Density estimators join stemplots and histograms as useful graphical tools for exploratory data analysis.

Density estimates can capture other unusual features of a distribution. Here is an example.

EXAMPLE

DATA FILE STUBHUB

1.45 StubHub! StubHub! is a Web site where fans can buy and sell tickets to sporting events. Ticket holders wanting to sell their tickets provide the location of their seats and the selling price. People wanting to buy tickets can choose from among the tickets offered for a given event.[34]

On Saturday October 18, 2008, the eleventh-ranked Missouri football team played number one Texas in Austin. On Thursday October 16, 2008, StubHub! listed 64 pairs of tickets for the game. One pair was offered at $883 per ticket. It was noted that these seats were in a suite and that food and bar were included. We discarded this outlier and examined the distribution of the price per ticket for the remaining 63 pairs of tickets. The histogram with a density estimate is given in Figure 1.33. The distribution has two peaks, one around $160 and another around $360. This is the identifying characteristic of a **bimodal distribution.** Since the stadium has upper and lower level seats, we suspect that differences in prices between these two types of seats is responsible for the two peaks. (Texas won 56 to 31.)

bimodal distribution

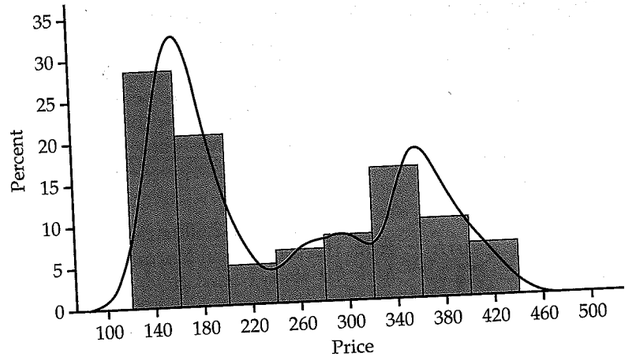

FIGURE 1.33 Histogram of StubHub! price per seat for tickets to the Missouri–Texas football game on October 18, 2008, with a density estimate, for Example 1.45. One outlier with a price per seat of $883 was deleted.

The previous exercise reminds of a continuing theme for data analysis. We looked at a histogram and a density estimate and saw something interesting. This led us to speculation. Additional data on the type and location of the seats may explain more about the prices than we see in Figure 1.33.

SECTION 1.3 Summary

The overall pattern of a distribution can often be described compactly by a **density curve.** A density curve has total area 1 underneath it. Areas under a density curve give proportions of observations for the distribution.

The **mean** μ (balance point), the **median** (equal-areas point), and the **quartiles** can be approximately located by eye on a density curve. The **standard deviation** σ cannot be located by eye on most density curves. The mean and median are equal for symmetric density curves, but the mean of a skewed curve is located farther toward the long tail than is the median.

The **Normal distributions** are described by bell-shaped, symmetric, unimodal density curves. The mean μ and standard deviation σ completely specify the Normal distribution $N(\mu, \sigma)$. The mean is the center of symmetry, and σ is the distance from μ to the change-of-curvature points on either side.

To **standardize** any observation x, subtract the mean of the distribution and then divide by the standard deviation. The resulting **z-score** $z = (x - \mu)/\sigma$ says how many standard deviations x lies from the distribution mean. All Normal distributions are the same when measurements are transformed to the standardized scale. In particular, all Normal distributions satisfy the **68–95–99.7 rule.**

If X has the $N(\mu, \sigma)$ distribution, then the standardized variable $Z = (X - \mu)/\sigma$ has the **standard Normal distribution** $N(0, 1)$. Proportions for any Normal distribution can be calculated by software or from the **standard Normal table** (Table A), which gives the **cumulative proportions** of $Z < z$ for many values of z.

The adequacy of a Normal model for describing a distribution of data is best assessed by a **Normal quantile plot,** which is available in most statistical software packages. A pattern on such a plot that deviates substantially from a straight line indicates that the data are not Normal.

SECTION 1.3 Exercises

For Exercises 1.101 and 1.02, see page 57; for Exercises 1.103 and 1.104, see page 58; for Exercises 1.105 and 1.106, see page 63; and for Exercises 1.107 and 1.108, see page 64.

1.109 Sketch some normal curves.

(a) Sketch a normal curve that has mean 10 and standard deviation 3.

(b) On the same x axis, sketch a normal curve that has mean 20 and standard deviation 3.

(c) How does the normal curve change when the mean is varied but the standard deviation stays the same?

1.110 The effect of changing the standard deviation.

(a) Sketch a normal curve that has mean 10 and standard deviation 3.

(b) On the same x axis, sketch a normal curve that has mean 10 and standard deviation 1.

(c) How does the normal curve change when the standard deviation is varied but the mean stays the same?

1.111 Know your density. Sketch density curves that might describe distributions with the following shapes:

(a) Symmetric, but with two peaks (that is, two strong clusters of observations).

(b) Single peak and skewed to the left.

1.112 Do women talk more? Conventional wisdom suggests that women are more talkative than men. One study designed to examine this stereotype collected data on the speech of 42 women and 37 men in the United States.[35] TALK

(a) The mean number of words spoken per day by the women was 14,297 with a standard deviation of 6441. Use the 68–95–99.7 rule to describe this distribution.

(b) Do you think that applying the rule in this situation is reasonable? Explain your answer.

(c) The men averaged 14,060 words per day with a standard deviation of 9065. Answer the questions in parts (a) and (b) for the men.

(d) Do you think that the data support the conventional wisdom? Explain your answer. Note that in Section 7.2 we will learn formal statistical methods to answer this type of question.

1.113 Data from Mexico. Refer to the previous exercise. A similar study in Mexico was conducted with

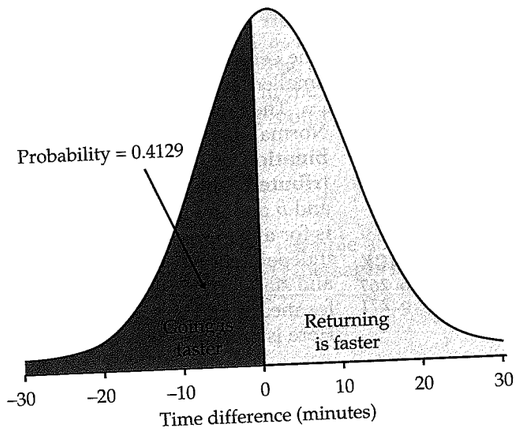

FIGURE 5.6 The Normal probability calculation for Example 5.9. The difference in times going to campus and returning from campus ($X - Y$) is Normal with mean 2 minutes and standard deviation 8.94 minutes.

Normal approximation for counts and proportions. This Normal approximation is just an example of the central limit theorem applied to these discrete random variables.

SECTION 5.1 Summary

The **sample mean** $\bar{x}$ of an SRS of size n drawn from a large population with mean μ and standard deviation σ has a sampling distribution with mean and standard deviation

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The sample mean $\bar{x}$ is therefore an unbiased estimator of the population mean μ and is less variable than a single observation.

Linear combinations of independent Normal random variables have Normal distributions. In particular, if the population has a Normal distribution, so does $\bar{x}$.

The **central limit theorem** states that for large n the sampling distribution of $\bar{x}$ is approximately $N(\mu, \sigma/\sqrt{n})$ for any population with mean μ and finite standard deviation σ.

SECTION 5.1 Exercises

For Exercise 5.1, see page 299; for Exercises 5.2 and 5.3, see pag 302; for Exercise 5.4, see page 303; for Exercise 5.5, see page 304; and for Exercise 5.6, see page 306.

5.7 What is wrong? Explain what is wrong in each of the following scenarios.

(a) If the variance of a population is 10, then the variance of the mean for an SRS of 30 observations from this population will be $10/\sqrt{30}$.

(b) When taking SRS's from a population, larger sample sizes will result in larger standard deviations of the sample mean.

(c) The mean of a sampling distribution of $\bar{x}$ changes when the sample size changes.

5.8 What is wrong? Explain what is wrong in each of the following statements.

(a) For large n, the distribution of observed values will be approximately Normal.

(b) The 68–95–99.7 rule says that $\bar{x}$ should be within $\mu \pm 2\sigma$ about 95% of the time.

(c) The central limit theorem states that for large n, μ is approximately Normal.

5.9 Generating a sampling distribution. Let's illustrate the idea of a sampling distribution in the case of a very small sample from a very small population. The population is the 10 scholarship players currently on your men's basketball team. For convenience, the 10 players have been labeled with the integers 0 to 9. For each player, the total amount of time spent (in minutes) on Facebook during the last month is recorded in the table below.

Player	0	1	2	3	4	5	6	7	8	9
Total Time (min)	370	290	358	366	323	319	358	309	327	368

The parameter of interest is the average amount of time on Facebook. The sample is an SRS of size $n = 3$ drawn from this population of players. Because the players are labeled 0 to 9, a single random digit from Table B chooses one player for the sample.

(a) Find the mean of the 10 players in the population. This is the population mean μ.

(b) Use Table B to draw an SRS of size 3 from this population (Note: you may sample the same player's time more than once). Write down the three times in your sample and calculate the sample mean $\bar{x}$. This statistic is an estimate of μ.

(c) Repeat this process 10 times using different parts of Table B. Make a histogram of the 10 values of $\bar{x}$. You are constructing the sampling distribution of $\bar{x}$.

(d) Is the center of your histogram close to μ? Would it get closer to μ the more times you repeated this sampling process? Explain.

5.10 Total sleep time of college students. In Example 5.1, the total sleep time per night among college students was approximately Normally distributed with mean $\mu = 7.02$ hours and standard deviation $\sigma = 1.15$ hours. Suppose you plan to take an SRS of size $n = 200$ and compute the average total sleep time.

(a) What is the standard deviation for the average time?

(b) Use the 95 part of the 68–95–99.7 rule to describe the variability of this sample mean.

(c) What is the probability that your average will be below 6.9 hours?

5.11 Determining sample size. Recall the previous exercise. Suppose you want to use a sample size such that about 95% of the averages fall within ±5 minutes of the true mean $\mu = 7.02$.

(a) Based on your answer to part (b) in Exercise 5.8, should the sample size be larger or smaller than 200? Explain.

(b) What standard deviation of the average do you need such that about 95% of all samples will have a mean within 5 minutes of μ?

(c) Using the standard deviation calculated in part (b), determine the number of students you need to sample.

5.12 Songs on an iPod. An iPod has about 10,000 songs. The distribution of the play time for these songs is highly skewed. Assume that the standard deviation for the population is 280 seconds.

(a) What is the standard deviation of the average time when you take an SRS of 10 songs from this population?

(b) How many songs would you need to sample if you wanted the standard deviation of $\bar{x}$ to be 15 seconds?

5.13 Bottling an energy drink. A bottling company uses a filling machine to fill cans with an energy drink. The cans are supposed to contain 250 milliliters (ml). The machine, however, has some variability, so the standard deviation of the size is $\sigma = 3$ ml. A sample of 6 cans is inspected each hour for process control purposes, and records are kept of the sample mean volume. If the process mean is exactly equal to the target value, what will be the mean and standard deviation of the numbers recorded?

5.14 Play times for songs on an iPod. Averages of several measurements are less variable than individual measurements. Suppose the true mean duration of the play time for the songs in the iPod of Exercise 5.12 is 350 seconds.

(a) Sketch on the same graph the two Normal curves, for sampling a single song and for the mean of 10 songs.

(b) What is the probability that the sample mean differs from the population mean by more than 19 seconds when only 1 song is sampled?

(c) How does the probability that you calculated in part (b) change for the mean of an SRS of 10 songs?

5.15 Can volumes. Averages are less variable than individual observations. Suppose that the can volumes in Exercise 5.13 vary according to a Normal distribution. In that case, the mean $\bar{x}$ of an SRS of cans also has a Normal distribution.