear that both the Poisson and the normal distributions can be used as approximations to binomial problems.
ANSWERS TO EXERCISES

Chapter 1

1. The set is \{0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,3\} so the mean is 11/19, the median is 1, and the mode is 0.

2. The median is \((11,187 + 9,333)/2 = 10,260\), the mean is 42,785.75, and the trimmed mean (with the upper and lower 25% removed) is 17,252.

3. From \((0.70 - 0.11 + 1.34 + 0.05 + x)/5 = 0.086\), the return was -1.55%.

4. a. 255 lb
   b. 265 lb
   c. 280.5 lb

5. The geometric mean of 31.07 organisms per deciliter does not exceed the EPA limit.

6. Fifty percent of the values are on either side of the mean. Since 3% + 20% + 21% = 44% while 3% + 20% + 21% + 23% = 67%, the median number of partners for men must be in the 5–10 category. Similarly, the median number of partners for women is in the 2–4 category. (The actual answers from the study are a median of 6 for men and 2 for women.)

Chapter 2

1. Mean is 147.55, range is 82, and standard deviation is 25.13.

2. a. 17.59, 18.41, 5.94
   b. 19.09, 18.41, 5.94; 20.23, 21.17, 6.83

3. $0.17, $0.19, $0.06

Chapter 3

1. 86%, 92%, 4%

2. a. \((221 - 206)/35 = 0.43\)
   \((216 - 206)/35 = 0.29\)
   \((168 - 206)/35 = 1.09\)
   b. 206 + 2.69(35) = 300
   206 + 0.28(35) = 216
   206 - 1.09(35) = 168

3. Fifth out of 20, 75%, \((14.23 - 18.87)/5.33 = -0.87\)

4. \((10.1 - 7.7)/2.5 = 0.96\)

5. \((2310 - 1751)/552 = 1.03\)

6. a. \(62.5 \pm 2(18.0) = 62.5 \pm 36.0\) or \((26.5, 98.5)\)
   b. The index will fall below 26.5 about 5%/2 = 2.5% of the time.

7. Use Chebyshev's theorem.
   a. \(475 \pm 2(55) = (365, 585)\)
   b. \((475 - 310)/55 = 3\) and \((640 - 475)/55 = 3; 1 - 1/3^2 = 88.9\%\)

Chapter 4

1.

2.
3. First calculate $\mu = 138.7$ and $\sigma = 41.0$.

4. First calculate $\mu = 54.83$ and $\sigma = 6.22$.

5. a.

<table>
<thead>
<tr>
<th></th>
<th>11</th>
<th>0 7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>6 2 7 8 2 7 5 7 8 5 0</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>3 7 7</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>9 7 8 1 6</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0 9 0 5 2 5 0 4 4</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>4 5 9 4 8 2 8 6</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>7 6 4 6</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>5 3 0 3</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>8</td>
</tr>
</tbody>
</table>

b. Arranging the set in numerical order gives (110, 117, 120, 122, 122, 125, 125, 126, 127, 127, 127, 128, 128, 128, 133, 137, 137, 137, 141, 146, 147, 148, 149, 150, 150, 150, 152, 154, 154, 155, 155, 159, 162, 164, 164, 165, 166, 166, 168, 168, 169, 174, 176, 176, 177, 189, 190, 193, 193, 195, 207, 219, 228). The smallest value is 110, the largest 228. The median is 153, the median of the top half 169, the median of the bottom 128.

6. a.
b.  
| 28 | 3 2 3 |
| 27 | 0 1 1 2 2 3 4 4 4 4 7 7 8 8 |
| 26 | 0 0 1 2 4 4 5 5 6 7 7 7 7 9 |
| 25 | 1 1 5 7 8 8 8 9 9 9 9 9 9 |
| 24 | 6 9 |

c. \( \mu = 266.4 \) and \( \sigma = 9.0 \)
d. 75%; 40 of 41, or 97.6%, are in (248.4, 284.4)
e. 68%, 95%, 99%, 70.7%, 97.6%, 100%

7. The median is less than the mean, so the responses are probably skewed to the right; there are a few high guesses with most of the responses on the lower end of the scale.

Chapter 5
1. Using the multiplicative rule, Michener calculated 13 \( \times \) 17 \( \times \) 23 \( \times \) 37 = 188,071.
2. Boston Chicken advertisers calculated the permutation \( P(16,3) = 3360 \), when they should have calculated the combination \( C(16,3) = 560 \).
3. Using the multiplicative rule gives \( 27 = 128 \).

Chapter 6
1. a. \( \frac{1}{4} + \frac{1}{7} - (\frac{1}{4})(\frac{1}{7}) = \frac{5}{14} \)
   b. \( \frac{1}{4} \)
2. Marilyn assumed independence of events (success for surgery A and surgery B), which is most likely wrong!
3. If \( p = 1/9,000,000,000 \), then \( P(0 \text{ errors}) = (1-p)^{20,000,000} = .99778 \) and \( P(\text{at least 1 error}) = 1 - .99778 = .00222 \).
4. \((10.15)^3 = .138 \)
   \[ 1 - (\frac{1}{.85})^3 = \frac{5}{15}.(6.85)^3 = .165 \]
   \[ 5.15(\frac{1}{.85})^3 + 10(1.15)^3(\frac{1}{.85})^3 = .973 \]
5. a. 0.5
   b. \(.5^8 = .0039 \)
6. \( 1 - .99975 \) = .0247

Chapter 7
1. \(.001(1,500,000) = 1500 \)
2. \(.03(1200) = 36 \)
3. \( \mu = 0(6) + 1000(3) + 10000(1) = 1300 \)
   \( \sigma^2 = (0 - 1300)^2(6) + (1000 - 1300)^2(3) + (10,000 - 1300)^2(1) = 8,610,000 \) and \( \sigma = 2934 \)
4. \( \mu = 23,400 \) and \( \sigma = 31,350 \)
5. a. \( 20,000(0.625) = 12,500 \)
   b. \( \mu = 12,500, \sigma^2 = 4687.5, \sigma = 68.5 \)
6. \( \mu = 250(0.79) = 18.25, \sigma = 2.22 \)

Chapter 8
1. \( P(0) = .45^3 = .091125, P(1) = 3(.55)(.45) = .334125, \)
   \( P(2) = 3(.55)^2(.45) = .408375, P(3) = (.55)^3 = .166375 \)
2. a. \(.210, .367, .275, .115, .029, .004, .000, .000 \)
   b. \(.029 + .004 = .033 \)
3. a. \(.005, .049, .181, .336, .312, .116 \)
   b. \( \mu = 5(0.05) = 3.25, \sigma = 1.067 \)
   c. 

Chapter 9
1. \( e^{-1.6} = .202, e^{-3.2} = .041, e^{-8.0} = .003 \)
2. \( e^8 = .003, (8^5/5!e^8 = .092, (8^8/8!e^8 = .140 \)
3. \( 1 - [e^{-4} + 4e^{-4} + (4^2/2)e^{-4} + (4^3/3!e^{-4}] = .567 \)
   \[ 1 - e^{-2} + 2e^{-2} + (2^2/2!e^{-2} + (2^3/3!e^{-2}) = .143 \]
4. \( e^2 + 2e^{-2} + (2^2/2!e^{-2} + (2^3/3!e^{-2})e^{-2} = .857 \)
   \[ e^{10} + 10e^{-10} + (10^2/2!e^{-10} + (10^3/3!e^{-10} = .010 \)