Chapter

6

Point Estimation

6.1 DESCRIPTIVE STATISTICS

In Chapter 2, we considered probability distributions of random variables whose space S contains a countable number of outcomes: either a finite number of outcomes or outcomes that can be put into a one-to-one correspondence with the positive integers. Such a random variable is said to be of the **discrete type**, and its distribution of probabilities is of the discrete type.

Of course, many experiments or observations of random phenomena do not have integers or other discrete numbers as outcomes, but instead are measurements selected from an interval of numbers. For example, you could find the length of time that it takes when waiting in line to buy frozen yogurt. Or the weight of a “1-pound” package of hot dogs could be any number between 0.94 pounds and 1.25 pounds. The weight of a miniature Baby Ruth candy bar could be any number between 20 and 27 grams. Even though such times and weights could be selected from an interval of values, times and weights are generally rounded off so that the data often look like discrete data. If, conceptually, the measurements could come from an interval of possible outcomes, we call them data from a distribution of the continuous type or, more simply, **continuous-type data**.

Given a set of continuous-type data, we shall group the data into classes and then construct a histogram of the grouped data. This will help us better visualize the data. The following guidelines and terminology will be used to group continuous-type data into classes of equal length (these guidelines can also be used for sets of discrete data that have a large range).

1. Determine the largest (maximum) and smallest (minimum) observations. The **range** is the difference, $R =$ maximum $-$ minimum.
2. In general, select from $k = 5$ to $k = 20$ classes, which are nonoverlapping intervals, usually of equal length. These classes should cover the interval from the minimum to the maximum.
3. Each interval begins and ends halfway between two possible values of the measurements, which have been rounded off to a given number of decimal places.

4. The first interval should begin about as much below the smallest value as the last interval ends above the largest.
5. The intervals are called **class intervals** and the boundaries are called **class boundaries**. We shall denote these k class intervals by

$$(c_0, c_1], (c_1, c_2], \ldots, (c_{k-1}, c_k].$$

6. The **class limits** are the smallest and the largest possible observed (recorded) values in a class.
7. The **class mark** is the midpoint of a class.

A frequency table is constructed that lists the class intervals, the class limits, a tabulation of the measurements in the various classes, the frequency f_i of each class, and the class marks. A column is sometimes used to construct a relative frequency (density) histogram. With class intervals of equal length, a frequency histogram is constructed by drawing, for each class, a rectangle having as its base the class interval and a height equal to the frequency of the class. For the relative frequency histogram, each rectangle has an **area** equal to the relative frequency f_i/n of the observations for the class. That is, the function defined by

$$h(x) = \frac{f_i}{(n)(c_i - c_{i-1})}, \qquad \text{for } c_{i-1} < x \le c_i, \quad i = 1, 2, \ldots, k,$$

is called a **relative frequency histogram** or **density histogram**, where f_i is the frequency of the ith class and n is the total number of observations. Clearly, if the class intervals are of equal length, the relative frequency histogram, $h(x)$, is proportional to the **frequency histogram** f_i, for $c_{i-1} < x \le c_i, i = 1, 2, \ldots, k$. The frequency histogram should be used only in those situations in which the class intervals are of equal length. A relative frequency histogram can be treated as an estimate of the underlying pdf.

Example 6.1-1

The weights in grams of 40 miniature Baby Ruth candy bars, with the weights ordered, are given in Table 6.1-1.

We shall group these data and then construct a histogram to visualize the distribution of weights. The range of the data is $R = 26.7 - 20.5 = 6.2$. The interval $(20.5, 26.7)$ could be covered with $k = 8$ classes of width 0.8 or with $k = 9$ classes of width 0.7. (There are other possibilities.) We shall use $k = 7$ classes of width 0.9. The first class interval will be $(20.45, 21.35)$ and the last class interval will be $(25.85, 26.75)$. The data are grouped in Table 6.1-2.

A relative frequency histogram of these data is given in Figure 6.1-1. Note that the total area of this histogram is equal to 1. We could also construct a frequency histogram in which the heights of the rectangles would be equal to the frequencies of the classes. The shape of the two histograms is the same. Later we will see

Table 6.1-1 Candy bar weights

20.5	20.7	20.8	21.0	21.0	21.4	21.5	22.0	22.1	22.5
22.6	22.6	22.7	22.7	22.9	22.9	23.1	23.3	23.4	23.5
23.6	23.6	23.6	23.9	24.1	24.3	24.5	24.5	24.8	24.8
24.9	24.9	25.1	25.1	25.2	25.6	25.8	25.9	26.1	26.7

Table 6.1-2 Frequency table of candy bar weights

Class Interval	Class Limits	Tabulation	Frequency (f_i)	$h(x)$	Class Marks
$(20.45, 21.35)$	20.5–21.3	𝍸	5	5/36	20.9
$(21.35, 22.25)$	21.4–22.2	||||	4	4/36	21.8
$(22.25, 23.15)$	22.3–23.1	𝍸 |||	8	8/36	22.7
$(23.15, 24.05)$	23.2–24.0	𝍸 ||	7	7/36	23.6
$(24.05, 24.95)$	24.1–24.9	𝍸 |||	8	8/36	24.5
$(24.95, 25.85)$	25.0–25.8	𝍸	5	5/36	25.4
$(25.85, 26.75)$	25.9–26.7	|||	3	3/36	26.3

the reason for preferring the relative frequency histogram. In particular, we will be superimposing on the relative frequency histogram the graph of a pdf. ■

Suppose that we now consider the situation in which we actually perform a certain random experiment n times, obtaining n observed values of the random variable—say, $x_1, x_2, \ldots, x_n$. Often the collection is referred to as a **sample**. It is possible that some of these values might be the same, but we do not worry about this at this time. We artificially create a probability distribution by placing the weight $1/n$ on each of these x-values. Note that these weights are positive and sum to 1, so we have a distribution we call the **empirical distribution**, since it is determined by the data $x_1, x_2, \ldots, x_n$. The mean of the empirical distribution is

$$\sum_{i=1}^{n} x_i \left(\frac{1}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

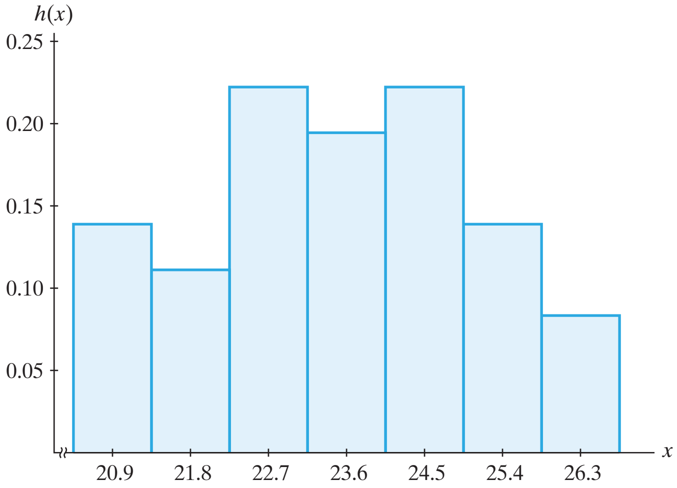


Figure 6.1-1 Relative frequency histogram of weights of candy bars

which is the arithmetic mean of the observations $x_1, x_2, \ldots, x_n$. We denote this mean by $\overline{x}$ and call it the **sample mean** (or mean of the sample $x_1, x_2, \ldots, x_n$). That is, the sample mean is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

which is, in some sense, an estimate of μ if the latter is unknown.

Likewise, the **variance of the empirical distribution** is

$$v = \sum_{i=1}^{n}(x_i - \overline{x})^2\left(\frac{1}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

which can be written as

$$v = \sum_{i=1}^{n} x_i^2\left(\frac{1}{n}\right) - \overline{x}^{\,2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^{\,2},$$

that is, the second moment about the origin minus the square of the mean. However, v is not called the sample variance, but

$$s^2 = \left[\frac{n}{n-1}\right]v = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

is, because we will see later that, in some sense, s^2 is a better estimate of an unknown σ^2 than is v. Thus, the **sample variance** is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

REMARK It is easy to expand the sum of squares; we have

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}.$$

Many find that the right-hand expression makes the computation easier than first taking the n differences, $x_i - \overline{x}$, $i = 1, 2, \ldots, n$; squaring them; and then summing. There is another advantage when $\overline{x}$ has many digits to the right of the decimal point. If that is the case, then $x_i - \overline{x}$ must be rounded off, and that creates an error in the sum of squares. In the easier form, that rounding off is not necessary until the computation is completed. Of course, if you are using a statistical calculator or statistics package on the computer, all of these computations are done for you. ■

The **sample standard deviation**, $s = \sqrt{s^2} \geq 0$, is a measure of how dispersed the data are from the sample mean. At this stage of your study of statistics, it is difficult to get a good understanding or meaning of the standard deviation s, but you can roughly think of it as the average distance of the values $x_1, x_2, \ldots, x_n$ from the mean $\overline{x}$. This is not true exactly, for, in general,

$$s \geq \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|,$$

but it is fair to say that s is somewhat larger, yet of the same magnitude, as the average of the distances of $x_1, x_2, \ldots, x_n$ from $\overline{x}$.

Example 6.1-2 Rolling a fair six-sided die five times could result in the following sample of $n = 5$ observations:

$$x_1 = 3,\;\; x_2 = 1,\;\; x_3 = 2,\;\; x_4 = 6,\;\; x_5 = 3.$$

In this case,

$$\bar{x} = \frac{3+1+2+6+3}{5} = 3$$

and

$$s^2 = \frac{(3-3)^2 + (1-3)^2 + (2-3)^2 + (6-3)^2 + (3-3)^2}{4} = \frac{14}{4} = 3.5.$$

It follows that $s = \sqrt{14/4} = 1.87$. We had noted that s can roughly be thought of as the average distance that the x-values are away from the sample mean $\bar{x}$. In this example, the distances from the sample mean, $\bar{x} = 3$, are 0, 2, 1, 3, 0, with an average of 1.2, which is less than $s = 1.87$. In general, s will be somewhat larger than this average distance. ■

There is an alternative way of computing s^2, because $s^2 = [n/(n-1)]v$ and

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

It follows that

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{n-1}.$$

Given a set of measurements, the sample mean is the center of the data such that the deviations from that center sum to zero; that is, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, where $x_1, x_2, \ldots, x_n$ and $\bar{x}$ are a given set of observations of $X_1, X_2, \ldots, X_n$ and $\overline{X}$. The sample standard deviation s, an observed value of S, gives a measure of how spread out the data are from the sample mean. If the histogram is "mound-shaped" or "bell-shaped," the following empirical rule gives rough approximations to the percentages of the data that fall between certain points. These percentages clearly are associated with the normal distribution.

Empirical Rule: Let $x_1, x_2, \ldots, x_n$ have a sample mean $\bar{x}$ and sample standard deviation s. If the histogram of these data is "bell-shaped," then, for large samples,

- approximately 68% of the data are in the interval $(\bar{x} - s, \bar{x} + s)$,
- approximately 95% of the data are in the interval $(\bar{x} - 2s, \bar{x} + 2s)$,
- approximately 99.7% of the data are in the interval $(\bar{x} - 3s, \bar{x} + 3s)$.

For the data in Example 6.1-1, the sample mean is $\bar{x} = 23.505$ and the standard deviation is $s = 1.641$. The number of weights that fall within one standard deviation of the mean, $(23.505 - 1.641, 23.505 + 1.641)$, is 27, or 67.5%. For these particular weights, 100% fall within two standard deviations of $\bar{x}$. Thus, the histogram is missing part of the "bell" in the tails in order for the empirical rule to hold.

When you draw a histogram, it is useful to indicate the location of $\bar{x}$, as well as that of the points $\bar{x} \pm s$ and $\bar{x} \pm 2s$.

There is a refinement of the relative frequency histogram that can be made when the class intervals are of equal length. The **relative frequency polygon** smooths out the corresponding histogram somewhat. To form such a polygon, mark the midpoints at the top of each "bar" of the histogram. Connect adjacent midpoints with straight-line segments. On each of the two end bars, draw a line segment from the top middle mark through the middle point of the outer vertical line of the bar. Of course, if the area underneath the tops of the relative frequency histogram is equal to 1, which it should be, then the area underneath the relative frequency polygon is also equal to 1, because the areas lost and gained cancel out by a consideration of congruent triangles. This idea is made clear in the next example.

Example 6.1-3

A manufacturer of fluoride toothpaste regularly measures the concentration of fluoride in the toothpaste to make sure that it is within the specification of 0.85 to 1.10 mg/g. Table 6.1-3 lists 100 such measurements.

The minimum of these measurements is 0.85 and the maximum is 1.06. The range is $1.06 - 0.85 = 0.21$. We shall use $k = 8$ classes of length 0.03. Note that $8(0.03) = 0.24 > 0.21$. We start at 0.835 and end at 1.075. These boundaries are the same distance below the minimum and above the maximum. In Table 6.1-4, we also give the values of the heights of each rectangle in the relative frequency histogram, so that the total area of the histogram is 1. These heights are given by the formula

$$h(x) = \frac{f_i}{(0.03)(100)} = \frac{f_i}{3}.$$

The plots of the relative frequency histogram and polygon are given in Figure 6.1-2.

If you are using a computer program to analyze a set of data, it is very easy to find the sample mean, the sample variance, and the sample standard deviation. However, if you have only grouped data or if you are not using a computer, you can obtain close approximations of these values by computing the mean $\overline{u}$ and

Table 6.1-3 Concentrations of fluoride in mg/g in toothpaste

0.98	0.92	0.89	0.90	0.94	0.99	0.86	0.85	1.06	1.01
1.03	0.85	0.95	0.90	1.03	0.87	1.02	0.88	0.92	0.88
0.88	0.90	0.98	0.96	0.98	0.93	0.98	0.92	1.00	0.95
0.88	0.90	1.01	0.98	0.85	0.91	0.95	1.01	0.88	0.89
0.99	0.95	0.90	0.88	0.92	0.89	0.90	0.95	0.93	0.96
0.93	0.91	0.92	0.86	0.87	0.91	0.89	0.93	0.93	0.95
0.92	0.88	0.87	0.98	0.98	0.91	0.93	1.00	0.90	0.93
0.89	0.97	0.98	0.91	0.88	0.89	1.00	0.93	0.92	0.97
0.97	0.91	0.85	0.92	0.87	0.86	0.91	0.92	0.95	0.97
0.88	1.05	0.91	0.89	0.92	0.94	0.90	1.00	0.90	0.93

Table 6.1-4 Frequency table of fluoride concentrations

Class Interval	Class Mark (u_i)	Tabulation	Frequency (f_i)	$h(x) = f_i/3$
$(0.835, 0.865)$	0.85	𝍸 ||	7	7/3
$(0.865, 0.895)$	0.88	𝍸 𝍸 𝍸 𝍸	20	20/3
$(0.895, 0.925)$	0.91	𝍸 𝍸 𝍸 𝍸 𝍸 ||	27	27/3
$(0.925, 0.955)$	0.94	𝍸 𝍸 𝍸 |||	18	18/3
$(0.955, 0.985)$	0.97	𝍸 𝍸 ||||	14	14/3
$(0.985, 1.015)$	1.00	𝍸 ||||	9	9/3
$(1.015, 1.045)$	1.03	|||	3	3/3
$(1.045, 1.075)$	1.06	||	2	2/3

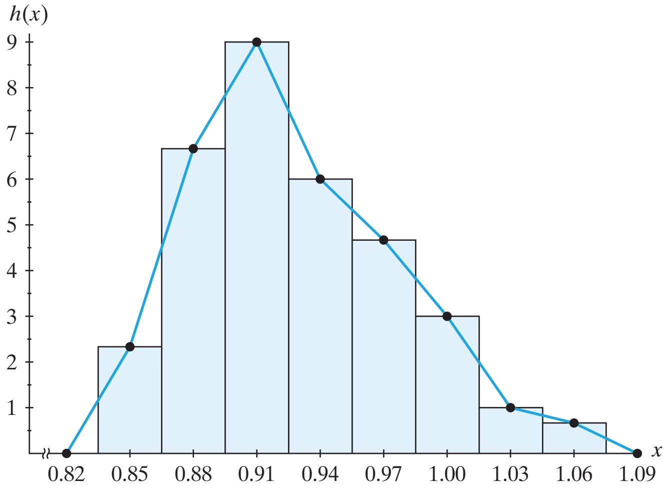


Figure 6.1-2 Concentrations of fluoride in toothpaste

variance s_u^2 of the grouped data, using the class marks weighted with their respective frequencies. We have

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{k} f_i u_i$$

$$= \frac{1}{100}\sum_{i=1}^{8} f_i u_i = \frac{92.83}{100} = 0.9283,$$

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(u_i - \overline{u})^2 = \frac{\sum_{i=1}^{k} f_i u_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} f_i u_i\right)^2}{n-1}$$

$$= \frac{0.237411}{99} = 0.002398.$$

Thus,

$$s_u = \sqrt{0.002398} = 0.04897.$$

These results compare rather favorably with $\overline{x} = 0.9293$ and $s_x = 0.04895$ of the original data. ■

In some situations, it is not necessarily desirable to use class intervals of equal widths in the construction of the frequency distribution and histogram. This is particularly true if the data are skewed with a very long tail. We now present an illustration in which it seems desirable to use class intervals of unequal widths; thus, we cannot use the relative frequency polygon.

Example 6.1-4 The following 40 losses, due to wind-related catastrophes, were recorded to the nearest \$1 million (these data include only losses of \$2 million or more; for convenience, they have been ordered and recorded in millions):

2	2	2	2	2	2	2	2	2	2
2	2	3	3	3	3	4	4	4	5
5	5	5	6	6	6	6	8	8	9
15	17	22	23	24	24	25	27	32	43

The selection of class boundaries is more subjective in this case. It makes sense to let $c_0 = 1.5$ and $c_1 = 2.5$ because only values of \$2 million or more are recorded and there are 12 observations equal to 2. We could then let $c_2 = 6.5$, $c_3 = 29.5$, and $c_4 = 49.5$, yielding the following relative frequency histogram:

$$h(x) = \begin{cases} \dfrac{12}{40}, & 1.5 < x \le 2.5, \\ \dfrac{15}{(40)(4)}, & 2.5 < x \le 6.5, \\ \dfrac{11}{(40)(23)}, & 6.5 < x \le 29.5, \\ \dfrac{2}{(40)(20)}, & 29.5 < x \le 49.5. \end{cases}$$

This histogram is displayed in Figure 6.1-3. It takes some experience before a person can display a relative frequency histogram that is most meaningful.

The areas of the four rectangles—0.300, 0.375, 0.275, and 0.050—are the respective relative frequencies. It is important to note in the case of unequal widths among class intervals that the *areas*, not the heights, of the rectangles are proportional to the frequencies. In particular, the first and second classes have frequencies $f_1 = 12$

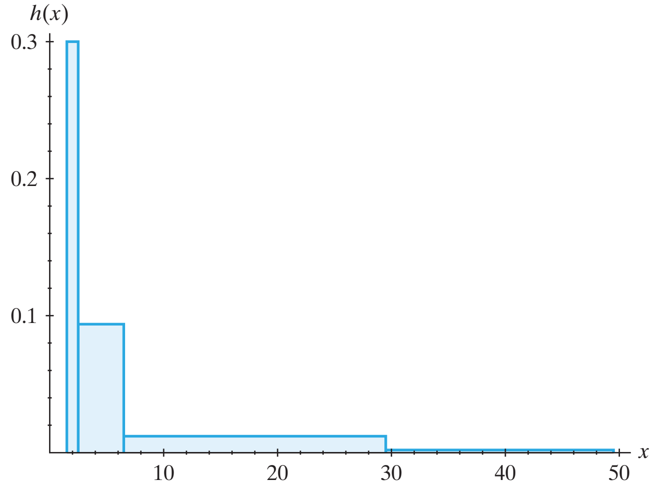


Figure 6.1-3 Relative frequency histogram of losses

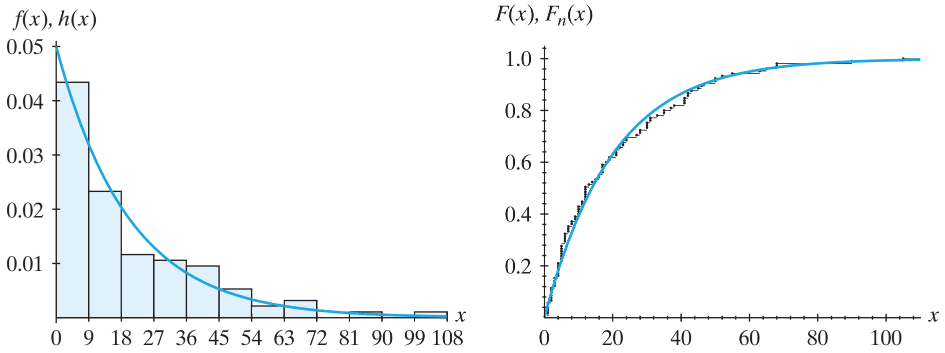
and $f_2 = 15$, yet the height of the first is greater than the height of the second, while here $f_1 < f_2$. If we have equal widths among the class intervals, then the heights are proportional to the frequencies. ■

For continuous-type data, the interval with the largest class height is called the **modal class** and the respective class mark is called the **mode**. Hence, in the last example, $x = 2$ is the mode and $(1.5, 2.5)$ the modal class.

Example 6.1-5 The following table lists 105 observations of X, the times in minutes between calls to 911:

30	17	65	8	38	35	4	19	7	14	12	4	5	4	2
7	5	12	50	33	10	15	2	10	1	5	30	41	21	31
1	18	12	5	24	7	6	31	1	3	2	22	1	30	2
1	3	12	12	9	28	6	50	63	5	17	11	23	2	46
90	13	21	55	43	5	19	47	24	4	6	27	4	6	37
16	41	68	9	5	28	42	3	42	8	52	2	11	41	4
35	21	3	17	10	16	1	68	105	45	23	5	10	12	17

To help determine visually whether the exponential model in Example 3.2-1 is perhaps appropriate for this situation, we shall look at two graphs. First, we have constructed a relative frequency histogram, $h(x)$, of these data in Figure 6.1-4(a), with $f(x) = (1/20)e^{-x/20}$ superimposed. Second, we have also constructed the empirical cdf of these data in Figure 6.1-4(b), with the theoretical cdf superimposed. Note



(a) Histogram and pdf of X

(b) Theoretical and empirical cdfs

Figure 6.1-4 Times between calls to 911

that $F_n(x)$, the **empirical cumulative distribution function**, is a step function with a vertical step of size $1/n$ at each observation of X. If k observations are equal, the step at that value is k/n.

STATISTICAL COMMENTS (**Simpson's Paradox**) While most of the first five chapters were about probability and probability distributions, we now mention some statistical concepts. The relative frequency, f/n, is called a **statistic** and is used to **estimate** a probability, p, which is usually unknown. For example, if a major league batter gets $f = 152$ hits in $n = 500$ official at bats during the season, then the relative frequency $f/n = 0.304$ is an estimate of his probability of getting a hit and is called his batting average for that season.

Once while speaking to a group of coaches, one of us (Hogg) made the comment that it would be possible for batter A to have a higher average than batter B for each season during their careers and yet B could have a better overall average at the end of their careers. While no coach spoke up, you could tell that they were thinking, "And that guy is supposed to know something about math."

Of course, the following simple example convinced them that the statement was true: Suppose A and B played only two seasons, with these results:

	Player A			Player B		
Season	AB	Hits	Average	AB	Hits	Average
1	500	126	0.252	300	75	0.250
2	300	90	0.300	500	145	0.290
Totals	800	216	0.270	800	220	0.275

Clearly, A beats B in the two individual seasons, but B has a better overall average. Note that during their better season (the second), B had more at bats than did A. This kind of result is often called **Simpson's paradox** and it can happen in real life. (See Exercises 6.1-10 and 6.1-11.)

Exercises

6.1-1. One characteristic of a car's storage console that is checked by the manufacturer is the time in seconds that it takes for the lower storage compartment door to open completely. A random sample of size $n = 5$ yielded the following times:

1.1	0.9	1.4	1.1	1.0

(a) Find the sample mean, $\overline{x}$.

(b) Find the sample variance, s^2.

(c) Find the sample standard deviation, s.

6.1-2. A leakage test was conducted to determine the effectiveness of a seal designed to keep the inside of a plug airtight. An air needle was inserted into the plug, which was then placed underwater. Next, the pressure was increased until leakage was observed. The magnitude of this pressure in psi was recorded for 10 trials:

3.1	3.5	3.3	3.7	4.5	4.2	2.8	3.9	3.5	3.3

Find the sample mean and sample standard deviation for these 10 measurements.

6.1-3. During the course of an internship at a company that manufactures diesel engine fuel injector pumps, a student had to measure the category "plungers that force the fuel out of the pumps." This category is based on a relative scale, measuring the difference in diameter (in microns or micrometers) of a plunger from that of an absolute minimum acceptable diameter. For 96 plungers randomly taken from the production line, the data are as follows:

17.1	19.3	18.0	19.4	16.5	14.4	15.8	16.6	18.5	14.9
14.8	16.3	20.8	17.8	14.8	15.6	16.7	16.1	17.1	16.5
18.8	19.3	18.1	16.1	18.0	17.2	16.8	17.3	14.4	14.1
16.9	17.6	15.5	17.8	17.2	17.4	18.1	18.4	17.8	16.7
17.2	13.7	18.0	15.6	17.8	17.0	17.7	11.9	15.9	17.8
15.5	14.6	15.6	15.1	15.4	16.1	16.6	17.1	19.1	15.0
17.6	19.7	17.1	13.6	15.6	16.3	14.8	17.4	14.8	14.9
14.1	17.8	19.8	18.9	15.6	16.1	15.9	15.7	22.1	16.1
18.9	21.5	17.4	12.3	20.2	14.9	17.1	15.0	14.4	14.7
15.9	19.0	16.6	15.3	17.7	15.8				

(a) Calculate the sample mean and the sample standard deviation of these measurements.

(b) Use the class boundaries $10.95, 11.95, \ldots, 22.95$ to construct a histogram of the data.

6.1-4. Ledolter and Hogg (see References) report that a manufacturer of metal alloys is concerned about customer complaints regarding the lack of uniformity in the melting points of one of the firm's alloy filaments. Fifty filaments are selected and their melting points determined. The following results were obtained:

320	326	325	318	322	320	329	317	316	331
320	320	317	329	316	308	321	319	322	335
318	313	327	314	329	323	327	323	324	314
308	305	328	330	322	310	324	314	312	318
313	320	324	311	317	325	328	319	310	324

(a) Construct a frequency distribution and display the histogram of the data.

(b) Calculate the sample mean and sample standard deviation.

(c) Locate $\overline{x}$ and $\overline{x} \pm s$, and $\overline{x} \pm 2s$ on your histogram. How many observations lie within one standard deviation of the mean? How many lie within two standard deviations of the mean?

6.1-5. In the casino game roulette, if a player bets \$1 on red, the probability of winning \$1 is 18/38 and the probability of losing \$1 is 20/38. Let X equal the number of successive \$1 bets that a player makes before losing \$5. One hundred observations of X were simulated on a computer, yielding the following data:

23	127	877	65	101	45	61	95	21	43
53	49	89	9	75	93	71	39	25	91
15	131	63	63	41	7	37	13	19	413
65	43	35	23	135	703	83	7	17	65
49	177	61	21	9	27	507	7	5	87
13	213	85	83	75	95	247	1815	7	13
71	67	19	615	11	15	7	131	47	25
25	5	471	11	5	13	75	19	307	33
57	65	9	57	35	19	9	33	11	51
27	9	19	63	109	515	443	11	63	9

(a) Find the sample mean and sample standard deviation of these data.

(b) Construct a relative frequency histogram of the data, using about 10 classes. The classes do not need to be of the same length.

(c) Locate $\overline{x}$, $\overline{x} \pm s$, $\overline{x} \pm 2s$, and $\overline{x} \pm 3s$ on your histogram.

(d) In your opinion, does the median or sample mean give a better measure of the center of these data?

6.1-6. An insurance company experienced the following mobile home losses in 10,000's of dollars for 50 catastrophic events:

1	2	2	3	3	4	4	5	5	5
5	6	7	7	9	9	9	10	11	12
22	24	28	29	31	33	36	38	38	38
39	41	48	49	53	55	74	82	117	134
192	207	224	225	236	280	301	308	351	527

(a) Using class boundaries 0.5, 5.5, 17.5, 38.5, 163.5, and 549.5, group these data into five classes.

(b) Construct a relative frequency histogram of the data.

(c) Describe the distribution of losses.

6.1-7. Ledolter and Hogg (see References) report 64 observations that are a sample of daily weekday afternoon (3 to 7 P.M.) lead concentrations (in micrograms per cubic meter, $\mu\text{g/m}^3$). The following data were recorded at an air-monitoring station near the San Diego Freeway in Los Angeles during the fall of 1976:

6.7	5.4	5.2	6.0	8.7	6.0	6.4	8.3	5.3	5.9	7.6
5.0	6.9	6.8	4.9	6.3	5.0	6.0	7.2	8.0	8.1	7.2
10.9	9.2	8.6	6.2	6.1	6.5	7.8	6.2	8.5	6.4	8.1
2.1	6.1	6.5	7.9	14.1	9.5	10.6	8.4	8.3	5.9	6.0
6.4	3.9	9.9	7.6	6.8	8.6	8.5	11.2	7.0	7.1	6.0
9.0	10.1	8.0	6.8	7.3	9.7	9.3	3.2	6.4		

(a) Construct a frequency distribution of the data and display the results in the form of a histogram. Is this distribution symmetric?

(b) Calculate the sample mean and sample standard deviation.

(c) Locate $\overline{x}$ and $\overline{x} \pm s$ on your histogram. How many observations lie within one standard deviation of the mean? How many lie within two standard deviations of the mean?

6.1-8. A small part for an automobile rearview mirror was produced on two different punch presses. In order to describe the distribution of the weights of those parts, a random sample was selected, and each piece was weighed in grams, resulting in the following data set:

3.968	3.534	4.032	3.912	3.572	4.014	3.682	3.608
3.669	3.705	4.023	3.588	3.945	3.871	3.744	3.711
3.645	3.977	3.888	3.948	3.551	3.796	3.657	3.667
3.799	4.010	3.704	3.642	3.681	3.554	4.025	4.079
3.621	3.575	3.714	4.017	4.082	3.660	3.692	3.905
3.977	3.961	3.948	3.994	3.958	3.860	3.965	3.592
3.681	3.861	3.662	3.995	4.010	3.999	3.993	4.004
3.700	4.008	3.627	3.970	3.647	3.847	3.628	3.646
3.674	3.601	4.029	3.603	3.619	4.009	4.015	3.615
3.672	3.898	3.959	3.607	3.707	3.978	3.656	4.027
3.645	3.643	3.898	3.635	3.865	3.631	3.929	3.635
3.511	3.539	3.830	3.925	3.971	3.646	3.669	3.931
4.028	3.665	3.681	3.984	3.664	3.893	3.606	3.699
3.997	3.936	3.976	3.627	3.536	3.695	3.981	3.587
3.680	3.888	3.921	3.953	3.847	3.645	4.042	3.692
3.910	3.672	3.957	3.961	3.950	3.904	3.928	3.984
3.721	3.927	3.621	4.038	4.047	3.627	3.774	3.983
3.658	4.034	3.778					

(a) Using about 10 (say, 8 to 12) classes, construct a frequency distribution.

(b) Draw a histogram of the data.

(c) Describe the shape of the distribution represented by the histogram.

6.1-9. Old Faithful is a geyser in Yellowstone National Park. Tourists always want to know when the next eruption will occur, so data have been collected to help make those predictions. In the following data set, observations were made on several consecutive days, and the data recorded give the starting time of the eruption (STE); the duration of the eruption, in seconds (DIS); the predicted time until the next eruption, in minutes (PTM); the actual time until the next eruption, in minutes (ATM); and the duration of the eruption, in minutes (DIM).

STE	DIS	PTM	ATM	DIM	STE	DIS	PTM	ATM	DIM
706	150	65	72	2.500	1411	110	55	65	1.833
818	268	89	88	4.467	616	289	89	97	4.817
946	140	65	62	2.333	753	114	58	52	1.900
1048	300	95	87	5.000	845	271	89	94	4.517
1215	101	55	57	1.683	1019	120	58	60	2.000
1312	270	89	94	4.500	1119	279	89	84	4.650
651	270	89	91	4.500	1253	109	55	63	1.817
822	125	59	51	2.083	1356	295	95	91	4.917
913	262	89	98	4.367	608	240	85	83	4.000
1051	95	55	59	1.583	731	259	86	84	4.317
1150	270	89	93	4.500	855	128	60	71	2.133
637	273	89	86	4.550	1006	287	92	83	4.783
803	104	55	70	1.733	1129	253	65	70	4.217
913	129	62	63	2.150	1239	284	89	81	4.733
1016	264	89	91	4.400	608	120	58	60	2.000
1147	239	82	82	3.983	708	283	92	91	4.717
1309	106	55	58	1.767	839	115	58	51	1.917
716	259	85	97	4.317	930	254	85	85	4.233
853	115	55	59	1.917	1055	94	55	55	1.567
952	275	89	90	4.583	1150	274	89	98	4.567
1122	110	55	58	1.833	1328	128	64	49	2.133
1220	286	92	98	4.767	557	270	93	85	4.500
735	115	55	55	1.917	722	103	58	65	1.717
830	266	89	107	4.433	827	287	89	102	4.783
1017	105	55	61	1.750	1009	111	55	56	1.850
1118	275	89	82	4.583	1105	275	89	86	4.583
1240	226	79	91	3.767	1231	104	55	62	1.733

(a) Construct a histogram of the durations of the eruptions, in seconds. Use 10 to 12 classes.

(b) Calculate the sample mean and locate it on your histogram. Does it give a good measure of the average length of an eruption? Why or why not?

(c) Construct a histogram of the lengths of the times between eruptions. Use 10 to 12 classes.

(d) Calculate the sample mean and locate it on your histogram. Does it give a good measure of the average length of the times between eruptions?

6.1-10. In 1985, Kent Hrbek of the Minnesota Twins and Dion James of the Milwaukee Brewers had the following numbers of hits (H) and official at bats (AB) on grass and artificial turf:

	Hrbek			James		
Playing Surface	AB	H	BA	AB	H	BA
Grass	204	50		329	93	
Artificial Turf	355	124		58	21	
Total	559	174		387	114	

(a) Find the batting average BA (namely, H/AB) of each player on grass.

(b) Find the BA of each player on artificial turf.

(c) Find the season batting averages for the two players.

(d) Interpret your results.

6.1-11. In 1985, Al Bumbry of the Baltimore Orioles and Darrell Brown of the Minnesota Twins had the following numbers of hits (H) and official at bats (AB) on grass and artificial turf:

	Bumbry			Brown		
Playing Surface	AB	H	BA	AB	H	BA
Grass	295	77		92	18	
Artificial Turf	49	16		168	53	
Total	344	93		260	71	

(a) Find the batting average BA (namely, H/AB) of each player on grass.

(b) Find the BA of each player on artificial turf.

(c) Find the season batting averages for the two players.

(d) Interpret your results.

6.2 EXPLORATORY DATA ANALYSIS

To explore the other characteristics of an unknown distribution, we need to take a sample of n observations, $x_1, x_2, \ldots, x_n$, from that distribution and often need to order them from the smallest to the largest. One convenient way of doing this is to use a stem-and-leaf display, a method that was started by John W. Tukey. [For more details, see the books by Tukey (1977) and Velleman and Hoaglin (1981).]

Possibly the easiest way to begin is with an example to which all of us can relate. Say we have the following 50 test scores on a statistics examination:

93	77	67	72	52	83	66	84	59	63
75	97	84	73	81	42	61	51	91	87
34	54	71	47	79	70	65	57	90	83
58	69	82	76	71	60	38	81	74	69
68	76	85	58	45	73	75	42	93	65

We can do much the same thing as a frequency table and histogram can, but keep the original values, through a **stem-and-leaf display**. For this particular data set, we could use the following procedure: The first number in the set, 93, is recorded by treating the 9 (in the tens place) as the stem and the 3 (in the units place) as the corresponding leaf. Note that this leaf of 3 is the first digit after the stem of 9 in Table 6.2-1. The second number, 77, is that given by the leaf of 7 after the stem of 7; the third number, 67, by the leaf of 7 after the stem of 6; the fourth number, 72, as the leaf of 2 after the stem of 7 (note that this is the second leaf on the 7 stem); and so on. Table 6.2-1 is an example of a stem-and-leaf display. If the leaves are carefully aligned vertically, this table has the same effect as a histogram, but the original numbers are not lost.

It is useful to modify the stem-and-leaf display by ordering the leaves in each row from smallest to largest. The resulting stem-and-leaf diagram is called an **ordered stem-and-leaf display**. Table 6.2-2 uses the data from Table 6.2-1 to produce an ordered stem-and-leaf display.

There is another modification that can also be helpful. Suppose that we want two rows of leaves with each original stem. We can do this by recording leaves 0, 1, 2, 3, and 4 with a stem adjoined with an asterisk ($*$) and leaves 5, 6, 7, 8, and 9 with

Table 6.2-1 Stem-and-leaf display of scores from 50 statistics examinations

Stems	Leaves	Frequency
3	4 8	2
4	2 7 5 2	4
5	2 9 1 4 7 8 8	7
6	7 6 3 1 5 9 0 9 8 5	10
7	7 2 5 3 1 9 0 6 1 4 6 3 5	13
8	3 4 4 1 7 3 2 1 5	9
9	3 7 1 0 3	5

Table 6.2-2 Ordered stem-and-leaf display of statistics examinations

Stems	Leaves	Frequency
3	4 8	2
4	2 2 5 7	4
5	1 2 4 7 8 8 9	7
6	0 1 3 5 5 6 7 8 9 9	10
7	0 1 1 2 3 3 4 5 5 6 6 7 9	13
8	1 1 2 3 3 4 4 5 7	9
9	0 1 3 3 7	5

a stem adjoined with a dot (•). Of course, in our example, by going from 7 original classes to 14 classes, we lose a certain amount of smoothness with this particular data set, as illustrated in Table 6.2-3, which is also ordered.

Tukey suggested another modification, which is used in the next example.

Table 6.2-3 Ordered stem-and-leaf display of statistics examinations

Stems	Leaves	Frequency
3∗	4	1
3•	8	1
4∗	2 2	2
4•	5 7	2
5∗	1 2 4	3
5•	7 8 8 9	4
6∗	0 1 3	3
6•	5 5 6 7 8 9 9	7
7∗	0 1 1 2 3 3 4	7
7•	5 5 6 6 7 9	6
8∗	1 1 2 3 3 4 4	7
8•	5 7	2
9∗	0 1 3 3	4
9•	7	1

Example 6.2-1 The following numbers represent ACT composite scores for 60 entering freshmen at a certain college:

26	19	22	28	31	29	25	23	20	33	23	26
30	27	26	29	20	23	18	24	29	27	32	24
25	26	22	29	21	24	20	28	23	26	30	19
27	21	32	28	29	23	25	21	28	22	25	24
19	24	35	26	25	20	31	27	23	26	30	29

An ordered stem-and-leaf display of these scores is given in Table 6.2-4, where leaves are recorded as zeros and ones with a stem adjoined with an asterisk ($*$), twos and threes with a stem adjoined with t, fours and fives with a stem adjoined with f, sixes and sevens with a stem adjoined with s, and eights and nines with a stem adjoined with a dot ($\bullet$).

There is a reason for constructing ordered stem-and-leaf diagrams. For a sample of n observations, $x_1, x_2, \ldots, x_n$, when the observations are ordered from smallest to largest, the resulting ordered data are called the **order statistics** of the sample. Statisticians have found that order statistics and certain of their functions are extremely valuable; we will provide some theory concerning them in Section 6.3. It is very easy to determine the values of the sample in order from an ordered stem-and-leaf display. As an illustration, consider the values in Table 6.2-2 or Table 6.2-3. The order statistics of the 50 test scores are given in Table 6.2-5.

Sometimes we give ranks to these order statistics and use the rank as the subscript on y. The first order statistic $y_1 = 34$ has rank 1; the second order statistic $y_2 = 38$ has rank 2; the third order statistic $y_3 = 42$ has rank 3; the fourth order statistic $y_4 = 42$ has rank 4, $\ldots$; and the 50th order statistic $y_{50} = 97$ has rank 50. It is also about as easy to determine these values from the ordered stem-and-leaf display. We see that $y_1 \le y_2 \le \cdots \le y_{50}$.

Table 6.2-4 Ordered stem-and-leaf display of 60 ACT scores

Stems	Leaves	Frequency
1$\bullet$	8 9 9 9	4
2$*$	0 0 0 0 1 1 1	7
2t	2 2 2 3 3 3 3 3 3	9
2f	4 4 4 4 4 5 5 5 5 5	10
2s	6 6 6 6 6 6 6 7 7 7 7	11
2$\bullet$	8 8 8 8 9 9 9 9 9 9	10
3$*$	0 0 0 1 1	5
3t	2 2 3	3
3f	5	1

Table 6.2-5 Order statistics of 50 exam scores

34	38	42	42	45	47	51	52	54	57
58	58	59	60	61	63	65	65	66	67
68	69	69	70	71	71	72	73	73	74
75	75	76	76	77	79	81	81	82	83
83	84	84	85	87	90	91	93	93	97

From either these order statistics or the corresponding ordered stem-and-leaf display, it is rather easy to find the **sample percentiles**. If $0 < p < 1$, then the $(100p)$th sample percentile has *approximately* np sample observations less than it and also $n(1-p)$ sample observations greater than it. One way of achieving this is to take the $(100p)$th sample percentile as the $(n+1)p$th order statistic, provided that $(n+1)p$ is an integer. If $(n+1)p$ is not an integer but is equal to r plus some proper fraction—say, a/b—use a weighted average of the rth and the $(r+1)$st order statistics. That is, define the $(100p)$th sample percentile as

$$\widetilde{\pi}_p = y_r + (a/b)(y_{r+1} - y_r) = (1 - a/b)y_r + (a/b)y_{r+1}.$$

Note that this formula is simply a linear interpolation between y_r and y_{r+1}. [If $p < 1/(n+1)$ or $p > n/(n+1)$, that sample percentile is not defined.]

As an illustration, consider the 50 ordered test scores. With $p = 1/2$, we find the 50th percentile by averaging the 25th and 26th order statistics, since $(n+1)p = (51)(1/2) = 25.5$. Thus, the 50th percentile is

$$\widetilde{\pi}_{0.50} = (1/2)y_{25} + (1/2)y_{26} = (71 + 71)/2 = 71.$$

With $p = 1/4$, we have $(n+1)p = (51)(1/4) = 12.75$, and the 25th sample percentile is then

$$\widetilde{\pi}_{0.25} = (1 - 0.75)y_{12} + (0.75)y_{13} = (0.25)(58) + (0.75)(59) = 58.75.$$

With $p = 3/4$, so that $(n+1)p = (51)(3/4) = 38.25$, the 75th sample percentile is

$$\widetilde{\pi}_{0.75} = (1 - 0.25)y_{38} + (0.25)y_{39} = (0.75)(81) + (0.25)(82) = 81.25.$$

Note that *approximately* 50%, 25%, and 75% of the sample observations are less than 71, 58.75, and 81.25, respectively.

Special names are given to certain percentiles. The 50th percentile is the **median** of the sample. The 25th, 50th, and 75th percentiles are, respectively, the **first, second**, and **third quartiles** of the sample. For notation, we let $\widetilde{q}_1 = \widetilde{\pi}_{0.25}$, $\widetilde{q}_2 = \widetilde{m} = \widetilde{\pi}_{0.50}$, and $\widetilde{q}_3 = \widetilde{\pi}_{0.75}$. The 10th, 20th, . . . , and 90th percentiles are the **deciles** of the sample, so note that the 50th percentile is also the median, the second quartile, and the fifth decile. With the set of 50 test scores, since $(51)(2/10) = 10.2$ and $(51)(9/10) = 45.9$, the second and ninth deciles are, respectively,

$$\widetilde{\pi}_{0.20} = (0.8)y_{10} + (0.2)y_{11} = (0.8)(57) + (0.2)(58) = 57.2$$

and

$$\widetilde{\pi}_{0.90} = (0.1)y_{45} + (0.9)y_{46} = (0.1)(87) + (0.9)(90) = 89.7.$$

The second decile is commonly called the 20th percentile, and the ninth decile is the 90th percentile.

Example 6.2-2

We illustrate the preceding ideas with the fluoride data given in Table 6.1-3. For convenience, we use 0.02 as the length of a class interval. The ordered stem-and-leaf display is given in Table 6.2-6.

This ordered stem-and-leaf diagram is useful for finding sample percentiles of the data. ■

We now find some of the sample percentiles associated with the fluoride data. Since $n = 100$, $(n+1)(0.25) = 25.25$, $(n+1)(0.50) = 50.5$, and $(n+1)(0.75) = 75.75$, so that the 25th, 50th, and 75th percentiles are, respectively,

$$\begin{aligned}
\widetilde{\pi}_{0.25} = (0.75)y_{25} + (0.25)y_{26} &= (0.75)(0.89) + (0.25)(0.89) &= 0.89,\\
\widetilde{\pi}_{0.50} = (0.50)y_{50} + (0.50)y_{51} &= (0.50)(0.92) + (0.50)(0.92) &= 0.92,\\
\widetilde{\pi}_{0.75} = (0.25)y_{75} + (0.75)y_{76} &= (0.25)(0.97) + (0.75)(0.97) &= 0.97.
\end{aligned}$$

These three percentiles are often called the **first quartile**, the **median** or **second quartile**, and the **third quartile**, respectively. Along with the smallest (the **minimum**) and largest (the **maximum**) values, they give the **five-number summary** of a set of data. Furthermore, the difference between the third and first quartiles is called the **interquartile range, IQR**. Here, it is equal to

$$\widetilde{q}_3 - \widetilde{q}_1 = \widetilde{\pi}_{0.75} - \widetilde{\pi}_{0.25} = 0.97 - 0.89 = 0.08.$$

Table 6.2-6 Ordered stem-and-leaf diagram of fluoride concentrations

Stems	Leaves	Frequency
0.8*f*	5 5 5 5	4
0.8*s*	6 6 6 7 7 7 7	7
0.8•	8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9	16
0.9∗	0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1	17
0.9*t*	2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3	19
0.9*f*	4 4 5 5 5 5 5 5 5	9
0.9*s*	6 6 7 7 7 7	6
0.9•	8 8 8 8 8 8 8 8 9 9	10
1.0∗	0 0 0 0 1 1 1	7
1.0*t*	2 3 3	3
1.0*f*	5	1
1.0*s*	6	1

One graphical means for displaying the five-number summary of a set of data is called a **box-and-whisker diagram**. To construct a horizontal box-and-whisker diagram, or, more simply, a **box plot**, draw a horizontal axis that is scaled to the data. Above the axis, draw a rectangular box with the left and right sides drawn at $\widetilde{q}_1$ and $\widetilde{q}_3$ and with a vertical line segment drawn at the median, $\widetilde{q}_2 = \widetilde{m}$. A left whisker is drawn as a horizontal line segment from the minimum to the midpoint of the left side of the box, and a right whisker is drawn as a horizontal line segment from the midpoint of the right side of the box to the maximum. Note that the length of the box is equal to the IQR. The left and right whiskers represent the first and fourth quarters of the data, while the two middle quarters of the data are represented, respectively, by the two sections of the box, one to the left and one to the right of the median line.

Example 6.2-3

Using the fluoride data shown in Table 6.2-6, we found that the five-number summary is given by

$$y_1 = 0.85,\ \widetilde{q}_1 = 0.89,\ \widetilde{q}_2 = \widetilde{m} = 0.92,\ \widetilde{q}_3 = 0.97,\ y_{100} = 1.06.$$

The box plot of these data is given in Figure 6.2-1. The fact that the long whisker is to the right and the right half of the box is larger than the left half of the box leads us to say that these data are slightly *skewed to the right*. Note that this skewness can also be seen in the histogram and in the stem-and-leaf diagram. ■

The next example illustrates how the box plot depicts data that are *skewed to the left*.

Example 6.2-4

The following data give the ordered weights (in grams) of 39 gold coins that were produced during the reign of Verica, a pre-Roman British king:

4.90	5.06	5.07	5.08	5.15	5.17	5.18	5.19	5.24	5.25
5.25	5.25	5.25	5.27	5.27	5.27	5.27	5.28	5.28	5.28
5.29	5.30	5.30	5.30	5.30	5.31	5.31	5.31	5.31	5.31
5.32	5.32	5.33	5.34	5.35	5.35	5.35	5.36	5.37	

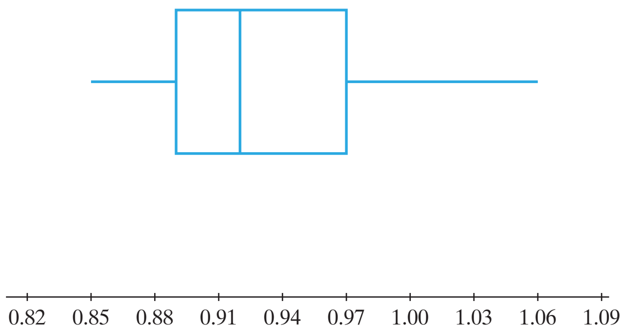


Figure 6.2-1 Box plot of fluoride concentrations

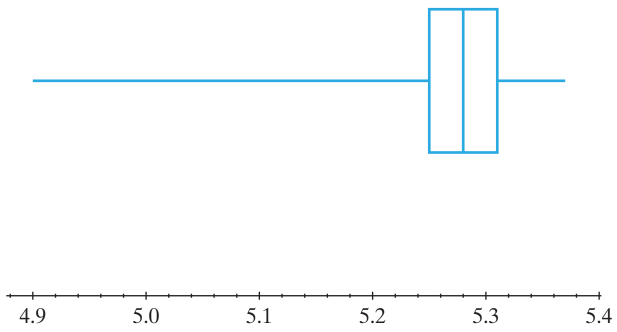


Figure 6.2-2 Box plot for weights of 39 gold coins

For these data, the minimum is 4.90 and the maximum is 5.37. Since

$$(39+1)(1/4) = 10, \qquad (39+1)(2/4) = 20, \qquad (39+1)(3/4) = 30,$$

we have

$$\begin{aligned}\widetilde{q}_1 = y_{10} &= 5.25,\\ \widetilde{m} = y_{20} &= 5.28,\\ \widetilde{q}_3 = y_{30} &= 5.31.\end{aligned}$$

Thus, the five-number summary is given by

$$y_1 = 4.90, \widetilde{q}_1 = 5.25, \widetilde{q}_2 = \widetilde{m} = 5.28, \widetilde{q}_3 = 5.31, y_{39} = 5.37.$$

The box plot associated with the given data is shown in Figure 6.2-2. Note that the box plot indicates that the data are skewed to the left. ■

Sometimes we are interested in picking out observations that seem to be much larger or much smaller than most of the other observations. That is, we are looking for outliers. Tukey suggested a method for defining outliers that is resistant to the effect of one or two extreme values and makes use of the IQR. In a box-and-whisker diagram, construct **inner fences** to the left and right of the box at a distance of 1.5 times the IQR. **Outer fences** are constructed in the same way at a distance of 3 times the IQR. Observations that lie between the inner and outer fences are called **suspected outliers**. Observations that lie beyond the outer fences are called **outliers**. The observations beyond the inner fences are denoted with a circle (•), and the whiskers are drawn only to the extreme values within or on the inner fences. When you are analyzing a set of data, suspected outliers deserve a closer look and outliers should be looked at very carefully. It does not follow that suspected outliers should be removed from the data, unless some error (such as a recording error) has been made. Moreover, it is sometimes important to determine the cause of extreme values, because outliers can often provide useful insights into the situation under consideration (such as a better way of doing things).

STATISTICAL COMMENTS There is a story that statisticians tell about Ralph Sampson, who was an excellent basketball player at the University of Virginia in the early 1980s and later was drafted by the Houston Rockets. He supposedly majored in communication studies at Virginia, and it is reported that the department there said that the average starting salary of their majors was much higher than those in the sciences; that happened because of Sampson's high starting salary with the Rockets. If this story is true, it would have been much more appropriate to report the median starting salary of majors and this median salary would have been much lower than the median starting salaries in the sciences. ■

Example 6.2-5

Continuing with Example 6.2-4, we find that the interquartile range is IQR $= 5.31 - 5.25 = 0.06$. Thus, the inner fences would be constructed at a distance of $1.5(0.06) = 0.09$ to the left and right of the box, and the outer fences would be constructed at a distance of $3(0.06) = 0.18$ to the left and right of the box. Figure 6.2-3 shows a box plot with the fences. Of course, since the maximum is 0.06 greater than $\widetilde{q}_3$, there are no fences to the right. From this box plot, we see that there are three suspected outliers and two outliers. (You may speculate as to why there are outliers with these data and why they fall to the left — that is, they are lighter than expected.) Note that many computer programs use an asterisk to plot outliers and suspected outliers, and do not print fences. ■

Some functions of two or more order statistics are quite important in modern statistics. We mention and illustrate one more, along with the range and the IQR, using the 100 fluoride concentrations shown in Table 6.2-6.

(a) **Midrange** = average of the extremes

$$= \frac{y_1 + y_n}{2} = \frac{0.85 + 1.06}{2} = 0.955.$$

(b) **Range** = difference of the extremes.

(c) **Interquartile range** = difference of third and first quartiles

$$= \widetilde{q}_3 - \widetilde{q}_1 = 0.97 - 0.89 = 0.08.$$

Thus, we see that the mean, the median, and the midrange are measures of the middle of the sample. In some sense, the standard deviation, the range, and the interquartile range provide measures of spread of the sample.

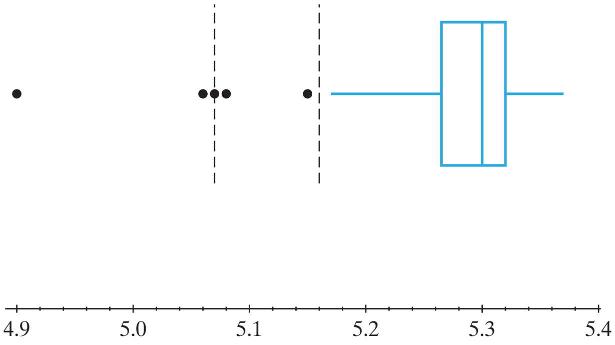


Figure 6.2-3 Box plot for weights of 39 gold coins with fences and outliers

Exercises

6.2-1. In Exercise 6.1-3, measurements for 96 plungers are given. Use those measurements to

(a) Construct a stem-and-leaf diagram using integer stems.

(b) Find the five-number summary of the data.

(c) Construct a box-and-whisker diagram. Are there any outliers?

6.2-2. When you purchase "1-pound bags" of carrots, you can buy either "baby" carrots or regular carrots. We shall compare the weights of 75 bags of each of these types of carrots. The following table gives the weights of the bags of baby carrots:

1.03	1.03	1.06	1.02	1.03	1.03	1.03	1.02	1.03	1.03
1.06	1.04	1.05	1.03	1.04	1.03	1.05	1.06	1.04	1.04
1.03	1.04	1.04	1.06	1.03	1.04	1.05	1.04	1.04	1.02
1.03	1.05	1.05	1.03	1.04	1.03	1.04	1.04	1.03	1.04
1.03	1.04	1.04	1.04	1.05	1.04	1.04	1.03	1.03	1.05
1.04	1.04	1.05	1.04	1.03	1.03	1.05	1.03	1.04	1.05
1.04	1.04	1.04	1.05	1.03	1.04	1.04	1.04	1.04	1.03
1.05	1.05	1.05	1.03	1.04					

This table gives the weights of the regular-sized carrots:

1.29	1.10	1.28	1.29	1.23	1.20	1.31	1.25	1.13	1.26
1.19	1.33	1.24	1.20	1.26	1.24	1.11	1.14	1.15	1.15
1.19	1.26	1.14	1.20	1.20	1.20	1.24	1.25	1.28	1.24
1.26	1.20	1.30	1.23	1.26	1.16	1.34	1.10	1.22	1.27
1.21	1.09	1.23	1.03	1.32	1.21	1.23	1.34	1.19	1.18
1.20	1.20	1.13	1.43	1.19	1.05	1.16	1.19	1.07	1.21
1.36	1.21	1.00	1.23	1.22	1.13	1.24	1.10	1.18	1.26
1.12	1.10	1.19	1.10	1.24					

(a) Calculate the five-number summary of each set of weights.

(b) On the same graph, construct box plots for each set of weights.

(c) If the carrots are the same price per package, which is the better buy? Which type of carrots would you select?

6.2-3. Here are underwater weights in kilograms for 82 male students:

3.7	3.6	4.0	4.3	3.8	3.4	4.1	4.0	3.7	3.4	3.5	3.8	3.7	4.9
3.5	3.8	3.3	4.8	3.4	4.6	3.5	5.3	4.4	4.2	2.5	3.1	5.2	3.8
3.3	3.4	4.1	4.6	4.0	1.4	4.3	3.8	4.7	4.4	5.0	3.2	3.1	4.2
4.9	4.5	3.8	4.2	2.7	3.8	3.8	2.0	3.4	4.9	3.3	4.3	5.6	3.2
4.7	4.5	5.2	5.0	5.0	4.0	3.8	5.3	4.5	3.8	3.8	3.4	3.6	3.3
4.2	5.1	4.0	4.7	6.5	4.4	3.6	4.7	4.5	2.3	4.0	3.7		

Here are underwater weights in kilograms for 100 female students:

2.0	2.0	2.1	1.6	1.9	2.0	2.0	1.3	1.3	1.2	2.3	1.9
2.1	1.2	2.0	1.6	1.1	2.2	2.2	1.4	1.7	2.4	1.8	1.7
2.0	2.1	1.6	1.7	1.8	0.7	1.9	1.7	1.7	1.1	2.0	2.3
0.5	1.3	2.7	1.8	2.0	1.7	1.2	0.7	1.1	1.1	1.7	1.7
1.2	1.2	0.7	2.3	1.7	2.4	1.0	2.4	1.4	1.9	2.5	2.2
2.1	1.4	2.4	1.8	2.5	1.3	0.5	1.7	1.9	1.8	1.3	2.0
2.2	1.7	2.0	2.5	1.2	1.4	1.4	1.2	2.2	2.0	1.8	1.4
1.9	1.4	1.3	2.5	1.2	1.5	0.8	2.0	2.2	1.8	2.0	1.6
1.5	1.6	1.5	2.6								

(a) Group each set of data into classes with a class width of 0.5 kilograms and in which the class marks are $0.5, 1.0, 1.5, \ldots$.

(b) Draw histograms of the grouped data.

(c) Construct box-and-whisker diagrams of the data and draw them on the same graph. Describe what this graph shows.

6.2-4. An insurance company experienced the following mobile home losses in 10,000's of dollars for 50 catastrophic events:

1	2	2	3	3	4	4	5	5	5
5	6	7	7	9	9	9	10	11	12
22	24	28	29	31	33	36	38	38	38
39	41	48	49	53	55	74	82	117	134
192	207	224	225	236	280	301	308	351	527

(a) Find the five-number summary of the data and draw a box-and-whisker diagram.

(b) Calculate the IQR and the locations of the inner and outer fences.

(c) Draw a box plot that shows the fences, suspected outliers, and outliers.

(d) Describe the distribution of losses. (See Exercise 6.1-6.)

6.2-5. In Exercise 6.1-5, data are given for the number of \$1 bets a player can make in roulette before losing \$5. Use those data to respond to the following:

(a) Determine the order statistics.

(b) Find the five-number summary of the data.

(c) Draw a box-and-whisker diagram.

(d) Find the locations of the inner and outer fences, and draw a box plot that shows the fences, the suspected outliers, and the outliers.

(e) In your opinion, does the median or sample mean give a better measure of the center of the data?

6.2-6. In the casino game roulette, if a player bets \$1 on red (or on black or on odd or on even), the probability of winning \$1 is 18/38 and the probability of losing \$1 is 20/38. Suppose that a player begins with \$5 and makes successive \$1 bets. Let Y equal the player's maximum capital before losing the \$5. One hundred observations of Y were simulated on a computer, yielding the following data:

25	9	5	5	5	9	6	5	15	45
55	6	5	6	24	21	16	5	8	7
7	5	5	35	13	9	5	18	6	10
19	16	21	8	13	5	9	10	10	6
23	8	5	10	15	7	5	5	24	9
11	34	12	11	17	11	16	5	15	5
12	6	5	5	7	6	17	20	7	8
8	6	10	11	6	7	5	12	11	18
6	21	6	5	24	7	16	21	23	15
11	8	6	8	14	11	6	9	6	10

(a) Construct an ordered stem-and-leaf display.

(b) Find the five-number summary of the data and draw a box-and-whisker diagram.

(c) Calculate the IQR and the locations of the inner and outer fences.

(d) Draw a box plot that shows the fences, suspected outliers, and outliers.

(e) Find the 90th percentile.

6.2-7. Let X denote the concentration of calcium carbonate ($\mathrm{CaCO_3}$) in milligrams per liter. Following are 20 observations of X:

130.8	129.9	131.5	131.2	129.5
132.7	131.5	127.8	133.7	132.2
134.8	131.7	133.9	129.8	131.4
128.8	132.7	132.8	131.4	131.3

(a) Construct an ordered stem-and-leaf display, using stems of 127, 128, . . . , 134.

(b) Find the midrange, range, interquartile range, median, sample mean, and sample variance.

(c) Draw a box-and-whisker diagram.

6.2-8. The weights (in grams) of 25 indicator housings used on gauges are as follows:

102.0	106.3	106.6	108.8	107.7
106.1	105.9	106.7	106.8	110.2
101.7	106.6	106.3	110.2	109.9
102.0	105.8	109.1	106.7	107.3
102.0	106.8	110.0	107.9	109.3

(a) Construct an ordered stem-and-leaf display, using integers as the stems and tenths as the leaves.

(b) Find the five-number summary of the data and draw a box plot.

(c) Are there any suspected outliers? Are there any outliers?

6.2-9. In Exercise 6.1-4, the melting points of a firm's alloy filaments are given for a sample of 50 filaments.

(a) Construct a stem-and-leaf diagram of those melting points, using as stems $30f, 30s, \ldots, 33f$.

(b) Find the five-number summary for these melting points.

(c) Construct a box-and-whisker diagram.

(d) Describe the symmetry of the data.

6.2-10. In Exercise 6.1-7, lead concentrations near the San Diego Freeway in 1976 are given. During the fall of 1977, the weekday afternoon lead concentrations (in $\mu\mathrm{g/m^3}$) at the measurement station near the San Diego Freeway in Los Angeles were as follows:

9.5	10.7	8.3	9.8	9.1	9.4	9.6	11.9	9.5	12.6	10.5
8.9	11.4	12.0	12.4	9.9	10.9	12.3	11.0	9.2	9.3	9.3
10.5	9.4	9.4	8.2	10.4	9.3	8.7	9.8	9.1	2.9	9.8
5.7	8.2	8.1	8.8	9.7	8.1	8.8	10.3	8.6	10.2	9.4
14.8	9.9	9.3	8.2	9.9	11.6	8.7	5.0	9.9	6.3	6.5
10.2	8.8	8.0	8.7	8.9	6.8	6.6	7.3	16.7		

(a) Construct a frequency distribution and display the results in the form of a histogram. Is this distribution symmetric?

(b) Calculate the sample mean and sample standard deviation.

(c) Locate $\bar{x}$, $\bar{x} \pm s$ on your histogram. How many observations lie within one standard deviation of the mean? How many lie within two standard deviations of the mean?

(d) Using the data from Exercise 6.1-7 and the data from this exercise, construct a back-to-back stem-and-leaf diagram with integer stems in the center and the leaves for 1976 going to the left and those for 1977 going to the right.

(e) Construct box-and-whisker displays of both sets of data on the same graph.

(f) Use your numerical and graphical results to interpret what you see.

REMARK In the spring of 1977, a new traffic lane was added to the freeway. This lane reduced traffic congestion but increased traffic speed. ■

6.3 ORDER STATISTICS

Order statistics are the observations of the random sample, arranged, or ordered, in magnitude from the smallest to the largest. In recent years, the importance of order statistics has increased owing to the more frequent use of nonparametric inferences and robust procedures. However, order statistics have always been prominent because, among other things, they are needed to determine rather simple statistics such as the sample median, the sample range, and the empirical cdf. Recall that in Section 6.2 we discussed observed order statistics in connection with descriptive and exploratory statistical methods. We will consider certain interesting aspects about their distributions in this section.

In most of our discussions about order statistics, we will assume that the n independent observations come from a continuous-type distribution. This means, among other things, that the probability of any two observations being equal is zero. That is, the probability is *1* that the observations can be ordered from smallest to largest without having two equal values. Of course, in practice, we do frequently observe *ties*; but if the probability of a tie is small, the distribution theory that follows will hold approximately. Thus, in the discussion here, we are assuming that the probability of a tie is zero.

Example 6.3-1 The values $x_1 = 0.62, x_2 = 0.98, x_3 = 0.31, x_4 = 0.81$, and $x_5 = 0.53$ are the $n = 5$ observed values of five independent trials of an experiment with pdf $f(x) = 2x$, $0 < x < 1$. The observed order statistics are

$$y_1 = 0.31 < y_2 = 0.53 < y_3 = 0.62 < y_4 = 0.81 < y_5 = 0.98.$$

Recall that the middle observation in the ordered arrangement, here $y_3 = 0.62$, is called the *sample median* and the difference of the largest and the smallest, here

$$y_5 - y_1 = 0.98 - 0.31 = 0.67,$$

is called the *sample range*. ■

If $X_1, X_2, \ldots, X_n$ are observations of a random sample of size n from a continuous-type distribution, we let the random variables

$$Y_1 < Y_2 < \cdots < Y_n$$

denote the order statistics of that sample. That is,

$$\begin{aligned}
Y_1 &= \text{smallest of } X_1, X_2, \ldots, X_n,\\
Y_2 &= \text{second smallest of } X_1, X_2, \ldots, X_n,\\
&\ \vdots\\
Y_n &= \text{largest of } X_1, X_2, \ldots, X_n.
\end{aligned}$$

There is a very simple procedure for determining the cdf of the rth order statistic, Y_r. This procedure depends on the binomial distribution and is illustrated in Example 6.3-2.

Example 6.3-2

Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ be the order statistics associated with n independent observations X_1, X_2, X_3, X_4, X_5, each from the distribution with pdf $f(x) = 2x$, $0 < x < 1$. Consider $P(Y_4 < 1/2)$. For the event $\{Y_4 < 1/2\}$ to occur, at least four of the random variables X_1, X_2, X_3, X_4, X_5 must be less than $1/2$, because Y_4 is the fourth smallest among the five observations. Thus, if the event $\{X_i < 1/2\}$, $i = 1, 2, \ldots, 5$, is called "success," we must have at least four successes in the five mutually independent trials, each of which has probability of success

$$P\left(X_i \le \frac{1}{2}\right) = \int_0^{1/2} 2x\,dx = \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

Hence,

$$P\left(Y_4 \le \frac{1}{2}\right) = \binom{5}{4}\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right) + \left(\frac{1}{4}\right)^5 = 0.0156.$$

In general, if $0 < y < 1$, then the cdf of Y_4 is

$$G(y) = P(Y_4 < y) = \binom{5}{4}(y^2)^4(1 - y^2) + (y^2)^5,$$

since this represents the probability of at least four "successes" in five independent trials, each of which has probability of success

$$P(X_i < y) = \int_0^y 2x\,dx = y^2.$$

For $0 < y < 1$, the pdf of Y_4 is therefore

$$\begin{aligned}
g(y) = G'(y) &= \binom{5}{4}4(y^2)^3(2y)(1 - y^2) + \binom{5}{4}(y^2)^4(-2y) + 5(y^2)^4(2y)\\
&= \frac{5!}{3!\,1!}(y^2)^3(1 - y^2)(2y), \qquad 0 < y < 1.
\end{aligned}$$

Note that in this example the cdf of each X is $F(x) = x^2$ when $0 < x < 1$. Thus,

$$g(y) = \frac{5!}{3!\,1!}[F(y)]^3\,[1 - F(y)]\,f(y), \qquad 0 < y < 1.$$ ■

The preceding example should make the following generalization easier to read: Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of n independent observations from a distribution of the continuous type with cdf $F(x)$ and pdf $F'(x) = f(x)$, where $0 < F(x) < 1$ for $a < x < b$ and $F(a) = 0$, $F(b) = 1$. (It is possible that $a = -\infty$ and/or $b = +\infty$.) The event that the rth order statistic Y_r is at most y, $\{Y_r \le y\}$, can occur if and only if at least r of the n observations are less than or equal to y. That is, here the probability of "success" on each trial is $F(y)$, and we must have at least r successes. Thus,

$$G_r(y) = P(Y_r \le y) = \sum_{k=r}^{n} \binom{n}{k} [F(y)]^k [1 - F(y)]^{n-k}.$$

Rewriting this slightly, we have

$$G_r(y) = \sum_{k=r}^{n-1} \binom{n}{k} [F(y)]^k [1 - F(y)]^{n-k} + [F(y)]^n.$$

Hence, the pdf of Y_r is

$$\begin{aligned} g_r(y) = G_r'(y) &= \sum_{k=r}^{n-1} \binom{n}{k} (k)[F(y)]^{k-1} f(y)[1 - F(y)]^{n-k} \\ &\quad + \sum_{k=r}^{n-1} \binom{n}{k} [F(y)]^k (n-k)[1 - F(y)]^{n-k-1}[-f(y)] \\ &\quad + n[F(y)]^{n-1} f(y). \end{aligned} \tag{6.3-1}$$

However, since

$$\binom{n}{k} k = \frac{n!}{(k-1)!\,(n-k)!} \qquad \text{and} \qquad \binom{n}{k}(n-k) = \frac{n!}{k!\,(n-k-1)!},$$

it follows that the pdf of Y_r is

$$g_r(y) = \frac{n!}{(r-1)!\,(n-r)!}\,[F(y)]^{r-1}[1 - F(y)]^{n-r} f(y), \qquad a < y < b,$$

which is the first term of the first summation in $g_r(y) = G_r'(y)$, Equation 6.3-1. The remaining terms in $g_r(y) = G_r'(y)$ sum to zero because the second term of the first summation (when $k = r + 1$) equals the negative of the first term in the second summation (when $k = r$), and so on. Finally, the last term of the second summation equals the negative of $n[F(y)]^{n-1} f(y)$. To see this clearly, the student is urged to write out a number of terms in these summations. (See Exercise 6.3-4.)

It is worth noting that the pdf of the smallest order statistic is

$$g_1(y) = n[1 - F(y)]^{n-1} f(y), \qquad a < y < b,$$

and the pdf of the largest order statistic is

$$g_n(y) = n[F(y)]^{n-1} f(y), \qquad a < y < b.$$

REMARK There is one quite satisfactory way to construct heuristically the expression for the pdf of Y_r. To do this, we must recall the multinomial probability and then consider the probability element $g_r(y)(\Delta y)$ of Y_r. If the length Δy is *very* small, $g_r(y)(\Delta y)$ represents approximately the probability

$$P(y < Y_r \le y + \Delta y).$$

Thus, we want the probability, $g_r(y)(\Delta y)$, that $(r-1)$ items fall less than y, that $(n-r)$ items are greater than $y + \Delta y$, and that one item falls between y and $y + \Delta y$. Recall that the probabilities on a single trial are

$$P(X \le y) = F(y),$$

$$P(X > y + \Delta y) = 1 - F(y + \Delta y) \;\approx\; 1 - F(y),$$

$$P(y < X \le y + \Delta y) \approx f(y)(\Delta y).$$

Thus, the multinomial probability is approximately

$$g_r(y)(\Delta y) = \frac{n!}{(r-1)!\,1!\,(n-r)!}\,[F(y)]^{r-1}[1 - F(y)]^{n-r}[f(y)(\Delta y)].$$

If we divide both sides by the length Δy, the formula for $g_r(y)$ results. ■

Example 6.3-3 Returning to Example 6.3-2, we shall now graph the pdfs of the order statistics $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ when sampling from a distribution with pdf $f(x) = 2x, 0 < x < 1$, and cdf $F(x) = x^2, 0 < x < 1$. These graphs are given in Figure 6.3-1. The respective pdfs and their means are as follows:

$$\begin{aligned}
g_1(y) &= 10y(1-y^2)^4, && 0 < y < 1; && \mu_1 = \frac{256}{693},\\
g_2(y) &= 40y^3(1-y^2)^3, && 0 < y < 1; && \mu_2 = \frac{128}{231},\\
g_3(y) &= 60y^5(1-y^2)^2, && 0 < y < 1; && \mu_3 = \frac{160}{231},\\
g_4(y) &= 40y^7(1-y^2), && 0 < y < 1; && \mu_4 = \frac{80}{99},\\
g_5(y) &= 10y^9, && 0 < y < 1; && \mu_5 = \frac{10}{11}.
\end{aligned}$$ ■

Recall that in Theorem 5.1-2 we proved that if X has a cdf $F(x)$ of the continuous type, then $F(X)$ has a uniform distribution on the interval from 0 to 1. If $Y_1 < Y_2 < \cdots < Y_n$ are the order statistics of n independent observations $X_1, X_2, \ldots, X_n$, then

$$F(Y_1) < F(Y_2) < \cdots < F(Y_n),$$

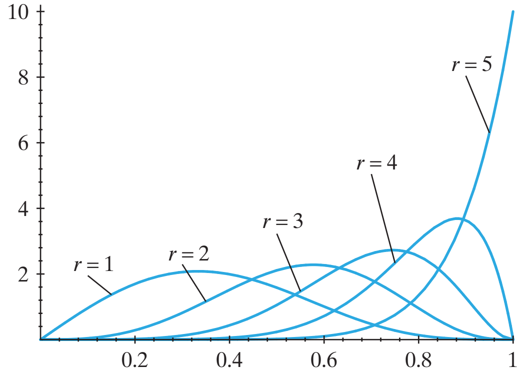


Figure 6.3-1 pdfs of order statistics, $f(x) = 2x, 0 < x < 1$

because F is a nondecreasing function and the probability of an equality is again zero. Note that this ordering could be looked upon as an ordering of the mutually independent random variables $F(X_1), F(X_2), \ldots, F(X_n)$, each of which is $U(0, 1)$. That is,

$$W_1 = F(Y_1) < W_2 = F(Y_2) < \cdots < W_n = F(Y_n)$$

can be thought of as the order statistics of n independent observations from that uniform distribution. Since the cdf of $U(0, 1)$ is $G(w) = w, 0 < w < 1$, the pdf of the rth order statistic, $W_r = F(Y_r)$, is

$$h_r(w) = \frac{n!}{(r-1)!\,(n-r)!}\, w^{r-1}(1-w)^{n-r}, \qquad 0 < w < 1.$$

Of course, the mean, $E(W_r) = E[F(Y_r)]$ of $W_r = F(Y_r)$, is given by the integral

$$E(W_r) = \int_0^1 w\, \frac{n!}{(r-1)!\,(n-r)!}\, w^{r-1}(1-w)^{n-r}\, dw.$$

This integral can be evaluated by integrating by parts several times, but it is easier to obtain the answer if we rewrite the integration as follows:

$$E(W_r) = \left(\frac{r}{n+1}\right) \int_0^1 \frac{(n+1)!}{r!\,(n-r)!} w^{r}(1-w)^{n-r}\, dw.$$

The integrand in this last expression can be thought of as the pdf of the $(r + 1)$st order statistic of $n + 1$ independent observations from a $U(0, 1)$ distribution. This is a beta pdf with $\alpha = r + 1$ and $\beta = n - r + 1$; hence, the integral must equal 1, and it follows that

$$E(W_r) = \frac{r}{n+1}, \qquad r = 1, 2, \ldots, n.$$

There is an extremely interesting interpretation of $W_r = F(Y_r)$. Note that $F(Y_r)$ is the cumulated probability up to and including Y_r or, equivalently, the area under $f(x) = F'(x)$ but less than Y_r. Consequently, $F(Y_r)$ can be treated as a random area. Since $F(Y_{r-1})$ is also a random area, $F(Y_r) - F(Y_{r-1})$ is the random area under $f(x)$ between Y_{r-1} and Y_r. The expected value of the random area between any two adjacent order statistics is then

$$\begin{aligned} E[F(Y_r) - F(Y_{r-1})] &= E[F(Y_r)] - E[F(Y_{r-1})] \\ &= \frac{r}{n+1} - \frac{r-1}{n+1} = \frac{1}{n+1}. \end{aligned}$$

Also, it is easy to show (see Exercise 6.3-6) that

$$E[F(Y_1)] = \frac{1}{n+1} \qquad \text{and} \qquad E[1 - F(Y_n)] = \frac{1}{n+1}.$$

That is, the order statistics $Y_1 < Y_2 < \cdots < Y_n$ partition the support of X into $n + 1$ parts and thus create $n + 1$ areas under $f(x)$ and above the x-axis. On the average, each of the $n + 1$ areas equals $1/(n + 1)$.

If we recall that the $(100p)$th percentile π_p is such that the area under $f(x)$ to the left of π_p is p, then the preceding discussion suggests that we let Y_r be an estimator of π_p, where $p = r/(n + 1)$. For this reason, we define the **$(100p)$th percentile of the sample** as Y_r, where $r = (n+1)p$. In case $(n+1)p$ is not an integer, we use a weighted average (or an average) of the two adjacent order statistics Y_r and Y_{r+1}, where r is the greatest integer $[(n + 1)p]$ (or, $\lfloor (n + 1)p \rfloor$) in $(n + 1)p$. In particular, the sample median is

$$\widetilde{m} = \begin{cases} Y_{(n+1)/2}, & \text{when } n \text{ is odd,} \\ \dfrac{Y_{n/2} + Y_{(n/2)+1}}{2}, & \text{when } n \text{ is even.} \end{cases}$$

Example 6.3-4 Let X equal the weight of soap in a "1000-gram" bottle. A random sample of $n = 12$ observations of X yielded the following weights, which have been ordered:

1013	1019	1021	1024	1026	1028
1033	1035	1039	1040	1043	1047

Since $n = 12$ is even, the sample median is

$$\widetilde{m} = \frac{y_6 + y_7}{2} = \frac{1028 + 1033}{2} = 1030.5.$$

The location of the 25th percentile (or first quartile) is

$$(n+1)(0.25) = (12+1)(0.25) = 3.25.$$

Thus, using a weighted average, we find that the first quartile is

$$\begin{aligned} \widetilde{q}_1 = y_3 + (0.25)(y_4 - y_3) &= (0.75)y_3 + (0.25)y_4 \\ &= (0.75)(1021) + (0.25)(1024) \ = \ 1021.75. \end{aligned}$$

Similarly, the 75th percentile (or third quartile) is

$$\begin{aligned} \widetilde{q}_3 = y_9 + (0.75)(y_{10} - y_9) &= (0.25)y_9 + (0.75)y_{10} \\ &= (0.25)(1039) + (0.75)(1040) \ = \ 1039.75, \end{aligned}$$

because $(12+1)(0.75) = 9.75$. Since $(12+1)(0.60) = 7.8$, the 60th percentile is

$$\widetilde{\pi}_{0.60} = (0.2)y_7 + (0.8)y_8 = (0.2)(1033) + (0.8)(1035) = 1034.6.$$ ■

The $(100p)$th percentile of a distribution is often called the quantile of order p. So if $y_1 \le y_2 \le \cdots \le y_n$ are the order statistics associated with the sample $x_1, x_2, \ldots, x_n$, then y_r is called the **sample quantile of order $r/(n+1)$** as well as the **$100r/(n+1)$th sample percentile**. Also, the percentile π_p of a theoretical distribution is the quantile of order p. Now, suppose the theoretical distribution is a good model for the observations. Then we plot (y_r, π_p), where $p = r/(n+1)$, for several values of r (possibly even for all r values, $r = 1, 2, \ldots, n$); we would expect these points (y_r, π_p) to lie close to a line through the origin with slope equal to 1 because $y_r \approx \pi_p$. If they are not close to that line, then we would doubt that the theoretical distribution is a good model for the observations. The plot of (y_r, π_p) for several values of r is called the **quantile–quantile plot** or, more simply, the ***q*–*q* plot**.

Given a set of observations of a random variable X, how can we decide, for example, whether or not X has an approximate normal distribution? If we have a large number of observations of X, a stem-and-leaf diagram or a histogram of the observations can often be helpful. (See Exercises 6.2-1 and 6.1-3, respectively.) For small samples, a q–q plot can be used to check on whether the sample arises from a normal distribution. For example, suppose the quantiles of a sample were plotted against the corresponding quantiles of a certain normal distribution and the pairs of points generated were on a straight line with slope 1 and intercept 0. Of course, we would then believe that we have an ideal sample from that normal distribution with that certain mean and standard deviation. Such a plot, however, requires that we know the mean and the standard deviation of this normal distribution, and we usually do not. However, since the quantile, q_p, of $N(\mu, \sigma^2)$ is related to the corresponding one, z_{1-p}, of $N(0, 1)$ by $q_p = \mu + \sigma z_{1-p}$, we can always plot the quantiles of the sample against the corresponding ones of $N(0, 1)$ and get the needed information. That is, if the sample quantiles are plotted as the x-coordinates of the pairs and the $N(0, 1)$ quantiles as the y-coordinates, and if the graph is almost a straight line, then it is reasonable to assume that the sample arises from a normal distribution. Moreover, the reciprocal of the slope of that straight line is a good estimate of the standard deviation σ because $z_{1-p} = (q_p - \mu)/\sigma$.

Example 6.3-5

In researching groundwater it is often important to know the characteristics of the soil at a certain site. Many of these characteristics, such as porosity, are at least partially dependent upon the grain size. The diameter of individual grains of soil can be measured. Here are the diameters (in mm) of 30 randomly selected grains:

1.24	1.36	1.28	1.31	1.35	1.20	1.39	1.35	1.41	1.31
1.28	1.26	1.37	1.49	1.32	1.40	1.33	1.28	1.25	1.39
1.38	1.34	1.40	1.27	1.33	1.36	1.43	1.33	1.29	1.34

For these data, $\bar{x} = 1.33$ and $s^2 = 0.0040$. May we assume that these are observations of a random variable X that is $N(1.33, 0.0040)$? To help answer this question, we shall construct a q–q plot of the standard normal quantiles that correspond to $p = 1/31, 2/31, \ldots, 30/31$ versus the ordered observations. To find these quantiles, it is helpful to use the computer.

k	Diameters in mm (x)	$p = k/31$	z_{1-p}	k	Diameters in mm (x)	$p = k/31$	z_{1-p}
1	1.20	0.0323	−1.85	16	1.34	0.5161	0.04
2	1.24	0.0645	−1.52	17	1.34	0.5484	0.12
3	1.25	0.0968	−1.30	18	1.35	0.5806	0.20
4	1.26	0.1290	−1.13	19	1.35	0.6129	0.29
5	1.27	0.1613	−0.99	20	1.36	0.6452	0.37
6	1.28	0.1935	−0.86	21	1.36	0.6774	0.46
7	1.28	0.2258	−0.75	22	1.37	0.7097	0.55
8	1.28	0.2581	−0.65	23	1.38	0.7419	0.65
9	1.29	0.2903	−0.55	24	1.39	0.7742	0.75
10	1.31	0.3226	−0.46	25	1.39	0.8065	0.86
11	1.31	0.3548	−0.37	26	1.40	0.8387	0.99
12	1.32	0.3871	−0.29	27	1.40	0.8710	1.13
13	1.33	0.4194	−0.20	28	1.41	0.9032	1.30
14	1.33	0.4516	−0.12	29	1.43	0.9355	1.52
15	1.33	0.4839	−0.04	30	1.49	0.9677	1.85

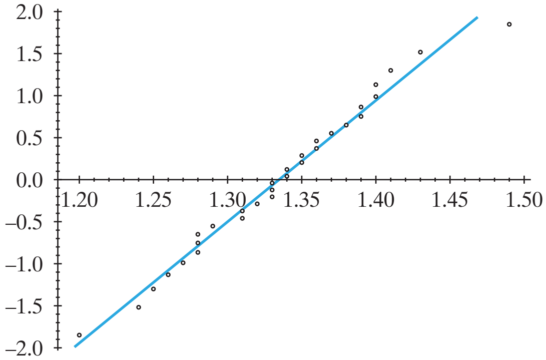


Figure 6.3-2 q–q plot, $N(0, 1)$ quantiles versus grain diameters

A q–q plot of these data is shown in Figure 6.3-2. Note that the points do fall close to a straight line, so the normal probability model seems to be appropriate on the basis of these few data. ■

Exercises

6.3-1. Some biology students were interested in analyzing the amount of time that bees spend gathering nectar in flower patches. Thirty-nine bees visited a high-density flower patch and spent the following times (in seconds) gathering nectar:

235	210	95	146	195	840	185	610	680	990
146	404	119	47	9	4	10	169	270	95
329	151	211	127	154	35	225	140	158	116
46	113	149	420	120	45	10	18	105	

(a) Find the order statistics.

(b) Find the median and 80th percentile of the sample.

(c) Determine the first and third quartiles (i.e., 25th and 75th percentiles) of the sample.

6.3-2. Let X equal the forced vital capacity (the volume of air a person can expel from his or her lungs) of a male freshman. Seventeen observations of X, which have been ordered, are

3.7	3.8	4.0	4.3	4.7	4.8	4.9	5.0	
5.2	5.4	5.6	5.6	5.6	5.7	6.2	6.8	7.6

(a) Find the median, the first quartile, and the third quartile.

(b) Find the 35th and 65th percentiles.

6.3-3. Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ be the order statistics of five independent observations from an exponential distribution that has a mean of $\theta = 3$.

(a) Find the pdf of the sample median Y_3.

(b) Compute the probability that Y_4 is less than 5.

(c) Determine $P(1 < Y_1)$.

6.3-4. In the expression for $g_r(y) = G'_r(y)$ in Equation 6.3-1, let $n = 6$, and $r = 3$, and write out the summations, showing that the "telescoping" suggested in the text is achieved.

6.3-5. Let $Y_1 < Y_2 < \cdots < Y_8$ be the order statistics of eight independent observations from a continuous-type distribution with 70th percentile $\pi_{0.7} = 27.3$.

(a) Determine $P(Y_7 < 27.3)$.

(b) Find $P(Y_5 < 27.3 < Y_8)$.

6.3-6. Let $W_1 < W_2 < \cdots < W_n$ be the order statistics of n independent observations from a $U(0, 1)$ distribution.

(a) Find the pdf of W_1 and that of W_n.

(b) Use the results of (a) to verify that $E(W_1) = 1/(n+1)$ and $E(W_n) = n/(n+1)$.

(c) Show that the pdf of W_r is beta.

6.3-7. Let $Y_1 < Y_2 < \cdots < Y_{19}$ be the order statistics of $n = 19$ independent observations from the exponential distribution with mean θ.

(a) What is the pdf of Y_1?

(b) Using integration, find the value of $E[F(Y_1)]$, where F is the cdf of the exponential distribution.

6.3-8. Let $W_1 < W_2 < \cdots < W_n$ be the order statistics of n independent observations from a $U(0,1)$ distribution.

(a) Show that $E(W_r^2) = r(r+1)/(n+1)(n+2)$, using a technique similar to that used in determining that $E(W_r) = r/(n+1)$.

(b) Find the variance of W_r.

6.3-9. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from an exponential distribution with pdf $f(x) = e^{-x}$, $0 < x < \infty$.

(a) Find the pdf of Y_r.

(b) Determine the pdf of $U = e^{-Y_r}$.

6.3-10. Use the heuristic argument to show that the joint pdf of the two order statistics $Y_i < Y_j$ is

$$\begin{aligned} g(y_i, y_j) = &\frac{n!}{(i-1)!(j-i-1)!(n-j)!} \\ &\times [F(y_i)]^{i-1}[F(y_j) - F(y_i)]^{j-i-1} \\ &\times [1 - F(y_j)]^{n-j} f(y_i) f(y_j), \ -\infty < y_i < y_j < \infty. \end{aligned}$$

6.3-11. Use the result of Exercise 6.3-10.

(a) Find the joint pdf of Y_1 and Y_n, the first and the nth order statistics of a random sample of size n from the $U(0,1)$ distribution.

(b) Find the joint and the marginal pdfs of $W_1 = Y_1/Y_n$ and $W_2 = Y_n$.

(c) Are W_1 and W_2 independent?

(d) Use simulation to confirm your theoretical results.

6.3-12. Nine measurements are taken on the strength of a certain metal. In order, they are 7.2, 8.9, 9.7, 10.5, 10.9, 11.7, 12.9, 13.9, 15.3, and these values correspond to the 10th, 20th, . . . , 90th percentiles of this sample. Construct a q–q plot of the measurements against the same percentiles of $N(0, 1)$. Does it seem reasonable that the underlying distribution of strengths could be normal?

6.3-13. Some measurements (in mm) were made on specimens of the spider *Sosippus floridanus*, which is native to Florida. Here are the lengths of nine female spiders and nine male spiders.

Female spiders	11.06	13.87	12.93	15.08	17.82
	14.14	12.26	17.82	20.17	
Male spiders	12.26	11.66	12.53	13.00	11.79
	12.46	10.65	10.39	12.26	

(a) Construct a q–q plot of the female spider lengths. Do they appear to be normally distributed?

(b) Construct a q–q plot of the male spider lengths. Do they appear to be normally distributed?

6.3-14. An interior automotive supplier places several electrical wires in a harness. A pull test measures the force required to pull spliced wires apart. A customer requires that each wire that is spliced into the harness withstand a pull force of 20 pounds. Let X equal the pull force required to pull a spliced wire apart. The following data give the values of a random sample of $n = 20$ observations of X:

28.8	24.4	30.1	25.6	26.4	23.9	22.1	22.5	27.6	28.1
20.8	27.7	24.4	25.1	24.6	26.3	28.2	22.2	26.3	24.4

(a) Construct a q–q plot, using the ordered array and the corresponding quantiles of $N(0,1)$.

(b) Does X appear to have a normal distribution?

6.4 MAXIMUM LIKELIHOOD ESTIMATION

In earlier chapters, we alluded to estimating characteristics of the distribution from the corresponding ones of the sample, hoping that the latter would be reasonably close to the former. For example, the sample mean $\overline{x}$ can be thought of as an estimate of the distribution mean μ, and the sample variance s^2 can be used as an estimate of the distribution variance σ^2. Even the relative frequency histogram associated with a sample can be taken as an estimate of the pdf of the underlying distribution. But how good are these estimates? What makes an estimate good? Can we say anything about the closeness of an estimate to an unknown parameter?

In this section, we consider random variables for which the functional form of the pmf or pdf is known, but the distribution depends on an unknown parameter (say, θ) that may have any value in a set (say, Ω) called the **parameter space**. For example, perhaps it is known that $f(x;\theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, and that $\theta \in \Omega = \{\theta : 0 < \theta < \infty\}$. In certain instances, it might be necessary for the experimenter to select precisely one member of the family $\{f(x,\theta), \theta \in \Omega\}$ as the most likely pdf of the random variable. That is, the experimenter needs a point estimate of the parameter θ, namely, the value of the parameter that corresponds to the selected pdf.

In one common estimation scenario, we take a random sample from the distribution to elicit some information about the unknown parameter θ. That is, we repeat the experiment n independent times, observe the sample, $X_1, X_2, \ldots, X_n$, and try to estimate the value of θ by using the observations $x_1, x_2, \ldots, x_n$. The function of $X_1, X_2, \ldots, X_n$ used to estimate θ—say, the statistic $u(X_1, X_2, \ldots, X_n)$—is called an **estimator** of θ. We want it to be such that the computed **estimate** $u(x_1, x_2, \ldots, x_n)$ is usually close to θ. Since we are estimating one member of $\theta \in \Omega$, such an estimator is often called a **point estimator**.

The following example should help motivate one principle that is often used in finding point estimates: Suppose that X is $b(1, p)$, so that the pmf of X is

$$f(x;p) = p^x(1-p)^{1-x}, \qquad x = 0, 1, \qquad 0 \le p \le 1.$$

We note that $p \in \Omega = \{p : 0 \le p \le 1\}$, where Ω represents the parameter space—that is, the space of all possible values of the parameter p. Given a random sample $X_1, X_2, \ldots, X_n$, the problem is to find an estimator $u(X_1, X_2, \ldots, X_n)$ such that $u(x_1, x_2, \ldots, x_n)$ is a good point estimate of p, where $x_1, x_2, \ldots, x_n$ are the observed values of the random sample. Now, the probability that $X_1, X_2, \ldots, X_n$ takes these particular values is (with $\Sigma\, x_i$ denoting $\sum_{i=1}^n x_i$)

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i},$$

which is the joint pmf of $X_1, X_2, \ldots, X_n$ evaluated at the observed values. One reasonable way to proceed toward finding a good estimate of p is to regard this probability (or joint pmf) as a function of p and find the value of p that maximizes it. That is, we find the p value most likely to have produced these sample values. The joint pmf, when regarded as a function of p, is frequently called the **likelihood function**. Thus, here the likelihood function is

$$\begin{aligned} L(p) &= L(p; x_1, x_2, \ldots, x_n) \\ &= f(x_1;p)f(x_2;p)\cdots f(x_n;p) \\ &= p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}, \qquad 0 \le p \le 1. \end{aligned}$$

If $\Sigma_{i=1}^n x_i = 0$, then $L(p) = (1-p)^n$, which is maximized over $p \in [0,1]$ by taking $\widehat{p} = 0$. If, on the other hand, $\Sigma_{i=1}^n x_i = n$, then $L(p) = p^n$ and this is maximized over $p \in [0,1]$ by taking $\widehat{p} = 1$. If $\Sigma_{i=1}^n x_i$ equals neither 0 nor n, then $L(0) = L(1) = 0$ while $L(p) > 0$ for all $p \in (0,1)$; thus, in this case it suffices to maximize $L(p)$ for $0 < p < 1$, which we do by standard methods of calculus. The derivative of $L(p)$ is

$$L'(p) = (\Sigma\, x_i)p^{\Sigma x_i - 1}(1-p)^{n-\Sigma x_i} - (n - \Sigma\, x_i)p^{\Sigma x_i}(1-p)^{n-\Sigma x_i - 1}.$$

Setting this first derivative equal to zero gives us, with the restriction that $0 < p < 1$,

$$p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}\left(\frac{\Sigma\, x_i}{p} - \frac{n - \Sigma\, x_i}{1-p}\right) = 0.$$

Since $0 < p < 1$, the preceding equation equals zero when

$$\frac{\Sigma\, x_i}{p} - \frac{n - \Sigma\, x_i}{1-p} = 0. \tag{6.4-1}$$

Multiplying each member of Equation 6.4-1 by $p(1-p)$ and simplifying, we obtain

$$\sum_{i=1}^{n} x_i - np = 0$$

or, equivalently,

$$p = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}.$$

It can be shown that $L''(\overline{x}) < 0$, so that $L(\overline{x})$ is a maximum. The corresponding statistic, namely, $(\sum_{i=1}^{n} X_i)/n = \overline{X}$, is called the **maximum likelihood estimator** and is denoted by $\widehat{p}$; that is,

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

When finding a maximum likelihood estimator, it is often easier to find the value of the parameter that maximizes the natural logarithm of the likelihood function rather than the value of the parameter that maximizes the likelihood function itself. Because the natural logarithm function is a strictly increasing function, the solutions will be the same. To see this, note that for $0 < p < 1$, the example we have been considering gives us

$$\ln L(p) = \left(\sum_{i=1}^{n} x_i\right)\ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p).$$

To find the maximum, we set the first derivative equal to zero to obtain

$$\frac{d\,[\ln L(p)]}{dp} = \left(\sum_{i=1}^{n} x_i\right)\left(\frac{1}{p}\right) + \left(n - \sum_{i=1}^{n} x_i\right)\left(\frac{-1}{1-p}\right) = 0,$$

which is the same as Equation 6.4-1. Thus, the solution is $p = \overline{x}$ and the maximum likelihood estimator for p is $\widehat{p} = \overline{X}$.

Motivated by the preceding example, we present the formal definition of maximum likelihood estimators (this definition is used in both the discrete and continuous cases).

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$ with pmf or pdf that is denoted by $f(x; \theta_1, \theta_2, \ldots, \theta_m)$. Suppose that $(\theta_1, \theta_2, \ldots, \theta_m)$ is restricted to a given parameter space Ω. Then the joint pmf or pdf of $X_1, X_2, \ldots, X_n$, namely,

$$\begin{aligned} L(\theta_1, \theta_2, \ldots, \theta_m) &= f(x_1; \theta_1, \ldots, \theta_m) f(x_2; \theta_1, \ldots, \theta_m) \\ &\quad \cdots f(x_n; \theta_1, \ldots, \theta_m), \qquad (\theta_1, \theta_2, \ldots, \theta_m) \in \Omega, \end{aligned}$$

when regarded as a function of $\theta_1, \theta_2, \ldots, \theta_m$, is called the **likelihood function**. Say

$$[u_1(x_1, \ldots, x_n), u_2(x_1, \ldots, x_n), \ldots, u_m(x_1, \ldots, x_n)]$$

is that m-tuple in Ω that maximizes $L(\theta_1, \theta_2, \ldots, \theta_m)$. Then

$$\begin{aligned}
\widehat{\theta_1} &= u_1(X_1, \ldots, X_n), \\
\widehat{\theta_2} &= u_2(X_1, \ldots, X_n), \\
&\vdots \\
\widehat{\theta_m} &= u_m(X_1, \ldots, X_n)
\end{aligned}$$

are **maximum likelihood estimators** of $\theta_1, \theta_2, \ldots, \theta_m$, respectively; and the corresponding observed values of these statistics, namely,

$$u_1(x_1, \ldots, x_n), u_2(x_1, \ldots, x_n), \ldots, u_m(x_1, \ldots, x_n),$$

are called **maximum likelihood estimates**. In many practical cases, these estimators (and estimates) are unique.

For many applications, there is just one unknown parameter. In these cases, the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Some additional examples will help clarify these definitions.

Example 6.4-1

Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential distribution with pdf

$$f(x; \theta) = \frac{1}{\theta}\, e^{-x/\theta}, \qquad 0 < x < \infty, \qquad \theta \in \Omega = \{\theta : 0 < \theta < \infty\}.$$

The likelihood function is given by

$$\begin{aligned}
L(\theta) &= L(\theta; x_1, x_2, \ldots, x_n) \\
&= \left(\frac{1}{\theta}\, e^{-x_1/\theta}\right)\left(\frac{1}{\theta}\, e^{-x_2/\theta}\right) \cdots \left(\frac{1}{\theta}\, e^{-x_n/\theta}\right) \\
&= \frac{1}{\theta^n} \exp\left(\frac{-\sum_{i=1}^{n} x_i}{\theta}\right), \qquad 0 < \theta < \infty.
\end{aligned}$$

The natural logarithm of $L(\theta)$ is

$$\ln L(\theta) = -(n)\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} x_i, \qquad 0 < \theta < \infty.$$

Thus,

$$\frac{d\,[\ln L(\theta)]}{d\theta} = \frac{-n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0.$$

The solution of this equation for θ is

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}.$$

Note that

$$\frac{d\,[\ln L(\theta)]}{d\theta} = \frac{1}{\theta}\left(-n + \frac{n\,\overline{x}}{\theta}\right) \begin{array}{ll} > 0, & \theta < \overline{x}, \\ = 0, & \theta = \overline{x}, \\ < 0, & \theta > \overline{x}. \end{array}$$

Hence, $\ln L(\theta)$ does have a maximum at $\overline{x}$, and it follows that the maximum likelihood estimator for θ is

$$\widehat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

■

Example 6.4-2 Let $X_1, X_2, \ldots, X_n$ be a random sample from the geometric distribution with pmf $f(x;p) = (1-p)^{x-1}p,\ x = 1, 2, 3, \ldots$. The likelihood function is given by

$$\begin{aligned} L(p) &= (1-p)^{x_1-1}p(1-p)^{x_2-1}p \cdots (1-p)^{x_n-1}p \\ &= p^n(1-p)^{\Sigma x_i - n}, \qquad 0 \le p \le 1. \end{aligned}$$

The natural logarithm of $L(p)$ is

$$\ln L(p) = n \ln p + \left(\sum_{i=1}^{n} x_i - n\right)\ln(1-p), \qquad 0 < p < 1.$$

Thus, restricting p to $0 < p < 1$, so as to be able to take the derivative, we have

$$\frac{d \ln L(p)}{dp} = \frac{n}{p} - \frac{\sum_{i=1}^{n} x_i - n}{1-p} = 0.$$

Solving for p, we obtain

$$p = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{x}},$$

and, by the second derivative test, this solution provides a maximum. So the maximum likelihood estimator of p is

$$\widehat{p} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}}.$$

This estimator agrees with our intuition because, in n observations of a geometric random variable, there are n successes in the $\sum_{i=1}^{n} x_i$ trials. Thus, the estimate of p is the number of successes divided by the total number of trials. ■

In the following important example, we find the maximum likelihood estimators of the parameters associated with the normal distribution.

Example 6.4-3 Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\theta_1, \theta_2)$, where

$$\Omega = \{(\theta_1, \theta_2) : -\infty < \theta_1 < \infty,\ 0 < \theta_2 < \infty\}.$$

That is, here we let $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. Then

$$L(\theta_1, \theta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{(x_i - \theta_1)^2}{2\theta_2}\right]$$

or, equivalently,

$$L(\theta_1,\theta_2) = \left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \exp\left[\frac{-\sum_{i=1}^n(x_i-\theta_1)^2}{2\theta_2}\right], \qquad (\theta_1,\theta_2)\in\Omega.$$

The natural logarithm of the likelihood function is

$$\ln L(\theta_1,\theta_2) = -\frac{n}{2}\ln(2\pi\theta_2) - \frac{\sum_{i=1}^n(x_i-\theta_1)^2}{2\theta_2}.$$

The partial derivatives with respect to θ_1 and θ_2 are

$$\frac{\partial\,(\ln L)}{\partial\,\theta_1} = \frac{1}{\theta_2}\sum_{i=1}^n(x_i-\theta_1)$$

and

$$\frac{\partial\,(\ln L)}{\partial\,\theta_2} = \frac{-n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^n(x_i-\theta_1)^2.$$

The equation $\partial\,(\ln L)/\partial\,\theta_1 = 0$ has the solution $\theta_1 = \overline{x}$. Setting $\partial\,(\ln L)/\partial\,\theta_2 = 0$ and replacing θ_1 by $\overline{x}$ yields

$$\theta_2 = \frac{1}{n}\sum_{i=1}^n(x_i-\overline{x})^2.$$

By considering the usual condition on the second-order partial derivatives, we see that these solutions do provide a maximum. Thus, the maximum likelihood estimators of $\mu = \theta_1$ and $\sigma^2 = \theta_2$ are

$$\widehat{\theta_1} = \overline{X} \qquad \text{and} \qquad \widehat{\theta_2} = \frac{1}{n}\sum_{i=1}^n(X_i-\overline{X})^2 = V.$$ ■

It is interesting to note that in our first illustration, where $\widehat{p} = \overline{X}$, and in Example 6.4-1, where $\widehat{\theta} = \overline{X}$, the expected value of the estimator is equal to the corresponding parameter. This observation leads to the following definition.

Definition 6.4-1
If $E[u(X_1,X_2,\ldots,X_n)] = \theta$, then the statistic $u(X_1,X_2,\ldots,X_n)$ is called an **unbiased estimator** of θ. Otherwise, it is said to be **biased**.

Example 6.4-4 Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample X_1,X_2,X_3,X_4 from a uniform distribution with pdf $f(x;\theta) = 1/\theta$, $0 < x \le \theta$. The likelihood function is

$$L(\theta) = \left(\frac{1}{\theta}\right)^4, \qquad 0 < x_i \le \theta,\ \ i = 1,2,3,4,$$

and equals zero if $\theta < x_i$ or if $x_i \le 0$. To maximize $L(\theta)$, we must make θ as small as possible; hence, the maximum likelihood estimator is

$$\widehat{\theta} = \max(X_i) = Y_4$$

because θ cannot be less than any X_i. Since $F(x;\theta) = x/\theta,\ 0 < x \le \theta$, the pdf of Y_4 is

$$g_4(y_4) = \frac{4!}{3!1!}\left(\frac{y_4}{\theta}\right)^3\left(\frac{1}{\theta}\right) = 4\,\frac{y_4^3}{\theta^4}, \qquad 0 < y_4 \le \theta.$$

Accordingly,

$$E(Y_4) = \int_0^\theta y_4 \cdot 4\,\frac{y_4^3}{\theta^4}\,dy_4 = \frac{4}{5}\,\theta$$

and Y_4 is a biased estimator of θ. However, $5Y_4/4$ is unbiased. ■

Example 6.4-5 We have shown that when sampling from $N(\theta_1 = \mu, \theta_2 = \sigma^2)$, one finds that the maximum likelihood estimators of μ and σ^2 are

$$\widehat{\theta_1} = \widehat{\mu} = \overline{X} \qquad \text{and} \qquad \widehat{\theta_2} = \widehat{\sigma^2} = \frac{(n-1)S^2}{n}.$$

Recalling that the distribution of $\overline{X}$ is $N(\mu, \sigma^2/n)$, we see that $E(\overline{X}) = \mu$; thus, $\overline{X}$ is an unbiased estimator of μ.

In Theorem 5.5-2, we showed that the distribution of $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$. Hence,

$$E(S^2) = E\left[\frac{\sigma^2}{n-1}\ \frac{(n-1)S^2}{\sigma^2}\right] = \frac{\sigma^2}{n-1}\,(n-1) = \sigma^2.$$

That is, the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an unbiased estimator of σ^2. Consequently, since

$$E(\widehat{\theta_2}) = \frac{n-1}{n}\,E(S^2) = \frac{n-1}{n}\,\sigma^2,$$

$\widehat{\theta_2}$ is a biased estimator of $\theta_2 = \sigma^2$. ■

Sometimes it is impossible to find maximum likelihood estimators in a convenient closed form, and numerical methods must be used to maximize the likelihood function. For example, suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a gamma distribution with parameters $\alpha = \theta_1$ and $\beta = \theta_2$, where $\theta_1 > 0, \theta_2 > 0$. It is difficult to maximize

$$L(\theta_1, \theta_2; x_1, \ldots, x_n) = \left[\frac{1}{\Gamma(\theta_1)\theta_2^{\theta_1}}\right]^n (x_1 x_2 \cdots x_n)^{\theta_1 - 1} \exp\left(-\sum_{i=1}^{n} x_i/\theta_2\right)$$

with respect to θ_1 and θ_2, owing to the presence of the gamma function $\Gamma(\theta_1)$. Thus, numerical methods must be used to maximize L once $x_1, x_2, \ldots, x_n$ are observed.

There are other ways, however, to easily obtain point estimates of θ_1 and θ_2. One of the early methods was to simply equate the first sample moment to the first theoretical moment. Next, if needed, the two second moments are equated, then the third moments, and so on, until we have enough equations to solve for the parameters. As an illustration, in the gamma distribution situation, let us simply equate the first two moments of the distribution to the corresponding moments of the empirical distribution. This seems like a reasonable way in which to find estimators, since the empirical distribution converges in some sense to the probability distribution, and hence corresponding moments should be about equal. In this situation, we have

$$\theta_1\theta_2 = \overline{X}, \qquad \theta_1\theta_2^2 = V,$$

the solutions of which are

$$\widetilde{\theta}_1 = \frac{\overline{X}^2}{V} \qquad \text{and} \qquad \widetilde{\theta}_2 = \frac{V}{\overline{X}}.$$

We say that these latter two statistics, $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$, are respective estimators of θ_1 and θ_2 found by the **method of moments**.

To generalize this discussion, let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution with pdf $f(x;\theta_1,\theta_2,\ldots,\theta_r)$, $(\theta_1,\ldots,\theta_r) \in \Omega$. The expectation $E(X^k)$ is frequently called the kth moment of the distribution, $k = 1,2,3,\ldots$. The sum $M_k = \sum_{i=1}^n X_i^k/n$ is the kth moment of the sample, $k = 1,2,3,\ldots$. The method of moments can be described as follows. Equate $E(X^k)$ to M_k, beginning with $k = 1$ and continuing until there are enough equations to provide unique solutions for $\theta_1,\theta_2,\ldots,\theta_r$ — say, $h_i(M_1,M_2,\ldots)$, $i = 1,2,\ldots,r$, respectively. Note that this could be done in an equivalent manner by equating $\mu = E(X)$ to $\overline{X}$ and $E[(X-\mu)^k]$ to $\sum_{i=1}^n (X_i - \overline{X})^k/n$, $k = 2,3$, and so on, until unique solutions for $\theta_1,\theta_2,\ldots,\theta_r$ are obtained. This alternative procedure was used in the preceding illustration. In most practical cases, the estimator $\widetilde{\theta}_i = h_i(M_1,M_2,\ldots)$ of θ_i, found by the method of moments, is an estimator of θ_i that in some sense gets close to that parameter when n is large, $i = 1,2,\ldots,r$.

The next two examples—the first for a one-parameter family and the second for a two-parameter family—illustrate the method-of-moments technique for finding estimators.

Example 6.4-6

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the distribution with pdf $f(x;\theta) = \theta x^{\theta-1}$, $0 < x < 1$, $0 < \theta < \infty$. Sketch the graphs of this pdf for $\theta = 1/4$, 1, and 4. Note that sets of observations for these three values of θ would look very different. How do we estimate the value of θ? The mean of this distribution is given by

$$E(X) = \int_0^1 x\,\theta\,x^{\theta-1}\,dx = \frac{\theta}{\theta+1}.$$

We shall set the distribution mean equal to the sample mean and solve for θ. We have

$$\overline{x} = \frac{\theta}{\theta+1}.$$

Solving for θ, we obtain the method-of-moments estimator,

$$\widetilde{\theta} = \frac{\overline{X}}{1-\overline{X}}.$$

Thus, an estimate of θ by the method of moments is $\overline{x}/(1-\overline{x})$. ■

Recall that in the method of moments, if two parameters have to be estimated, the first two sample moments are set equal to the first two distribution moments that are given in terms of the unknown parameters. These two equations are then solved simultaneously for the unknown parameters.

Example 6.4-7 Let the distribution of X be $N(\mu, \sigma^2)$. Then

$$E(X) = \mu \qquad \text{and} \qquad E(X^2) = \sigma^2 + \mu^2.$$

For a random sample of size n, the first two moments are given by

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{and} \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

We set $m_1 = E(X)$ and $m_2 = E(X^2)$ and solve for μ and σ^2. That is,

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \mu \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma^2 + \mu^2.$$

The first equation yields $\overline{x}$ as the estimate of μ. Replacing μ^2 with $\overline{x}^2$ in the second equation and solving for σ^2, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n} = v$$

as the solution of σ^2. Thus, the method-of-moments estimators for μ and σ^2 are $\widetilde{\mu} = \overline{X}$ and $\widetilde{\sigma^2} = V$, which are the same as the maximum likelihood estimators. Of course, $\widetilde{\mu} = \overline{X}$ is unbiased, whereas $\widetilde{\sigma^2} = V$ is biased. ■

In Example 6.4-5, we showed that $\overline{X}$ and S^2 are unbiased estimators of μ and σ^2, respectively, when one is sampling from a normal distribution. This is also true when one is sampling from any distribution with a finite variance σ^2. That is, $E(\overline{X}) = \mu$ and $E(S^2) = \sigma^2$, provided that the sample arises from a distribution with variance $\sigma^2 < \infty$. (See Exercise 6.4-11.) Although S^2 is an unbiased estimator of σ^2, S is a biased estimator of σ. In Exercise 6.4-14, you are asked to show that, when one is sampling from a normal distribution, cS is an unbiased estimator of σ, where

$$c = \frac{\sqrt{n-1}\,\Gamma\left(\dfrac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\dfrac{n}{2}\right)}.$$

REMARK Later we show that S^2 is an unbiased estimator of σ^2, provided it exists, for every distribution, not just the normal. ■

Exercises

6.4-1. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, where the mean $\theta = \mu$ is such that $-\infty < \theta < \infty$ and σ^2 is a known positive number. Show that the maximum likelihood estimator for θ is $\widehat{\theta} = \overline{X}$.

6.4-2. A random sample $X_1, X_2, \ldots, X_n$ of size n is taken from $N(\mu, \sigma^2)$, where the variance $\theta = \sigma^2$ is such that $0 < \theta < \infty$ and μ is a known real number. Show that the maximum likelihood estimator for θ is $\widehat{\theta} = (1/n)\sum_{i=1}^{n}(X_i - \mu)^2$ and that this estimator is an unbiased estimator of θ.

6.4-3. A random sample $X_1, X_2, \ldots, X_n$ of size n is taken from a Poisson distribution with a mean of λ, $0 < \lambda < \infty$.

(a) Show that the maximum likelihood estimator for λ is $\widehat{\lambda} = \overline{X}$.

(b) Let X equal the number of flaws per 100 feet of a used computer tape. Assume that X has a Poisson distribution with a mean of λ. If 40 observations of X yielded 5 zeros, 7 ones, 12 twos, 9 threes, 5 fours, 1 five, and 1 six, find the maximum likelihood estimate of λ.

6.4-4. For determining half-lives of radioactive isotopes, it is important to know what the background radiation is in a given detector over a specific period. The following data were taken in a γ-ray detection experiment over 98 ten-second intervals:

58	50	57	58	64	63	54	64	59	41	43	56	60	50
46	59	54	60	59	60	67	52	65	63	55	61	68	58
63	36	42	54	58	54	40	60	64	56	61	51	48	50
60	42	62	67	58	49	66	58	57	59	52	54	53	53
57	43	73	65	45	43	57	55	73	62	68	55	51	55
53	68	58	53	51	73	44	50	53	62	58	47	63	59
59	56	60	59	50	52	62	51	66	51	56	53	59	57

Assume that these data are observations of a Poisson random variable with mean λ.

(a) Find the values of $\overline{x}$ and s^2.

(b) What is the value of the maximum likelihood estimator of λ?

(c) Is S^2 an unbiased estimator of λ?

(d) Which of $\overline{x}$ and s^2 would you recommend for estimating λ? Why? You could compare the variance of $\overline{X}$ with the variance of S^2, which is

$$\mathrm{Var}(S^2) = \frac{\lambda(2\lambda n + n - 1)}{n(n-1)}.$$

6.4-5. Let $X_1, X_2, \ldots, X_n$ be a random sample from distributions with the given probability density functions. In each case, find the maximum likelihood estimator $\widehat{\theta}$.

(a) $f(x;\theta) = (1/\theta^2)\, x\, e^{-x/\theta}, \quad 0 < x < \infty, \quad 0 < \theta < \infty.$

(b) $f(x;\theta) = (1/2\theta^3)\, x^2\, e^{-x/\theta}, \quad 0 < x < \infty, \quad 0 < \theta < \infty.$

(c) $f(x;\theta) = (1/2)\, e^{-|x-\theta|}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$

Hint: Finding θ involves minimizing $\sum |x_i - \theta|$, which is a difficult problem. When $n = 5$, do it for $x_1 = 6.1$, $x_2 = -1.1$, $x_3 = 3.2$, $x_4 = 0.7$, and $x_5 = 1.7$, and you will see the answer. (See also Exercise 2.2-8.)

6.4-6. Find the maximum likelihood estimates for $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ if a random sample of size 15 from $N(\mu, \sigma^2)$ yielded the following values:

31.5	36.9	33.8	30.1	33.9
35.2	29.6	34.4	30.5	34.2
31.6	36.7	35.8	34.5	32.7

6.4-7. Let $f(x;\theta) = \theta x^{\theta-1}$, $0 < x < 1$, $\theta \in \Omega = \{\theta : 0 < \theta < \infty\}$. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from this distribution.

(a) Sketch the pdf of X for **(i)** $\theta = 1/2$, **(ii)** $\theta = 1$, and **(iii)** $\theta = 2$.

(b) Show that $\widehat{\theta} = -n/\ln\left(\prod_{i=1}^{n} X_i\right)$ is the maximum likelihood estimator of θ.

(c) For each of the following three sets of 10 observations from the given distribution, calculate the values of the maximum likelihood estimate and the method-of-moments estimate of θ:

(i)	0.0256	0.3051	0.0278	0.8971	0.0739
	0.3191	0.7379	0.3671	0.9763	0.0102
(ii)	0.9960	0.3125	0.4374	0.7464	0.8278
	0.9518	0.9924	0.7112	0.2228	0.8609
(iii)	0.4698	0.3675	0.5991	0.9513	0.6049
	0.9917	0.1551	0.0710	0.2110	0.2154

6.4-8. Let $f(x;\theta) = (1/\theta)x^{(1-\theta)/\theta}$, $0 < x < 1$, $0 < \theta < \infty$.

(a) Show that the maximum likelihood estimator of θ is $\widehat{\theta} = -(1/n)\sum_{i=1}^{n} \ln X_i$.

(b) Show that $E(\widehat{\theta}) = \theta$ and thus that $\widehat{\theta}$ is an unbiased estimator of θ.

6.4-9. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the exponential distribution whose pdf is $f(x;\theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, $0 < \theta < \infty$.

(a) Show that $\overline{X}$ is an unbiased estimator of θ.

(b) Show that the variance of $\overline{X}$ is θ^2/n.

(c) What is a good estimate of θ if a random sample of size 5 yielded the sample values 3.5, 8.1, 0.9, 4.4, and 0.5?

6.4-10. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a geometric distribution for which p is the probability of success.

(a) Use the method of moments to find a point estimate for p.

(b) Explain intuitively why your estimate makes good sense.

(c) Use the following data to give a point estimate of p:

3	34	7	4	19	2	1	19	43	2
22	4	19	11	7	1	2	21	15	16

6.4-11. Let $X_1, X_2, \dots, X_n$ be a random sample from a distribution having finite variance σ^2. Show that

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$$

is an unbiased estimator of σ^2. HINT: Write

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$$

and compute $E(S^2)$.

6.4-12. Let $X_1, X_2, \dots, X_n$ be a random sample from $b(1,p)$ (i.e., n Bernoulli trials). Thus,

$$Y = \sum_{i=1}^{n} X_i \text{ is } b(n,p).$$

(a) Show that $\overline{X} = Y/n$ is an unbiased estimator of p.

(b) Show that $\text{Var}(\overline{X}) = p(1-p)/n$.

(c) Show that $E[\overline{X}(1-\overline{X})/n] = (n-1)[p(1-p)/n^2]$.

(d) Find the value of c so that $c\overline{X}(1-\overline{X})$ is an unbiased estimator of $\text{Var}(\overline{X}) = p(1-p)/n$.

6.4-13. Let $X_1, X_2, \dots, X_n$ be a random sample from a uniform distribution on the interval $(\theta - 1, \theta + 1)$.

(a) Find the method-of-moments estimator of θ.

(b) Is your estimator in part (a) an unbiased estimator of θ?

(c) Given the following $n = 5$ observations of X, give a point estimate of θ:

6.61	7.70	6.98	8.36	7.26

(d) The method-of-moments estimator actually has greater variance than the maximum likelihood estimator of θ, *namely* $[\min(X_i) + \max(X_i)]/2$. Compute the value of the latter estimator for the $n = 5$ observations in (c).

6.4-14. Let $X_1, X_2, \dots, X_n$ be a random sample of size n from a normal distribution.

(a) Show that an unbiased estimator of σ is cS, where

$$c = \frac{\sqrt{n-1}\,\Gamma\left(\dfrac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\dfrac{n}{2}\right)}.$$

HINT: Recall that the distribution of $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$.

(b) Find the value of c when $n = 5$; when $n = 6$.

(c) Graph c as a function of n. What is the limit of c as n increases without bound?

6.4-15. Given the following 25 observations from a gamma distribution with mean $\mu = \alpha\theta$ and variance $\sigma^2 = \alpha\theta^2$, use the method-of-moments estimators to find point estimates of α and θ:

6.9	7.3	6.7	6.4	6.3	5.9	7.0	7.1	6.5	7.6	7.2	7.1	6.1
7.3	7.6	7.6	6.7	6.3	5.7	6.7	7.5	5.3	5.4	7.4	6.9	

6.4-16. An urn contains 64 balls, of which N_1 are orange and N_2 are blue. A random sample of $n = 8$ balls is selected from the urn without replacement, and X is equal to the number of orange balls in the sample. This experiment was repeated 30 times (the 8 balls being returned to the urn before each repetition), yielding the following data:

3	0	0	1	1	1	1	3	1	1	2	0	1	3	1
0	1	0	2	1	1	2	3	2	2	4	3	1	1	2

Using these data, guess the value of N_1 and give a reason for your guess.

6.4-17. Let the pdf of X be defined by

$$f(x) = \begin{cases} \left(\dfrac{4}{\theta^2}\right)x, & 0 < x \le \dfrac{\theta}{2}, \\ -\left(\dfrac{4}{\theta^2}\right)x + \dfrac{4}{\theta}, & \dfrac{\theta}{2} < x \le \theta, \\ 0, & \text{elsewhere,} \end{cases}$$

where $\theta \in \Omega = \{\theta : 0 < \theta \le 2\}$.

(a) Sketch the graph of this pdf when $\theta = 1/2, \theta = 1$, and $\theta = 2$.

(b) Find an estimator of θ by the method of moments.

(c) For the following observations of X, give a point estimate of θ:

0.3206	0.2408	0.2577	0.3557	0.4188
0.5601	0.0240	0.5422	0.4532	0.5592

6.4-18. Let independent random samples, each of size n, be taken from the k normal distributions with means $\mu_j = c + d[j - (k+1)/2]$, $j = 1, 2, \dots, k$, respectively, and common variance σ^2. Find the maximum likelihood estimators of c and d.

6.4-19. Let the independent normal random variables $Y_1, Y_2, \dots, Y_n$ have the respective distributions $N(\mu, \gamma^2 x_i^2)$, $i = 1, 2, \dots, n$, where $x_1, x_2, \dots, x_n$ are known but not all the same and no one of which is equal to zero. Find the maximum likelihood estimators for μ and γ^2.

6.5 A SIMPLE REGRESSION PROBLEM

There is often interest in the relation between two variables—for example, the temperature at which a certain chemical reaction is performed and the yield of a chemical compound resulting from the reaction. Frequently, one of these variables, say, x, is known in advance of the other, so there is interest in predicting a future random variable Y. Since Y is a random variable, we cannot predict its future observed value $Y = y$ with certainty. Let us first concentrate on the problem of estimating the mean of Y—that is, $E(Y\,|\,x)$. Now, $E(Y\,|\,x)$ is usually a function of x. For example, in our illustration with the yield, say Y, of the chemical reaction, we might expect $E(Y\,|\,x)$ to increase with increasing temperature x. Sometimes $E(Y\,|\,x) = \mu(x)$ is assumed to be of a given form, such as linear, quadratic, or exponential; that is, $\mu(x)$ could be assumed to be equal to $\alpha + \beta x$, $\alpha + \beta x + \gamma x^2$, or $\alpha e^{\beta x}$. To estimate $E(Y\,|\,x) = \mu(x)$, or, equivalently, the parameters α, β, and γ, we observe the random variable Y for each of n possibly different values of x—say, $x_1, x_2, \ldots, x_n$. Once the n independent experiments have been performed, we have n pairs of known numbers $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. These pairs are then used to estimate the mean $E(Y\,|\,x)$. Problems like this are often classified under **regression** because $E(Y\,|\,x) = \mu(x)$ is frequently called a regression curve.

REMARK A model for the mean that is of the form $\alpha + \beta x + \gamma x^2$ is called a linear model because it is linear in the parameters, α, β, and γ. Note, however, that a plot of this model versus x is not a straight line unless $\gamma = 0$. Thus, a linear model may be nonlinear in x. On the other hand, $\alpha e^{\beta x}$ is not a linear model, because it is not linear in α and β. ■

Let us begin with the case in which $E(Y\,|\,x) = \mu(x)$ is a linear function of x. The data points are $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, so the first problem is that of fitting a straight line to the set of data. (See Figure 6.5-1.) In addition to assuming that the mean of Y is a linear function, we assume that, for a particular value of x, the value of Y will differ from its mean by a random amount ε. We further assume that the distribution of ε is $N(0, \sigma^2)$. So we have, for our linear model,

$$Y_i = \alpha_1 + \beta x_i + \varepsilon_i,$$

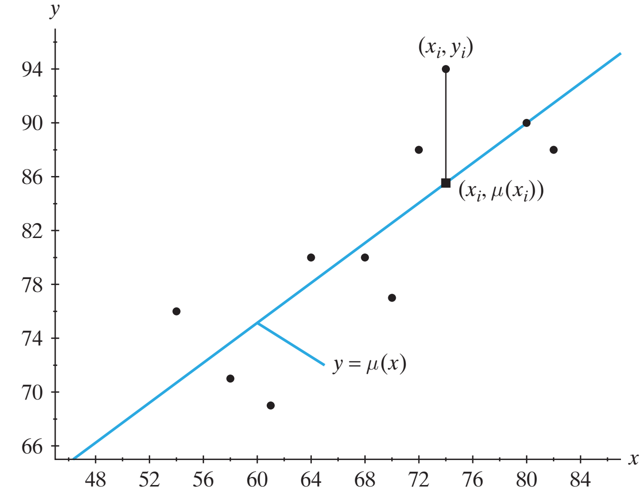


Figure 6.5-1 Scatter plot and the line $y = \mu(x)$

where ε_i, for $i = 1, 2, \ldots, n$, are independent and $N(0, \sigma^2)$. The unknown parameters α_1 and β are the Y-intercept and slope, respectively, of the line $\mu(x) = \alpha_1 + \beta x$.

We shall now find point estimates, specifically maximum likelihood estimates, for α_1, β, and σ^2. For convenience, we let $\alpha_1 = \alpha - \beta\,\overline{x}$, so that

$$Y_i = \alpha + \beta(x_i - \overline{x}) + \varepsilon_i, \text{ where } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Then Y_i is equal to a nonrandom quantity, $\alpha + \beta(x_i - \overline{x})$, plus a mean-zero normal random variable ε_i. Hence, $Y_1, Y_2, \ldots, Y_n$ are mutually independent normal variables with respective means $\alpha + \beta(x_i - \overline{x})$, $i = 1, 2, \ldots, n$, and unknown variance σ^2. Their joint pdf is therefore the product of the individual probability density functions; that is, the likelihood function equals

$$L(\alpha, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{[y_i - \alpha - \beta(x_i - \overline{x})]^2}{2\sigma^2}\right\}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \overline{x})]^2}{2\sigma^2}\right\}.$$

To maximize $L(\alpha, \beta, \sigma^2)$ or, equivalently, to minimize

$$-\ln L(\alpha, \beta, \sigma^2) = \frac{n}{2}\ln(2\pi\sigma^2) + \frac{\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \overline{x})]^2}{2\sigma^2},$$

we must select α and β to minimize

$$H(\alpha, \beta) = \sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \overline{x})]^2.$$

Since $|y_i - \alpha - \beta(x_i - \overline{x})| = |y_i - \mu(x_i)|$ is the vertical distance from the point (x_i, y_i) to the line $y = \mu(x)$, we note that $H(\alpha, \beta)$ represents the sum of the squares of those distances. Thus, selecting α and β so that the sum of the squares is minimized means that we are fitting the straight line to the data by the **method of least squares**. Accordingly, the maximum likelihood estimates of α and β are also called **least squares estimates.**

To minimize $H(\alpha, \beta)$, we find the two first-order partial derivatives

$$\frac{\partial H(\alpha, \beta)}{\partial\alpha} = 2\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \overline{x})](-1)$$

and

$$\frac{\partial H(\alpha, \beta)}{\partial\beta} = 2\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \overline{x})][-(x_i - \overline{x})].$$

Setting $\partial H(\alpha, \beta)/\partial\alpha = 0$, we obtain

$$\sum_{i=1}^{n} y_i - n\alpha - \beta\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

Since

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

we have

$$\sum_{i=1}^{n} y_i - n\alpha = 0;$$

thus,

$$\widehat{\alpha} = \overline{Y}.$$

With α replaced by $\overline{y}$, the equation $\partial H(\alpha, \beta)/\partial\beta = 0$ yields

$$\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) - \beta\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0$$

or, equivalently,

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} Y_i(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Standard methods of multivariate calculus can be used to show that this solution obtained by equating the first-order partial derivatives of $H(\alpha, \beta)$ to zero is indeed a point of minimum. Hence, the line that best estimates the mean line, $\mu(x) = \alpha + \beta(x_i - \overline{x})$, is $\widehat{\alpha} + \widehat{\beta}(x_i - \overline{x})$, where

$$\widehat{\alpha} = \overline{y} \tag{6.5-1}$$

and

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} y_i(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} x_i\right)^2}. \tag{6.5-2}$$

To find the maximum likelihood estimator of σ^2, consider the partial derivative

$$\frac{\partial[-\ln L(\alpha, \beta, \sigma^2)]}{\partial(\sigma^2)} = \frac{n}{2\sigma^2} - \frac{\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \overline{x})]^2}{2(\sigma^2)^2}.$$

Setting this equal to zero and replacing α and β by their solutions $\widehat{\alpha}$ and $\widehat{\beta}$, we obtain

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \widehat{\alpha} - \widehat{\beta}(x_i - \overline{x})]^2. \tag{6.5-3}$$

A formula that is useful in calculating $\widehat{n\sigma^2}$ is

$$\widehat{n\sigma^2} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 - \widehat{\beta}\sum_{i=1}^{n} x_i y_i + \widehat{\beta}\left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right). \tag{6.5-4}$$

Note that the summand in Equation 6.5-3 for $\widehat{\sigma^2}$ is the square of the difference between the value of Y_i and the estimated mean of Y_i. Let $\widehat{Y_i} = \widehat{\alpha} + \widehat{\beta}(x_i - \overline{x})$, the estimated mean value of Y_i, given x. The difference

$$Y_i - \widehat{Y_i} = Y_i - \widehat{\alpha} - \widehat{\beta}(x_i - \overline{x})$$

Table 6.5-1 Calculations for test score data

x	y	x^2	xy	y^2	$\widehat{y}$	$y-\widehat{y}$	$(y-\widehat{y})^2$
70	77	4,900	5,390	5,929	82.561566	−5.561566	30.931016
74	94	5,476	6,956	8,836	85.529956	8.470044	71.741645
72	88	5,184	6,336	7,744	84.045761	3.954239	15.636006
68	80	4,624	5,440	6,400	81.077371	−1.077371	1.160728
58	71	3,364	4,118	5,041	73.656395	−2.656395	7.056434
54	76	2,916	4,104	5,776	70.688004	5.311996	28.217302
82	88	6,724	7,216	7,744	91.466737	−3.466737	12.018265
64	80	4,096	5,120	6,400	78.108980	1.891020	3.575957
80	90	6,400	7,200	8,100	89.982542	0.017458	0.000305
61	69	3,721	4,209	4,761	75.882687	−6.882687	47.371380
683	813	47,405	56,089	66,731	812.999999	0.000001	217.709038

is called the ith **residual**, $i = 1, 2, \ldots, n$. The maximum likelihood estimate of σ^2 is then the sum of the squares of the residuals divided by n. It should always be true that the sum of the residuals is equal to zero. However, in practice, due to rounding off, the sum of the observed residuals, $y_i - \widehat{y_i}$, sometimes differs slightly from zero. A graph of the residuals plotted as a scatter plot of the points $x_i, y_i - \widehat{y_i}$, $i = 1, 2, \ldots, n$, can show whether or not linear regression provides the best fit.

Example 6.5-1

The data plotted in Figure 6.5-1 are 10 pairs of test scores of 10 students in a psychology class, x being the score on a preliminary test and y the score on the final examination. The values of x and y are shown in Table 6.5-1. The sums that are needed to calculate estimates of the parameters are also given. Of course, the estimates of α and β have to be found before the residuals can be calculated.

Thus, $\widehat{\alpha} = 813/10 = 81.3$, and

$$\widehat{\beta} = \frac{56,089 - (683)(813)/10}{47,405 - (683)(683)/10} = \frac{561.1}{756.1} = 0.742.$$

Since $\overline{x} = 683/10 = 68.3$, the least squares regression line is

$$\widehat{y} = 81.3 + (0.742)(x - 68.3).$$

The maximum likelihood estimate of σ^2 is

$$\widehat{\sigma^2} = \frac{217.709038}{10} = 21.7709.$$

A plot of the residuals for these data is shown in Figure 6.5-2. ■

We shall now consider the problem of finding the distributions of $\widehat{\alpha}$, $\widehat{\beta}$, and $\widehat{\sigma^2}$ (or distributions of functions of these estimators). We would like to be able to

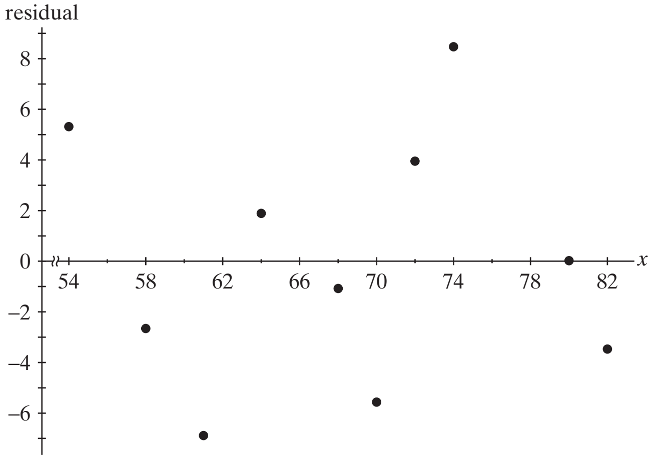


Figure 6.5-2 Residuals plot for data in Table 6.5-1

say something about the error of the estimates to find confidence intervals for the parameters.

The preceding discussion treated $x_1, x_2, \ldots, x_n$ as nonrandom constants. Of course, many times they can be set by the experimenter; for example, an experimental chemist might produce a compound at many different temperatures. But these numbers might instead be observations on an earlier random variable, such as an SAT score or a preliminary test grade (as in Example 6.5-1). Nevertheless, we consider the problem on the condition that the x-values are given in either case. Thus, in finding the distributions of $\widehat{\alpha}$, $\widehat{\beta}$, and $\widehat{\sigma^2}$, the only random variables are $Y_1, Y_2, \ldots, Y_n$.

Since $\widehat{\alpha}$ is a linear function of independent and normally distributed random variables, $\widehat{\alpha}$ has a normal distribution with mean

$$E(\widehat{\alpha}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i)$$
$$= \frac{1}{n}\sum_{i=1}^{n} [\alpha + \beta(x_i - \overline{x})] = \alpha$$

and variance

$$\mathrm{Var}(\widehat{\alpha}) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \mathrm{Var}(Y_i) = \frac{\sigma^2}{n}.$$

The estimator $\widehat{\beta}$ is also a linear function of $Y_1, Y_2, \ldots, Y_n$ and hence has a normal distribution with mean

$$E(\widehat{\beta}) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})E(Y_i)}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$
$$= \frac{\sum_{i=1}^{n} (x_i - \overline{x})[\alpha + \beta(x_i - \overline{x})]}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$
$$= \frac{\alpha \sum_{i=1}^{n} (x_i - \overline{x}) + \beta \sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \beta$$

and variance

$$\operatorname{Var}(\widehat{\beta}) = \sum_{i=1}^{n} \left[\frac{x_i - \overline{x}}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right]^2 \operatorname{Var}(Y_i)$$

$$= \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\left[\sum_{i=1}^{n} (x_i - \overline{x})^2\right]^2} \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

STATISTICAL COMMENTS We now give an illustration (see Ledolter and Hogg in the References) using data from the *Challenger* explosion on January 28, 1986. It would not be appropriate to actually carry out an analysis of these data using the regression methods introduced in this section, for they require the variables to be continuous while in this case the Y variable is discrete. Rather, we present the illustration to make the point that it can be very important to examine the relationship between two variables, and to do so using all available data.

The *Challenger* space shuttle was launched from Cape Kennedy in Florida on a very cold January morning. Meteorologists had forecasted temperatures (as of January 27) in the range of 26°–29° Fahrenheit. The night before the launch there was much debate among engineers and NASA officials whether a launch under such low-temperature conditions would be advisable. Several engineers advised against a launch because they thought that O-ring failures were related to temperature. Data on O-ring failures experienced in previous launches were available and were studied the night before the launch. There were seven previous incidents of known distressed O-rings. Figure 6.5-3(a) displays this information; it is a simple scatter plot of the number of distressed rings per launch against temperature at launch.

From this plot alone, there does not seem to be a strong relationship between the number of O-ring failures and temperature. On the basis of this information, along with many other technical and political considerations, it was decided to launch the *Challenger* space shuttle. As you all know, the launch resulted in disaster: the loss of seven lives and billions of dollars, and a serious setback to the space program.

One may argue that engineers looked at the scatter plot of the number of failures against temperature but could not see a relationship. However, this argument misses the fact that engineers did not display *all the data that were relevant to the question*. They looked only at instances in which there were failures; they ignored

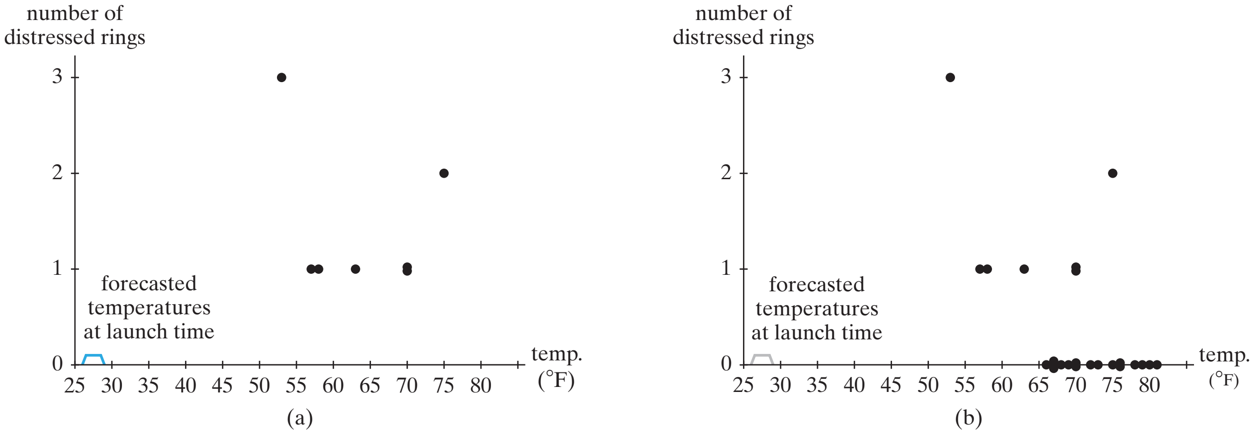


Figure 6.5-3 Number of distressed rings per launch versus temperature

the cases where there were no failures. In fact, there were 17 previous launches in which no failures occurred. A scatter plot of the number of distressed O-rings per launch against temperature using data from all previous shuttle launches is given in Figure 6.5-3(b).

It is difficult to look at these data and not see a relationship between failures and temperature. Moreover, one recognizes that an extrapolation is required and that an inference about the number of failures outside the observed range of temperature is needed. The actual temperature at launch was 31°F, while the lowest temperature recorded at a previous launch was 53°F. It is always very dangerous to extrapolate inferences to a region for which one does not have data. If NASA officials had looked at this plot, certainly the launch would have been delayed. This example shows why it is important to have statistically minded engineers involved in important decisions.

These comments raise two interesting points: (1) It is important to produce a scatter plot of one variable against another. (2) It is also important to plot *relevant data*. Yes, it is true that some data were used in making the decision to launch the *Challenger*. But not all the relevant data were utilized. To make good decisions, it takes knowledge of statistics as well as subject knowledge, common sense, and an ability to question the relevance of information. ■

Exercises

6.5-1. Show that the residuals, $Y_i - \widehat{Y}_i$ $(i = 1, 2, \ldots, n)$, from the least squares fit of the simple linear regression model sum to zero.

6.5-2. In some situations where the regression model is useful, it is known that the mean of Y when $X = 0$ is equal to 0, i.e., $Y_i = \beta x_i + \varepsilon_i$ where ε_i for $i = 1, 2, \ldots, n$ are independent and $N(0, \sigma^2)$.

(a) Obtain the maximum likelihood estimators, $\widehat{\beta}$ and $\widehat{\sigma^2}$, of β and σ^2 under this model.

(b) Find the distributions of $\widehat{\beta}$ and $\widehat{\sigma^2}$. (You may use, without proof, the fact that $\widehat{\beta}$ and $\widehat{\sigma^2}$ are independent, together with Theorem 9.3-1.)

6.5-3. The midterm and final exam scores of 10 students in a statistics course are tabulated as shown.

(a) Calculate the least squares regression line for these data.

(b) Plot the points and the least squares regression line on the same graph.

(c) Find the value of $\widehat{\sigma^2}$.

Midterm	Final	Midterm	Final
70	87	67	73
74	79	70	83
80	88	64	79
84	98	74	91
80	96	82	94

6.5-4. The final grade in a calculus course was predicted on the basis of the student's high school grade point average in mathematics, Scholastic Aptitude Test (SAT) score in mathematics, and score on a mathematics entrance examination. The predicted grades x and the earned grades y for 10 students are given (2.0 represents a C, 2.3 a C+, 2.7 a B–, etc.).

(a) Calculate the least squares regression line for these data.

(b) Plot the points and the least squares regression line on the same graph.

(c) Find the value of $\widehat{\sigma^2}$.

x	y	x	y
2.0	1.3	2.7	3.0
3.3	3.3	4.0	4.0
3.7	3.3	3.7	3.0
2.0	2.0	3.0	2.7
2.3	1.7	2.3	3.0

6.5-5. A student who considered himself to be a "car guy" was interested in how the horsepower and weight of a car affected the time that it takes the car to go from 0 to 60 mph. The following table gives, for each of 14 cars, the horsepower, the time in seconds to go from 0 to 60 mph, and the weight in pounds:

Horsepower	0–60	Weight	Horsepower	0–60	Weight
230	8.1	3516	282	6.2	3627
225	7.8	3690	300	6.4	3892
375	4.7	2976	220	7.7	3377
322	6.6	4215	250	7.0	3625
190	8.4	3761	315	5.3	3230
150	8.4	2940	200	6.2	2657
178	7.2	2818	300	5.5	3518

(a) Calculate the least squares regression line for "0–60" versus horsepower.

(b) Plot the points and the least squares regression line on the same graph.

(c) Calculate the least squares regression line for "0–60" versus weight.

(d) Plot the points and the least squares regression line on the same graph.

(e) Which of the two variables, horsepower or weight, has the most effect on the "0–60" time?

6.5-6. Let x and y equal the ACT scores in social science and natural science, respectively, for a student who is applying for admission to a small liberal arts college. A sample of $n = 15$ such students yielded the following data:

x	y	x	y	x	y
32	28	30	27	26	32
23	25	17	23	16	22
23	24	20	30	21	28
23	32	17	18	24	31
26	31	18	18	30	26

(a) Calculate the least squares regression line for these data.

(b) Plot the points and the least squares regression line on the same graph.

(c) Find point estimates for α, β, and σ^2.

6.5-7. The Federal Trade Commission measured the number of milligrams of tar and carbon monoxide (CO) per cigarette for all domestic cigarettes. Let x and y equal the measurements of tar and CO, respectively, for 100-millimeter filtered and mentholated cigarettes. A sample of 12 brands yielded the following data:

Brand	x	y	Brand	x	y
Capri	9	6	Now	3	4
Carlton	4	6	Salem	17	18
Kent	14	14	Triumph	6	8
Kool Milds	12	12	True	7	8
Marlboro Lights	10	12	Vantage	8	13
Merit Ultras	5	7	Virginia Slims	15	13

(a) Calculate the least squares regression line for these data.

(b) Plot the points and the least squares regression line on the same graph.

(c) Find point estimates for α, β, and σ^2.

6.5-8. The data in the following table, part of a set of data collected by Ledolter and Hogg (see References), provide the number of miles per gallon (mpg) for city and highway driving of 2007 midsize-model cars, as well as the curb weight of the cars:

Type	mpg City	mpg Hwy	Curb Weight
Ford Fusion V6 SE	20	28	3230
Chevrolet Sebring Sedan Base	24	32	3287
Toyota Camry Solara SE	24	34	3240
Honda Accord Sedan	20	29	3344
Audi A6 3.2	21	29	3825
BMW 5-series 525i Sedan	20	29	3450
Chrysler PT Cruiser Base	22	29	3076
Mercedes E-Class E350 Sedan	19	26	3740
Volkswagen Passat Sedan 2.0T	23	32	3305
Nissan Altima 2.5	26	35	3055
Kia Optima LX	24	34	3142

(a) Find the least squares regression line for highway mpg (y) and city mpg (x).

(b) Plot the points and the least squares regression line on the same graph.

(c) Repeat parts (a) and (b) for the regression of highway mpg (y) on curb weight (x).

6.5-9. Using an Instron 4204, rectangular strips of Plexiglas® were stretched to failure in a tensile test. The following data give the change in length, in millimeters (mm), before breaking (x) and the cross–sectional area in square millimeters (mm^2) (y):

(5.28, 52.36)	(5.40, 52.58)	(4.65, 51.07)	(4.76, 52.28)	(5.55, 53.02)
(5.73, 52.10)	(5.84, 52.61)	(4.97, 52.21)	(5.50, 52.39)	(6.24, 53.77)

(a) Find the equation of the least squares regression line.

(b) Plot the points and the line on the same graph.

(c) Interpret your output.

6.5-10. The "golden ratio" is $\phi = (1 + \sqrt{5})/2$. John Putz, a mathematician who was interested in music, analyzed Mozart's sonata movements, which are divided into two distinct sections, both of which are repeated in performance (see References). The length of the "Exposition" in measures is represented by a and the length of the "Development and Recapitulation" is represented by b. Putz's conjecture was that Mozart divided his movements close to the golden ratio. That is, Putz was interested in studying whether a scatter plot of $a + b$ against b not only would be linear, but also would actually fall along the line $y = \phi x$. Here are the data in tabular form, in which the first column identifies the piece and movement by the Köchel cataloging system:

Köchel	a	b	$a+b$	Köchel	a	b	$a+b$
279, I	38	62	100	279, II	28	46	74
279, III	56	102	158	280, I	56	88	144
280, II	24	36	60	280, III	77	113	190
281, I	40	69	109	281, II	46	60	106
282, I	15	18	33	282, III	39	63	102
283, I	53	67	120	283, II	14	23	37
283, III	102	171	273	284, I	51	76	127
309, I	58	97	155	311, I	39	73	112
310, I	49	84	133	330, I	58	92	150
330, III	68	103	171	332, I	93	136	229
332, III	90	155	245	333, I	63	102	165
333, II	31	50	81	457, I	74	93	167
533, I	102	137	239	533, II	46	76	122
545, I	28	45	73	547a, I	78	118	196
570, I	79	130	209				

(a) Make a scatter plot of the points $a + b$ against the points b. Is this plot linear?

(b) Find the equation of the least squares regression line. Superimpose it on the scatter plot.

(c) On the scatter plot, superimpose the line $y = \phi x$. Compare this line with the least squares regression line (graphically if you wish).

(d) Find the sample mean of the points $(a + b)/b$. Is the mean close to ϕ?

6.6* ASYMPTOTIC DISTRIBUTIONS OF MAXIMUM LIKELIHOOD ESTIMATORS

Let us consider a distribution of the continuous type with pdf $f(x;\theta)$ such that the parameter θ is not involved in the support of the distribution. Moreover, we want $f(x;\theta)$ to possess a number of mathematical properties that we do not list here. However, in particular, we want to be able to find the maximum likelihood estimator $\widehat{\theta}$ by solving

$$\frac{\partial[\ln L(\theta)]}{\partial\theta} = 0,$$

where here we use a partial derivative sign because $L(\theta)$ involves $x_1, x_2, \ldots, x_n$, too.

That is,

$$\frac{\partial[\ln L(\widehat{\theta})]}{\partial\theta} = 0,$$

where now, with $\widehat{\theta}$ in this expression, $L(\widehat{\theta}) = f(X_1;\widehat{\theta})f(X_2;\widehat{\theta})\cdots f(X_n;\widehat{\theta})$. We can approximate the left-hand member of this latter equation by a linear function found from the first two terms of a Taylor's series expanded about θ, namely,

$$\frac{\partial[\ln L(\theta)]}{\partial\theta} + (\widehat{\theta} - \theta)\frac{\partial^2[\ln L(\theta)]}{\partial\theta^2} \approx 0,$$

when $L(\theta) = f(X_1;\theta)f(X_2;\theta)\cdots f(X_n;\theta)$.

Obviously, this approximation is good enough only if $\widehat{\theta}$ is close to θ, and an adequate mathematical proof involves those conditions, which we have not given here. (See Hogg, McKean, and Craig, 2013.) But a heuristic argument can be made by solving for $\widehat{\theta} - \theta$ to obtain

$$\widehat{\theta} - \theta = \frac{\dfrac{\partial[\ln L(\theta)]}{\partial\theta}}{-\dfrac{\partial^2[\ln L(\theta)]}{\partial\theta^2}}. \tag{6.6-1}$$

Recall that

$$\ln L(\theta) = \ln f(X_1;\theta) + \ln f(X_2;\theta) + \cdots + \ln f(X_n;\theta)$$

and

$$\frac{\partial \ln L(\theta)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial[\ln f(X_i;\theta)]}{\partial\theta}, \tag{6.6-2}$$

which is the numerator in Equation 6.6-1. However, Equation 6.6-2 gives the sum of the n independent and identically distributed random variables

$$Y_i = \frac{\partial[\ln f(X_i;\theta)]}{\partial\theta}, \qquad i = 1, 2, \ldots, n,$$

and thus, by the central limit theorem, has an approximate normal distribution with mean (in the continuous case) equal to

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{\partial[\ln f(x;\theta)]}{\partial\theta} f(x;\theta)\,dx &= \int_{-\infty}^{\infty} \frac{\partial[f(x;\theta)]}{\partial\theta}\frac{f(x;\theta)}{f(x;\theta)}\,dx\\ &= \int_{-\infty}^{\infty} \frac{\partial[f(x;\theta)]}{\partial\theta}\,dx\\ &= \frac{\partial}{\partial\theta}\left[\int_{-\infty}^{\infty} f(x;\theta)\,dx\right]\\ &= \frac{\partial}{\partial\theta}[1]\\ &= 0.\end{aligned}$$

Clearly, we need a certain mathematical condition that makes it permissible to interchange the operations of integration and differentiation in those last steps. Of course, the integral of $f(x;\theta)$ is equal to 1 because it is a pdf.

Since we now know that the mean of each Y is

$$\int_{-\infty}^{\infty} \frac{\partial[\ln f(x;\theta)]}{\partial\theta} f(x;\theta)\,dx = 0,$$

let us take derivatives of each member of this equation with respect to θ, obtaining

$$\int_{-\infty}^{\infty} \left\{\frac{\partial^2[\ln f(x;\theta)]}{\partial\theta^2} f(x;\theta) + \frac{\partial[\ln f(x;\theta)]}{\partial\theta}\frac{\partial[f(x;\theta)]}{\partial\theta}\right\} dx = 0.$$

However,

$$\frac{\partial[f(x;\theta)]}{\partial\theta} = \frac{\partial[\ln f(x;\theta)]}{\partial\theta} f(x;\theta);$$

so

$$\int_{-\infty}^{\infty} \left\{\frac{\partial[\ln f(x;\theta)]}{\partial\theta}\right\}^2 f(x;\theta)\,dx = -\int_{-\infty}^{\infty} \frac{\partial^2[\ln f(x;\theta)]}{\partial\theta^2} f(x;\theta)\,dx.$$

Since $E(Y) = 0$, this last expression provides the variance of $Y = \partial[\ln f(X;\theta)]/\partial\theta$. Then the variance of the sum in Equation 6.6-2 is n times this value, namely,

$$-nE\left\{\frac{\partial^2[\ln f(X;\theta)]}{\partial\theta^2}\right\}.$$

Let us rewrite Equation 6.6-1 as

$$\frac{\sqrt{n}\,(\widehat{\theta}-\theta)}{\left(\dfrac{1}{\sqrt{-E\{\partial^2[\ln f(X;\theta)]/\partial\theta^2\}}}\right)} = \frac{\left(\dfrac{\partial[\ln L(\theta)]/\partial\theta}{\sqrt{-nE\{\partial^2[\ln f(X;\theta)]/\partial\theta^2\}}}\right)}{\left(\dfrac{-\dfrac{1}{n}\dfrac{\partial^2[\ln L(\theta)]}{\partial\theta^2}}{E\{-\partial^2[\ln f(X;\theta)]/\partial\theta^2\}}\right)}. \tag{6.6-3}$$

Since it is the sum of n independent random variables (see Equation 6.6-2),

$$\partial[\ln f(X_i;\theta)]/\partial\theta, \qquad i = 1, 2, \ldots, n,$$

the numerator of the right-hand member of Equation 6.6-3 has an approximate $N(0,1)$ distribution, and the aforementioned unstated mathematical conditions require, in some sense, that

$$-\frac{1}{n}\frac{\partial^2[\ln L(\theta)]}{\partial\theta^2} \qquad \text{converge to} \qquad E\{-\partial^2[\ln f(X;\theta)]/\partial\theta^2\}.$$

Accordingly, the ratios given in Equation 6.6-3 must be approximately $N(0,1)$. That is, $\widehat{\theta}$ has an approximate normal distribution with mean θ and standard deviation

$$\frac{1}{\sqrt{-nE\{\partial^2[\ln f(X;\theta)]/\partial\theta^2\}}}.$$

Example 6.6-1

(Continuation of Example 6.4-1.) With the underlying exponential pdf

$$f(x;\theta) = \frac{1}{\theta}e^{-x/\theta}, \qquad 0 < x < \infty, \qquad \theta \in \Omega = \{\theta : 0 < \theta < \infty\},$$

$\overline{X}$ is the maximum likelihood estimator. Since

$$\ln f(x;\theta) = -\ln\theta - \frac{x}{\theta}$$

and

$$\frac{\partial[\ln f(x;\theta)]}{\partial\theta} = -\frac{1}{\theta} + \frac{x}{\theta^2} \qquad \text{and} \qquad \frac{\partial^2[\ln f(x;\theta)]}{\partial\theta} = \frac{1}{\theta^2} - \frac{2x}{\theta^3},$$

we have

$$-E\left[\frac{1}{\theta^2}-\frac{2X}{\theta^3}\right]=-\frac{1}{\theta^2}+\frac{2\theta}{\theta^3}=\frac{1}{\theta^2},$$

because $E(X)=\theta$. That is, $\overline{X}$ has an approximate normal distribution with mean θ and standard deviation $\theta/\sqrt{n}$. Thus, the random interval $\overline{X}\pm 1.96(\theta/\sqrt{n})$ has an approximate probability of 0.95 that it covers θ. Substituting the observed $\overline{x}$ for θ, as well as for $\overline{X}$, we say that $\overline{x}\pm 1.96\,\overline{x}/\sqrt{n}$ is an approximate 95% confidence interval for θ. ■

While the development of the preceding result used a continuous-type distribution, the result holds for the discrete type also, as long as the support does not involve the parameter. This is illustrated in the next example.

Example 6.6-2 (Continuation of Exercise 6.4-3.) If the random sample arises from a Poisson distribution with pmf

$$f(x;\lambda)=\frac{\lambda^x e^{-\lambda}}{x!},\qquad x=0,1,2,\ldots;\qquad \lambda\in\Omega=\{\lambda: 0<\lambda<\infty\},$$

then the maximum likelihood estimator for λ is $\widehat{\lambda}=\overline{X}$. Now,

$$\ln f(x;\lambda)=x\ln\lambda-\lambda-\ln x!.$$

Also,

$$\frac{\partial[\ln f(x;\lambda)]}{\partial\lambda}=\frac{x}{\lambda}-1\qquad\text{and}\qquad\frac{\partial^2[\ln f(x;\lambda)]}{\partial\lambda^2}=-\frac{x}{\lambda^2}.$$

Thus,

$$-E\left(-\frac{X}{\lambda^2}\right)=\frac{\lambda}{\lambda^2}=\frac{1}{\lambda},$$

and $\widehat{\lambda}=\overline{X}$ has an approximate normal distribution with mean λ and standard deviation $\sqrt{\lambda/n}$. Finally, $\overline{x}\pm 1.645\sqrt{\overline{x}/n}$ serves as an approximate 90% confidence interval for λ. With the data in Exercise 6.4-3, $\overline{x}=2.225$, and it follows that this interval ranges from 1.837 to 2.613. ■

It is interesting that there is another theorem which is somewhat related to the preceding result in that the variance of $\widehat{\theta}$ serves as a lower bound for the variance of every unbiased estimator of θ. Thus, we know that if a certain unbiased estimator has a variance equal to that lower bound, we cannot find a better one, and hence that estimator is the best in the sense of being the minimum-variance unbiased estimator. So, in the limit, the maximum likelihood estimator is this type of best estimator.

We describe this **Rao–Cramér inequality** here without proof. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution of the continuous type with pdf $f(x;\theta)$, $\theta\in\Omega=\{\theta: c<\theta<d\}$, where the support of X does not depend upon θ, so that we can differentiate, with respect to θ, under integral signs like that in the following integral:

$$\int_{-\infty}^{\infty} f(x;\theta)\,dx=1.$$

If $Y = u(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of θ, then

$$\begin{aligned}\mathrm{Var}(Y) &\geq \frac{1}{n\int_{-\infty}^{\infty}\{[\partial \ln f(x;\theta)/\partial\theta]\}^2 f(x;\theta)\,dx}\\ &= \frac{-1}{n\int_{-\infty}^{\infty}[\partial^2 \ln f(x;\theta)/\partial\theta^2]\, f(x;\theta)\,dx}.\end{aligned}$$

Note that the integrals in the denominators are, respectively, the expectations

$$E\left\{\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^2\right\} \qquad \text{and} \qquad E\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right];$$

sometimes one is easier to compute than the other. Note also that although the Rao–Cramér lower bound has been stated only for a continuous-type distribution, it is also true for a discrete-type distribution, with summations replacing integrals.

We have computed this lower bound for each of two distributions: exponential with mean θ and Poisson with mean λ. Those respective lower bounds were θ^2/n and λ/n. (See Examples 6.6-1 and 6.6-2.) Since, in each case, the variance of $\overline{X}$ equals the lower bound, then $\overline{X}$ is the minimum-variance unbiased estimator.

Let us consider another example.

Example 6.6-3

(Continuation of Exercise 6.4-7.) Let the pdf of X be given by

$$f(x;\theta) = \theta x^{\theta-1}, \qquad 0 < x < 1, \qquad \theta \in \Omega = \{\theta : 0 < \theta < \infty\}.$$

We then have

$$\ln f(x;\theta) = \ln\theta + (\theta - 1)\ln x,$$

$$\frac{\partial \ln f(x;\theta)}{\partial\theta} = \frac{1}{\theta} + \ln x,$$

and

$$\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2} = -\frac{1}{\theta^2}.$$

Since $E(-1/\theta^2) = -1/\theta^2$, the greatest lower bound of the variance of every unbiased estimator of θ is θ^2/n. Moreover, the maximum likelihood estimator $\widehat{\theta} = -n/\ln\prod_{i=1}^n X_i$ has an approximate normal distribution with mean θ and variance θ^2/n. Thus, in a limiting sense, $\widehat{\theta}$ is the minimum variance unbiased estimator of θ. ■

To measure the value of estimators, their variances are compared with the Rao–Cramér lower bound. The ratio of the Rao–Cramér lower bound to the actual variance of any unbiased estimator is called the **efficiency** of that estimator. An estimator with an efficiency of, say, 50%, means that $1/0.5 = 2$ times as many sample observations are needed to do as well in estimation as can be done with the minimum variance unbiased estimator (the 100% efficient estimator).

Exercises

6.6-1. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\theta, \sigma^2)$, where σ^2 is known.

(a) Show that $Y = (X_1 + X_2)/2$ is an unbiased estimator of θ.

(b) Find the Rao–Cramér lower bound for the variance of an unbiased estimator of θ for a general n.

(c) What is the efficiency of Y in part (a)?

6.6-2. Let $X_1, X_2, \ldots, X_n$ denote a random sample from $b(1, p)$. We know that $\overline{X}$ is an unbiased estimator of p and that $\text{Var}(\overline{X}) = p(1-p)/n$. (See Exercise 6.4-12.)

(a) Find the Rao–Cramér lower bound for the variance of every unbiased estimator of p.

(b) What is the efficiency of $\overline{X}$ as an estimator of p?

6.6-3. (Continuation of Exercise 6.4-2.) In sampling from a normal distribution with known mean μ, the maximum likelihood estimator of $\theta = \sigma^2$ is $\widehat{\theta} = \sum_{i=1}^{n}(X_i - \mu)^2/n$.

(a) Determine the Rao–Cramér lower bound.

(b) What is the approximate distribution of $\widehat{\theta}$?

(c) What is the exact distribution of $n\widehat{\theta}/\theta$, where $\theta = \sigma^2$?

6.6-4. Find the Rao–Cramér lower bound, and thus the asymptotic variance of the maximum likelihood estimator $\widehat{\theta}$, if the random sample $X_1, X_2, \ldots, X_n$ is taken from each of the distributions having the following pdfs:

(a) $f(x;\theta) = (1/\theta^2)\, x\, e^{-x/\theta}, \quad 0 < x < \infty, \quad 0 < \theta < \infty.$

(b) $f(x;\theta) = (1/2\theta^3)\, x^2\, e^{-x/\theta}, \quad 0 < x < \infty, \quad 0 < \theta < \infty.$

(c) $f(x;\theta) = (1/\theta)\, x^{(1-\theta)/\theta}, \quad 0 < x < 1, \quad 0 < \theta < \infty.$

6.7 SUFFICIENT STATISTICS

We first define a sufficient statistic $Y = u(X_1, X_2, \ldots, X_n)$ for a parameter, using a statement that, in most books, is given as a necessary and sufficient condition for sufficiency, namely, the well-known Fisher–Neyman factorization theorem. We do this because we find that readers at the introductory level can apply such a definition easily. However, using this definition, we shall note, by examples, its implications, one of which is also sometimes used as the definition of sufficiency. An understanding of Example 6.7-3 is most important in an appreciation of the value of sufficient statistics.

Definition 6.7-1

(Factorization Theorem) Let $X_1, X_2, \ldots, X_n$ denote random variables with joint pdf or pmf $f(x_1, x_2, \ldots, x_n; \theta)$, which depends on the parameter θ. The statistic $Y = u(X_1, X_2, \ldots, X_n)$ is sufficient for θ if and only if

$$f(x_1, x_2, \ldots, x_n; \theta) = \phi[u(x_1, x_2, \ldots, x_n); \theta]h(x_1, x_2, \ldots, x_n),$$

where ϕ depends on $x_1, x_2, \ldots, x_n$ only through $u(x_1, \ldots, x_n)$ and $h(x_1, \ldots, x_n)$ does not depend on θ.

Let us consider several important examples and consequences of this definition. We first note, however, that in all instances in this book the random variables $X_1, X_2, \ldots, X_n$ will be of a random sample, and hence their joint pdf or pmf will be of the form

$$f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta).$$

Example 6.7-1

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Poisson distribution with parameter $\lambda > 0$. Then

$$f(x_1;\lambda)f(x_2;\lambda)\cdots f(x_n;\lambda) = \frac{\lambda^{\Sigma x_i}e^{-n\lambda}}{x_1!x_2!\cdots x_n!} = (\lambda^{n\bar{x}}e^{-n\lambda})\left(\frac{1}{x_1!x_2!\cdots x_n!}\right),$$

where $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$. Thus, from the factorization theorem (Definition 6.7-1), it is clear that the sample mean $\overline{X}$ is a sufficient statistic for λ. It can easily be shown that the maximum likelihood estimator for λ is also $\overline{X}$, so here the maximum likelihood estimator is a function of a sufficient statistic. ■

In Example 6.7-1, if we replace $n\bar{x}$ by $\sum_{i=1}^{n} x_i$, it is quite obvious that the sum $\sum_{i=1}^{n} X_i$ is also a sufficient statistic for λ. This certainly agrees with our intuition, because if we know one of the statistics $\overline{X}$ and $\sum_{i=1}^{n} X_i$, we can easily find the other. If we generalize this idea, we see that if Y is sufficient for a parameter θ, then every single-valued function of Y not involving θ, but with a single-valued inverse, is also a sufficient statistic for θ. The reason is that if we know either Y or that function of Y, we know the other. More formally, if $W = v(Y) = v[u(X_1, X_2, \ldots, X_n)]$ is that function and $Y = v^{-1}(W)$ is the single-valued inverse, then the factorization theorem can be written as

$$f(x_1, x_2, \ldots, x_n; \theta) = \phi[v^{-1}\{v[u(x_1, x_2, \ldots, x_n)]\}; \theta]\, h(x_1, x_2, \ldots, x_n).$$

The first factor of the right-hand member of this equation depends on $x_1, x_2, \ldots, x_n$ through $v[u(x_1, x_2, \ldots, x_n)]$, so $W = v[u(X_1, X_2, \ldots, X_n)]$ is a sufficient statistic for θ. We illustrate this fact and the factorization theorem with an underlying distribution of the continuous type.

Example 6.7-2

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, 1)$, $-\infty < \mu < \infty$. The joint pdf of these random variables is

$$\begin{aligned}\frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right] &= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}[(x_i-\bar{x})+(\bar{x}-\mu)]^2\right] \\ &= \left\{\exp\left[-\frac{n}{2}(\bar{x}-\mu)^2\right]\right\}\left\{\frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\right\}.\end{aligned}$$

From the factorization theorem, we see that $\overline{X}$ is sufficient for μ. Now, $\overline{X}^3$ is also sufficient for μ, because knowing $\overline{X}^3$ is equivalent to having knowledge of the value of $\overline{X}$. However, $\overline{X}^2$ does not have this property, and it is not sufficient for μ. ■

One extremely important consequence of the sufficiency of a statistic Y is that the conditional probability of any given event A in the support of $X_1, X_2, \ldots, X_n$, given that $Y = y$, does not depend on θ. This consequence is sometimes used as the definition of sufficiency and is illustrated in the next example.

Example 6.7-3

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pmf

$$f(x;p) = p^x(1-p)^{1-x}, \qquad x = 0, 1,$$

where the parameter p is between 0 and 1. We know that

$$Y = X_1 + X_2 + \cdots + X_n$$

is $b(n,p)$ and Y is sufficient for p because the joint pmf of $X_1, X_2, \ldots, X_n$ is

$$p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n} = [p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}](1),$$

where $\phi(y;p) = p^y(1-p)^{n-y}$ and $h(x_1, x_2, \ldots, x_n) = 1$. What, then, is the conditional probability $P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y)$, where $y = 0, 1, \ldots, n-1$, or n? Unless the sum of the nonnegative integers $x_1, x_2, \ldots, x_n$ equals y, this conditional probability is obviously equal to zero, which does not depend on p. Hence, it is interesting to consider the solution only when $y = x_1 + \cdots + x_n$. From the definition of conditional probability, we have

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y) &= \frac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(Y = y)} \\ &= \frac{p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n}}{\binom{n}{y} p^y (1-p)^{n-y}} \\ &= \frac{1}{\binom{n}{y}}, \end{aligned}$$

where $y = x_1 + \cdots + x_n$. Since y equals the number of ones in the collection $x_1, x_2, \ldots, x_n$, this answer is only the probability of selecting a particular arrangement, namely, $x_1, x_2, \ldots, x_n$, of y ones and $n - y$ zeros, and does not depend on the parameter p. That is, given that the sufficient statistic $Y = y$, the conditional probability of $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ does not depend on the parameter p. ■

It is interesting to observe that the underlying pdf or pmf in Examples 6.7-1, 6.7-2, and 6.7-3 can be written in the exponential form

$$f(x;\theta) = \exp[K(x)p(\theta) + S(x) + q(\theta)],$$

where the support is free of θ. That is, we have, respectively,

$$\frac{e^{-\lambda}\lambda^x}{x!} = \exp\{x \ln\lambda - \ln x! - \lambda\}, \qquad x = 0, 1, 2, \ldots,$$

$$\frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} = \exp\left\{x\mu - \frac{x^2}{2} - \frac{\mu^2}{2} - \frac{1}{2}\ln(2\pi)\right\}, \qquad -\infty < x < \infty,$$

and

$$p^x(1-p)^{1-x} = \exp\left\{x \ln\left(\frac{p}{1-p}\right) + \ln(1-p)\right\}, \qquad x = 0, 1.$$

In each of these examples, the sum $\sum_{i=1}^{n} X_i$ of the observations of the random sample is a sufficient statistic for the parameter. This idea is generalized by Theorem 6.7-1.

Theorem 6.7-1 Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with a pdf or pmf of the exponential form

$$f(x;\theta) = \exp[K(x)p(\theta) + S(x) + q(\theta)]$$

on a support free of θ. Then the statistic $\sum_{i=1}^{n} K(X_i)$ is sufficient for θ.

Proof The joint pdf (pmf) of $X_1, X_2, \ldots, X_n$ is

$$\exp\left[p(\theta)\sum_{i=1}^{n} K(x_i) + \sum_{i=1}^{n} S(x_i) + nq(\theta)\right]$$

$$= \left\{\exp\left[p(\theta)\sum_{i=1}^{n} K(x_i) + nq(\theta)\right]\right\}\left\{\exp\left[\sum_{i=1}^{n} S(x_i)\right]\right\}.$$

In accordance with the factorization theorem, the statistic $\sum_{i=1}^{n} K(X_i)$ is sufficient for θ. □

In many cases, Theorem 6.7-1 permits the student to find a sufficient statistic for a parameter with very little effort, as shown in the next example.

Example 6.7-4 Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with pdf

$$f(x;\theta) = \frac{1}{\theta}e^{-x/\theta} = \exp\left[x\left(-\frac{1}{\theta}\right) - \ln\theta\right], \qquad 0 < x < \infty,$$

provided that $0 < \theta < \infty$. Here, $K(x) = x$. Thus, $\sum_{i=1}^{n} X_i$ is sufficient for θ; of course, $\overline{X} = \sum_{i=1}^{n} X_i/n$ is also sufficient. ■

Note that if there is a sufficient statistic for the parameter under consideration and if the maximum likelihood estimator of this parameter is unique, then the maximum likelihood estimator is a function of the sufficient statistic. To see this heuristically, consider the following: If a sufficient statistic exists, then the likelihood function is

$$L(\theta) = f(x_1, x_2, \ldots, x_n; \theta) = \phi[u(x_1, x_2, \ldots, x_n); \theta]\, h(x_1, x_2, \ldots, x_n).$$

Since $h(x_1, x_2, \ldots, x_n)$ does not depend on θ, we maximize $L(\theta)$ by maximizing $\phi[u(x_1, x_2, \ldots, x_n); \theta]$. But ϕ is a function of $x_1, x_2, \ldots, x_n$ only through the statistic $u(x_1, x_2, \ldots, x_n)$. Thus, if there is a unique value of θ that maximizes ϕ, then it must be a function of $u(x_1, x_2, \ldots, x_n)$. That is, $\widehat{\theta}$ is a function of the sufficient statistic $u(X_1, X_2, \ldots, X_n)$. This fact was alluded to in Example 6.7-1, but it could be checked with the use of other examples and exercises.

In many cases, we have two (or more) parameters—say, θ_1 and θ_2. All of the preceding concepts can be extended to these situations. For example, Definition 6.7-1 (the factorization theorem) becomes the following in the case of two parameters: If

$$f(x_1, \ldots, x_n; \theta_1, \theta_2) = \phi[u_1(x_1, \ldots, x_n), u_2(x_1, \ldots, x_n); \theta_1, \theta_2]h(x_1, \ldots, x_n),$$

where ϕ depends on $x_1, x_2, \ldots, x_n$ only through $u_1(x_1, \ldots, x_n)$, $u_2(x_1, \ldots, x_n)$, and $h(x_1, x_2, \ldots, x_n)$ does not depend upon θ_1 or θ_2, then $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ and $Y_2 = u_2(X_1, X_2, \ldots, X_n)$ are **jointly sufficient statistics** for θ_1 and θ_2.

Example 6.7-5 Let $X_1, X_2, \ldots, X_n$ denote a random sample from a normal distribution $N(\theta_1 = \mu, \theta_2 = \sigma^2)$. Then

$$\prod_{i=1}^{n} f(x_i;\theta_1,\theta_2) = \left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \exp\left[-\sum_{i=1}^{n}(x_i-\theta_1)^2 \Big/ 2\theta_2\right]$$

$$= \exp\left[\left(-\frac{1}{2\theta_2}\right)\sum_{i=1}^{n} x_i^2 + \left(\frac{\theta_1}{\theta_2}\right)\sum_{i=1}^{n} x_i - \frac{n\theta_1^2}{2\theta_2} - n\ln\sqrt{2\pi\theta_2}\right]\cdot(1).$$

Thus,

$$Y_1 = \sum_{i=1}^{n} X_i^2 \qquad \text{and} \qquad Y_2 = \sum_{i=1}^{n} X_i$$

are joint sufficient statistics for θ_1 and θ_2. Of course, the single-valued functions of Y_1 and Y_2, namely,

$$\overline{X} = \frac{Y_2}{n} \qquad \text{and} \qquad S^2 = \frac{Y_1 - Y_2^2/n}{n-1},$$

are also joint sufficient statistics for θ_1 and θ_2. ■

Actually, we can see from Definition 6.7-1 and Example 6.7-5 that if we can write the pdf in the exponential form, it is easy to find joint sufficient statistics. In that example,

$$f(x;\theta_1,\theta_2) = \exp\left(\frac{-1}{2\theta_2}x^2 + \frac{\theta_1}{\theta_2}x - \frac{\theta_1^2}{2\theta_2} - \ln\sqrt{2\pi\theta_2}\right);$$

so

$$Y_1 = \sum_{i=1}^{n} X_i^2 \qquad \text{and} \qquad Y_2 = \sum_{i=1}^{n} X_i$$

are joint sufficient statistics for θ_1 and θ_2. A much more complicated illustration is given if we take a random sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ from a bivariate normal distribution with parameters $\theta_1 = \mu_X$, $\theta_2 = \mu_Y$, $\theta_3 = \sigma_X^2$, $\theta_4 = \sigma_Y^2$, and $\theta_5 = \rho$. In Exercise 6.7-3, we write the bivariate normal pdf $f(x,y;\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)$ in exponential form and see that $Z_1 = \sum_{i=1}^n X_i^2$, $Z_2 = \sum_{i=1}^n Y_i^2$, $Z_3 = \sum_{i=1}^n X_iY_i$, $Z_4 = \sum_{i=1}^n X_i$, and $Z_5 = \sum_{i=1}^n Y_i$ are joint sufficient statistics for θ_1, θ_2, θ_3, θ_4, and θ_5. Of course, the single-valued functions

$$\overline{X} = \frac{Z_4}{n}, \qquad \overline{Y} = \frac{Z_5}{n}, \qquad S_X^2 = \frac{Z_1 - Z_4^2/n}{n-1},$$

$$S_Y^2 = \frac{Z_2 - Z_5^2/n}{n-1}, \qquad R = \frac{(Z_3 - Z_4Z_5/n)/(n-1)}{S_XS_Y}$$

are also joint sufficient statistics for those parameters.

The important point to stress for cases in which sufficient statistics exist is that once the sufficient statistics are given, there is no additional information about the parameters left in the remaining (conditional) distribution. That is, all statistical inferences should be based upon the sufficient statistics. To help convince the reader of this in point estimation, we state and prove the well-known **Rao–Blackwell theorem**.

Theorem 6.7-2 Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf or pmf $f(x;\theta)$, $\theta \in \Omega$. Let $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ be a sufficient statistic for θ, and let $Y_2 = u_2(X_1, X_2, \ldots, X_n)$ be an unbiased estimator of θ, where Y_2 is not a function of Y_1 alone. Then $E(Y_2 \,|\, y_1) = u(y_1)$ defines a statistic $u(Y_1)$, a function of the sufficient statistic Y_1, which is an unbiased estimator of θ, and its variance is less than that of Y_2.

Proof Let $g(y_1, y_2; \theta)$ be the joint pdf or pmf of Y_1 and Y_2. Let $g_1(y_1;\theta)$ be the marginal of Y_1; thus,

$$\frac{g(y_1, y_2; \theta)}{g_1(y_1; \theta)} = h(y_2 \,|\, y_1)$$

is the conditional pdf or pmf of Y_2, given that $Y_1 = y_1$. This equation does not depend upon θ, since Y_1 is a sufficient statistic for θ. Of course, in the continuous case,

$$u(y_1) = \int_{S_2} y_2 h(y_2 \,|\, y_1)\, dy_2 = \int_{S_2} y_2 \frac{g(y_1, y_2; \theta)}{g_1(y_1; \theta)} dy_2$$

and

$$\begin{aligned} E[u(Y_1)] &= \int_{S_1} \left(\int_{S_2} y_2 \frac{g(y_1, y_2; \theta)}{g_1(y_1; \theta)} dy_2 \right) g_1(y_1; \theta)\, dy_1 \\ &= \int_{S_1} \int_{S_2} y_2\, g(y_1, y_2; \theta)\, dy_2\, dy_1 = \theta, \end{aligned}$$

because Y_2 is an unbiased estimator of θ. Thus, $u(Y_1)$ is also an unbiased estimator of θ.

Now, consider

$$\begin{aligned} \text{Var}(Y_2) = E[(Y_2 - \theta)^2] \;&=\; E[\{Y_2 - u(Y_1) + u(Y_1) - \theta\}^2] \\ &= E[\{Y_2 - u(Y_1)\}^2] + E[\{u(Y_1) - \theta\}^2] + 2E[\{Y_2 - u(Y_1)\}\{u(Y_1) - \theta\}]. \end{aligned}$$

But the latter expression (i.e., the third term) is equal to

$$2 \int_{S_1} [u(y_1) - \theta] \left\{ \int_{S_2} [y_2 - u(y_1)]\, h(y_2 \,|\, y_1)\, dy_2 \right\} g(y_1; \theta)\, dy_1 = 0,$$

because $u(y_1)$ is the mean $E(Y_2 \,|\, y_1)$ of Y_2 in the conditional distribution given by $h(y_2 \,|\, y_1)$. Thus,

$$\text{Var}(Y_2) = E[\{Y_2 - u(Y_1)\}^2] + \text{Var}[u(Y_1)].$$

However, $E[\{(Y_2 - u(Y_1)\}^2 \geq 0$, as it is the expected value of a positive expression. Therefore,

$$\text{Var}(Y_2) \geq \text{Var}[u(Y_1)]. \qquad \square$$

The importance of this theorem is that it shows that for every other unbiased estimator of θ, we can always find an unbiased estimator based on the sufficient statistic that has a variance at least as small as the first unbiased estimator. Hence, in that sense, the one based upon the sufficient statistic is at least as good as the first one. More importantly, we might as well begin our search for an unbiased estimator with the smallest variance by considering only those unbiased estimators based upon the sufficient statistics. Moreover, in an advanced course we show that if the underlying distribution is described by a pdf or pmf of the exponential form, then, if an unbiased estimator exists, there is only one function of the sufficient statistic that is unbiased. That is, that unbiased estimator is unique. (See Hogg, McKean, and Craig, 2013.)

There is one other useful result involving a sufficient statistic Y for a parameter θ, particularly with a pdf of the exponential form. It is that if another statistic Z has a distribution that is free of θ, then Y and Z are independent. This is the reason $Z = (n-1)S^2$ is independent of $Y = \overline{X}$ when the sample arises from a distribution that is $N(\theta, \sigma^2)$. The sample mean is a sufficient statistic for θ, and

$$Z = (n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$$

has a distribution that is free of θ. To see this, we note that the mgf of Z, namely, $E(e^{tZ})$, is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left[t\sum_{i=1}^{n}(x_i-\overline{x})^2\right]\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{n}\exp\left[-\frac{\sum(x_i-\theta)^2}{2\sigma^2}\right]dx_1dx_2\ldots dx_n.$$

Changing variables by letting $x_i - \theta = w_i$, $i = 1, 2, \ldots, n$, the preceding expression becomes

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left[t\sum_{i=1}^{n}(w_i-\overline{w})^2\right]\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{n}\exp\left[-\frac{\sum w_i^2}{2\sigma^2}\right]dw_1dw_2\ldots dw_n,$$

which is free of θ.

An outline of the proof of this result is given by noting that

$$\int_y [h(z\,|\,y) - g_2(z)]\,g_1(y;\theta)\,dy = g_2(z) - g_2(z) = 0$$

for all $\theta \in \Omega$. However, $h(z\,|\,y)$ is free of θ due to the hypothesis of sufficiency; so $h(z\,|\,y) - g_2(z)$ is free of θ, since Z has a distribution that is free of θ. Since $N(\theta, \sigma^2)$ is of the exponential form, $Y = \overline{X}$ has a pdf $g_1(y\,|\,\theta)$ that requires $h(z\,|\,y) - g_2(z)$ to be equal to zero. That is,

$$h(z\,|\,y) = g_2(z),$$

which means that Z and Y are independent. This proves the independence of $\overline{X}$ and S^2, which was stated in Theorem 5.5-2.

Example 6.7-6

Let $X_1, X_2, \ldots, X_n$ be a random sample from a gamma distribution with α (given) and $\theta > 0$, which is of exponential form. Now, $Y = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ, since the gamma pdf is of the exponential form. Clearly, then,

$$Z = \frac{\sum_{i=1}^{n} a_iX_i}{\sum_{i=1}^{n} X_i},$$

where not all constants $a_1, a_2, \ldots, a_n$ are equal, has a distribution that is free of the spread parameter θ because the mgf of Z, namely,

$$E(e^{tZ}) = \int_0^{\infty}\int_0^{\infty}\cdots\int_0^{\infty}\frac{e^{t\Sigma a_iX_i/\Sigma X_i}}{[\Gamma(\alpha)]^n\theta^{n\alpha}}(x_1x_2\cdots x_n)^{\alpha-1}e^{-\Sigma x_i/\theta}dx_1dx_2\ldots dx_n,$$

and does not depend upon θ, as is seen by the transformation $w_i = x_i/\theta$, $i = 1, 2, \ldots, n$. So Y and Z are independent statistics. ■

This special case of the independence of Y and Z concerning one sufficient statistic Y and one parameter θ was first observed by Hogg (1953) and then generalized to several sufficient statistics for more than one parameter by Basu (1955) and is usually called **Basu's theorem**.

Due to these results, sufficient statistics are extremely important and estimation problems are based upon them when they exist.

Exercises

6.7-1. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(0, \sigma^2)$.

(a) Find a sufficient statistic Y for σ^2.

(b) Show that the maximum likelihood estimator for σ^2 is a function of Y.

(c) Is the maximum likelihood estimator for σ^2 unbiased?

6.7-2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with mean $\lambda > 0$. Find the conditional probability $P(X_1 = x_1, \ldots, X_n = x_n \,|\, Y = y)$, where $Y = X_1 + \cdots + X_n$ and the nonnegative integers $x_1, x_2, \ldots, x_n$ sum to y. Note that this probability does not depend on λ.

6.7-3. Write the bivariate normal pdf $f(x, y; \theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ in exponential form and show that $Z_1 = \sum_{i=1}^n X_i^2$, $Z_2 = \sum_{i=1}^n Y_i^2$, $Z_3 = \sum_{i=1}^n X_i Y_i$, $Z_4 = \sum_{i=1}^n X_i$, and $Z_5 = \sum_{i=1}^n Y_i$ are joint sufficient statistics for θ_1, θ_2, θ_3, θ_4, and θ_5.

6.7-4. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x; \theta) = \theta x^{\theta - 1}$, $0 < x < 1$, where $0 < \theta$.

(a) Find a sufficient statistic Y for θ.

(b) Show that the maximum likelihood estimator $\widehat{\theta}$ is a function of Y.

(c) Argue that $\widehat{\theta}$ is also sufficient for θ.

6.7-5. Let $X_1, X_2, \ldots, X_n$ be a random sample from a gamma distribution with $\alpha = 1$ and $1/\theta > 0$. Show that $Y = \sum_{i=1}^n X_i$ is a sufficient statistic, Y has a gamma distribution with parameters n and $1/\theta$, and $(n-1)/Y$ is an unbiased estimator of θ.

6.7-6. Let $X_1, X_2, \ldots, X_n$ be a random sample from a gamma distribution with known parameter α and unknown parameter $\theta > 0$.

(a) Show that $Y = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.

(b) Show that the maximum likelihood estimator of θ is a function of Y and is an unbiased estimator of θ.

6.7-7. Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution with pmf $f(x; p) = p(1-p)^{x-1}$, $x = 1, 2, 3, \ldots$, where $0 < p \le 1$.

(a) Show that $Y = \sum_{i=1}^n X_i$ is a sufficient statistic for p.

(b) Find a function of Y that is an unbiased estimator of $\theta = 1/p$.

6.7-8. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(0, \theta)$, where $\sigma^2 = \theta > 0$ is unknown. Argue that the sufficient statistic $Y = \sum_{i=1}^n X_i^2$ for θ and $Z = \sum_{i=1}^n a_i X_i / \sum_{i=1}^n X_i$ are independent. Hint: Let $x_i = \theta w_i$, $i = 1, 2, \ldots, n$, in the multivariate integral representing $E[e^{tZ}]$.

6.7-9. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\theta_1, \theta_2)$. Show that the sufficient statistics $Y_1 = \overline{X}$ and $Y_2 = S^2$ are independent of the statistic

$$Z = \sum_{i=1}^{n-1} \frac{(X_{i+1} - X_i)^2}{S^2}$$

because Z has a distribution that is free of θ_1 and θ_2.

Hint: Let $w_i = (x_i - \theta_1)/\sqrt{\theta_2}$, $i = 1, 2, \ldots, n$, in the multivariate integral representing $E[e^{tZ}]$.

6.7-10. Find a sufficient statistic for θ, given a random sample, $X_1, X_2, \ldots, X_n$, from a distribution with pdf $f(x; \theta) = \{\Gamma(2\theta)/[\Gamma(\theta)]^2\} x^{\theta-1}(1-x)^{\theta-1}$, $0 < x < 1$.

6.7-11. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x; \theta) = (1/2)\theta^3 x^2 e^{-\theta x}$, $0 < x < \infty$. Show that $Y = \sum_{i=1}^n X_i$ and $Z = (X_1 + X_2)/Y$ are independent.

6.7-12. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(0, \sigma^2)$, where n is odd. Let Y and Z be the mean and median of the sample. Argue that Y and $Z - Y$ are independent so that the variance of Z is $\text{Var}(Y) + \text{Var}(Z - Y)$. We know that $\text{Var}(Y) = \sigma^2/n$, so that we could estimate the $\text{Var}(Z - Y)$ by Monte Carlo. This might be more efficient than estimating $\text{Var}(Z)$ directly since $\text{Var}(Z - Y) \le \text{Var}(Z)$. This scheme is often called the **Monte Carlo Swindle**.

6.8 BAYESIAN ESTIMATION

We now describe another approach to estimation that is used by a group of statisticians who call themselves Bayesians. To understand their approach fully would require more text than we can allocate to this topic, but let us begin this brief introduction by considering a simple application of the theorem of the Reverend Thomas Bayes. (See Section 1.5.)

Example 6.8-1 Suppose we know that we are going to select an observation from a Poisson distribution with mean λ equal to 2 or 4. Moreover, prior to performing the experiment, we believe that $\lambda = 2$ has about four times as much chance of being the parameter as does $\lambda = 4$; that is, the prior probabilities are $P(\lambda = 2) = 0.8$ and $P(\lambda = 4) = 0.2$. The experiment is now performed and we observe that $x = 6$. At this point, our intuition tells us that $\lambda = 2$ seems less likely than before, as the observation $x = 6$ is much more probable with $\lambda = 4$ than with $\lambda = 2$, because, in an obvious notation,

$$P(X = 6 \,|\, \lambda = 2) = 0.995 - 0.983 = 0.012$$

and

$$P(X = 6 \,|\, \lambda = 4) = 0.889 - 0.785 = 0.104,$$

from Table III in Appendix B. Our intuition can be supported by computing the conditional probability of $\lambda = 2$, given that $X = 6$:

$$\begin{aligned} P(\lambda = 2 \,|\, X = 6) &= \frac{P(\lambda = 2, X = 6)}{P(X = 6)} \\ &= \frac{P(\lambda = 2)P(X = 6 \,|\, \lambda = 2)}{P(\lambda = 2)P(X = 6 \,|\, \lambda = 2) + P(\lambda = 4)P(X = 6 \,|\, \lambda = 4)} \\ &= \frac{(0.8)(0.012)}{(0.8)(0.012) + (0.2)(0.104)} = 0.316. \end{aligned}$$

This conditional probability is called the posterior probability of $\lambda = 2$, given the single data point (here, $x = 6$). In a similar fashion, the posterior probability of $\lambda = 4$ is found to be 0.684. Thus, we see that the probability of $\lambda = 2$ has decreased from 0.8 (the prior probability) to 0.316 (the posterior probability) with the observation of $x = 6$. ■

In a more practical application, the parameter, say, θ can take many more than two values as in Example 6.8-1. Somehow Bayesians must assign prior probabilities to this total parameter space through a prior pdf $h(\theta)$. They have developed procedures for assessing these prior probabilities, and we simply cannot do justice to these methods here. Somehow $h(\theta)$ reflects the prior weights that the Bayesian wants to assign to the various possible values of θ. In some instances, if $h(\theta)$ is a constant and thus θ has the uniform prior distribution, we say that the Bayesian has a **noninformative** prior. If, in fact, some knowledge of θ exists in advance of experimentation, noninformative priors should be avoided if at all possible.

Also, in more practical examples, we usually take several observations, not just one. That is, we take a random sample, and there is frequently a good statistic, say, Y, for the parameter θ. Suppose we are considering a continuous case and the pdf of Y, say, $g(y;\theta)$, can be thought of as the conditional pdf of Y, given θ. [Henceforth in this section, we write $g(y;\theta) = g(y \,|\, \theta)$.] Thus, we can treat

$$g(y \,|\, \theta)h(\theta) = k(y, \theta)$$

as the joint pdf of the statistic Y and the parameter. Of course, the marginal pdf of Y is

$$k_1(y) = \int_{-\infty}^{\infty} h(\theta)g(y\,|\,\theta)\,d\theta.$$

Consequently,

$$\frac{k(y,\theta)}{k_1(y)} = \frac{g(y\,|\,\theta)h(\theta)}{k_1(y)} = k(\theta\,|\,y)$$

would serve as the conditional pdf of the parameter, given that $Y = y$. This formula is essentially Bayes' theorem, and $k(\theta\,|\,y)$ is called the **posterior pdf of** θ, given that $Y = y$.

Bayesians believe that everything which needs to be known about the parameter is summarized in this posterior pdf $k(\theta\,|\,y)$. Suppose, for example, that they were pressed into making a point estimate of the parameter θ. They would note that they would be guessing the value of a random variable, here θ, given its pdf $k(\theta\,|\,y)$. There are many ways that this could be done: The mean, the median, or the mode of that distribution would be reasonable guesses. However, in the final analysis, the best guess would clearly depend upon the penalties for various errors created by incorrect guesses. For instance, if we were penalized by taking the square of the error between the guess, say, $w(y)$, and the real value of the parameter θ, clearly we would use the conditional mean

$$w(y) = \int_{-\infty}^{\infty} \theta k(\theta\,|\,y)\,d\theta$$

as our Bayes estimate of θ. The reason is that, in general, if Z is a random variable, then the function of b, $E[(Z-b)^2]$, is minimized by $b = E(Z)$. (See Example 2.2-4.) Likewise, if the penalty (loss) function is the absolute value of the error, $|\theta - w(y)|$, then we use the median of the distribution, because with any random variable Z, $E[\,|Z-b|\,]$ is minimized when b equals the median of the distribution of Z. (See Exercise 2.2-8.)

Example 6.8-2

Suppose that Y has a binomial distribution with parameters n and $p = \theta$. Then the pmf of Y, given θ, is

$$g(y\,|\,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \qquad y = 0, 1, 2, \ldots, n.$$

Let us take the prior pdf of the parameter to be the beta pdf:

$$h(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad 0 < \theta < 1.$$

Such a prior pdf provides a Bayesian a great deal of flexibility through the selection of the parameters α and β. Thus, the joint probabilities can be described by a product of a binomial pmf with parameters n and θ and this beta pdf, namely,

$$k(y,\theta) = \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1},$$

on the support given by $y = 0, 1, 2, \ldots, n$ and $0 < \theta < 1$. We find

$$\begin{aligned} k_1(y) &= \int_0^1 k(y, \theta)\, d\theta \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \frac{\Gamma(\alpha+y)\Gamma(n+\beta-y)}{\Gamma(n+\alpha+\beta)} \end{aligned}$$

on the support $y = 0, 1, 2, \ldots, n$ by comparing the integral with one involving a beta pdf with parameters $y+\alpha$ and $n-y+\beta$. Therefore,

$$\begin{aligned} k(\theta \,|\, y) &= \frac{k(y, \theta)}{k_1(y)} \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+y)\Gamma(n+\beta-y)}\, \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}, \qquad 0 < \theta < 1, \end{aligned}$$

which is a beta pdf with parameters $y+\alpha$ and $n-y+\beta$. With the squared error loss function we must minimize, with respect to $w(y)$, the integral

$$\int_0^1 [\theta - w(y)]^2\, k(\theta \,|\, y)\, d\theta,$$

to obtain the Bayes estimator. But, as noted earlier, if Z is a random variable with a second moment, then $E[(Z-b)^2]$ is minimized by $b = E(Z)$. In the preceding integration, θ is like the Z with pdf $k(\theta \,|\, y)$, and $w(y)$ is like the b, so the minimization is accomplished by taking

$$w(y) = E(\theta \,|\, y) = \frac{\alpha+y}{\alpha+\beta+n},$$

which is the mean of the beta distribution with parameters $y+\alpha$ and $n-y+\beta$. (See Exercise 5.2-8.) It is instructive to note that this Bayes estimator can be written as

$$w(y) = \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right),$$

which is a weighted average of the maximum likelihood estimate y/n of θ and the mean $\alpha/(\alpha+\beta)$ of the prior pdf of the parameter. Moreover, the respective weights are $n/(\alpha+\beta+n)$ and $(\alpha+\beta)/(\alpha+\beta+n)$. Thus, we see that α and β should be selected so that not only is $\alpha/(\alpha+\beta)$ the desired prior mean, but also the sum $\alpha+\beta$ plays a role corresponding to a sample size. That is, if we want our prior opinion to have as much weight as a sample size of 20, we would take $\alpha+\beta = 20$. So if our prior mean is 3/4, we select $\alpha = 15$ and $\beta = 5$. That is, the prior pdf of θ is beta(15, 5). If we observe $n = 40$ and $y = 28$, then the posterior pdf is beta($28+15 = 43$, $12+5 = 17$). The prior and posterior pdfs are shown in Figure 6.8-1. ■

In Example 6.8-2, it is quite convenient to note that it is not really necessary to determine $k_1(y)$ to find $k(\theta \,|\, y)$. If we divide $k(y, \theta)$ by $k_1(y)$, we get the product of a factor that depends on y but does *not* depend on θ—say, $c(y)$—and we have

$$\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}.$$

That is,

$$k(\theta \,|\, y) = c(y)\, \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}, \qquad 0 < \theta < 1.$$

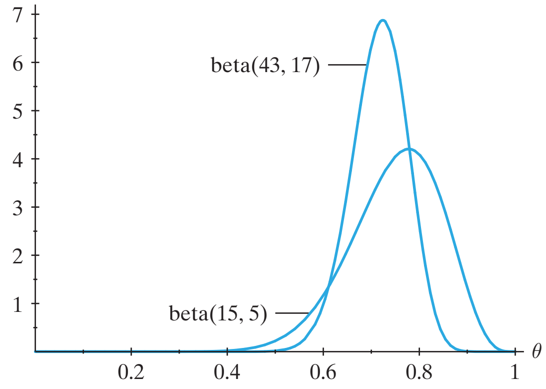


Figure 6.8-1 Beta prior and posterior pdfs

However, $c(y)$ must be that "constant" needed to make $k(\theta \,|\, y)$ a pdf, namely,

$$c(y) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}.$$

Accordingly, Bayesians frequently write that $k(\theta \,|\, y)$ is proportional to $k(y, \theta) = g(y \,|\, \theta)h(\theta)$; that is,

$$k(\theta \,|\, y) \propto g(y \,|\, \theta)\, h(\theta).$$

Then, to actually form the pdf $k(\theta \,|\, y)$, they simply find the "constant" (which is, of course, actually some function of y) such that the expression integrates to 1.

Example 6.8-3 Suppose that $Y = \overline{X}$ is the mean of a random sample of size n that arises from the normal distribution $N(\theta, \sigma^2)$, where σ^2 is known. Then $g(y \,|\, \theta)$ is $N(\theta, \sigma^2/n)$. Suppose further that we are able to assign prior weights to θ through a prior pdf $h(\theta)$ that is $N(\theta_0, \sigma_0^2)$. Then we have

$$k(\theta \,|\, y) \propto \frac{1}{\sqrt{2\pi}\,(\sigma/\sqrt{n})} \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left[-\frac{(y-\theta)^2}{2(\sigma^2/n)} - \frac{(\theta-\theta_0)^2}{2\sigma_0^2} \right].$$

If we eliminate all constant factors (including factors involving y only), then

$$k(\theta \,|\, y) \propto \exp\left[-\frac{(\sigma_0^2 + \sigma^2/n)\theta^2 - 2(y\sigma_0^2 + \theta_0\sigma^2/n)\theta}{2(\sigma^2/n)\sigma_0^2} \right].$$

This expression can be simplified by completing the square, to read (after eliminating factors not involving θ)

$$k(\theta \,|\, y) \propto \exp\left\{ -\frac{[\theta - (y\sigma_0^2 + \theta_0\sigma^2/n)/(\sigma_0^2 + \sigma^2/n)]^2}{[2(\sigma^2/n)\sigma_0^2]/[\sigma_0^2 + (\sigma^2/n)]} \right\}.$$

That is, the posterior pdf of the parameter is obviously normal with mean

$$\frac{y\sigma_0^2 + \theta_0\sigma^2/n}{\sigma_0^2 + \sigma^2/n} = \left(\frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n} \right) y + \left(\frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n} \right) \theta_0$$

and variance $(\sigma^2/n)\sigma_0^2/(\sigma_0^2 + \sigma^2/n)$. If the squared error loss function is used, then this posterior mean is the Bayes estimator. Again, note that it is a weighted average of the maximum likelihood estimate $y = \bar{x}$ and the prior mean θ_0. The Bayes estimator $w(y)$ will always be a value between the prior judgment and the usual estimate. Note also, here and in Example 6.8-2, that the Bayes estimator gets closer to the maximum likelihood estimate as n increases. Thus, the Bayesian procedures permit the decision maker to enter his or her prior opinions into the solution in a very formal way so that the influence of those prior notions will be less and less as n increases. ■

In Bayesian statistics, all the information is contained in the posterior pdf $k(\theta \,|\, y)$. In Examples 6.8-2 and 6.8-3, we found Bayesian point estimates with the use of the squared error loss function. Note that if the loss function is the absolute value of the error, $|w(y) - \theta|$, then the Bayes estimator would be the median of the posterior distribution of the parameter, which is given by $k(\theta \,|\, y)$. Hence, the Bayes estimator changes—as it should—with different loss functions.

Finally, if an interval estimate of θ is desired, we would find two functions of y—say, $u(y)$ and $v(y)$—such that

$$\int_{u(y)}^{v(y)} k(\theta \,|\, y)\, d\theta = 1 - \alpha,$$

where α is small—say, $\alpha = 0.05$. Then the observed interval from $u(y)$ to $v(y)$ would serve as an interval estimate for the parameter in the sense that the posterior probability of the parameter's being in that interval is $1-\alpha$. In Example 6.8-3, where the posterior pdf of the parameter was normal, the interval

$$\frac{y\sigma_0^2 + \theta_0\sigma^2/n}{\sigma_0^2 + \sigma^2/n} \pm 1.96\sqrt{\frac{(\sigma^2/n)\sigma_0^2}{\sigma_0^2 + \sigma^2/n}}$$

serves as an interval estimate for θ with posterior probability of 0.95.

In closing this short section on Bayesian estimation, note that we could have begun with the sample observations $X_1, X_2, \ldots, X_n$, rather than some statistic Y. Then, in our discussion, we would replace $g(y \,|\, \theta)$ by the likelihood function

$$L(\theta) = f(x_1 \,|\, \theta) f(x_2 \,|\, \theta) \cdots f(x_n \,|\, \theta),$$

which is the joint pdf of $X_1, X_2, \ldots, X_n$, given θ. Thus, we find that

$$k(\theta \,|\, x_1, x_2, \ldots, x_n) \propto h(\theta) f(x_1 \,|\, \theta) f(x_2 \,|\, \theta) \cdots f(x_n \,|\, \theta) = h(\theta) L(\theta).$$

Now, $k(\theta \,|\, x_1, x_2, \ldots, x_n)$ contains all the information about θ, given the data. Thus, depending on the loss function, we would choose our Bayes estimate of θ as some characteristic of this posterior distribution, such as the mean or the median. It is interesting to observe that if the loss function is zero for some small neighborhood about the true parameter θ and is some large positive constant otherwise, then the Bayes estimate, $w(x_1, x_2, \ldots, x_n)$, is essentially the mode of this conditional pdf, $k(\theta \,|\, x_1, x_2, \ldots, x_n)$. The reason for this is that we want to take the estimate so that it has as much posterior probability as possible in a small neighborhood around it. Finally, note that if $h(\theta)$ is a constant (a noninformative prior), then this Bayes estimate using the mode is exactly the same as the maximum likelihood estimate. More generally, if $h(\theta)$ is not a constant, then the Bayes estimate using the mode can be thought of as a weighted maximum likelihood estimate in which the weights reflect prior opinion about θ. That is, that value of θ which maximizes $h(\theta)L(\theta)$ is the mode of the posterior distribution of the parameter given the data and can be used as the Bayes estimate associated with the appropriate loss function.

Example 6.8-4 Let us consider again Example 6.8-2, but now say that $X_1, X_2, \ldots, X_n$ is a random sample from the Bernoulli distribution with pmf

$$f(x\,|\,\theta) = \theta^x(1-\theta)^{1-x}, \qquad x = 0, 1.$$

With the same prior pdf of θ, the joint distribution of $X_1, X_2, \ldots, X_n$ and θ is given by

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}, \qquad 0 < \theta < 1,\ \ x_i = 0, 1.$$

Of course, the posterior pdf of θ, given that $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, is such that

$$k(\theta\,|\,x_1, x_2, \ldots, x_n) \propto \theta^{\sum_{i=1}^n x_i + \alpha - 1}(1-\theta)^{n-\sum_{i=1}^n x_i + \beta - 1}, \qquad 0 < \theta < 1,$$

which is beta with $\alpha^* = \sum x_i + \alpha$, $\beta^* = n - \sum x_i + \beta$. The conditional mean of θ is

$$\frac{\sum_{i=1}^n x_i + \alpha}{n+\alpha+\beta} = \left(\frac{n}{n+\alpha+\beta}\right)\left(\frac{\sum_{i=1}^n x_i}{n}\right) + \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)\left(\frac{\alpha}{\alpha+\beta}\right),$$

which, with $y = \sum x_i$, is exactly the same result as that of Example 6.8-2. ■

Exercises

6.8-1. Let Y be the sum of the observations of a random sample from a Poisson distribution with mean θ. Let the prior pdf of θ be gamma with parameters α and β.

(a) Find the posterior pdf of θ, given that $Y = y$.

(b) If the loss function is $[w(y) - \theta]^2$, find the Bayesian point estimate $w(y)$.

(c) Show that $w(y)$ found in (b) is a weighted average of the maximum likelihood estimate y/n and the prior mean $\alpha\beta$, with respective weights of $n/(n + 1/\beta)$ and $(1/\beta)/(n + 1/\beta)$.

6.8-2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a gamma distribution with known α and with $\theta = 1/\tau$. Say τ has a prior pdf that is gamma with parameters α_0 and θ_0, so that the prior mean is $\alpha_0\theta_0$.

(a) Find the posterior pdf of τ, given that $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

(b) Find the mean of the posterior distribution found in part (a), and write it as a function of the sample mean $\overline{X}$ and $\alpha_0\theta_0$.

(c) Explain how you would find a 95% interval estimate of τ if $n = 10$, $\alpha = 3$, $\alpha_0 = 10$, and $\theta_0 = 2$.

6.8-3. In Example 6.8-2, take $n = 30$, $\alpha = 15$, and $\beta = 5$.

(a) Using the squared error loss, compute the expected loss (risk function) associated with the Bayes estimator $w(Y)$.

(b) The risk function associated with the usual estimator Y/n is, of course, $\theta(1-\theta)/30$. Find those values of θ for which the risk function in part (a) is less than $\theta(1-\theta)/30$. In particular, if the prior mean $\alpha/(\alpha+\beta) = 3/4$ is a reasonable guess, then the risk function in part (a) is the better of the two (i.e., is smaller in a neighborhood of $\theta = 3/4$) for what values of θ?

6.8-4. Consider a random sample $X_1, X_2, \ldots, X_n$ from a distribution with pdf

$$f(x\,|\,\theta) = 3\theta x^2 e^{-\theta x^3}, \qquad 0 < x < \infty.$$

Let θ have a prior pdf that is gamma with $\alpha = 4$ and the usual $\theta = 1/4$. Find the conditional mean of θ, given that $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

6.8-5. In Example 6.8-3, suppose the loss function $|\theta - w(Y)|$ is used. What is the Bayes estimator $w(Y)$?

6.8-6. Let Y be the largest order statistic of a random sample of size n from a distribution with pdf $f(x\,|\,\theta) = 1/\theta$, $0 < x < \theta$. Say θ has the prior pdf

$$h(\theta) = \beta\alpha^\beta/\theta^{\beta+1}, \qquad \alpha < \theta < \infty,$$

where $\alpha > 0$, $\beta > 0$.

(a) If $w(Y)$ is the Bayes estimator of θ and $[\theta - w(Y)]^2$ is the loss function, find $w(Y)$.

(b) If $n = 4$, $\alpha = 1$, and $\beta = 2$, find the Bayesian estimator $w(Y)$ if the loss function is $|\theta - w(Y)|$.

6.8-7. Refer to Example 6.8-3. Suppose we select $\sigma_0^2 = d\sigma^2$, where σ^2 is known in that example. What value do we assign to d so that the variance of the posterior pdf of the parameter is two thirds of the variance of $Y = \overline{X}$, namely, σ^2/n?

6.8-8. Consider the likelihood function $L(\alpha, \beta, \sigma^2)$ of Section 6.5. Let α and β be independent with priors $N(\alpha_1, \sigma_1^2)$ and $N(\beta_0, \sigma_0^2)$. Determine the posterior mean of $\alpha + \beta(x - \overline{x})$.

6.9* MORE BAYESIAN CONCEPTS

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf (pmf) $f(x\,|\,\theta)$, and let $h(\theta)$ be the prior pdf. Then the distribution associated with the marginal pdf of $X_1, X_2, \ldots, X_n$, namely,

$$k_1(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{\infty} f(x_1\,|\,\theta) f(x_2\,|\,\theta) \cdots f(x_n\,|\,\theta)\, h(\theta)\, d\theta,$$

is called the **predictive distribution** because it provides the best description of the probabilities on $X_1, X_2, \ldots, X_n$. Often this creates some interesting distributions. For example, suppose there is only one X with the normal pdf

$$f(x\,|\,\theta) = \frac{\sqrt{\theta}}{\sqrt{2\pi}} e^{-(\theta x^2)/2}, \qquad -\infty < x < \infty.$$

Here, $\theta = 1/\sigma^2$, the inverse of the variance, is called the **precision** of X. Say this precision has the gamma pdf

$$h(\theta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\, \theta^{\alpha-1} e^{-\theta/\beta}, \qquad 0 < \theta < \infty.$$

Then the predictive pdf is

$$\begin{aligned} k_1(x) &= \int_0^\infty \frac{\theta^{\alpha+\frac{1}{2}-1} e^{-\left(\frac{x^2}{2}+\frac{1}{\beta}\right)\theta}}{\Gamma(\alpha)\beta^\alpha\sqrt{2\pi}}\, d\theta \\ &= \frac{\Gamma(\alpha+1/2)}{\Gamma(\alpha)\beta^\alpha\sqrt{2\pi}}\, \frac{1}{(1/\beta + x^2/2)^{\alpha+1/2}}, \qquad -\infty < x < \infty. \end{aligned}$$

Note that if $\alpha = r/2$ and $\beta = 2/r$, where r is a positive integer, then

$$k_1(x) \propto \frac{1}{(1 + x^2/r)^{(r+1)/2}}, \qquad -\infty < x < \infty,$$

which is a t pdf with r degrees of freedom. So if the inverse of the variance—or precision θ—of a normal distribution varies as a gamma random variable, a generalization of a t distribution has been created that has heavier tails than the normal distribution. This **mixture** of normals (different from a mixed distribution) is attained by weighing with the gamma distribution in a process often called **compounding**.

Another illustration of compounding is given in the next example.

Example 6.9-1 Suppose X has a gamma distribution with the two parameters k and θ^{-1}. (That is, the usual α is replaced by k and θ by its reciprocal.) Say $h(\theta)$ is gamma with parameters α and β, so that

$$\begin{aligned} k_1(x) &= \int_0^\infty \frac{\theta^k x^{k-1} e^{-\theta x}}{\Gamma(k)} \frac{1}{\Gamma(\alpha)\beta^\alpha} \theta^{\alpha-1} e^{-\theta/\beta}\, d\theta \\ &= \int_0^\infty \frac{x^{k-1}\theta^{k+\alpha-1} e^{-\theta(x+1/\beta)}}{\Gamma(k)\Gamma(\alpha)\beta^\alpha} d\theta \\ &= \frac{\Gamma(k+\alpha)x^{k-1}}{\Gamma(k)\Gamma(\alpha)\beta^\alpha} \frac{1}{(x+1/\beta)^{k+\alpha}} \\ &= \frac{\Gamma(k+x)\beta^k x^{k-1}}{\Gamma(k)\Gamma(\alpha)(1+\beta x)^{k+\alpha}}, \qquad 0 < x < \infty. \end{aligned}$$

Of course, this is a generalization of the F distribution, which we obtain by letting $\alpha = r_2/2$, $k = r_1/2$, and $\beta = r_1/r_2$. ■

Note how well the prior $h(\theta)$ "fits" with $f(x \,|\, \theta)$ or $f(x_1 \,|\, \theta)f(x_2 \,|\, \theta) \cdots f(x_n \,|\, \theta)$ in all of our examples, and the posterior distribution is of exactly the same form as the prior. In Example 6.8-2, both the prior and the posterior were beta. In Example 6.8-3, both the prior and posterior were normal. In Example 6.9-1, both the prior and the posterior (if we had found it) were gamma. When this type of pairing occurs, we say that that class of prior pdfs (pmfs) is a **conjugate family of priors**. Obviously, this makes the mathematics easier, and usually the parameters in the prior distribution give us enough flexibility to obtain good fits.

Example 6.9-2

(Berry, 1996) This example deals with *predictive probabilities*, and it concerns the breakage of glass panels in high-rise buildings. One such case involved 39 panels, and of the 39 panels that broke, it was known that 3 broke due to nickel sulfide (NiS) stones found in them. Loss of evidence prevented the causes of breakage of the other 36 panels from being known. So the court wanted to know whether the manufacturer of the panels or the builder was at fault for the breakage of these 36 panels.

From expert testimony, it was thought that usually about 5% breakage is caused by NiS stones. That is, if this value of p is selected from a beta distribution, we have

$$\frac{\alpha}{\alpha+\beta} = 0.05. \tag{6.9-1}$$

Moreover, the expert thought that if two panels from the same lot break and one breakage was caused by NiS stones, then, due to the pervasive nature of the manufacturing process, the probability of the second panel breaking due to NiS stones increases to about 95%. Thus, the posterior estimate of p (see Example 6.8-2) with one "success" after one trial is

$$\frac{\alpha+1}{\alpha+\beta+1} = 0.95. \tag{6.9-2}$$

Solving Equations 6.9-1 and 6.9-2 for α and β, we obtain

$$\alpha = \frac{1}{360} \qquad \text{and} \qquad \beta = \frac{19}{360}.$$

Now updating the posterior probability with 3 "successes" out of 3 trials, we obtain the posterior estimate of p:

$$\frac{\alpha+3}{\alpha+\beta+3} = \frac{1/360+3}{20/360+3} = \frac{1081}{1100} = 0.983.$$

Of course, the court that heard the case wanted to know the expert's opinion about the probability that all of the remaining 36 panels broke because of NiS stones. Using updated probabilities after the third break, then the fourth, and so on, we obtain the product

$$\left(\frac{1/360+3}{20/360+3}\right)\left(\frac{1/360+4}{20/360+4}\right)\left(\frac{1/360+5}{20/360+5}\right)\cdots\left(\frac{1/360+38}{20/360+38}\right) = 0.8664.$$

That is, the expert held that the probability that all 36 breakages were caused by NiS stones was about 87%, which is the needed value in the court's decision. ■

We now look at a situation in which we have two unknown parameters; we will use, for convenience, what is called a noninformative prior, which usually puts uniform distributions on the parameters. Let us begin with a random sample $X_1, X_2, \ldots, X_n$ from the normal distribution $N(\theta_1, \theta_2)$, and suppose we have little prior knowledge about θ_1 and θ_2. We then use the noninformative prior that θ_1 and $\ln \theta_2$ are uniform and independent; that is,

$$h_1(\theta_1)h_2(\theta_2) \propto \frac{1}{\theta_2}, \qquad -\infty < \theta_1 < \infty,\ \ 0 < \theta_2 < \infty.$$

Of course, we immediately note that we cannot find a constant c such that c/θ_2 is a joint pdf on that support. That is, this noninformative prior pdf is not a pdf at all; hence, it is called an **improper** prior. However, we use it anyway, because it will be satisfactory when multiplied by the joint pdf of $X_1, X_2, \ldots, X_n$. We have the product

$$\left(\frac{1}{\theta_2}\right)\left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \exp\left[-\sum_{i=1}^{n}\frac{(x_i-\theta_1)^2}{2\theta_2}\right].$$

Thus,

$$k_{12}(\theta_1, \theta_2 \,|\, x_1, x_2, \ldots, x_n) \propto \left(\frac{1}{\theta_2}\right)^{\frac{n}{2}+1} \exp\left[-\frac{1}{2}\left\{(n-1)s^2 + n(\bar{x}-\theta_1)^2\right\}/\theta_2\right]$$

since $\sum_{i=1}^{n}(x_i-\theta_1)^2 = (n-1)s^2 + n(\bar{x}-\theta_1)^2 = D$. It then follows that

$$k_1(\theta_1 \,|\, x_1, x_2, \ldots, x_n) \propto \int_0^\infty k_{12}(\theta_1, \theta_2 \,|\, x_1, x_2, \ldots, x_n)\, d\theta_2.$$

Changing variables by letting $z = 1/\theta_2$, we obtain

$$\begin{aligned} k_1(\theta_1 \,|\, x_1, x_2, \ldots, x_n) &\propto \int_0^\infty \frac{z^{n/2+1}}{z^2} e^{-\frac{1}{2}Dz}\, dz \\ &\propto D^{-n/2} = \left[(n-1)s^2 + n(\bar{x}-\theta_1)^2\right]^{-n/2}. \end{aligned}$$

To get this pdf in a more familiar form, let $t = (\theta_1 - \bar{x})/(s/\sqrt{n}\,)$, with Jacobian $s/\sqrt{n}$, to yield

$$k(t \,|\, x_1, x_2, \ldots, x_n) \propto \frac{1}{[1 + t^2/(n-1)]^{[(n-1)+1]/2}}, \qquad -\infty < t < \infty.$$

That is, the conditional pdf of t, given $x_1, x_2, \ldots, x_n$, is Student's t with $n-1$ degrees of freedom. Thus, a $(1-\alpha)$ **probability interval** for θ_1 is given by

$$-t_{\alpha/2} < \frac{\theta_1 - \bar{x}}{s/\sqrt{n}} < t_{\alpha/2},$$

or

$$\bar{x} - t_{\alpha/2}\, s/\sqrt{n} < \theta_1 < \bar{x} + t_{\alpha/2}\, s/\sqrt{n}.$$

The reason we get the same answer in this case is that we use a noninformative prior. Bayesians do not like to use a noninformative prior if they really know something about the parameters. For example, say they believe that the precision $1/\theta_2$ has a gamma distribution with parameters α and β instead of the noninformative prior. Then finding the conditional pdf of θ_1 becomes a much more difficult integration. However, it can be done, but we leave it to a more advanced course. (See Hogg, McKean, and Craig, 2013.)

Example 6.9-3 (Johnson and Albert, 1999) The data in this example, a sample of $n = 13$ measurements of the National Oceanographic and Atmospheric Administration (NOAA)/Environmental Protection Agency (EPA) ultraviolet (UV) index taken in Los Angeles, were collected from archival data of every Sunday in October during the years 1995–1997 in a database maintained by NOAA. The 13 UV readings are

$$7,\ 6,\ 5,\ 5,\ 3,\ 6,\ 5,\ 5,\ 3,\ 5,\ 5,\ 4,\ 4,$$

and, although they are integer values, we assume that they are taken from a $N(\mu, \sigma^2)$ distribution.

The Bayesian analysis, using a noninformative prior in the preceding discussion, implies that, with $\mu = \theta_1$,

$$\frac{\mu - 4.846}{0.317}, \qquad \text{where} \qquad \bar{x} = 4.846 \qquad \text{and} \qquad \frac{s}{\sqrt{n}} = 0.317,$$

has a posterior t distribution with $n - 1 = 12$ degrees of freedom. For example, a posterior 95% probability interval for μ is

$$(4.846 - [t_{0.025}(12)][0.317],\ 4.846 + [t_{0.025}(12)][0.317]) = (4.155,\ 5.537).$$ ■

Example 6.9-4 Tsutakawa et. al. (1985) discuss mortality rates from stomach cancer over the period 1972–1981 in males aged 45–64 in 84 cities in Missouri. Ten-year observed mortality rates in 20 of these cities are listed in Table 6.9-1, where y_i represents the number of deaths due to stomach cancer among this subpopulation in city i from 1972–1981, and n_i is the estimated size of this subpopulation in city i at the beginning of 1977 (estimated by linear interpolation from the 1970 and 1980 U.S. Census figures). Let p_i, $i = 1, 2, \ldots, 20$, represent the corresponding probabilities of death due to stomach cancer, and assume that $p_1, p_2, \ldots, p_{20}$ are taken independently from a beta distribution with parameters α and β. Then the posterior mean of p_i is

$$\widehat{p_i}\left(\frac{n_i}{n_i + \alpha + \beta}\right) + \left(\frac{\alpha}{\alpha + \beta}\right)\left(\frac{\alpha + \beta}{n_i + \alpha + \beta}\right), \quad i = 1, 2, \ldots, 20,$$

where $\widehat{p_i} = y_i/n_i$. Of course, the parameters α and β are unknown, but we have assumed that $p_1, p_2, \ldots, p_{20}$ arose from a similar distribution for these cities in Missouri; that is, we assume that our prior knowledge concerning the proportions is *exchangeable*. So it would be reasonable to estimate $\alpha/(\alpha + \beta)$, the prior mean of a proportion, with the formula

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_{20}}{n_1 + n_2 + \cdots + n_{20}} = \frac{71}{71{,}478} = 0.000993,$$

Table 6.9-1 Cancer mortality rates

y_i	n_i	$\widehat{p_i}$	Posterior Estimate	y_i	n_i	$\widehat{p_i}$	Posterior Estimate
0	1083	0	0.00073	0	855	0	0.00077
2	3461	0.00058	0.00077	0	657	0	0.00081
1	1208	0.00083	0.00095	1	1025	0.00098	0.00099
0	527	0	0.00084	2	1668	0.00120	0.00107
1	583	0.00172	0.00111	3	582	0.00515	0.00167
0	917	0	0.00076	1	857	0.00117	0.00103
1	680	0.00147	0.00108	1	917	0.00109	0.00102
54	53637	0.00101	0.00101	0	874	0	0.00077
0	395	0	0.00088	1	581	0.00172	0.00111
3	588	0.00510	0.00167	0	383	0	0.00088

for the data given in Table 6.9-1. Thus, the posterior estimate of p_i is found by *shrinking* $\widehat{p_i}$ toward the pooled estimate of the mean $\alpha/(\alpha+\beta)$—namely, $\overline{y}$. That is, the posterior estimate is

$$\widehat{p_i}\left(\frac{n_i}{n_i+\alpha+\beta}\right)+\overline{y}\left(\frac{\alpha+\beta}{n_i+\alpha+\beta}\right).$$

The only question remaining is how much weight should be given to the prior, represented by $\alpha+\beta$, relative to $n_1, n_2, \ldots, n_{20}$. Considering the sizes of the samples from the various cities, we selected $\alpha+\beta = 3000$ (which means that the prior is worth about a sample of size 3000), which resulted in the posterior probabilities given in Table 6.9-1. Note how this type of shrinkage tends to pull the posterior estimates much closer to the average, particularly those associated with small sample sizes. Baseball fans might try this type of shrinkage in predicting some of the final batting averages of the better batters about a quarter of the way through the season. ■

It is clear that difficult integration caused Bayesians great problems until very recent times, in which advances in computer methods "solved" many of these problems. As a simple illustration, suppose the pdf of a statistic Y is $f(y\,|\,\theta)$ and the prior pdf $h(\theta)$ is such that

$$k(\theta\,|\,y)=\frac{f(y\,|\,\theta)\,h(\theta)}{\int_{-\infty}^{\infty} f(y\,|\,\tau)\,h(\tau)\,d\tau}$$

is not a nice pdf with which to deal. In particular, say that we have a squared error loss and we wish to determine $E(\theta\,|\,y)$, namely,

$$\delta(y)=\frac{\int_{-\infty}^{\infty}\theta f(y\,|\,\theta)\,h(\theta)\,d\theta}{\int_{-\infty}^{\infty} f(y\,|\,\theta)\,h(\theta)\,d\theta},$$

but cannot do it easily. Let $f(y\,|\,\theta)=w(\theta)$. Then we wish to evaluate the ratio

$$\frac{E[\theta\, w(\theta)]}{E[w(\theta)]},$$

where y is given and the expected values are taken with respect to θ. To do so, we simply generate a number of θ values, say, $\theta_1, \theta_2, \ldots, \theta_m$ (where m is large), from the distribution given by $h(\theta)$. Then we estimate the numerator and denominator of the desired ratio by

$$\sum_{i=1}^{m} \frac{\theta_i\, w(\theta_i)}{m} \qquad \text{and} \qquad \sum_{i=1}^{m} \frac{w(\theta_i)}{m},$$

respectively, to obtain

$$\tau = \frac{\sum_{i=1}^{m} \theta_i\, w(\theta_i)/m}{\sum_{i=1}^{m} w(\theta_i)/m}.$$

In addition to this simple Monte Carlo procedure, there are additional ones that are extremely useful in Bayesian inferences. Two of these are the **Gibbs sampler** and the **Markov chain Monte Carlo (MCMC)**. The latter is used in **hierarchical Bayes models** in which the prior has another parameter that has its own prior (called the **hyperprior**). That is, we have

$$f(y \,|\, \theta), \qquad h(\theta \,|\, \tau), \qquad \text{and} \qquad g(\tau).$$

Hence,

$$k(\theta, \tau \,|\, y) = \frac{f(y \,|\, \theta)\, h(\theta \,|\, \tau)\, g(\tau)}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y \,|\, \eta)\, h(\eta \,|\, v)\, g(v)\, d\eta\, dv}$$

and

$$k_1(\theta \,|\, y) = \int_{-\infty}^{\infty} k(\theta, \tau \,|\, y)\, d\tau.$$

Thus, a Bayes estimator, for a squared error loss, is

$$\int_{-\infty}^{\infty} \theta\, k_1(\theta \,|\, y)\, d\theta.$$

Using the Gibbs sampler, we can generate a stream of values $(\theta_1, \tau_1), (\theta_2, \tau_2), \ldots$ that allows us to estimate $k(\theta, \tau \,|\, y)$ and $\int_{-\infty}^{\infty} \theta\, k_1(\theta \,|\, y)\, d\theta$. These procedures are the MCMC procedures. (For additional references, see Hogg, McKean, and Craig, 2013.)

Exercises

6.9-1. Let X have a Poisson distribution with parameter θ. Let θ be $\Gamma(\alpha, \beta)$. Show that the marginal pmf of X (the compound distribution) is

$$k_1(x) = \frac{\Gamma(\alpha + x)\, \beta^x}{\Gamma(\alpha)\, x!\, (1 + \beta)^{\alpha + x}}, \qquad x = 0, 1, 2, 3, \ldots,$$

which is a generalization of the negative binomial distribution.

6.9-2. Suppose X is $b(n, \theta)$ and θ is beta(α, β). Show that the marginal pdf of X (the compound distribution) is

$$k_1(x) = \frac{n!\, \Gamma(\alpha + \beta)\, \Gamma(x + \alpha)\, \Gamma(n - x + \beta)}{x!\, (n - x)!\, \Gamma(\alpha)\, \Gamma(\beta)\, \Gamma(n + \alpha + \beta)},$$

for $x = 0, 1, 2, \ldots, n$.

6.9-3. Let X have the geometric pmf $\theta(1 - \theta)^{x-1}$, $x = 1, 2, 3, \ldots$, where θ is beta with parameters α and β. Show that the compound pmf is

$$\frac{\Gamma(\alpha + \beta)\, \Gamma(\alpha + 1)\, \Gamma(\beta + x - 1)}{\Gamma(\alpha)\, \Gamma(\beta)\, \Gamma(\alpha + \beta + x)}, \qquad x = 1, 2, 3, \ldots.$$

With $\alpha = 1$, this is one form of **Zipf's law**,

$$\frac{\beta}{(\beta + x)(\beta + x - 1)}, \qquad x = 1, 2, 3, \ldots.$$

6.9-4. Let X have the pdf

$$f(x \,|\, \theta) = \theta \tau x^{\tau - 1} e^{-\theta x^{\tau}}, \qquad 0 < x < \infty,$$

where the distribution of θ is $\Gamma(\alpha, \beta)$. Find the compound distribution of X, which is called the **Burr distribution**.

6.9-5. Let $X_1, X_2, \ldots, X_n$ be a random sample from a gamma distribution with $\alpha = 1, \theta$. Let $h(\theta) \propto 1/\theta$, $0 < \theta < \infty$, be an improper noninformative prior.

(a) Find the posterior pdf of θ.

(b) Change variables by letting $z = 1/\theta$, and show that the posterior distribution of Z is $\Gamma(n, 1/y)$, where $y = \sum_{i=1}^{n} x_i$.

(c) Use $2yz$ to obtain a $(1 - \alpha)$ probability interval for z and, of course, for θ.

6.9-6. Let X_1, X_2 be a random sample from the Cauchy distribution with pdf

$$f(x \,|\, \theta_1, \theta_2) = \frac{1}{\pi} \frac{\theta_2}{\theta_2^2 + (x - \theta_1)^2},$$

$$-\infty < x < \infty, \quad -\infty < \theta_1 < \infty, \quad 0 < \theta_2 < \infty.$$

Consider the noninformative prior $h(\theta_1, \theta_2) \propto 1$ on that support. Obtain the posterior pdf (except for constants) of θ_1, θ_2 if $x_1 = 3$ and $x_2 = 7$. For estimates, find θ_1, θ_2 that maximizes this posterior pdf; that is, find the mode of that posterior. (This might require some reasonable "trial and error" or an advanced method of maximizing a function of two variables.)

HISTORICAL COMMENTS

When a statistician thinks of estimation, he or she recalls R. A. Fisher's contributions to many aspects of the subject: maximum likelihood, estimation, efficiency, and sufficiency. Of course, many more statisticians have contributed to that discipline since the 1920s. It would be an interesting exercise for the reader to go through the tables of contents of the *Journal of the American Statistical Association*, the *Annals of Statistics*, and related journals to observe how many articles are about estimation. Often our friends ask, "What is there left to do in mathematics?" University libraries are full of expanding journals of new mathematics, including statistics.

We must observe that most maximum likelihood estimators have approximate normal distributions for large sample sizes, and we give a heuristic proof of it in this chapter. These estimators are of what is called the *regular cases*—in particular, those cases in which the parameters are not in the endpoints of the support of X. Abraham de Moivre proved this theorem for $\widehat{p}$ of the binomial distribution, and Laplace and Gauss did so for $\overline{X}$ in a number of other distributions. This is the real reason the normal distribution is so important: Most estimators of parameters have approximate normal distributions, allowing us to construct confidence intervals (see Chapter 7) and perform tests (see Chapter 8) with such estimates.

The Neo-Bayesian movement in America really started with J. Savage in the 1950s. Initially, Bayesians were limited in their work because it was extremely difficult to compute certain distributions, such as the conditional one, $k(\theta \,|\, x_1, x_2, \ldots, x_n)$. However, toward the end of the 1970s, computers were becoming more useful and thus computing was much easier. In particular, the Bayesians developed Gibbs sampling and Markov chain Monte Carlo (MCMC). It is our opinion that the Bayesians will continue to expand and Bayes methods will be a major approach to statistical inferences, possibly even dominating professional applications. This is difficult for three fairly classical (non-Bayesian) statisticians (as we are) to admit, but, in all fairness, we cannot ignore the strong trend toward Bayesian methods.