1.3 Graphical Summaries of Data

In the previous section we discussed numerical summaries of either a sample or a data. In this section, we look at commonly used graphical representations of both numerical data and categorical data.

Stem-and-leaf Plots

A **stem-and-leaf** plot provides a convenient way to display the distribution of a numerical data. The display consists of two parts: A column for the stems and a column for the leaves. Each number in the data set is partitioned into a stem and a leaf. A good feature of stem-and-leaf plots is that they display all the sample values. One can reconstruct the sample in its entirety from a stem-and-leaf plot.

Example 1.3.1

Construct a stem-and-leaf plot for the following set of data where the stems are the tens digits of a data and the 1's digits are the leaves.

18 19 21 22 28 29 29 32 33 38 39 40 41 56 57 64.

Solution

We have

Stem	Leaf
1	8 9
2	1 2 8 9 9
3	2 3 8 9
4	0 1
5	6 7
6	4 ■

Note that in each split stem row the leaves for heights are arranged in increasing order away from the stem (from left to right).

If you wish two compare two distributions then a **back-to-back stem-and-leaf** can be used. For this plot, the same stem is used for the leaves of both plots. We illustrate this idea in the next example.

Example 1.3.2

The following are the grades of 20 students on two different exams:

Exam 1 :61 62 62 65 66 73 75 77 77 80 81 82 84 84 85 85 87 87 92 92 92 93 96 96.

Exam 2 :61 61 62 67 67 68 70 72 75 75 78 78 81 82 82 86 87 87 88 88 91 91 91 93

Make a back-to-back stem-and-leaf plot of these data.

Solution
The back-to-back stem-and-leaf plot is shown below.

Exam1										Exam2								
				6	5	2	2	1	6	1	1	2	7	7	8			
					7	7	5	3	7	0	2	5	5	8	8			
7	7	5	5	4	4	2	1	0	8	1	2	2	6	7	7	8	8	
			6	6	3	2	2	2	9	1	1	1	3					

Dot or Line plots

A **dot plot** is a graph that can be used to give a rough impression of the shape of a sample. It is useful when the sample size is not too large and when the sample contains some repeated values. A dotplot gives a good indication of where the sample values are concentrated and where the gaps are.

Example 1.3.3
Suppose thirty people live in an apartment building. These are the following ages:

58	30	37	36	34	49	35	40	47	47
39	54	47	48	54	50	35	40	38	47
48	34	40	46	49	47	35	48	47	46

Make a dot plot of the ages.

Solution.
The line plot is given in Figure 1.3.1

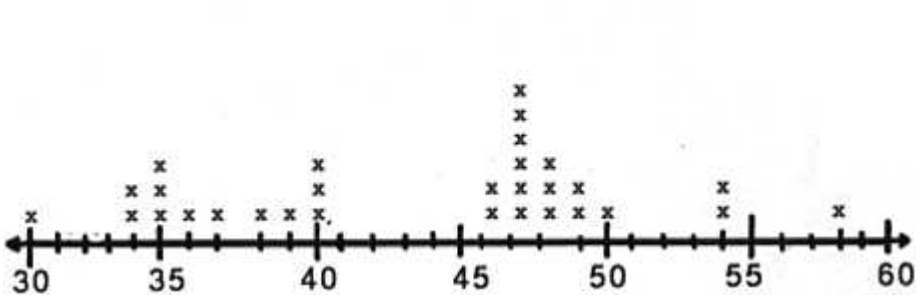


Figure 1.3.1

This graph shows all the ages of the people who live in the apartment building. It shows the youngest person is 30, and the oldest is 58. Most people in

the building are over 46 years of age. The most common age is 47 ■

Stem-and-leaf plots and dot plots are rarely used in formal presentations, however. Graphics more commonly used in formal presentations include the histogram and the boxplot, which we will next discuss.

Frequency Distributions and Histograms

When we deal with large sets of data, a good overall picture and sufficient information can be often conveyed by distributing the data into a number of **classes** or **class intervals** and to determine the number of elements belonging to each class, called **class frequency.** For instance, the following table shows some test scores from a math class.

65	91	85	76	85	87	79	93
82	75	100	70	88	78	83	59
87	69	89	54	74	89	83	80
94	67	77	92	82	70	94	84
96	98	46	70	90	96	88	72

It's hard to get a feel for this data in this format because it is unorganized. To construct a frequency distribution,

- Compute the class width $CW = \frac{\text{Largest data value} - \text{smallest data value}}{\text{Desirable number of classes}}$.

- Round CW to the next highest whole number so that the classes cover the whole data.

Thus, if we want to have 6 class intervals then $CW = \frac{100-46}{6} = 9$. The low number in each class is called the **lower class limit**, and the high number is called the **upper class limit.**

With the above information we can construct the following table called **frequency distribution**.

Class	Frequency
$40- < 50$	1
$50- < 60$	2
$60- < 70$	6
$70- < 80$	8
$80- < 90$	14
$90 - 100$	9

There is no hard-and-fast rule as to how to choose the endpoints of the class intervals. In general, it is good to have more intervals rather than fewer, but it is also good to have large numbers of sample points in the intervals. Striking the proper balance is a matter of judgment and of trial and error.
Once frequency distributions are constructed, it is usually advisable to present them graphically. The most common form of graphical representation is the **histogram**.
In a histogram, each of the classes in the frequency distribution is represented by a vertical bar whose height is the class frequency of the interval. The horizontal endpoints of each vertical bar correspond to the class endpoints.
A histogram of the math scores is given in Figure 1.3.2.

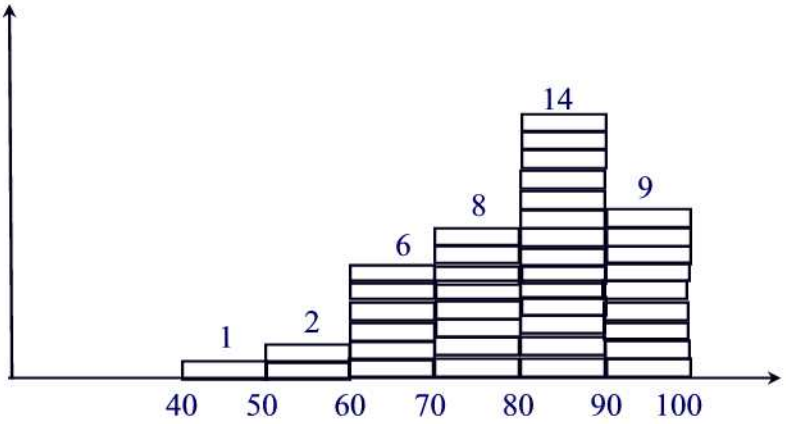


Figure 1.3.2

One advantage to the stem-and-leaf plot over the histogram is that the stem-and-leaf plot displays not only the frequency for each interval, but also displays all of the individual values within that interval.

Histograms with Unequal Class Widths
When drawing histograms it is possible that the intervals will not have the same width. To create such a histogram, we find the class **density** given by the formula

$$\text{density} = \frac{\text{relative frequency}}{\text{class width}}.$$

Then the heights of the rectangles must be set equal to the densities.

Example 1.3.4
Construct a histogram for the data given in the table below.

Class Interval (g/gal)	Frequency	Relative Frequency	Density
1–< 3	12	0.1935	0.0968
3–< 5	11	0.1774	0.0887
5–< 7	18	0.2903	0.1452
7–< 9	9	0.1452	0.0726
9–< 11	5	0.0806	0.0403
11–< 15	3	0.0484	0.0121
15–< 25	4	0.0645	0.0065

Solution.
The histogram is shown in Figure 1.3.3 ■

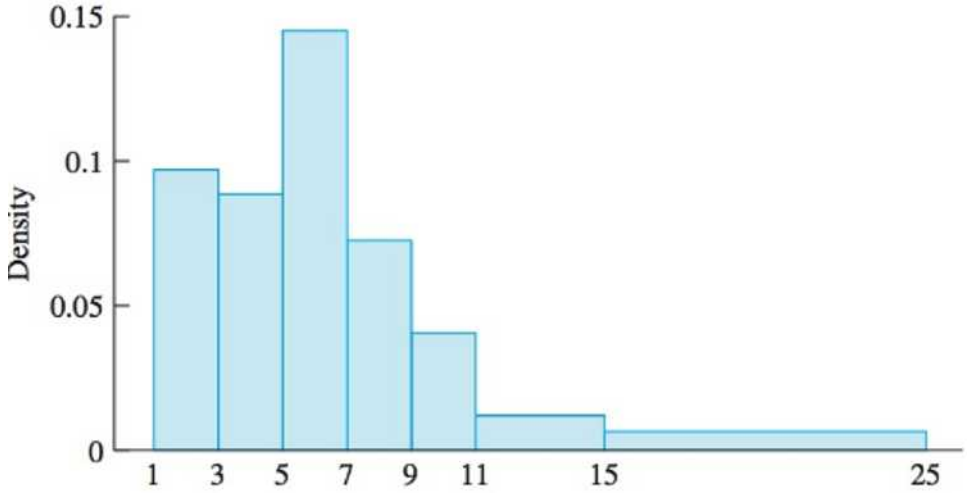


Figure 1.3.3

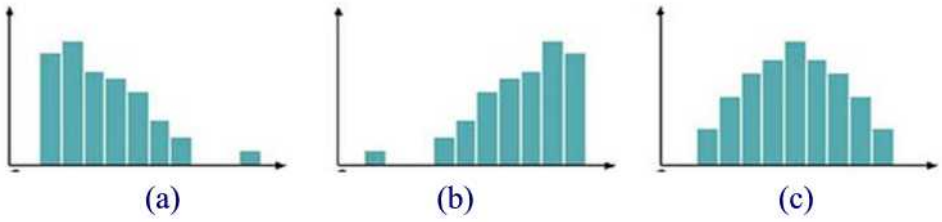
Symmetry and Skewness of a Histogram

The shape of a histogram is often described by its **symmetry** or non-symmetry (also called **skewness**). In a skewed histogram, one side, or tail, is longer than the other. A histogram with a long right-hand tail is said to be **skewed to the right**, or **positively skewed**. Positively skewed data generally have the mean to the right of the median.

A histogram with a long left-hand tail is said to be **skewed to the left**, or **negatively skewed**. Negatively skewed data generally have the mean to

the left of the median.
For a symmetric histogram, the mean and median have approximately the same value.

Example 1.3.5
Determine which histogram is symmetric, positively skewed or negatively skewed.



Solution.
(a) Positively skewed (b) Negatively skewed (c) Symmetric (or no skew) ■

Unimodal and Bimodal Histograms
Recall that the mode of a set of data is the value with the highest frequency. On a histogram, the mode is the local maximum. When a histogram has only one peak it is called **unimodal**. A histogram with two peaks is called a **bimodal**. Example of a bimodal histogram is shown in Figure 1.3.4.

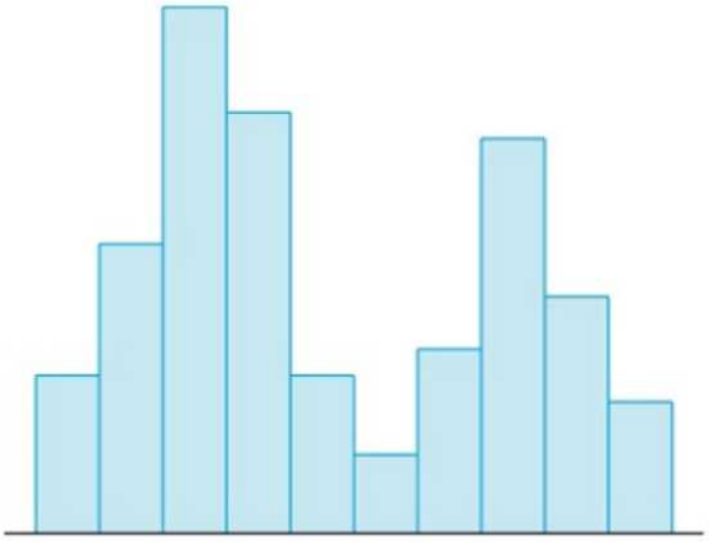


Figure 1.3.4

Boxplot

A **boxplot**, also known as a box and whisker diagram, is another graphical way of displaying the distribution of data. A boxplot is created as follows:

• Compute the median and the first and third quartiles of the sample. Indicate these with horizontal lines. Draw vertical lines to complete the box. The height of the box is called the **interquartile range** and is denoted by IQR = thirdquartile − firstquartile = $Q_3 - Q_1$. Note that since 75% of the data is less than the third quartile, and 25% of the data is less than the first quartile, it follows that 50%, or half, of the data are between the first and third quartiles. The interquartile range is therefore the distance needed to span the middle half of the data.

• Find the largest sample value that is less than or equal to $Q_3 + 1.5$IQR, and the smallest sample value that is greater than or equal to $Q_1 - 1.5$IQR. Extend vertical lines (**whiskers**) from the quartile lines to these points.

• Points whose numerical values are greater than $Q_3 + 1.5$IQR or less than $Q_1 - 1.5$IQR, are designated as outliers. Plot each outlier individually.

Figure 1.3.5 illustrates the sketch of a boxplot.

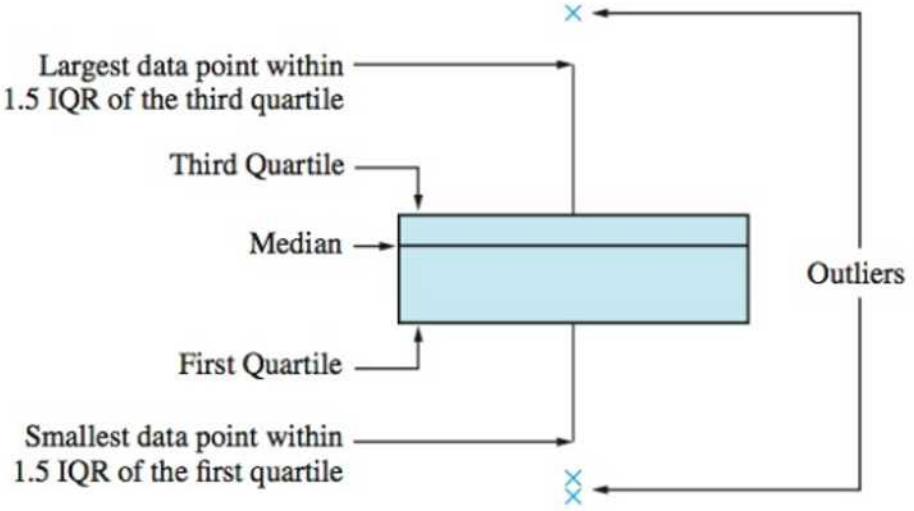


Figure 1.3.5

Apart from any outliers, a boxplot can be thought of as having four pieces: the two parts of the box separated by the median line, and the two whiskers. Again apart from outliers, each of these four parts represents one-quarter of the data. The boxplot therefore indicates how large an interval is spanned

by each quarter of the data, and in this way it can be used to determine the regions in which the sample values are more densely crowded and the regions in which they are more sparse.

Example 1.3.6
Below is a random sample of 20 concentrations of a chemical in milligrams per liter.

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

Construct the boxplot of these data.

Solution.
In this case, the interquartile range is IQR $= 132.7 - 130.35 = 2.35$. Therefore, the lower limit is calculated as $Q_1 - 1.5\text{IQR} = 130.35 - 1.5(2.35) = 126.825$. Therefore, the lower adjacent value is the same as the minimum value, 127.8, because 127.8 is lowest observation still inside the region defined by the lower bound of 126.825. The upper limit is calculated as $Q_3 + 1.5\text{IQR} = 132.7 + 1.5(2.35) = 136.225$. Therefore, the upper adjacent value is the same as the maximum value, 134.8, because 134.8 is the highest observation still inside the region defined by the upper bound of 136.225. The boxplot is shown in Figure 1.3.6 ■

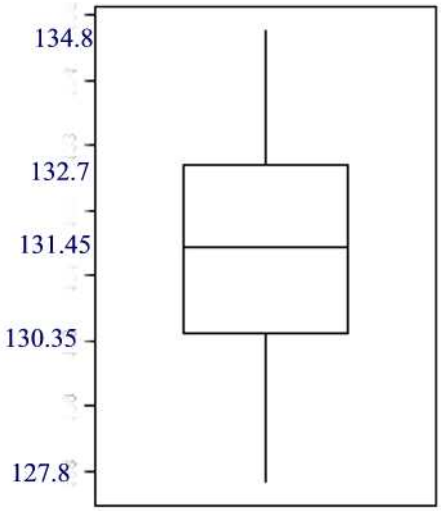


Figure 1.3.6

Scatterplot

Data for which each item consists of more than one value are called **multivariate data**. When each item is a pair of values, the data are said to be **bivariate**. One of the most useful graphical summaries for numerical bivariate data is the **scatterplot**. In the simple case where each item is in the form (x, y), a scatterplot is obtained by simply plotting the points in the Cartesian coordinate system.

Example 1.3.7

Create a scatter plot of for the following data:

x	1.4	2.4	4.0	4.9	5.7	6.3	7.8	9.0	9.3	11.0
y	2.3	3.7	5.7	9.9	6.9	15.8	15.4	36.9	34.6	53.2

Solution.

The scatterplot is shown in Figure 1.3.7 ■

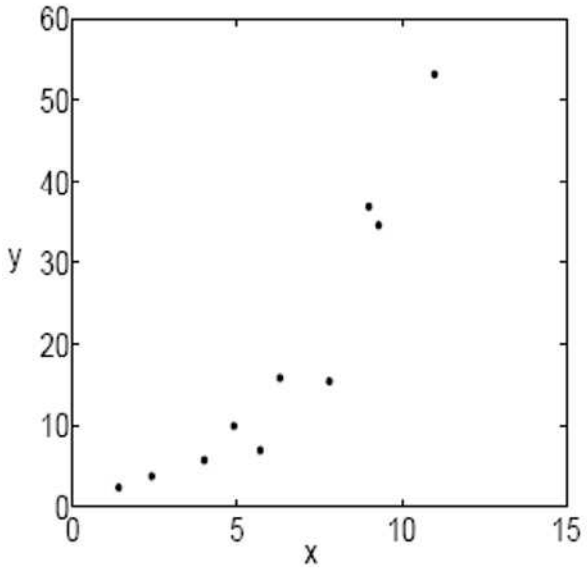


Figure 1.3.7