CHAPTER
4

Moments and the Shape of Histograms

4.1 What You Will Learn in This Chapter

In Chapter 3, we discovered the essential role played by the shape of histograms in summarizing the properties of a statistical experiment. This chapter builds on that beginning. Pictures are well and good, but we need precise measurements of the characteristics of shape; this is the subject of this chapter. We will discover that almost all the information we might require can be captured by calculating just four numbers, called *moments*. **Moments** are merely the averages of powers of the variable values. In the process, we will refine the notions of the location, spread, symmetry, and peakedness of a histogram as measures of the characteristics of shape. We will recognize that once location and spread have been determined, it is more informative to look at standardized variables, or standardized moments, to measure the remaining shape characteristics.

4.2 Introduction

We saw in Chapter 3 that different types of experiments produce different shapes of histogram. Although the visual impression we have obtained from the histograms has been informative, transmitting this information to someone else is neither obvious nor easy to do. We need a procedure that will enable us to express shape in a precise way and not have to rely on visual impressions, useful as they have been. Our new objective is to develop expressions that will enable us to describe shape succinctly and to communicate our findings to others in an unambiguous manner.

A corollary benefit of this approach is that it will enable us to compress all the information contained in the data into a very small number of expressions. Indeed, in most cases we will be able to compress thousands of data points into only four numbers! If we succeed, then this will be a truly impressive result.

4.3 The Mean, a Measure of Location

In Chapter 3 we reduced shape to four characteristics: location, spread, peakedness, and skewness. Our easiest measure of shape is location, so let us begin with it. We already

have a measure of location, the median. But the median is not very sensitive to changes in the values of data points. This can be an advantage, but for now we are more interested in reflecting changes in shape to changes in the data. Consider the following five numbers: 1 2 3 4 5.

The center of location as indicated by the median is clearly at 3. Now consider these five numbers: 1 2 3 4 8.

The median still insists that the center of location is at 3. But if we look at the data in a different light, the center of location should be more to the right. The data are plotted on a line as shown in Figure 4.1. While 3 is the observation in the middle, we are ignoring the size of the last digit. For example, what if the last digit instead of 8 were 30? The median would still insist that the center of location is 3! We need a new measure that will overcome the fact that the median does not reflect the magnitudes of individual observations, only the number that is greater or smaller than the median. We need a measure that in some sense "balances" the smaller and the larger observations; we need to allow for a very large observation to offset many small ones.

If we look at just two numbers, an obvious measure of the center of location is halfway between; that is, the center is given by $(a + b)/2$, where a and b are any two numbers. What if we had three numbers? Might not a useful definition be $(a + b + c)/3$? Try $a = 3$, $b = 6$, and $c = 9$. Is not 6 a reasonable choice for the center of location? Try plotting these three numbers. This approach to the definition of the center of location is called the **(arithmetic) mean,** or just the mean. For any given set of data, we calculate the mean by adding up all the values and dividing that sum by the number of data points added.

Let us consider some examples from our previous work. We listed the means and medians of some of the distributions from Chapter 3 in Table 4.1. Reexamine the corresponding histograms carefully and relate the difference between the median and the mean to the shape of the histogram. Note particularly whether the mean is bigger than

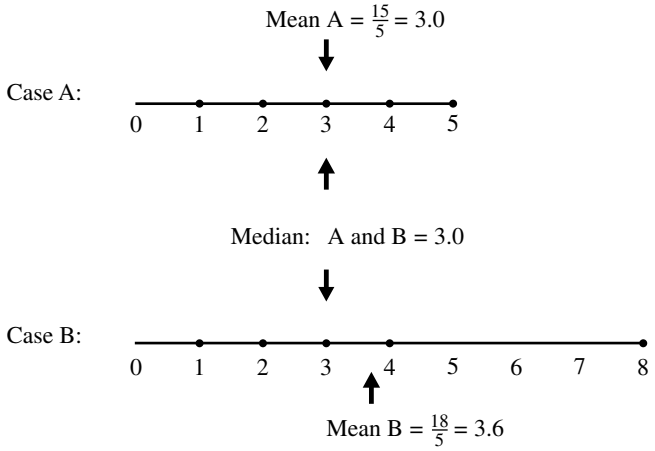

Figure 4.1 Illustrations of the difference between the median and the mean

Table 4.1 **List of Medians and Means for Chapter 3 Histograms**

Figure	Subject	Median	Mean
3.5	Tosses of an eight-sided die	5.00	4.80
3.9	Student final grades	67.00	67.30
3.11	Film revenues	15.00	26.50
3.12	Gaussian distributions	.05	.01
3.13	Weibull distributions	.95	.97
3.14	Beta (9,2) distribution	.90	.83
3.14	Arc sine distribution	.60	.49
3.14	Uniform distribution	.50	.52
3.14	Lognormal distribution	1.10	1.81

the median, or vice versa, when the histogram is asymmetric and how close they are when the histogram seems to be symmetric, as it is for the first two examples. When the distributions have long right-hand tails, for example, film revenues and the lognormal distribution, the mean is bigger than the median; the opposite is the case when there is a left-hand tail, as in the beta distribution.

An Aside on Notation

We need to be a little more formal in our statements. In the future, we will represent our observations by lowercase letters, especially from the end of the alphabet (e.g., *x, y, w, v, z*). Each lowercase letter will represent one variable, or the outcome from one experiment or one survey. To represent each observation on each variable, we will write a subscript to the variable. For example, if our data are from a variable called "wins," then its values

3, 1, 5, 8, 5, 1, 1, 0

could be represented by

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$$

or more succinctly by

$$x_i, i = 1, \ldots, 8$$

where $x_1 = 3$, $x_2 = 1$, $x_3 = 5$, and so on. Usually, the index refers to the order in which the data were recorded.

Another variable might be called y, with values indicated by y_i. The number of data points is traditionally labeled N. Sometimes we have different numbers of values for each variable, so we have to distinguish the number of variable values to add; we can do this by indexing N. N_1 or N_2 might represent 50 and 340 values of two different variables, respectively. A general statement for listing a variable's values is

$$x_i, i = 1, \ldots, N_1$$

or, for another variable

$$z_j, j = 1, \ldots, N_2$$

or, for yet another variable

$$y_k, k = 1, \ldots, N_3$$

These examples show the flexibility of the system of notation. In these examples, there are three variables, *x, z,* and *y*. They each have N_1, N_2, and N_3 observations. The indexing subscripts *i, j,* and *k* indicate the individual observations in each set of numbers.

One last notational convenience is the symbol $\sum$, which indicates that what follows is to be added. For example

$$3 + 5 + 7 + 9 + 14$$

can be represented by

$$\sum x_i$$

where $x_i, i = 1, \ldots, 5$ represents the numbers 3,5,7,9,14; so

$$\sum x_i = 38$$

the sum of the numbers 3,5,7,9,14.

Usually, the limits of the summation are clear, start at 1 and go to *N*. But where this is not so, the limits of summation will be indicated as follows:

$$\sum_{1}^{N} x_i$$

A more useful example is

$$\sum_{2}^{(N-1)} x_i$$

which indicates that only the numbers from 2 to $N - 1$ are to be added.

We can now use our new system of notation to reexpress the mean in a more succinct manner:

$$\begin{aligned} \bar{x} &= \frac{\left(\sum x_i\right)}{N} \\ &= N^{-1} \sum x_i \end{aligned} \tag{4.1}$$

where $\bar{x}$ is a symbol that represents the mean and N^{-1} means divide by *N*. Equation 4.1 represents the operation of adding up *N* numbers and then dividing that sum by *N*.

Let us try another example to illustrate our new notation:

15 10 12 3 8 21

The mean of these six numbers is 11.5. *N* takes the value 6; the sum of these six numbers is 69, or $\sum x_i = 69$, where $x_1 = 15, x_2 = 10, \ldots, x_6 = 21$.

Without doing any formal calculations, quickly estimate the means the following:

1. {4, 8}
2. {3, 7, 5}

Table 4.2 **Blood Pressure Readings of Young Drug Users 17–24 Years Old**

Blood Pressure	Cell Mark	Absolute Frequency
85–90	87.5	1
90–95	92.5	0
95–100	97.5	1
100–105	102.5	6
105–110	107.5	9
110–115	112.5	12
115–120	117.5	16
120–125	122.5	14
125–130	127.5	14
130–135	132.5	12
135–140	137.5	6
140–145	142.5	4
145–150	147.5	2
150–155	152.5	1
	Total	98

3. {1, 3, 20}
4. {−12, 2}

Draw a line on a scrap of paper and place the numbers that you are averaging and your estimate of the average on it to visualize the process. Add the median and the formal calculation of the mean, and compare the results.

Averaging Grouped Data

These calculations are easy enough, but what if the data we have are like those in Table 4.2? Here the data are in terms of absolute frequencies; we do not have a set of values to add up and divide by the number of entries. We have lost information because we do not have the original data that went into making the tables. But we would still like to discover the center of location. Consider for example the fourth cell in Table 4.2, the one that lies between the boundaries 100 and 105. Observe that there are 6 data points, and the class (cell) mark has the value 102.5. When we created cells for continuous data in Chapter 3, we chose the cell mark as the midpoint of the cell because that point was the best choice to represent the cell. With 6 data points in a cell, the value of the cell mark repeated 6 times represents the 6 unknown values in this cell. So it is with all the other cells. The cell mark represents each of the unknown entries in that cell; the number of representations is given by the absolute frequency in that cell.

Consequently, we may approximate the value of the actual mean that would be given by $N^{-1} \sum x_i$, if we had the actual entries, x_i, by the following approach:

1. In each cell, multiply the absolute frequency by the cell mark to get an approximation to the true sum in that cell.
2. Add up the values obtained over all cells.
3. Divide the final total by N, the total number of observations in the data set.

Table 4.3 **Cumulative Frequencies for the Data in Table 4.2**

C_j	F_j	$F_j\,C_j$
87.5	1	87.5
92.5	0	0.0
97.5	1	97.5
102.5	6	615.0
107.5	9	967.5
112.5	12	1350.0
117.5	16	1880.0
122.5	14	1715.0
127.5	14	1785.0
132.5	12	1590.0
137.5	6	825.0
142.5	4	570.0
147.5	2	295.0
152.5	1	152.5
Total	98	11,930.0

This somewhat lengthy expression is very simple as will become clear once we re-express it as follows:

$$\bar{x} \approx N^{-1} \sum_{1}^{k} F_j C_j$$

where "≈" means that the left-hand side of the expression is only approximated by the expression on the right-hand side; k is the number of cells into which the data are placed; F_j is the absolute frequency in the jth cell, $j = 1, \ldots k$; and C_j is the cell mark for the jth cell. In Table 4.2, showing blood pressure for young drug users, there are 14 cells, $k = 14$; the total number of observations, N, is 98.

Let us list F_j, C_j, and the products F_jC_j, $j = 1, \ldots 14$ in Table 4.3:

$$N = \sum F_j = 98; \sum F_j C_j = 11{,}930$$

We conclude that the mean of the blood pressure data is approximately $11{,}930/98 = 121.7$; that is, for these data, using y to represent the blood pressure readings, we obtain:

$$\bar{y} \approx \frac{\left(\sum F_j C_j\right)}{N}$$
$$= 121.7$$

Recall that f_j, the relative frequency, is given by F_j/N; so the preceding expression can be rewritten as

$$\bar{y} \approx \sum_{j=1}^{k} f_j C_j = 121.7$$

Let's consider another example. Look at the data in Table 4.4, which shows household income in 1979. These data have 10 cells, so $k = 10$. Let w represent the variable "household income" and $\bar{w}$ its mean. The total number of observations

Table 4.4 **Household Income in 1979 for the United States**

Income in 1979	Cell Mark	Number of Households (millions)	Percent of Total
	C_i	F_i	$100 \cdot f_i$
Less Than \$ 7,500	\$ 3,750	17.1	21.2
\$ 7,500–\$ 14,999	\$ 11,250	18.7	23.2
\$ 15,000–\$ 19,999	\$ 17,500	11.4	14.2
\$ 20,000–\$ 24,999	\$ 22,500	10.0	12.4
\$ 25,000–\$ 29,999	\$ 27,500	7.4	9.2
\$ 30,000–\$ 34,999	\$ 32,500	5.2	6.5
\$ 35,000–\$ 39,999	\$ 37,500	3.4	4.2
\$ 40,000–\$ 49,999	\$ 45,500	3.6	4.5
\$ 50,000–\$ 74,999	\$ 62,500	2.6	3.2
\$ 75,000–\$149,999	\$112,500	1.1	1.4

is 80.5 million households, so $N = 80.5$ million. The values of $f_j = F_j/N$ and C_j are listed in Table 4.5; $\bar{w} \approx \sum f_j C_j = \$20{,}520.0$.

It is important to remember that these last expressions for the mean in terms of the frequencies, relative or absolute, are only approximations to the true value of the mean that would be obtained by $N^{-1} \sum x_i$ if we had the orginal data used to make the frequencies. These original data are usually referred to as *raw,* as in "uncooked," data, and the data that are in cells are called *grouped* data.

Interpreting the Mean

We now have another possible answer to our question about which film studio we should invest in—that with the largest mean, see Table 4.6. Studio P has the largest mean with a value of \$38.3 million, and the next largest is studio W with a revenue of \$30 million. In contrast, the largest median was for studio B with a value of \$23 million, and the next largest was for studio T with a value of \$22.1 million. Reexamine the box-and-whisker plots in Figure 3.3 to put these results into perspective. The mean values are very sensitive to the large values in the "tails of the distributions." Based on

Table 4.5 **Household Income in the United States in 1979**

$f_1 = .212$	$C_1 = \$3{,}750$	$f_1 \cdot C_1 = \$795.0$
$f_2 = .232$	$C_2 = 11{,}250$	$f_2 \cdot C_2 = 2{,}610.0$
$f_3 = .142$	$C_3 = 17{,}500$	$f_3 \cdot C_3 = 2{,}485.0$
$f_4 = .124$	$C_4 = 22{,}500$	$f_4 \cdot C_4 = 2{,}790.0$
$f_5 = .092$	$C_5 = 27{,}500$	$f_5 \cdot C_5 = 2{,}530.0$
$f_6 = .065$	$C_6 = 32{,}500$	$f_6 \cdot C_6 = 2{,}112.5$
$f_7 = .042$	$C_7 = 37{,}500$	$f_7 \cdot C_7 = 1{,}575.0$
$f_8 = .045$	$C_8 = 45{,}500$	$f_8 \cdot C_8 = 2{,}047.5$
$f_9 = .032$	$C_9 = 62{,}500$	$f_9 \cdot C_9 = 2{,}000.0$
$f_{10} = .014$	$C_{10} = 112{,}500$	$f_{10} \cdot C_{10} = 1{,}575.0$
$\sum f_j = 1.000$		$\sum f_j C_j = \$20{,}520.0$

Table 4.6 **Mean and Median Revenues ($millions) for Film Studios**

Film Studio	Mean	Median
B	28.0	23.0
C	25.5	12.9
F	21.2	12.1
M	17.2	7.3
O	24.8	16.7
P	38.3	18.1
T	27.1	22.1
U	25.3	12.8
W	30.0	20.1
All	26.5	15.0

a comparison of the means, you might be tempted to choose studio P as your best choice.

But is this really the best choice? What if the two histograms looked like those shown in Figure 4.2, which shows two smoothed histograms—one with a larger mean and a larger spread. (For the moment, pay attention only to the shape of the histograms; the notations m_1 and m_2 will be explained in the next section.) With this larger spread, you can get larger revenues and a larger mean revenue, but you can also get smaller revenues. The choice is now not as clear as it seemed to be. Alternatively, we could think about a case in which one distribution has a larger mean but a smaller degree of spread than the other. Which is better in this case, and why?

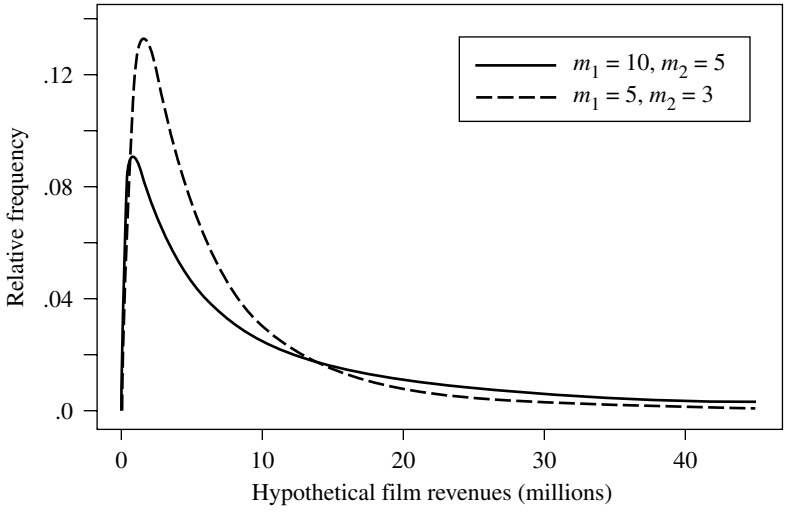

Figure 4.2 Comparison of two hypothetical smoothed histograms

However we choose to answer this question, we will need a measure of spread that is sensitive to the variations in the data about the mean. Neither the range nor the interquartile range is sensitive to variations in the data except for the extremes and for designating the percentage of terms that lie in the interquartile range.

4.4 The Second Moment as a Measure of Spread

As long as we do not change the value of the largest and the smallest observations, the value of the range stays the same. It is not very sensitive to changes in the data. For our purposes, however, it is important that we have information on values that are large, but still less than the largest. We will need more than one very large revenue producing film! So we need a measure of spread that will be sensitive to the value of each observation. Our examination of Figures 3.3 and 4.2 showed that it was the spread around the measure of location that seemed to be important, so it would appear to be sensible to measure spread about the mean.

Let us try the average difference between the data points and the mean, which we define by

$$m_1 = \frac{\sum(x_i - \bar{x})}{N} \tag{4.2}$$

In this expression we are adding up the differences between the x_i and their means and dividing by the number of additions. What would we get with actual data? Apply the expression to the film revenue data for studio B shown in Table 3.3. The mean for studio B is 28.02353. The 17 values for the differences between the observed revenues and the mean are

$$-19.72, 34.08, -4.42, -5.02, 2.98, 3.28, -13.82, -14.82, -17.42, -18.92, -23.92, 43.18, 32.18, 24.18, -4.42, -7.12, -10.22$$

The sum of the differences is -0.00001, not exactly equal to zero because of rounding. Now try another data set, say studio C. The result is again approximately zero.

This does not seem to be a very useful measure for the spread of a histogram. But why? Rewrite Equation 4.2 as follows:

$$\begin{aligned} m_1 &= \frac{\sum(x_i - \bar{x})}{N} \\ &= \frac{\left[\left(\sum x_i\right) - \left(\sum \bar{x}\right)\right]}{N} \\ &= \frac{\left[\left(\sum x_i\right) - \left(\sum x_i\right)\right]}{N} \\ &= 0 \end{aligned} \tag{4.3}$$

Remember that in these expressions we are adding N terms, so that $\sum \bar{x}$ is merely N lots of $\bar{x}$, which in turn is simply:

$$N\bar{x} = N\frac{\sum x_i}{N} = \sum x_i$$

We see from this expression that m_1, the average sum of the differences, is identically zero; that is, it will be zero for any set of data.

Adding differences obviously does not work. The problem is that the positive differences above the mean offset the negative differences below the mean. This result is another way of saying that the mean is a "center of location." We could have defined the mean as that value such that the differences between the variable values and the mean add up to zero.

We need to consider an alternative approach. If we square the differences and then get the average squared difference, then we will not get zero identically—that is, zero for all possible variables. (There is one very special case for which the sum of squared "differences" is zero; all the values are the same.)

Let us write down the general expression:

$$m_2 = \frac{\sum (x_i - \bar{x})^2}{N} \tag{4.4}$$

where x_i represents the values of the variable, $\bar{x}$ represents the mean, and the whole expression means that we are "averaging" the squared differences between the x_i and $\bar{x}$ the mean.

Remember that we are using the word *averaging* to mean the simple operation of adding up all the terms and then dividing by the number of additions.

Let us try this operation on a few simple numbers. Consider:

$$\{1, 2, 3, 4\}$$

$$\bar{x} = 2.5$$

which you should verify for yourself. The individual differences are

$$\begin{aligned} \{(x_i - \bar{x})\} &= \{(1 - 2.5), (2 - 2.5), (3 - 2.5), (4 - 2.5)\} \\ &= \{-1.5, -0.5, 0.5, 1.5\} \\ \{(x_i - \bar{x})^2\} &= \{2.25, 0.25, 0.25, 2.25\} \\ \sum (x_i - \bar{x})^2 &= 5.0 \\ \frac{\sum (x_i - \bar{x})^2}{N} &= 1.25 \end{aligned}$$

because $N = 4$ in this example.

Let us carry out these calculations on a few examples from Chapter 3. The results are listed in Table 4.7 in the column under the heading "m_2." Our new measure of spread is called the **second moment**, and we give it the symbol "m_2." It is the average squared deviation of the variable's values from the mean; m_2 is obtained from a set of data using Equation 4.4.

We now have a measure of spread that is sensitive to the values taken by all the variables. Suppose that the second of the four previously listed values is 2.1 instead of 2.0; the mean is now 2.525 instead of 2.5. The squared measure of spread, the second moment, is now 1.227 not 1.25. A 5% change in one of the four variable values changes the total by only 1%, the mean by 1%, and the second moment by 1.8%.

Table 4.7 **Lists of Means and Second Moments for Chapter 3 Histograms**

Figure	Subject	Mean	m_2	$\sqrt{m_2}$
3.5	Tosses of eight-sided die	4.80	6.00	2.50
3.9	Student final grades	67.30	132.00	11.50
3.11	Film revenues	$26.50	$1050.50	$32.41
3.12	Gaussian	0.01	0.84	0.92
3.13	Weibull	0.97	0.02	0.14
3.14	Beta (9,2)	0.83	0.01	0.11
3.14	Arc sine	0.49	0.12	0.34
3.14	Uniform	0.52	0.09	0.30
3.14	Lognormal	1.81	6.66	2.60

There is one minor problem with using the second moment; the units are squared. For example, if we are observing film revenues in millions of dollars, the second moment will be in the units of millions of dollars squared. Or if we are observing household incomes in thousands of dollars, the second moment will be in thousands of dollars squared. If nothing else, these are large numbers. This is inconvenient, especially as we want to use the measure of spread to put the value of the mean into context. The way out is straightforward; take the square root of the second moment to get a measure of spread that has the correct units of measurement. The change in the square root of the second moment resulting from the 5% change in one of the data entries is now only 0.9%.The relationship of the square root of the second moment to the size of the mean is illustrated in Table 4.7. We need a name for the square root of the second moment, which is a mouthful for anyone. Let us define the **standard deviation**, albeit temporarily, as the square root of the second moment. We will have many occasions to use this term.

You may recall that when we only had frequencies we were able to approximate the mean by the expression:

$$\frac{\sum_{1}^{k} F_j C_j}{N} = \sum f_j C_j$$

where there are k cells of data with a total of N observations.

We can also approximate the second moment in a similar manner. With N observations on a variable x in k cells of data, we can approximate

$$m_2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

by

$$\begin{aligned} m_2 &\cong \frac{\left[\sum_{1}^{k} F_j (C_j - \bar{x})^2\right]}{N} \\ &= \sum_{1}^{k} f_j (C_j - \bar{x})^2 \end{aligned} \tag{4.5}$$

Table 4.8 **Entries from Table 4.2**

C_j	$(C_j - \bar{y})^2$	F_j	$F_j(Cj - \bar{y})^2$
87.5	$(-34.2)^2$	1	1169.64
92.5	$(-29.2)^2$	0	0.00
97.5	$(-24.2)^2$	1	585.64
102.5	$(-19.2)^2$	6	2211.84
107.5	$(-14.2)^2$	9	1814.76
112.5	$(-9.2)^2$	12	1015.68
117.5	$(-4.2)^2$	16	282.24
122.5	$(0.8)^2$	14	8.96
127.5	$(5.8)^2$	14	470.96
132.5	$(10.8)^2$	12	1399.68
137.5	$(15.8)^2$	6	1497.84
142.5	$(20.8)^2$	4	1730.56
147.5	$(25.8)^2$	2	1331.68
152.5	$(30.8)^2$	1	948.64
Total		98	12,969.88

where F_j is the absolute frequency in cell j, f_j is the relative frequency in cell j, f_j is F_j/N, and C_j is the class mark in cell j.

F_j represents the number of occurrences in cell j. C_j represents "the value taken" by the variable in cell j. Because we are looking at squared differences, we square the difference between C_j, the representative value, and $\bar{x}$, the mean.

An example of the second moment using the blood pressure data listed appears in Table 4.8. The values for C_j and F_j shown in Table 4.8 are the same as we used for calculating the approximation to the mean, which we calculated as $\bar{y} = 121.7$.

The sum of the weighted squares, $\sum F_j(C_j - \bar{y})^2$, is 12,969.88 and $\sum F_j$ is 98. So for these grouped data the approximate value for the second moment is 132.35; that is, $12{,}969.88/98 = 132.35$. The value for the square root of the second moment is 11.5.

4.5 General Definition of Moments

In working these examples you have noticed a similarity in the procedures for finding the **mean**, m_1', and our measure of spread, m_2. In both cases we "averaged"; that is, we added something up and divided by the number of additions. This suggests an immediate generalization. If we can average the x_i and if we can average the squared differences, we can average any power of the data! This insight suggests a clever way of describing all these averages in a way that these similarities are stressed, so that we can take advantage of knowing one general procedure—"one expression fits all."

However, there is one difference between the mean and our measure of spread; for the mean we merely averaged, for the spread we averaged squared differences from the mean. With this in mind, let us define **moments**.

The **first moment about the origin** is the mean. The phrase "about the origin" merely means that we are looking at differences between the x_i and the origin, zero; that is, $x_i - 0$ is nothing more than x_i. We now generalize the idea:

$$\begin{aligned} m'_1 &= \frac{\left(\sum x_i\right)}{N} \\ m'_2 &= \frac{\left(\sum x_i^2\right)}{N} \\ m'_3 &= \frac{\left(\sum x_i^3\right)}{N} \\ m'_4 &= \frac{\left(\sum x_i^4\right)}{N} \end{aligned} \tag{4.6}$$

and so on. The symbol m'_1 is called the first moment (about the origin) and is nothing more than our old friend the mean, $\bar{x}$; m'_2 is called the second moment about the origin; and m'_3 is the third moment about the origin.

Our measure of spread is also a moment, but this is a moment about the mean as we showed in Equation 4.4. We can generalize this idea too. Consider:

$$\begin{aligned} m_1 &= \frac{\sum(x_i - \bar{x})}{N} \\ m_2 &= \frac{\sum(x_i - \bar{x})^2}{N} \\ m_3 &= \frac{\sum(x_i - \bar{x})^3}{N} \\ m_4 &= \frac{\sum(x_i - \bar{x})^4}{N} \end{aligned} \tag{4.7}$$

and so on. These moments are called **moments about the mean**. They are just the averaged values of the powers of the differences of the x_i from the mean. Compare the expressions for the two sets of moments carefully: m'_1 is the first moment about the origin and is simply the mean; m_1 is the first moment about the mean, and as we saw it is identically zero. We also saw that m_1 can be used as the definition of the mean. The symbol m_2, or its square root, $\sqrt{m_2}$, is our new measure of spread; we now see that m_2 is also called "the second moment about the mean." The symbol m_3 is the third moment about the mean, and so on.

So far we have seen a use for m'_1, the first moment about the origin, or the mean, and for m_2, or its square root, $\sqrt{m_2}$, the second moment about the mean. We will now see if the other moments will prove to be of any help. But do we use m'_r or m_r, $r = 1, 2, 3, \ldots$? Do we take moments about the origin, or about the mean?

One thought is that given that we have already discovered the location of the data, we are no longer interested in the first moment, so we can ignore it in our further examination of the data. In the hope that this will prove to be a good idea, let us proceed

by looking only at m_r, the rth moment about the mean, where $m_r = N^{-1} \sum (x_i - \bar{x})^r$; that is, we look at averaged powers of differences of the raw data from the mean. We have in fact "subtracted out" the effect of the mean.

If you recall, we decided that there are four important indicators of the shape of a histogram: location, spread, skewness, and peakedness. Symmetry is the absence of skewness, and flatness is the absence of peakedness. We already have measures for the first two, location and spread; the measures are m_1', the mean, and m_2, the second moment about the mean. Let us now have a look at the property of skewness.

Before we investigate the third and fourth moments in detail, let us consider a potential application of the higher moments. You may be aware, especially over the past decade, that there is considerable controversy in the United States about "income inequality." This discussion is definitely about the shape of the income distribution. Although there has also been some concern about the overall level of the income distribution, the main issue is whether the richest quintile has improved relative to the lowest quintile; a quintile is one-fifth of a distribution. The mean can tell us about the level of the overall distribution, and we can ask whether the mean of the income distribution did, or did not, increase over the past decade. We might also examine the second sample moment about the mean for a measure of the spread in income and ask whether it has changed over the last decade. But neither of these measures gets at the heart of the real question. It has been well known for a very long time that income distributions are highly skewed to the right, as is shown for example in the lognormal distribution (Figure 3.14). One important question is whether the distribution of income has become even more skewed to the right, or is it less skewed? A related question is whether the amount of the distribution in the tails of the distribution has changed, even if the measure of spread may not have changed very much. All these questions are about the *shape* of the distribution of income, not about its location or spread. To answer these questions we need new measures of shape, and to these topics we now turn.

The Third Moment as a Measure of Skewness

Reexamine Figures 3.12 to 3.14. Figure 3.12, at least for the large sample sizes, and the uniform and arc sine distributions in Figure 3.14 represent symmetric histograms. The beta distribution in Figure 3.14 represents a histogram skewed to the left; that is, it has a tail to the left. Figure 3.13 and the lognormal distribution in Figure 3.14 represent histograms that are skewed to the right; that is, they have tails to the right. Left-skewed distributions have observations that extend much farther to the left of the mean than the corresponding observations to the right of the mean. The opposite is the case for right-skewed distributions. With symmetric distributions the two sides balance.

We have discovered that the mean and the median are approximately equal when the distribution appears to be symmetric. We also saw that for distributions that are skewed to the right, the mean is bigger than the median and that for distributions skewed to the left the mean is less than the median. This observation does provide some idea of the sign of the skewness, positive or negative, right or left, but it is not a useful measure of the extent of the skewness. We need a measure that will be sensitive to all the observations.

Table 4.9 **The Third and Fourth Moments about the Mean for Chapter 3 Histograms**

Figure	Subject	m_3	m_4
3.5	Tosses of eight-sided die	−2.700	62.900
3.9	Student final grades	−555.100	57,366.000
3.11	Film revenues	101,008.000	15,404,336.000
3.12	Gaussian	0.070	1.800
3.13	Weibull	−0.002	0.001
3.14	Beta (9,2)	−0.001	0.001
3.14	Arc sine	0.002	0.022
3.14	Uniform	0.006	0.013
3.14	Lognormal	91.500	1787.000

We should also require of our measure of skewness that symmetric histograms have a measure of skewness that is zero, left-tailed histograms a negative measure, and right-tailed histograms a positive measure to reflect our observations about skewed histograms. These remarks lead to the following idea: Third powers of differences between x_i and $\bar{x}$ will certainly be negative for the x_i below the mean and positive for the x_i above the mean, and it is a good bet that the sum of third powers will be zero for symmetric histograms. This means that we should look at m_3.

The only way to find out is to try our new measure on some simple examples. Consider:

$$x_i = -2, -1, 0, 1, 2$$
$$y_i = 0, 1, 2, 3, 14$$
$$w_i = -12, -1, 0, 1, 2$$

The means of *x, y,* and *w* are, in turn: 0, 4, and –2. The second moments are, in turn: 2, 26, and 26. Now *x, y,* and *w* are respectively symmetric, skewed to the right, and skewed to the left. You can verify this by plotting each of the five points on a line. The cubed differences are

$$\begin{array}{lccccc} x_i: & (-2-0)^3, & (-1-0)^3, & (0-0)^3, & (1-0)^3, & (2-0)^3 \\ & -8, & -1, & 0, & 1, & 8 \\ y_i: & (0-4)^3, & (1-4)^3, & (2-4)^3, & (3-4)^3, & (14-4)^3 \\ & -64, & -27, & -8, & -1, & 1000 \\ {}_i: & (-12+2)^3, & (-1+2)^3, & (0+2)^3, & (1+2)^3, & (2+2)^3 \\ w_i & -1000, & 1, & 8, & 27, & 64 \end{array}$$

The sum of the cubed differences are for *x, y,* and *w*: 0, 900, and −900, respectively. Dividing by five in each case gives us the third moment. We get for *x, y,* and *w* the values 0, 180, and −180, respectively. At least the third moment gets the signs correct; that is, the third moment is zero for symmetric distributions, negative for left-tailed distributions, and positive for right-tailed distributions.

Using our new tool, let us examine once again the data from some Chapter 3 figures. The results are presented in the column headed "m_3" in Table 4.9. The signs seem to be correct, except for the Weibull, which is a small puzzle. The values for Figures 3.11 to 3.14, except for the film revenues and the lognormal, seem to be essentially zero. But

one obvious fact should impress you; the numbers go from very small to huge—this looks like a problem we will have to address.

We now have three useful measures of shape: m_1' for location, m_2 for spread, and m_3 for skewness. If m_3 is negative, then the histogram has a left-hand tail; if positive, it has a right-hand tail; and if zero, it is symmetric about the mean. However, the units of measurement for the third moment, m_3, are in terms of the units for the original data cubed; this may present a problem.

The Fourth Moment as a Measure of Peakedness, or "Fat Tails"

Our last shape characteristic is peakedness. An alternative, but less elegant, description of what is measured by the fourth moment is "fat tails." A single-peaked histogram with a lot of distributional weight in the center and in the tails, but not in the shoulders is said to have "fat tails"; or it is "highly peaked." Recall that if one area of a histogram has more weight, somewhere else must have less, because the relative frequencies must sum to one.

Our approach using moments has been successful, so let us continue it. We will reexamine the same data by calculating this time m_4. Let us consider three simple examples:

x_i : $-2, -1, 0, 1, 2$
y_i : $-2, 0, 0, 0, 2$
w_i : $-2, -2, 0, 2, 2$

Once again, plot these three sets of five numbers on a piece of paper to get some idea of the shape of these very simple distributions.

The means are zero in each case. The second moments are 2, 1.6, and 3.2. The fourth powers of the differences are

x_i : $-2^4, -1^4, 0^4, 1^4, 2^4$
$16, 1, 0, 1, 16$
y_i : $-2^4, 0, 0, 0, 2^4$
$16, 0, 0, 0, 16$
w_i : $-2^4, -2^4, 0, 2^4, 2^4$
$16, 16, 0, 16, 16$

These differences raised to the fourth power and averaged for *x, y,* and *w* give 6.8, 6.4, and 12.8. But is 12.8 a large number? Is 6.8 a small number? At the moment we cannot tell.

Perhaps, if we examined more realistic examples of distributions we would be able to get a better idea. Let us list our figures from Chapter 3 in the rough order from flattest to most peaked. This gives us the order: arc sine, whose U-shape is actually "anti-peaked"; the uniform; Figures 3.9 and 3.12, the beta distribution; film revenues; the Weibull; and the lognormal. Looking carefully at these figures we see that the most peaked histograms also tend to have fat tails. This is because the total area under a histogram is one, so that if the middle of the histogram is characterized by a narrow peak (high frequency), the tails must be thin (low frequency) and spread out. A peaked histogram, relative to a flat histogram, has a lot of observations near the mean, and a few observations very far away from the mean. Note that a U-shaped distribution has very few observations near the mean, and most of them in the tails.

Table 4.10 **Comparison of Moments in Different Units of Measurement**

Moment	Height (in.)	Height (ft)
m_1'	65.37 in.	5.450 ft
m_2	11.24 in.^{2}	0.078 ft^2
m_3	−3.83 in.^{3}	−0.002 ft^3
m_4	283.20 in.^{4}	0.014 ft^4

If we raise the differences between the x_i and $\bar{x}$ to a high power, the "big" numbers will have a much bigger effect on the sum of the powers than the more numerous observations in the middle that produce small differences. When we calculate m_4, and if our intuition has been reliable, we should see the values of m_4 increase as we examine the distributions in the order we specified.

The results of these calculations are in Table 4.9. Unfortunately, the results do not match our expectations. The fourth moments of the distributions that we thought would be the smallest are not so, nor are the fourth moments we thought would be the largest so. We need to go back to the drawing board.

4.6 Standardized Moments

You may have noticed another problem with the third and fourth moments: what's large? It is not very helpful to say that m_4 is large for highly peaked distributions if we do not know what constitutes "large." Our cavalier attitude needs reassessment.

Even for m_3 we have a problem. If we want to compare two histograms that are both skewed to the left, for example, can we unambiguously say that one of them is more skewed than the other on the basis of the values that we obtained for m_3? Yet another problem is raised if we try the next experiment.

In Table 4.10, we recorded the first four moments on the heights of enrollees in a fitness class measured in inches. We converted these numbers into feet. We are essentially dealing with the same histogram of heights, so our conversion to feet was merely a re-scaling of the data; we should conclude that we have the same shape of histogram.

The mean is easy enough to understand; m_1' in feet is just one-twelfth of m_1' in inches. But the second moment is not so easy to interpret, and except for signs, m_3 and m_4 appear to be completely different between the two measurements. This is not desirable at all; our results should not depend on our choice of the units of measurement if we are to create a generally useful measure of shape.

Now suppose that we measure heights not from zero as we have done so far, but from 60 inches; the idea is that we are interested mainly in the deviations of actual heights from 5 feet. If we now recalculate all our moments, we get

$$m_1' = \quad 5.37 \text{ in.}$$
$$m_2 = \quad 11.24 \text{ in.}^2$$
$$m_3 = -3.83 \text{ in.}^3$$
$$m_4 = 283.20 \text{ in.}^4$$

Comparing the means we get what we might have expected: m'_1 for the data after subtracting 60 inches is just our original mean value of 65.37 inches less 60 inches. But what may be surprising is that m_2, m_3, m_4 are all exactly the same as the moments we obtained from the original data!

A moment's reflection may help you to see why this is so. Look at the definition of m_2, for example:

$$m_2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

Define a new variable y_i by

$$y_i = x_i - 60$$

because we subtracted 60 from the height data in inches.

To emphasize the variable whose moment is being taken, let us change notation slightly to $m_r(y)$, $m_r(x)$, or $m_r(w)$ to represent the rth moment about the mean for the variables y, x, and w respectively.

Now calculate the mean for the new variable y:

$$\begin{aligned} m'_1(y) &= m'_1(x_i - 60) \\ &= m'_1(x_i) - m'_1(60) \\ &= 65.37 - 60 = 5.37 \end{aligned}$$

If you have any trouble following these steps, a quick look at Appendix A will soon solve your difficulty.

Now we calculate $m_2(y)$.

$$\begin{aligned} m_2(y) &= \frac{\sum (y_i - \bar{y})^2}{N} \\ &= \frac{\sum \left[(x_i - 60) - (m'_1(x) - 60) \right]^2}{N} \end{aligned}$$

where we have substituted

$$x_i - 60 \text{ for } y_i; \; m'_1(x) - 60 \text{ for } m'_1(y)$$

but,

$$\begin{aligned} &\frac{\sum \left[(x_i - 60) - (m'_1(x) - 60) \right]^2}{N} \\ &= \frac{\sum (x_i - \bar{x})^2}{N} \end{aligned}$$

because the two "60s" cancel in the previous line: $-60 - (-60) = -60 + 60 = 0$. We have shown that the second moments for the variables x and y, where y is given by $y = x - 60$, are the same.

This result will hold for all our moments about the mean. This property is called *invariance;* the moments $m_2, m_3, m_4 \ldots$ are all said to be **invariant** to changes in the

origin. This means that you can add or subtract any value whatsoever from the variable and the calculated value of all the moments about the mean will be unchanged.

The qualification "about the mean" is crucial; this is not true for the moments about the origin.

We have solved one problem; we now know that, beyond the first moment, we need to look at moments that are about the mean and so invariant to changes in the origin. But the problem of interpreting the moments, detected with the heights, goes further than a change of origin. When we changed from inches to feet, we did not change the origin, it is still zero, but we did change the scale of the variable. But "scale" is like spread; a bigger scale will produce a larger spread and consequently a larger value for m_2, our measure of spread. Further, our measure of spread is squared. Reconsider the expression for m_2:

$$m_2(x) = \frac{\sum(x_i - \bar{x})^2}{N}$$

and consider changing from measuring x_i in inches to a variable y_i that is height measured in feet. We do this by dividing the x_i entries by 12:

$$\begin{aligned}\frac{\sum(y_i - \bar{y})^2}{N} &= \frac{\sum\left(\frac{x_i}{12} - \frac{\bar{x}}{12}\right)^2}{N} \\ &= \frac{\left[\frac{\sum(x_i - \bar{x})^2}{N}\right]}{144}\end{aligned}$$

where $144 = 12^2$. Or more generally, if $y_i = b \times x_i$, for any constant b, then

$$m_2(y) = b^2 \times m_2(x)$$

Even more generally, if $y_i = a + (b \times x_i)$, for any constants a and b,

$$m_2(y) = b^2 \times m_2(x)$$

If we now apply this same approach to the higher moments, we will discover that, whenever

$$\begin{aligned}y_i &= a + (b \times x_i) \\ m_2(y) &= b^2 \times m_2(x) \\ m_3(y) &= b^3 \times m_3(x) \\ m_4(y) &= b^4 \times m_4(x)\end{aligned}$$

and so on.

This business of changing variables by adding constants and multiplying by constants may still seem a little mysterious, but a familiar example will help. You know that temperature can be measured either in degrees Fahrenheit or in degrees Celsius. The temperature of your sick sister is the same whatever units of measurement you use; that is, how much fever your sister has is a given, but how you measure that temperature, or how you record it, does depend on your choice of measuring instrument.

To reexpress the idea, your choice of the units of measurement alters the measurement, but it clearly does not alter the degree of fever.

One choice is to measure in degrees Fahrenheit. Suppose that you observe a measure of 102 degrees Fahrenheit. That observation is equivalent to a measured temperature of 38.9 degrees Celsius. As you may remember from your high school physics, degrees Fahrenheit are related to degrees Celsius by

$$\text{deg.F} = 32 \text{ deg.F} + \left(\frac{9}{5}\right) \times \text{deg.C}$$

but this is just like

$$\text{deg.F} = a + b \times \text{deg.C}$$

where $a = 32$ deg.F and $b = \frac{9}{5}$.

The origin for degrees Fahrenheit, relative to degrees Celsius, is 32°F and the scale adjustment is to multiply by $\frac{9}{5}$. So, if you were interested in the shape of histograms of temperatures, you would not want your results to depend on how you measured the data [except, of course, for measures of location (that depend directly on origin and scale) and measures of scale or spread that should depend only on scale, not on the choice of origin.]

We can reexpress our results so far by saying that m_1' indicates the chosen origin for the units of measurement and that $\sqrt{m_2}$ indicates the chosen scale of measurement; "$\sqrt{\ }$" is the traditional square root sign and means take the square root of its argument.

We now have the answer to our problem of trying to decide whether a third or a fourth moment is large or small. We also have our answer to the question of how to ensure that our measures of shape do not depend on the way in which we have measured the data. If we divide the third and fourth moments about the mean by the appropriate power of m_2, we will have a set of moments that will be invariant to changes in scale. This is equivalent to picking an arbitrary value for b in the equation $y_i = a + (b \times x_i)$.

We conclude that to discuss shape beyond location and scale in unambiguous terms, one has to measure shape in terms of expressions that are invariant to changes in origin or in scale. As we saw, any change in scale in the variable x is raised to the power 3 in the third moment and to the power 4 in the fourth moment. Thus an easy way to overcome the effects of scale is to divide m_3 by $(m_2)^{3/2}$ and m_4 by $(m_2)^2$.

Define $\hat{\alpha}_1$ and $\hat{\alpha}_2$, the **standardized** third and fourth **moments**, by

$$\begin{aligned} \hat{\alpha}_1 &= \frac{m_3}{(m_2)^{3/2}} \\ \hat{\alpha}_2 &= \frac{m_4}{(m_2)^2} \end{aligned} \tag{4.8}$$

Our rationale for picking this "peculiar" notation, $\hat{\alpha}_1$, $\hat{\alpha}_2$, will be apparent in a few chapters. For the moment we need labels and $\hat{\alpha}_1$, $\hat{\alpha}_2$ are as good as any.

Consider any arbitrary change of origin and scale of any variable; that is, if x_i is the variable of interest, look at $y_i = a + (b \times x_i)$, for any values of a and b,

$$\begin{aligned}\hat{\alpha}_1(y) &= \frac{m_3(y)}{(m_2(y))^{3/2}} \\ &= \frac{b^3 m_3(x)}{(b^2 m_2(x))^{3/2}} \\ &= \frac{m_3(x)}{(m_2(x))^{3/2}} \\ &= \hat{\alpha}_1(x)\end{aligned}$$

$$\begin{aligned}\hat{\alpha}_2(y) &= \frac{m_4(y)}{(m_2(y))^2} \\ &= \frac{b^4 m_4(x)}{(b^2 m_2(x))^2} \\ &= \frac{m_4(x)}{(m_2(x))^2} \\ &= \hat{\alpha}_2(x)\end{aligned}$$

These two new measures, $\hat{\alpha}_1$ and $\hat{\alpha}_2$, are invariant to changes in both scale and origin and so may correctly be called measures of shape. Before looking at some practical uses of our new tools, consider the following set of simple examples that will illustrate the ideas involved.

A key concept involved in any discussion of the higher moments is that, because a measure of the center of location and of spread, or scale, have already been determined, their confounding effects should be removed from the calculation of the higher moments. For example, the value of the third moment about the origin will reflect the effects of the degree of asymmetry, the center of location, and the spread. However, until the latter two effects are allowed for, you cannot distinguish the effect of asymmetry on the third moment.

We define four simple variables to illustrate. The four variables are x_a, x_b, x_c and x_d:

$x_a = 1, 2, 3$
$x_b = 1, 2, 9$
$x_c = -3, 1, 2$
$x_d = -1, 0(10 \text{ times}), 1$

Before beginning, sketch the frequencies as a line chart on any scrap of paper. We now calculate the four moments for each variable as well as the standardized third and fourth moments:

$$m'_1(x_a) = 2 \qquad m_2(x_a) = \frac{2}{3}$$

$$m_3(x_a) = 0 \qquad m_4(x_a) = \frac{2}{3}$$

$$\hat{\alpha}_1(x_a) = 0 \qquad \hat{\alpha}_2(x_a) = \frac{3}{2} = 1.5$$

$$m'_1(x_b) = 4 \qquad m_2(x_b) = \frac{38}{3} = 12.66$$

$$m_3(x_b) = 30 \qquad m_4(x_b) = \frac{722}{3} = 240.66$$

$$\hat{\alpha}_1(x_b) = \frac{30}{45.1} = 0.67 \quad \hat{\alpha}_2(x_b) = \frac{240.66}{160.4} = 1.5$$

$$m'_1(x_c) = 0 \qquad m_2(x_c) = \frac{14}{3} = 4.66$$

$$m_3(x_c) = -6 \qquad m_4(x_c) = \frac{98}{3} = 32.66$$

$$\hat{\alpha}_1(x_c) = \frac{-6}{10.1} = -0.59 \quad \hat{\alpha}_2(x_c) = \frac{32.66}{21.78} = 1.5$$

$$m'_1(x_d) = 0 \qquad m_2(x_d) = \frac{2}{12} = 0.166$$

$$m_3(x_d) = 0 \qquad m_4(x_d) = \frac{2}{12} = 0.166$$

$$\hat{\alpha}_1(x_d) = 0 \qquad \hat{\alpha}_2(x_d) = \frac{12}{2} = 6$$

The variables x_a and x_d are symmetric, so the third moment is zero, and we do not have to worry about the scaling problem. But is x_b five times more asymmetric than x_c as the raw third moment values would indicate? The standardized third moments are 0.67 and −0.59, which seems to be much more reasonable in light of our drawings of the distributions. The unstandardized third moments are so different because the second moment of x_b is nearly three times greater than that of x'_c; that is, the unstandardized third moments have compounded the effects of the asymmetry with the differences in the values of the second moments, which indicate the degree of spread.

Looking at the fourth unstandardized moments, we would be misled into thinking that x_d has the smallest value for peakedness, that the value for peakedness for x_b is the largest by far, and that the value for x_c is much greater than that for x_a. All of these conclusions are wrong as we can see by examining our $\hat{\alpha}_2$ values, the values of the standardized fourth moments. Variable x_d has the largest value for peakedness as we might suspect if we look carefully at the relative frequency line chart. Interestingly, the values for peakedness for all the other variables are identical; again, we might suspect that fact from a glance at the frequency line charts.

Similarly, reconsider Figures 3.5 to 3.14 by examining the $\hat{\alpha}_1$ and $\hat{\alpha}_2$ values shown in Table 4.11 and the unstandardized moments in Table 4.9. Consider, for example, the $\hat{\alpha}_1$ values for Figure 3.11 and the lognormal in Figure 3.14; the latter is greater than the former, but for the unstandardized moments shown in Table 4.9 the opposite is true. This is due to the differences in the second moments. We noted previously that the lognormal distribution is a model for income distributions, and Figure 3.11 is the graph of the distribution for the film revenues. If these data are to be believed, we might wonder whether film revenues are less or more asymmetric than incomes.

Table 4.11 **List of Means, Second, and Standardized Third and Fourth Moments for Chapter 3 Histograms**

Figure	Subject	m_1	m_2	$\sqrt{m_2}$	$\hat{\alpha}_1$	$\hat{\alpha}_2$
3.5	Die toss	4.80	6.00	2.50	−0.18	1.70
3.9	Final grades	67.30	132.00	11.50	−0.37	3.30
3.11	Film revenues ($mm)	26.50	1050.50 ($mm)^2	32.41	3.00	14.00
3.12	Gaussian	0.01	0.84	0.92	0.09	2.60
3.13	Weibull	0.97	0.02	0.13	−0.80	4.10
3.14	Beta (9,2)	0.83	0.01	0.11	−1.10	4.10
3.14	Arc sine	0.49	0.12	0.34	0.04	1.60
3.14	Uniform	0.52	0.09	0.30	0.25	1.80
3.14	Lognormal	1.81	6.66	2.60	5.32	40.30

Using the data in Table 4.4, we can calculate that the second moment of household income is 3.1 times 10^8 dollars, and the unstandardized third moment is 13.1 times 10^{12} dollars. However, the standardized third moment for household income is 2.4 and the standardized third moment for the film revenue data is 3.0, which is only a little larger.

Recall that the household incomes data in Table 4.4 involve some considerable approximation, especially in that the very highest incomes, although of very low relative frequency, were not recorded in the table. We expect that the actual standardized third moment is greater than that calculated. In any event, we can conclude that film revenues are not substantially more skewed than incomes generally.

What of peakedness for these data? The fourth moment for the household income data is 114.2 times 10^{16} dollars, but the standardized fourth moment is 11.9. The corresponding values for the standardized fourth moments shown in Table 4.11 are 13.9 for the film revenue and 40.3 for the lognormal distribution. Recalling once again that the nature of our approximations for the household income data is likely to underestimate the fourth moment, we can put the film revenue results into some perspective; they are at least not substantially more peaked than the household income data.

Three distributions in Table 4.11 seem to have similar values for the standardized fourth moments. Figure 3.5 was for the die-tossing experiment, and we would expect intuitively that the distribution of outcomes would be flat and not peaked. The uniform distribution is the ultimate in flat distributions. The arc sine distribution is U-shaped; such a distribution might be thought of as anti-peaked, so it should have a very low value for the standardized fourth moment. The $\hat{\alpha}_2$ values are, respectively, 1.7, 1.8, and 1.6. The Weibull and beta distributions shown in Figures 3.13 and 3.14, respectively, have about the same standardized fourth moments as we would expect from looking at the figures; both have $\hat{\alpha}_2$ values of 4.1.

Some Practical Uses for Higher Moments

At last we are ready to compare our film revenue data to see which film studios we want to back. Before looking at the listing of all four moments for the nine different film studios shown in Table 4.12, it would be helpful for you to refresh your memory

Table 4.12 **The First Four Moments of Film Revenues ($ millions) for Nine Film Studios**

Film Studio	m_1'	m_2	$\sqrt{m_2}$	$\hat{\alpha}_1$	$\hat{\alpha}_2$
B	28.0	407.0	20.2	0.9	2.3
C	25.5	1578.8	39.7	3.6	16.7
F	21.2	457.7	21.4	1.5	4.3
M	17.2	549.4	23.4	2.9	13.2
O	24.8	748.7	27.4	2.5	9.6
P	38.3	2309.6	48.1	1.8	5.4
T	27.1	782.6	28.0	3.0	13.7
U	25.3	1120.1	33.5	3.8	19.9
W	30.0	775.5	27.8	1.9	7.8
All	26.5	1050.5	32.4	3.0	14.0

by reexamining Figure 3.3, which shows the box-and-whisker plots of the film revenue data.

Now let us see what we can learn from Table 4.12. With respect to the means, one procedure might be to look first at the studios that exceed the overall mean. With this criterion, we should look at B, P, T, and W, although C and U are very close. P has the biggest mean return by far, so one might be tempted to pick it. But might not this result be from the fact that P had a couple of very big wins? Remember the mean is sensitive to all the data values and P's median value is about the same as that of the others.

If we are to consider the variability of the returns, we will have to think of how to trade off large means against large second moments, on the presumption that a larger second moment is all things considered not a good idea. One way to do this is to plot the means and the square roots of the second moments so that both returns and spread, or "variability," are in the same units; see for example Figure 4.3. This figure plots the means against the square roots of the second moments of the film revenues.

If you now imagine that you have an innate sense of your willingness to trade off returns against variability, then you want to be on the lowest line that passes through at least two of the alternatives and leaves all the other points above and to the left of the line. Restricting yourself to such a line means that you can choose between the alternatives on the line using whatever other criteria you wish, but that you will not be able to find a larger mean without having to pay a bigger price in variability; or re-expressed, you cannot find a smaller variability without having to accept a smaller mean.

For our film revenues, the line joining studios B and P is just such a line. W is so close to that line that one should probably include it as an alternative as well. Studio W is an example of a studio with the biggest return for a given value for the second moment, or for a given measure of spread, relative to studios T and O. Studio B is an example of a studio with the minimum second moment for a given mean return. Studio P has the largest mean and the largest variation.

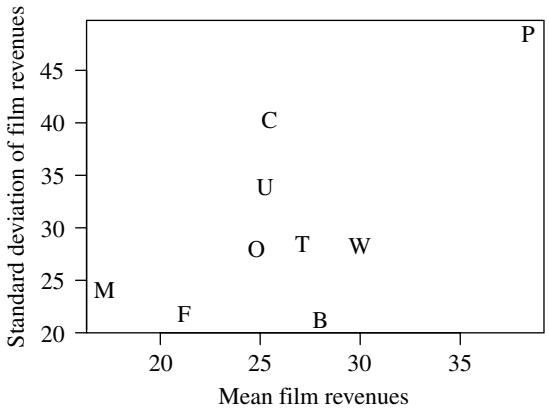

Figure 4.3 Means and standard deviations of film revenues ($millions)

Which studio you choose, or better still which combination of studios you choose, depends on your personal feelings about the trade-off between return and the variability of returns. You can get a specific combination of return and variability on the line joining B and P by putting part of your money into B and part into P.

The return–variability trade-off that we outlined in Figure 4.3 makes most sense when the distributions are nearly symmetric. But when the distributions are skewed to the right, this comparison loses a lot of its charm. One might want to consider the degree of asymmetry that is involved in the choices. For example, it happens that B, P, and W, which lie on our return–variability trade-off line, all have relatively small standardized third moments; those for C and U are the highest. If the measure of skewness is large, that implies that for a given mean and a given degree of variability, the value taken by the average return depends to a substantial extent on a relatively small number of very large returns; a small value for the standardized third moment implies the opposite. For a given variance, the distribution with the smaller standardized third moment is in a sense "less risky" in that your average return over a long period of time will depend less on the rare, but very large, return. In this particular example, you might well decide that the small size of the third moments for the three studios B, W, and P enhances your decision to look at these three.

This idea is more easily seen if we consider the comparison of two distributions of returns, one of which has a negative standardized third moment, the other a positive standardized third moment, with equal means and equal second moments. For the former return distribution, a large number of small positive returns is offset by an unusual, but very large negative return of, say, bankruptcy status. Merely to consider this choice is to recognize the importance of examining the third moment. The alternative distribution of returns is positively skewed so that the mean return is achieved by the averaging of a large number of very small—even negative—returns, with a few very large returns. As you contemplate these alternative distributions, you will recognize that your personal reaction to risk is affected by the presence of nonzero third moments.

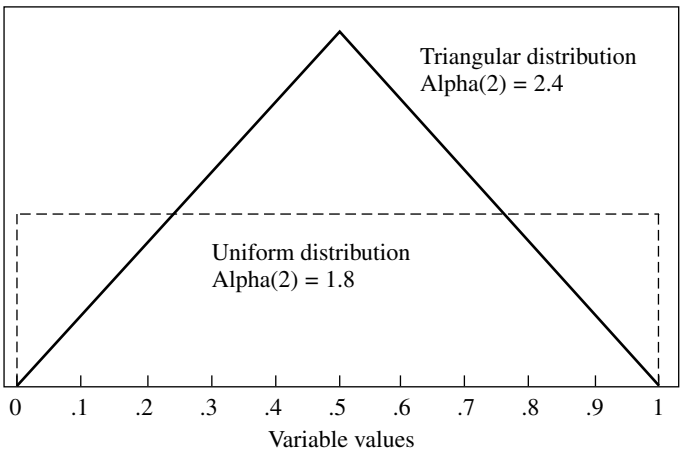

Figure 4.4 Comparison of two distributions with different alpha(2) values. The area under each curve represents probability.

We should now examine the usefulness of the standardized fourth moment in making our decision. As we have said, the fourth moment provides information about the peakedness of the distribution, or the existence of fat tails of the distribution—that is, the concentration of data about the mean as opposed to the tails of the distribution. The fourth moment is easiest to interpret when the distribution is symmetric; otherwise, one has to disentangle the effects of asymmetry from those due to the peakedness or fat tails problem. The standardized fourth moment was smallest when the distribution was flat, or anti-peaked as was the arc sine distribution. It is very large when the distribution has a narrow "spike" in the middle and large tails.

Figure 4.4 provides a comparison that is easier to visualize because the range of the data in both cases is restricted to the interval [0,1]. Two distributions are compared, the uniform that we have seen before in Tables 4.7 and 4.11 and in Figure 3.14, and a new one, the triangular distribution, for which the name is obvious. In interpreting Figure 4.4 remember that the areas under histograms add up to one, so that what is gained in relative frequency in one region has to be compensated for elsewhere. Remember also that you have to compare standardized fourth moments; that is, you have to allow for differences in the values of the second moments. The second moment for the triangular distribution is $(\frac{1}{3})(\frac{1}{2})^3$ and that for the uniform is $(\frac{1}{3})(\frac{1}{2})^2$; so the second moment for the triangular distribution is one-half that for the uniform. The triangular distribution's standardized fourth moment, $\hat{\alpha}_2 = 2.4$, is greater than that for the uniform, $\hat{\alpha}_2 = 1.8$, because the former distribution has fatter tails than the uniform, but only after allowing for the difference in the second moments.

We can get an even clearer picture of the role played by the standardized fourth moment if we examine Figure 4.5. In this figure I have listed the raw data on the extreme left and have plotted the standardized data in the middle of the figure; the

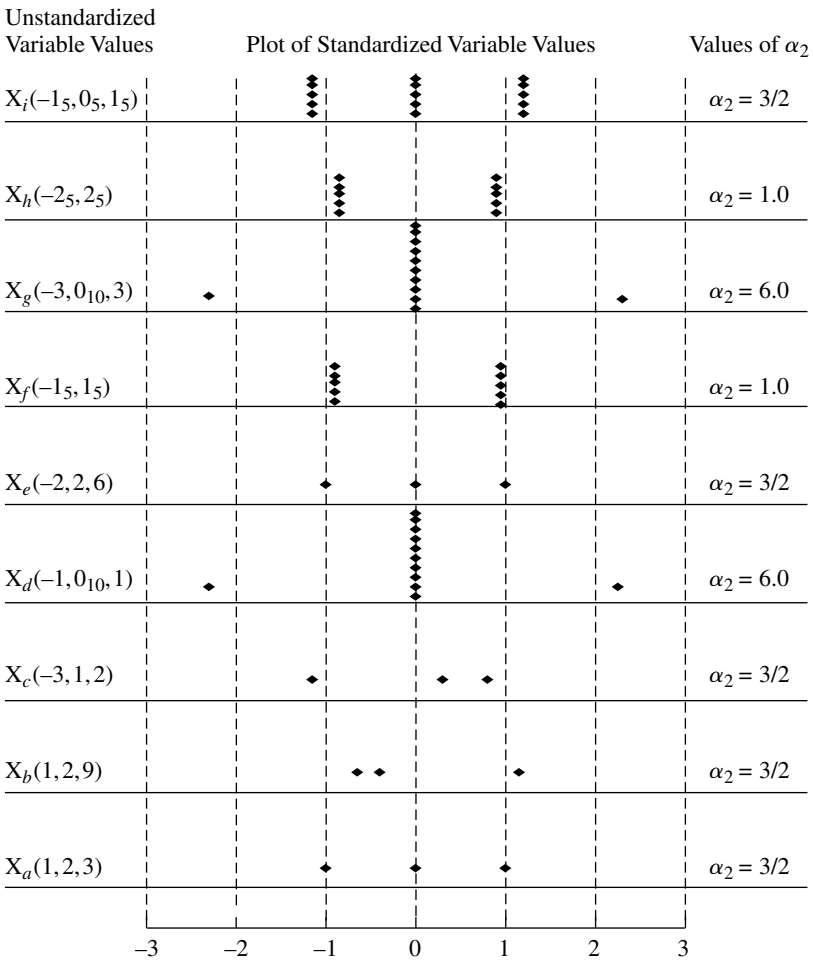

Figure 4.5 A comparison of alpha(2) values for standardized variables. Each diamond represents an observation.

values of $\hat{\alpha}_2$ are shown at the far right. The calculated values of $\hat{\alpha}_2$ vary from a low of 1.0 to a high of 6.0; uniformly distributed data have an $\hat{\alpha}_2$ value of $\frac{3}{2}$. Both x_f and x_h are U-shaped distributions and have $\hat{\alpha}_2$ values of only 1.0; x_d and x_g illustrate distributions that yield large values for $\hat{\alpha}_2$. It is worth some time comparing distributions and the corresponding values for the standardized fourth moment. Compare the distributions that have the same $\hat{\alpha}_2$ values, and compare x_f, x_i, and x_g with $\hat{\alpha}_2$ values ranging from 1 to 6.

We can illustrate these ideas with the film revenue data. Studios C, M, T, and U have the biggest values for $\hat{\alpha}_2$, whereas B, F, P, and W have the smallest. Of these, we have seen from Figure 4.3 that F is dominated by B in that B has a much larger mean revenue and a smaller second moment. P has a surprisingly low value for its

standardized fourth moment given that it has the greatest range of all the alternatives. Somewhat surprisingly in this example, the examination of the standardized fourth moments also confirms that a choice between B, W, and P is best. In terms of determining the optimal combination of weights for choosing an optimal mix of the three studios, one might well place a higher weight on B because it has the smallest value for $\hat{\alpha}_2$. Other things equal, studio B is less risky than the other "mean variance" optimal alternatives, W and P.

You now have several ways to choose between the alternative studios. Clearly, choosing on the basis of the mean alone is not enough; some recognition of the variability of the data is important. The second moment measures the extent of that variability, but the third and the fourth moments help you to describe the nature of that variability. We have seen that the distribution of revenues can be quite different, even between distributions that have the same degree of variability as measured by the second moment. Just as we can consider our implicit and intuitive trade-off between return and variability, so we can consider our implicit and intuitive trade-off between distributions with different degrees of asymmetry and different degrees of peakedness. Normally, we might expect that people would prefer less asymmetry, given a level of return and degree of variability. Similarly, we might expect them to prefer less peakedness and hence thinner tails—that is, less reliance for a given average return on large, but low-frequency, occurrences. The choice now is up to the individual to evaluate the trade-offs in these various characteristics of the film revenue distributions.

4.7 Standardization of Variables

Before we began looking closely at our film revenue data, we recognized the importance of the effects of origin and scale in the measurement of our variables. As a consequence we had to define $\hat{\alpha}_1$ and $\hat{\alpha}_2$ to obtain scale and origin invariant measures of shape. This technique is very useful and provides a lot of simplification. It is easier to handle variables for which the mean is zero and for which the second moment is one. The process is known as *standardization of variables*. Standardization is a simple procedure: subtract the mean and divide by the square root of the second moment; the reason for the square root will be clear soon, if it is not already. Define the variable y_i by

$$y_i = \frac{(x_i - \bar{x})}{(\sqrt{m_2}(x))}$$

which is the general statement for "subtract the mean and divide by the square root of m_2."

The transformed variable, y_i, has a mean of zero and a second moment of one! Let's check this. Rewrite the equation expressing y_i as

$$y_i = \left[\frac{-\bar{x}}{(\sqrt{m_2}(x))}\right] + \left[(\sqrt{m_2}(x))^{-1}\right] \times x_i$$

but this is just

$$y_i = a + b \times x_i$$

where

$$a = \frac{-\bar{x}}{(\sqrt{m_2})} \text{ and } b = (\sqrt{m_2})^{-1}$$

If we now calculate the first two moments of y, we get

$$\begin{aligned} y &= \frac{\left(\sum y_i\right)}{N} \\ &= \left(\frac{-\bar{x}}{\sqrt{m_2}(x)}\right) + \frac{(\sqrt{m_2}(x))^{-1} \times \sum x_i}{N} \\ &= \left[\frac{-\bar{x}}{(\sqrt{m_2}(x))}\right] + \frac{\bar{x}}{\sqrt{m_2}(x)} \\ &= 0 \end{aligned}$$

$$\frac{\sum (y_i - \bar{y})^2}{N} = (\sqrt{m_2}(x))^{-2} \times \frac{\sum (x_i - \bar{x})^2}{N}$$

Because

$$y = a + b \times x = 0 \text{ and } (\sqrt{m_2}(x))^{-2} = (m_2(x))^{-1}$$

we get

$$\frac{m_2(x)}{m_2(x)} = 1$$

so, we conclude

$$\frac{\sum (y_i - \bar{y})^2}{N} = 1$$

Even more interesting is that we get yet another simplification:

$$\hat{\alpha}_1(y) = \frac{m_3(y)}{(m_2(y)^{3/2})} = m_3(y)$$

$$\hat{\alpha}_2(y) = \frac{m_4(y)}{(m_2(y)^2)} = m_4(y)$$

because $m_2(y) = 1$.

The variable y is called "x standardized." Standardized variables are very convenient in that their means are zero, their second moments are one, and their standardized third and fourth moments, $\hat{\alpha}_1$ and $\hat{\alpha}_2$, are easily calculated by the ordinary third and fourth moments of the standardized variables.

We get the same result for $\hat{\alpha}_1$ and $\hat{\alpha}_2$ whether we first standardize the variable and calculate the third and fourth moments or we calculate the third and fourth moments of the original variable and then divide by the second moment.

The Higher Moments about the Origin

All of this time we have ignored the higher moments about the origin, m'_2, m'_3, m'_4. This is because the moments about the mean are more useful for understanding the

shapes of distributions. But the moments about the origin come into their own when we want an efficient way to calculate the actual values of moments about the mean. For example, for any variable x:

$$\begin{aligned} m_2(x) &= \frac{\sum (x_i - \bar{x})^2}{N} \\ &= \frac{\sum \left[(x_i^2 - 2 \times x_i \times \bar{x} + (\bar{x})^2 \right]}{N} \\ &= \frac{\left[\sum x_i^2 - 2 \times \sum x_i \times \bar{x} + \sum (\bar{x})^2 \right]}{N} \\ &= \frac{\sum x_i^2}{N} - (\bar{x})^2 \end{aligned}$$

where

$$\frac{\sum x_i \bar{x}}{N} = \bar{x}^2, \text{ and } \frac{\sum \bar{x}^2}{N} = \bar{x}^2$$

so

$$m_2(x) = m_2'(x) - (m_1'(x))^2 \tag{4.9}$$

This relationship between the second moments is very useful and we will use it a lot, so Equation 4.9 is a useful one to remember. Similar relationships hold for the rest of the moments, but we will not need them for awhile. However, examples are provided in the exercises to this chapter.

Higher Moments and Grouped Data

We can also calculate the grouped approximations to the third and higher moments. However, remembering that using the grouped data involves an approximation, we can easily see that raising the differences between the actual and the approximation to higher and higher powers will soon lead to huge differences between the true values of the moments and the approximations. The more the cell mark, C_j, differs from the actual values of the data within the cell, the greater the difference between the true moment values and the group approximations.

4.8 Summary

This chapter's objective has been to convert the visual ideas of shape discussed in Chapter 3 into a precise mathematical formulation. We implemented this idea by defining moments. *Moments* are simply averaged powers of the values of a variable. Four moments are all that are needed to characterize the shape of most histograms you will need to use. Moments can be defined in terms of *moments about the mean* or as *moments about the origin*. It is convenient to carry out all the discussion of the use of moments in terms of the moments about the mean for all moments after the first one. However, to calculate moments about the mean it is convenient to use moments about the origin.

The *mean* was defined in Equation 4.1, and its approximation using cell data was defined subsequently. The mean is a measure of location, and it is obtained by averaging the observed data. The mean is the first moment about the origin.

The general moments about the origin were defined in Equations 4.6 and moments about the mean in Equations 4.7. The *second moment about the mean* is particularly important as it is a measure of spread of the variable values. The second moment and its approximation using data in cells is presented in Equations 4.4 and 4.5. We temporarily defined the term *standard deviation* as the square root of the second moment; the standard deviation is important as it is the measure of spread that has the same units of measurement as the original data, whereas the second moments units are squared.

To capture the effects of skewness and peakedness, we discovered that we had to look at the standardized values of the third and fourth moments; these *standardized moments* are called $\hat{\alpha}_1$ and $\hat{\alpha}_2$, and their expressions are given in Equations 4.8.

Standardization is an important tool that simplifies much analysis. Standardization of any variable is achieved by subtracting the mean and dividing the result by the square root of the second moment. Standardized variables by their construction have zero means and second moments equal to one.

The second moment about the mean can usefully be reexpressed as the difference between the second moment about the origin and the square of the first moment as shown in Equation 4.9.

Case Study

Was There Age Discrimination in a Public Utility?

In Chapter 3, we examined some of the distributions of the data in this case to get a feel for what might be involved. In this chapter, we will use the tools that we have just developed to explore in more depth the issues previously raised. The first question to resolve concerns the shape of the distributions. We can use the menu command Moments in S-Plus to calculate the moments of the distributions and compare these calculations to the shapes of the observed histograms. Recall Figures 3.19 and 3.20.

However, before we begin calculating moments, recall the warnings given in Chapter 3 that these data contain coding errors. You have to decide how to handle this difficulty. One way is to recognize that although there are errors their effects on the calculations are minimal and so can be ignored, but note the problem for the reader of your work. An alternative is to run the analysis without the controversial data and tell the reader about your decision. In the following analysis, we will ignore the coding errors because they have little effect on our results at this stage.

We obtain using the menu command Moments on the age variable the following output for all individuals in the Discdata.xls file:

```
Age of all individuals in the file
First four moments about the mean
No. of observations = 479
```

continues on next page

(Continued)

First = 46.225
Second = 78.016
Standard deviation = 8.833
Third =–146.452
Fourth = 16664.98
Standardized moments beginning with the third = –0.213 2.738

We observe that the mean of the overall distribution of ages is 46 years with a standard deviation of about nine years. The distribution is nearly symmetric, but with a small left-hand tail, $\hat{\alpha}_1 = -0.21$. The amount of peakedness is modest at $\hat{\alpha}_2 = 2.7$. We know from Figure 3.15 that we would discover very little difference in these calculations if we were to separate men and women.

Of greater importance is to examine the difference in moments for the subgroups created by querying whether the person was a former employee(internal) or external.

Age of former employees
First four moments about the mean
No. of observations = 375
First = 47.259
Second = 75.637
Standard deviation = 8.697
Third = –132.341
Fourth = 15737.85
Standardized moments beginning with the third = –0.201 2.751

Age of external applicants
First four moments about the mean
No. of observations = 104
First = 42.5
Second = 68.865
Standard deviation = 8.299
Third = −225.288
Fourth = 10944.6
Standardized moments beginning with the third = −0.394 2.308

As we might guess from the box-and-whisker plots, the mean age and standard deviation of the former employees are greater than that for the external applicants, but the differences are not very large. Both distributions are skewed left, and there is a modest amount of peakedness.

Another comparison is between those hired, not hired, and those that did not apply as they transferred or were near retirement (see Figure 3.18). How do these distributions differ?

Age of former employees that were not hired
First four moments about the mean
No. of observations = 52
First = 49.058
Second = 31.093
Standard deviation = 5.576
Third = –1.862
Fourth = 2495.982
Standardized moments beginning with the third = –0.011 2.582

Age of former employees that were hired
First four moments about the mean
No. of observations = 204
First = 43.931
Second = 57.427
Standard deviation = 7.578
Third = –213.539
Fourth = 9958.013
Standardized moments beginning with the third = –0.491 3.020

Age of former employees that did not apply
First four moments about the mean
No. of observations = 119
First = 52.176
Second = 81.742
Standard deviation = 9.041
Third = –405.121
Fourth = 15071.32
Standardized moments beginning with the third = –0.548 2.256

Our first observation is that across these divisions, there seems to be little difference

continues on next page

(Continued)
between the three distributions with respect to shape as measured by the standardized third and fourth moments; all have very modest amounts of left skewness and moderate amounts of peakedness. The difference in ages between those hired and not hired does reflect the union's charges; those hired are on average younger than those not hired. However, the difference seems to be very small relative to the degree of dispersion in the age data. This conclusion is amply born out looking at the age distribution of those who did not apply; the mean and the degree of variation are far greater as one would expect.

Our overall conclusions on the age variable are that for all the subdistributions that we have examined, the shapes of the distributions are very similar. Indeed, the difference between our sample distributions and the normal distribution in Chapter 3 are very slight, but our distributions show some evidence of having a left-hand tail. With respect to the complaint, although we do observe that the ages of the internal applicants relative to the external applicants and the ages of those not hired relative to those hired are both older than the comparison group, the differences are very small, especially in light of the size of the second moments that we calculated. As an aside, we also observe that the second moments of those hired and those who did not apply were larger than those not hired, but the differences do not appear to be large.

In Chapter 3, we examined the breakdown of salaries by the same categories. Let us examine how the calculation of moments helps. The overall distribution of salaries (see Figure 3.20) has moments:

```
Salaries of all individuals on the
  Discdata.xls file
First four moments about the mean
No. of observations = 479
First = 61457.64
Second = 213976391
Standard deviation = 14627.93
Third = 256386072762
Fourth = 1.387142e+017
Standardized moments beginning with
  the third = 0.082 3.030
```

We immediately note striking differences with respect to our previous distributions. Here, the measures of shape indicate clearly that the distribution of salaries—although having moderate, but higher, levels of peakedness relative to the distribution of age—are in this case skewed to the right. We note that the mean for all internal employees is about $61,500 with a standard deviation of about $14,600. The salaries of most employees in this firm lie in a region from about $32,300 to $90,700. We now investigate how this overall distribution is altered by considering the various subcategories.

First, we ask what is the difference between the salaries of those who were hired from the former employees and those who were not?

```
Salaries of former employees that were
  not hired
First four moments about the mean
No. of observations = 52
First = 60026
Second = 238852724
Standard deviation = 15454.86
Third = 1546485329952
Fourth = 1.868219e+017
Standardized moments beginning with
  the third = 0.419 3.275

Salaries of former employees that were
  hired
First four moments about the mean
No. of observations = 219
First = 61803.43
Second = 222177292
Standard deviation = 14905.61
Third = 277031532737
```

continues on next page

(Continued)

```
Fourth = 1.459941e+017
Standardized moments beginning with
  the third = 0.084 2.958
```

From these figures we see that the average salaries of those hired were slightly greater than for those not hired. This result does not lend support to the union's contention that the firm was trying to save money by letting go the higher-priced employees.

```
Salaries of former employees that did
  not apply
First four moments about the mean
No. of observations = 121
First = 61758.62
Second = 166257253
Standard deviation = 12894.08
Third = –969433408053
Fourth = 8.044588e+016
Standardized moments beginning with
  the third = –0.452 2.910
```

The mean for those who did not apply is surprisingly very close to that for those hired, but the standard deviation is much less. Interestingly and somewhat surprisingly, this income distribution has a left-hand tail, whereas all the others have a right-hand tail. The reason presumably is that there are included in the group "did not apply" a number of very low-salaried people. Indeed, an examination of the corresponding box-and-whisker plots (see Figure 3.17) indicates that this is the case.

One cannot calculate the salary breakdown by the distinction between internal and external applicant as we have no information on the salaries of the external candidates.

Overall, the variations across the subgroups were with one exception very small. Indeed, the average salaries of those previous employees hired were greater by a small amount than the average salaries of those not hired. More surprisingly, the average income of those who retired was nearly equal to that of those hired. There are differences across the various groups in terms of the shape of the distributions, but they are all very small in extent.

Nevertheless, we have already learned a fair amount about the facts of this case. As we proceed, we shall learn more still.

Exercises

If you have not yet read the "Addendum to the Reader" at the end of Chapter 1, you will really need to do so for the exercises in this chapter.

Calculation Practice

Exercises 4.1 through 4.4 provide some basic practice on the use of summation notation and the elementary algebra that is needed for the rest of the book. See Appendix A, "Mathematics Appendix" for a more intensive review of these ideas.

For questions 4.1 to 4.4, use the following set of numbers:

$$\{x_i\} = \{1, 2, 3, 5, 7, 8, 9, 12\}, i = 1, 8$$

4.1 Use a hand calculator to calculate each of the following expressions. Indicate the relationships between your results.

a. $\sum x_i$

b. $\sum x_i^2$

c. $(\sum x_i)^2$

d. $3\sum x_i$

e. $\sum 18x_i$

f. $\sum x_i^3$

g. $(\sum x_i)^3$

4.2 Use a hand calculator to calculate each of the following expressions. Note that you can use the results from Exercise 4.1. Indicate which are moments, and define the moment.

a. $\sum x_i/N$

b. $\sum x_i^2/N$

c. $(\sum x_i)^2/N$

d. $\sum x_i^3/N$

e. $(\sum x_i)^3/N$

f. $\sum x_i^4/N$

g. $(\sum x_i)^4/N$

4.3 Use a hand calculator to calculate each of the following expressions. Indicate the relationships between these results.

a. $\sum(x_i - (\sum x_i/N))$

b. $\sum(x_i - \bar{x})$

c. $\sum(x_i - \bar{x})/N$

4.4 Use a hand calculator to calculate each of the following expressions. Indicate the relationships between these results.

a. $\sum(x_i - \bar{x})^2$

b. $\sum(x_i - \bar{x})^2/N$

c. $(\sum(x_i - \bar{x}))^2/N$

d. $(\sum x_i^2/N) - \bar{x}^2$

e. $(\sum x_i^2)/N - (\sum x_i)^2/N$

4.5 ***Worked.*** *Objective: To link the preceding algebraic exercises to the use of the computer in "adding up."* Exercises 4.5 through 4.7 provide more practice on the use of summation notation and elementary algebra. For these exercises, refer to the following set of numbers: $\{x_i\} = \{-4, 2, -1, 0, 5, 7, -2, 8\}$. Please calculate the following:

a. $\sum x_i$

b. $\sum x_i^2$

c. $(\sum x_i)^2$

d. $-2\sum x_i$

e. $\sum 64x_i$

f. $\sum x_i^3$

g. $(\sum x_i)^3$

4.6 Confirm by using a hand calculator that you get the same results as in the following computer method.

[***Computer directions:*** To calculate the seven exercises in 4.5, the set of numbers x_i has been placed in a file Test.xls. Use the following directions:

Start S-Plus. On the menu bar, click on File, Import Data, From File, folder Xfiles, subfolder Misc. Click on file Test.xls. Click Open. On the menu bar, click on Data, Transform. In the Transform dialog in Expression:

a. Key in sum(*x*). Click Apply. The answer will be in column V1 in the Test data frame (answer = 15).

b. Key in sum(*x*^2). Click Apply. The answer will appear in column V2 (answer = 163).

c. Key in sum(*x*)^2. Click Apply (answer = 255).

d. Key in −2*sum(*x*). Click Apply (answer = −30).

e. Key in sum(64**x*). Click Apply (answer = 960).

f. Key in sum(*x*^3). Click Apply (answer = 915).

g. Key in sum(*x*)^3. Click Apply (answer = 3375).]

The answers from the computer should confirm your hand calculations.

Summarize your calculations by stating some general rules of summation that are illustrated by these questions.

4.7 Now try calculating the following expressions by hand using the set of numbers in Exercise 4.5.

a. $\sum x_i/N$

b. $\sum x_i^2/N$

c. $(\sum x_i)^2/N$

d. $\sum x_i^3/N$

e. $(\sum x_i)^3/N$

f. $\sum x_i^4/N$

g. $(\sum x_i)^4/N$

h. $\sum(x_i - (\sum x_i/N))$

i. $\sum(x_i - \bar{x})$

j. $\sum(x_i - \bar{x})/N$

k. Which of these are moments? Distinguish moments about the origin and moments about the mean.

l. Complete the exercise by confirming that the use of the computer gives the same results. If you get a different result, you should first check your use of parentheses to make sure that you are performing the same calculations by hand as by computer.

4.8 Using the set of numbers in Exercise 4.5 calculate:

a. $\sum(x_i - \bar{x})^2$

b. $\sum(x_i - \bar{x})^2/N$

c. $(\sum(x_i - \bar{x}))^2/N$

d. $(\sum x_i^2)/N - \bar{x}^2$

e. $(\sum x_i^2)/N - (\sum x_i)^2/N$

f. Summarize these calculations by stating some general rules that would apply to any data set.

g. Complete the exercise by confirming that use of the computer gives the same results.

Exploring the Tools

4.9 Moments about the mean can be reexpressed in terms of moments about the origin, and vice versa. For the first three moments derive the relationship between each moment about the mean and the moments about the origin. Similarly, derive the relationship between each moment about the origin and the moments about the mean, plus the mean itself. Can you detect a general principle?

4.10 As practice and as a reminder for the next few exercises, consider the following relationships. Try to follow the algebra for each statement, and verify the results when $a = 3$, $b = 2$. Then confirm the result when $a = 7$ and $b = 3$.

$$\begin{aligned}(a-b)^3 &= (a-b)^2(a-b)\\ &= (a^2 - 2ab + b^2)(a-b)\\ &= a^3 - 3a^2b + 3ab^2 - b^3\end{aligned}$$

$$\begin{aligned}(a-b)^4 &= (a-b)^2(a-b)^2\\ &= (a^2 - 2ab + b^2)(a^2 - 2ab + b^2)\\ &= a^4 - 4a^3b + 6a^2b^2 - 4ab^3 + b^4\end{aligned}$$

4.11 Using the numbers $\{x_i\} = \{-2, -1, 0, 1, 2, 6\}$, $i = 1, 6$, calculate with a hand calculator:

a. $\sum x_i/N$

b. $\sum(x_i - \bar{x})^2/N$

c. $\sum x_i^2/N - (\sum x_i)^2/N$

d. $\sum(x_i - \bar{x})^3/N$

e. $\sum x_i^3/N - 3(\sum x_i^2/N)\bar{x} + 3\bar{x}^3 - \bar{x}^3$

f. Using your results, algebraically express the second and third moments about the mean in terms of moments about the origin.

g. Complete the exercise by confirming that use of the computer gives the same results.

4.12 Using the same numbers listed in Exercise 4.11, calculate:

a. $\sum(x_i - \bar{x})^4/N$

b. $\sum x_i^4/N - 4(\sum x_i^3/N)\bar{x} + 6(\sum x_i^2/N)\bar{x}^2 - 4\bar{x}^4 + \bar{x}^4$

c. $\sum x_i^4/N - 4(\sum x_i^3/N)\bar{x} + 6(\sum x_i^2/N)\bar{x}^2 - 3\bar{x}^4$

d. Using your results, reexpress the fourth moment about the mean in terms of moments about the origin.

e. Complete the exercise by confirming that use of the computer gives the same results.

4.13 Under what circumstances is the second moment about the mean zero? When is the fourth moment about the mean zero? The first moment about the mean is identically zero. Explain and illustrate. How does this fact aid your understanding of the "mean"?

4.14 The second moment is an average squared deviation. This can be seen by noting that in the formula for a second moment, the sum of the deviations squared is divided by the number of observations. If there are N observations, there are N deviations from the mean. Let $N = 4$. Write out the formula for the second moment in longhand (do not use summation notation). For any N if possible, but for $N = 4$ if you must, prove that only $N - 1$ of the N deviations are independent (i.e., prove that if you know what $N - 1$ of the deviations are, you know what the Nth deviation is as well).

4.15 You observe your mother's temperature three times a day for many days while she is sick. Your thermometer is calibrated in degrees Fahrenheit. The attending nurse has also been taking your mother's temperature, except that her thermometer is calibrated in degrees Celsius. Explain how you would begin to check whether the distribution of the nurse's temperature readings is similar to yours. How would you specify "similar"?

4.16 Why does it make sense that skewness should be measured by a deviation raised to an odd power, as opposed to an even one?

a. Which observations in a data set do the most to increase the third moment?

b. Which observations in a data set do the most to increase the fourth moment?

c. If distribution A is shaped like a triangle and distribution B is shaped like a U, which distribution will have the larger value for the standardized fourth moment given the second moments are the same?

d. If distribution C is very flat and D has a single hump in the middle of the distribution, which has the larger standardized fourth moment given the second moments are the same?

e. What do the mean and the square root of the second moment have in common that the mean and the second moment itself do not?

f. A brand-new statistical software package freshly installed on your computer has calculated various statistics about the age and income distributions of a populous but less-developed country. You discover that the third moments of both variables are negative. Without hesitation, you call the software company and tell them you want your money back. Why?

g. Write out the expressions for $\hat{\alpha}_1$ and $\hat{\alpha}_2$. After examining these expressions specify the units in which they are measured.

h. Prove that the mean of a standardized variable is 0.

4.17 When do the average, $\bar{x}$, and median differ? If you were examining the following sets of data, which measure of location would you use and why? When might you use both? If you examine both, which do you think is larger and why?

a. income of households

b. length of fishing poles

c. number of family members

d. duration of unemployment

e. size of earthquakes

f. wind strength of hurricanes

g. number of felonies committed in a year

4.18 ***Worked.*** *Objective: To illustrate the use of the Moments menu command.* Recall the data on 72 NYU student grades, two midterms and a final, that we discussed in Chapter 3 (see folder Xfiles, subfolder Misc, file Grades.xls).

a. For the variable *midterm1*, calculate the first four moments using the following two computer procedures and compare your answers.

[***Computer directions:*** Start S-Plus. On the menu bar, click on File, Data Import, From File, folder Xfiles, subfolder Misc. Double click on file Grades.xls.

Procedure 1: To calculate the moments step by step: on the menu bar, click on Data, Transform.

(1) To calculate the number of observations, *n*: In the New Column Name box key in *n*. In the expressions box, key in length (midterm1). Click Apply. The answer will be in the *n* column (the answer will be repeated throughout the column) in the Grades data frame (answer = 72).

(2) To calculate *m*1, the mean, key in mean for the New Column Name. In Expression, key in sum(midterm1)/*n*. Click Apply (answer = 72.31).

(3) To calculate *m*2, key in name m2 and sum((midterm1-mean) ^2)/n. Click Apply (answer = 196.52).

(4) To calculate *m*3, key in name m3 and sum((midterm1-mean) ^3)/n. Click Apply (answer = −824.28).

(5) To calculate *m*4, key in name m4 and sum((midterm1-mean) ^4)/n. Click Apply (answer = 90505.14).

Procedure 2: To calculate all the moments: On the menu bar, click on Text Routines, Moments. In the dialog, select Data Frame, Grades. In Variable, select "midterm1." Click OK.] Compare the results from procedures 1 and 2. They will be the same. If not, you made a mistake; recheck your entries.

b. Repeat procedures 1 and 2 using the variable *final* in Grades.xls.

Now that you have had practice calculating moments step by step on the computer, you can use the Moments menu command as in procedure 2 whenever you need to calculate moments. Better still after this exercise, you will have no doubt about what it is that the computer is calculating on your behalf.

4.19 An absentminded professor has calculated the mean, second moment, and third and fourth moments for his economics class' final exam. The grades for his 24 students are as follows: 52 92 86 80 88 64 92 76 92 68 68 64 70 68 68 88 88 94 64 56 84 84 74 62

a. What are the results of the professor's calculations?

[***Computer hint:*** In S-Plus, on the menu bar, click on Data, New Data Object, data.frame. Click OK. Key in the grades into the first column. On the menu bar, click on Text Routines, Moments. In the dialog, select Data Frame, SDF#. In Variable, select V1. Click OK. The plots in (*b*) can be created by clicking Graph, 2D-Plot on the menu bar and selecting the graph type desired.]

b. Create the box-and-whisker plot and a histogram for the professor's grades. Answer all the following questions *before* doing any calculations on the computer; use the computer merely to check your intuition.

(1) Having gone through all the work, the professor relaxes by drinking tea and petting his cat. Lifting the cat, he finds six ungraded and fur-covered exams. Suddenly he remembers that there are 30, not 24, students in his class. After grading the newfound exams, he finds that each of the six earned the mean score. Which of his previous calculations now need to be redone? Which can be left unchanged?

(2) Consider each of the moment calculations that must be changed. Without recalculating, in which direction will each result change? Remember that the number of grades included in the calculations has also changed.

(3) Add the six "mean" exams to the histogram. Does the histogram reflect what you said would happen in question (2)?

(4) Do you need to redraw the box-and-whisker plot?

4.20 More reflections on the absentminded professor. Do *not* use the computer for this exercise. Before he had found the last six exams and viewed the distribution of grades, he was thinking of changing his teaching, or at least his grading, methods. Why do you think he was considering this, and what do you think he should have changed? After considering the additional six exams, do you think it is still so imperative for him to change? Why or why not?

Late at night, just before falling asleep, the absentminded professor remembers that he has been grading so far on a scale of 50 to 100, not on a scale of 0 to 100. He wants to make the scores in 4.19a comparable to his other exams, and he wants the information on the moments to be comparable as well. What does he have to do to fix things? Be nice and do it for him; he's sleepy.

4.21 Worked. *Objective: To indicate the loss of information in using grouped data*. In the text it was explained that it is not very useful to use grouped or cell data to calculate the higher moments; even the second moment is not very accurate. This exercise will help illustrate this fact. The mathematics and verbal scores for both the PSAT and the SAT are recorded in the file Psatsat.xls.

a. For each score catagory calculate all four moments and the standardized third and fourth moments using the observed data.

[***Computer directions:*** Start S-Plus. On the menu bar, click on File, Import Data, From File, folder Xfiles, subfolder Testscor. Click on file Psatsat.xls. Click Open. On the menu bar, click on Text Routines, Moments. In the dialog, select Data Frame, Psatsat. In Variable, select "psatmath." Click OK. Repeat the Moments command for each variable.]

b. Using the procedure developed in Chapter 3, the PSAT and the SAT data have been grouped into cells. A file, Psatcell.xls, has been created with the number of observations by cell (*pmobs*, *pvobs*, etc.) and the cell marks (*pmmark*, *pvmark*, etc.) for each of the variables in file Psatsat.xls. Use the following directions to calculate the moments from the cell data.

[***Computer directions:*** On the menu bar, click on File, Import Data, From File. Select folder Xfiles, subfolder Testscor. Double click on the file Psatcell.xls. On the menu bar, click on Data, Transform.

(1) To calculate the number of observations, *n*: In New Column Name, key in n. In Expression, key in sum(pmobs). Click Apply. The answer will be in the Psatcell data frame in column *n* (answer = 520).

(2) To calculate *m*1, the mean: key in mean for the name. In Expression, key in sum(pmmark* pmobs)/n. Click Apply (answer = 45.75).

(3) To calculate *m*2, key in m2 for the name. In Expression, key in sum(pmobs*(pmmark-mean)^2)/n. Click Apply (answer = 135.78).

(4) To calculate *m*3, key in m3 for the name. In Expression, key in sum(pmobs*(pmmark-mean)^3)/n. Click Apply (answer = 149.83).

(5) To calculate *m*4, key in m4 for the name. In Expression, key in sum(pmobs*(pmmark-mean)^4)/n. Click Apply (answer = 48495.76).]

c. Calculate the moments for the other three variables from the cell data, and compare your answers with the moments calculated using the observed data in step (a).

4.22 ***Worked.*** *Objective: To illustrate the relationship between the shape of a distribution and its moments*. In Chapter 3, we mentioned a variety of different distributions that were generated by different types of experiments. In this chapter, we have shown that as the shape of the distribution changes, so do the moments. This exercise will illustrate this

essential idea in terms of histograms from computer experiments and from idealized shapes of distributions (that is, in terms of histograms that are very smooth, such as might be generated by a very large set of data).

[***Computer directions:*** Start S-Plus. Click on Labs, How Are Populations Distributed? In the dialog box in Lab Option, click Simulated Data. Click on each listed Distribution Family. Click Apply.]

Print your graphs so that you can examine them closely. Compare the shapes of the distributions to the values of the moments shown on the graph. Try to draw out general statements that you can make about the relationship between the values taken by the four moments and the observed distributions. Notice that the unstandardized third and fourth moments are not very informative, whereas the standardized third and fourth moments are very informative. Observe the extent to which the histograms you observe do, or do not, match the idealized distributions.

This is one of the most important exercises in this chapter. It is vital that you learn at an intuitive level the relationship between moments and the shapes of distributions.

4.23 ***Worked.*** *Objective: To relate moments to the shape of histograms.*

a. In Chapter 3, in Exercise 3.18, you generated a number of histograms from various distributions or experiments. In this exercise, we will calculate the moments from similar types of histograms. Compare your moment calculations, especially the standardized third and fourth moments, with the shapes of the histograms that you will generate in this exercise. First, generate six sets of 400 random numbers each based on a different distribution.

[***Computer directions:*** Open S-Plus. On the menu bar, click on Data, Random Numbers.

(1) In the Random Number Generation dialog, in Sample Size, key in 400. In Distribution scroll to normal (it is the first entry in the list). In the Save As box, key in Norm.den as a file name.* Uncheck Print Results. Click Apply.

(2) Follow the directions in (1). In the Distribution box scroll to LogNormal. Select a new file name.

(3) Follow the directions in (1). In the Distribution box scroll to Uniform. Select a new file name.

(4) Follow the directions in (1). In the Distribution box scroll to Chi-square. In the Deg. of Freedom 1 box, key in 5. Select a new file name.

(5) Follow the directions in (1). In the Distribution box scroll to Weibull. In the Scale box, key in 3. In the Shape 1 box, key in 2. Select a new file name.

(6) Follow the direction in (1). In the Distribution box scroll to Beta. In the Shape 1 box, key in 2. In the Shape 2 box, key in 5. Select a new file name. (*Note: To open the files created, open the Object Browser and double click on the data frame names.)]

b. For each of these six data sets, generate the histograms, give them a descriptive title, and print out the results. Calculate the first four moments, including the standardized moments.

[***Computer directions:*** Highlight the data column in the first dataset. On the menu bar, click on Graph, 2D-Plot, Histogram. Click on OK. To title the graph, on the menu bar, click Insert, Titles, Main. In the area that is highlighted on the graph, key in a title such as Normal Density: 400 Obs. Print the graph. Repeat for all six data sets.

To calculate the moments that correspond to the histograms: On the menu bar, click on Text Routines, Moments. In the dialog box, select Data Frame, Norm.den. In Variable, select "sample." Click OK. Repeat for each data set substituting the appropriate file name for Norm.den.]

Compare the variation in the moments, especially the standardized moments, with the variation in the shapes of the histograms.

c. Repeat the previous exercises with 200 observations.

d. Repeat the previous exercises with 600 observations.

e. As the number of observations increases what implications do you observe?

This exercise illustrates the point that each unique experiment has a corresponding distribution and that as the number of observations increases the smoothness and persistence of shape over repeated samplings is enhanced.

Applications

In all the following exercises, use the "Moments" function where appropriate.

4.24 In Exercise 3.23, we examined the distribution of the elderly across states; see folder Xfiles, subfolder Misc, file Age65.xls. Calculate the moments to provide more precise information about the distribution of the elderly. Are there any surprises in the results? The values that you are measuring are the percentages of the population in each state that are greater than 65 years old (*pct65*); so the mean, for example, is the mean percentage of those over 65. The mean is averaged over the percentages recorded in each state.

4.25 Using the data on weight and fat (see folder Xfiles, subfolder Misc, file Coles.xls), examine the differences that the weight loss program might have made to weight and fat content for the subjects involved. Was the program successful? How are you defining "successful"? The main question is, How different are the two distributions? Explain how you would characterize "different."

4.26 In Exercise 3.22, we examined the distributions of arms shipments by the United States and the U.S.S.R. (see folder Xfiles, subfolder Misc, file Arms.xls). Let us now consider calculating both sets of moments for the two distributions. What conclusions can we draw about the differences between U.S. and U.S.S.R. policies from examining these moments? Pay particular attention to the higher moments about the mean, as the first moment in particular is, perhaps, intuitively obvious.

4.27 In the folder Xfiles, subfolder Gnp, subfolder Intlgnp, there are three files of GNP data: Swegnp.xls, Ukgnp.xls and Usgnp.xls for Sweden, the United Kingdom, and the United States, respectively. Clearly the means and variances of these very different countries will be different, even after conversion to a common monetary unit. However, it is not clear that the shapes of the distributions of GNP components will be different. Using both box-and-whisker plots and calculations of the first four standardized moments, explore the differences in the shapes of the distributions among the three countries with respect to the three variables "priceind," "consump," and "invest." What policy implications do you draw from your analysis?

4.28 In the folder Xfiles, subfolder Testscor, a file on scholastic aptitude tests, Psatsat.xls, contains four variables: PSAT scores and SAT scores for the mathematics and verbal tests.

a. Compare the shapes of the box-and-whisker plots of the math scores between the PSAT and SAT, recognizing that the two scores are measured on different scales.

b. Compare the shapes of the verbal scores between the PSAT and SAT, recognizing that the two scores are measured on different scales. Calculate the standardized moments.

c. Similarly, compare the verbal and math scores for the PSAT in the same way.

d. What conclusions about the performance of students on these two sets of tests do you draw?

4.29 In the folder Xfiles, subfolder Misc, file Geyser1.xls are recorded the durations of and the intervals between eruptions of Old Faithful in

Yellowstone National Park. By calculating the four moments of both variables, what conclusions can you draw about the eruptions of Old Faithful? What questions are you stimulated to ask from examining these data?

4.30 Recall the discussion in Chapter 2 on IQ and height. Using the variables on IQ and height in the folder Xfiles, subfolder Misc, file Psychol.xls, explore the relevance to the interpretation of the moments of each variable. Recall that IQ has no natural origin, but that height clearly does. If you were to shift your attention to IQ differences, relative to some benchmark individual, what would be the change in your interpretation of the moments?

4.31 Explain the difference between examining the moments of the variable sales rank (*slsrank*) in the file Rock.xls in the folder Xfiles, subfolder Misc and the corresponding moments that you would observe from the actual sales figures, if they were available. (*Hint:* Recall from Chapter 2 the difference between a measurement, such as "sales" which will have units of measurements attached, and the ranking of such a variable).

4.32 Obtain the data on stock prices in NASDAQ and NYSE from the folder Xfiles, subfolder Misc, file Nyseotc.xls. Calculate the first four moments, and indicate how they help you to answer the following questions:

a. Which market do you prefer if you are a broker?

b. Which market do you prefer if you are buying stocks for your retirement?

c. Which market do you prefer if you have recieved a small inheritance, and you are already well off?

4.33 Using the data in the file Cardata.xls from the folder Xfiles, subfolder Misc, calculate the first four moments of the variables *mpg, displace, horsepower, weight,* and *price*. Given the wide disparity in the definitions of these variables and the differences in the associated units of measurement, presumably comparing the means and the second moments will be of little interest, but what of the standardized sample third and fourth moments, $\hat{\alpha}_1$ and $\hat{\alpha}_2$? What general conclusions about the characteristics of cars can you draw from a comparison of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ values across the variables that were recorded?

4.34 Case Study: Was There Age Discrimination in a Public Utility?

Along with the age discrimination charges, the plaintiff's attorney is interested in discovering evidence of differential treatment between male and female employees to enrich his case by adding gender discrimination charges. In response to this concern, you are asked to estimate the four moments of the salary for

a. all individuals in the Discdata.xls file under 40 years old by gender.

b. all individuals in the Discdata.xls file 40 years old and older by gender.

Write a short report for your supervisor explaining the practical importance of your findings for this case.

The data are stored in Xfiles in the folder Agedisc. **(See the Computer Hint for Exercise 3.30.)**

4.35 Case Study: Was There Age Discrimination in a Public Utility? The plaintiff's attorney suspects that hiring practices having a disparate impact on a protected class of people (like those of age 40 and over) can be detected by comparing the differences in the age distribution across various groups of employees or potential employees. He asked your Forensic Economics Consulting firm to compare the age distribution of older individuals (ages 40 and over) among (1) all individuals on Discdata.xls file, (2) all former employees(internal), and (3) all external applicants to the distribution for those employees that were hired within each one of these groups. You will make three sets of comparisons.

a. You are asked to estimate for your supervisor the four moments for the age variable for

(1) all individuals in the file 40 and over and among all individuals 40 and over, those who were hired

(2) all former employees (internal) 40 and over and among former employees 40 and over, those who were hired

(3) all external applicants 40 and over and among external applicants 40 and over, those who were hired.

b. You are also asked to discuss your understanding of the information provided by the data. Prepare a comprehensive statistical analysis report for your supervisor. The data are stored in Xfiles in the folder Agedisc.

[***Computer hint:*** See Exercise 3.30. Use Data, Subset and Data, Split to create needed data frames.]

4.36 Case Study: Was There Age Discrimination in a Public Utility? The labor union asked your consulting firm to investigate the presence of a "wage mobility discrimination" problem within this age discrimination case. Mobility discrimination is said to exist if certain employees are prevented from advancing, even though everyone is paid equally for equivalent work.

a. You are asked to prepare for the investigation team the following estimates for their next meeting:

(1) the four moments for salary of all individuals on the Discdata.xls file

(2) the four moments for salary of all females

(3) the four moments for salary of all males

b. Write a short, summary report with your findings. In your report indicate the limitations of your analysis. The data are stored in Xfiles in the folder Agedisc.