CHAPTER 2

Fundamentals of Statistics

This chapter is a brief review for readers with some prior experience with quantitative analysis of data. Readers without such experience, or those who prefer more thorough coverage of the material, may refer to the textbooks by Devore (2004) or Mendenhall et al. (2006).

2.1 STATISTICAL THINKING

Statistics is a branch of mathematics, but it is not an axiomatic science as are many other of its branches (where facts are concluded from predetermined axioms). In statistics, the translation of reality to a statistical problem is a mix of art and science, and there are often many possible solutions, each with a variety of possible interpretations.

The science of statistics can be divided into two major branches—descriptive statistics and inferential statistics. Descriptive statistics describes samples or populations by using numerical summaries or graphs. No probabilistic models are needed for descriptive statistics. On the other hand, in inferential statistics, we draw conclusions about a population based on a sample. Here we build a probabilistic model describing the population of interest, and then draw information about the model from the sample. When analyzing data, we often start with descriptive statistics, but most practical applications will require the use of inferential statistics. This book is primarily about inferential statistics.

In Chapter 1, we emphasized that variability is everywhere, and we need to utilize statistical thinking to deal with it. In order to assess the variability, we first need to define precisely what we are trying to measure, or observe. We can then collect the data and analyze them. Let us describe that process, and on the way, introduce definitions of some important concepts in statistics.

Statistics for Imaging, Optics, and Photonics, Peter Bajorski.

Definition 2.1. A *measurement* is a value that is observed or measured.

Definition 2.2. An *experimental unit* is an object on which a measurement is obtained.

Definition 2.3. A *population* is often defined as a set of experimental units of interest to the investigator. Sometimes, we take repeated measurements of one characteristic of a single experimental unit. In that case, a *population* would be a set of all such possible measurements of that experimental unit, both the actual measurements taken and those that can be taken hypothetically in the future.

Definition 2.4. A *sample* is a subset selected from the population of interest.

When designing a study, one should specify the population that addresses the question of interest. For example, when investigating the color of nominally red plastic part #ACME-454, we could define a population of experimental units as all parts #ACME-454 produced in the past and those that will be produced in the future at a given plant of ACME Labs.

We can say that this population is hypothetical because it includes objects not existing at the time. It is often convenient to think that the population is infinite. This approach is especially useful when dealing with repeated measurements of the same object. Infinite populations are also used as approximations of populations consisting of a large number of experimental units. As you can see, defining a population is not always exact science.

Once we know the population of interest, we can identify a suitable sampling method, which describes how the sample will be selected from the population. Our goal is to make the sample to be representative of the population, that is, it should look like the population, except for being smaller. The closer we get to this ideal, the more precise are our conclusions from the sample to the population. There are whole books describing how to select samples (see Thompson (2002), Lohr (2009), Scheaffer et al. (2011), and Levy and Lemeshow (2009)).

If a data set was given to you, you need to find out how the data were collected, so that you can identify the population it represents. The less we know about the sampling procedure used, the less useful the sample is. In extreme cases, it might be prudent to use the old adage "garbage in–garbage out," and try to collect new data instead of using unreliable data.

Let's say, you were given data on color measurements of 10 parts #ACME-454 that were taken from the current production process. However, there is no information about the process of selecting the 10 parts. They all might have been taken from one batch produced within 1 h or each part might have been produced on a different day. They could also be rejects from the process. In this case, it would be more productive to design a new study of those parts in order to collect new data.

The purpose of this section is to give the reader a general overview of the principles of statistical thinking and a sense of the nuances associated with statistics. If reading it

led you to having even more questions than you started with, then continue to the following sections and chapters, where you will find many answers.

2.2 DATA FORMAT

Data are often organized in a way that is convenient for data collection. In order to implement statistical thinking and better understand the data, we usually find it convenient to organize the data into the format of a traditional statistical database. The format consists of a spreadsheet, where *observations* are placed in rows and *variables* are placed in columns. Example 2.1 illustrates this traditional formatting technique.

Example 2.1 Optical fibers permit transmission of signals over longer distances and at higher bandwidths than other forms of communication. An experiment was performed in order to find out how much power is lost when sending signals through optical fiber. Five pieces of 100 m length of optical fiber were tested. A laser light signal was sent from one end through each piece of optical fiber, and the output power was measured at the other end. The power of the laser source was 80 mW. The results are shown in Table 2.1, where each row represents a set of results for a single piece of optical fiber. Each unique optical fiber is identified by a number recorded in the first column of the table. The remaining columns contain the variables from the experiment. The Input Power (P_{in}) is the nominal value of 80 mW, which is the same for all observations. The Output Power (P_{out}) given in the next column is a quantity that was measured in the experiment. The Power Loss $\left(L_{\text{power}}\right)$ in the last column was calculated in decibels (dB) according to the following formula:

$$\text{Power Loss (dB)} = 10 \log_{10} \frac{\text{Output Power}}{\text{Input Power}}. \tag{2.1}$$

Organized in this way, the data are easily analyzed. For a small data set like this one, we can often draw some conclusions directly from the table, but for larger data sets, we will need some summary statistics and graphs to understand the data.

Since the Power Loss is calculated from the Output Power (with constant Input Power), the two variables convey the same information (within this data set). So, if we

Table 2.1 Experimental Results on Five Pieces of Optical Fiber

Optical Fiber Number	Input Power (mW)	Output Power (mW)	Power Loss (dB)
1	80	72.8	−0.4096
2	80	70.0	−0.5799
3	80	72.0	−0.4576
4	80	68.8	−0.6550
5	80	73.6	−0.3621

Negative dB means that there is loss of power.

are trying to characterize a typical fiber based on the five pieces, which of the two variables should we use? This question will be addressed in the next section on descriptive statistics. □

The data are not always as neatly organized as those in Table 2.1. At the same time, it is not always necessary to have an actual statistical database in the Table 2.1 format. However, in the process of statistical thinking, we want to identify what the observations and variables are in a given context, since this will be crucial in our statistical analysis.

2.3 DESCRIPTIVE STATISTICS

When dealing with data, especially with large amounts of data, we find it useful to summarize them with some appropriately chosen summary (or descriptive) statistics. We will now concentrate on the values of one variable and will denote the n observations of that variable by $x_1, x_2, \ldots, x_n$. Note that the subscript index does not imply any particular order in those values. The first step in understanding the data is to describe the magnitude of the observations. When we think of data as numbers on the number axis, the magnitude will tell us a general location of the data on the axis. In the following subsection, we discuss various statistics for describing the data location.

2.3.1 Measures of Location

The most popular descriptive statistic is the *sample mean* defined by

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.2}$$

which describes the general (on average) location of the data. One appealing property of the sample mean is a physical property that it is the balance point for a system of equal weights placed at the points x_i, $i = 1, \ldots, n$, on the number axis. Figure 2.1 shows an example of five data points with equal weights, which are balanced at the $\overline{x}$ point.

Example 2.1 (cont.) For the data in Table 2.1, we can calculate the sample means of all three variables. For the Input Power variable, we get its sample mean $\overline{P}_{\text{in}} = 80\,\text{mW}$, of course. For Output Power, we obtain $\overline{P}_{\text{out}} = 71.44\,\text{mW}$, and for the Power Loss, $\overline{L}_{\text{power}} = -0.4928$. The means are supposed to represent a typical or an average optical fiber. Let us assume that an optical fiber regarded as average has the

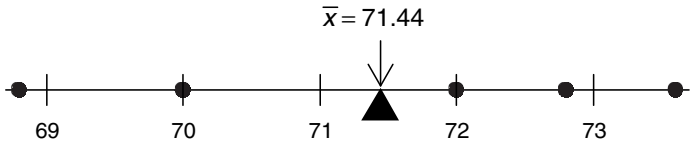

Figure 2.1 Five Output Power values balanced at the sample mean point (see Example 2.1).

Output Power value of $P_{\text{out}} = 71.44$ mW, that is, the same as the previously calculated mean. According to formula (2.1), its power loss would be described as -0.4915 dB, which is different from the previously calculated average Power Loss of $\overline{L}_{\text{power}} = -0.4928$. The question is which of the two values should be regarded as a typical power loss value. There is an easy mathematical explanation for why the two numbers differ. Let us say that a variable y is calculated as a function of another variable x, that is, $y = f(x)$. In this case, Power Loss is calculated as a function of Output Power. This means that for observations x_i, $i = 1, \ldots, n$, we have $y_i = f(x_i)$, $i = 1, \ldots, n$. What we have just observed in our calculations simply means that $\overline{y} \neq f(\overline{x})$. In other words, a transformation of the mean is not necessarily the same as the mean of the transformed values. A special case is when the function f is linear, and we do get an equality $\overline{y} = f(\overline{x})$, that is, for $y_i = ax_i + b$, we have $\overline{y} = a\overline{x} + b$.

Despite the above explanation, we still do not know which of the two power loss values we should regard as typical for the type of optical fiber used in the experiment. The answer will depend on how such a number would be used. Here we give two possible interpretations. If the five measurements were performed on the same piece of optical fiber, then the sample mean $\overline{P}_{\text{out}}$ would estimate the "true" output power of the fiber. The true power loss for that fiber should then be calculated as $10 \log_{10}\left(\overline{P}_{\text{out}}/80\right) = -0.4915$ dB. An alternative scenario would be when the five different pieces tested in the experiment represent an optical fiber used in an existing communication network, and we are trying to characterize a typical network power loss (over 100 m). In this case, it would be more appropriate to use the value of $\overline{L}_{\text{power}} = -0.4928$. To understand this point, imagine the five pieces being connected into one 500 m optical fiber. Its power loss would then be calculated as the sum of the five power loss values in Table 2.1, resulting in the total power loss of -2.4642 dB. The same value (up to the round-off error) can be obtained by multiplying the typical value of $\overline{L}_{\text{power}} = -0.4928$ by 5.

We now need to introduce the concept of ordered statistics. Let's say we have n observations x_i, $i = 1, \ldots, n$, of a given variable. We order those numbers from the smallest to the largest, and call the smallest one the value of the first-order statistic denoted by $x_{(1)}$. The second smallest value becomes the second-order statistic denoted by $x_{(2)}$, and so on until the largest value becomes the nth-order statistic denoted by $x_{(n)}$. We can now introduce the *sample median*, which is the middle value in the data set defined as

$$\tilde{x} = \begin{cases} x_{(k)} & \text{for odd } n = 2k-1, \\ \left(x_{(k)} + x_{(k+1)}\right)/2 & \text{for even } n = 2k. \end{cases} \tag{2.3}$$

In Example 2.1, $n = 5$ is odd, hence $k = 3$, and for the Output Power variable, we have $\tilde{x} = x_{(3)} = 72$. The sample median is called a robust statistic because it is not impacted by unusual observations called outliers. It is also useful for skewed data, where the mean is pulled away from the bulk of data because of being influenced by a few large values. Figure 2.2 shows an example where the bulk of the data is in the range between 0 and 2, but the sample mean is above 2 because of two outliers.

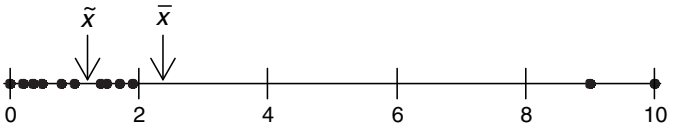

Figure 2.2 A data set skewed to the right due to two outliers. The sample mean does not represent the bulk of data as well as the sample median does.

The sample median can be regarded as too robust in the sense that it depends only on the ordered statistics in the middle of the data. As a compromise between the mean and the median, we can define a trimmed mean, where a certain percent of the lowest and highest values are removed, and the mean is calculated from the remaining values. Note that the median is an extreme case of the trimmed mean, where the same number of the lowest and highest values are removed until only one or two observations are left.

The sample median divides the data set into two halves. For a more detailed description of the data distribution, we can divide data into one hundred parts and describe the position (or location) of each part. To this end, we can define a *sample*$(100p)$*th percentile*, where p is a fraction $(0 \leq p \leq 1)$, as a number x such that approximately $(100p)\%$ of data is below x and the remaining $(100(1-p))\%$ of data is above x. A $(100p)$th percentile is also called a pth quantile. Percentiles are often used in reporting results of standardized tests, because they tell us how a person performed in relation to all other test takers. Of course, it is not always possible to divide the data into an arbitrary fraction, so we need a more formal definition. We first assign the kth-order statistic $x_{(k)}$ as the $(k-1)/(n-1)$ quantile. When a different-level quantile is needed, it is interpolated from the two nearest quantiles previously calculated as the ordered statistics. The sample percentiles are best calculated for large samples, but here we give an educational example for the five observations of the Output Power variable in Example 2.1. For $n = 5$, the five ordered statistics are assigned as 0th, 25th, 50th, 75th, and 100th percentiles. A 90th percentile is calculated by a linear interpolation as the weighted average of the two ordered statistics, that is,

$$\frac{100-90}{100-75}x_{(4)} + \frac{90-75}{100-75}x_{(5)}, \tag{2.4}$$

which gives 73.28 for the Output Power variable (given as Problem 2.1). There are many other ways of calculating percentiles, and the best way may depend on the context of data. For large n, all methods give similar results.

It is easy to see that the sample median is the 50th percentile. We also define the *first* and *third quartiles* as the 25th and 75th percentiles, respectively. The two quartiles together with the median, which is also the second quartile, divide the data set into four parts with approximately even counts of points.

2.3.2 Measures of Variability

In the previous subsection, we discussed the location aspect of data. Another important feature of data is their variability. The simplest measure of variability is

the *range*, which is defined as the difference between the maximum and minimum values, that is, $x_{(n)}-x_{(1)}$ for a sample of size n. A significant disadvantage of the range is its dependence on the two most extreme observations, which makes it sensitive to outliers.

A different way to describe variability is to use deviations from a central point, such as the mean. The deviations from the mean, defined as $d_i = x_i-\overline{x}$, have the property that they sum up to zero (see Problem 2.2). Hence, the measures of variability typically consider magnitudes of deviations and ignore their signs. The most popular measures of variability are the *sample variance* defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} d_i^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i-\overline{x})^2 \tag{2.5}$$

and the associated *sample standard deviation* defined as $s = \sqrt{s^2}$. They both convey the equivalent information, but the advantage of the standard deviation is that it is expressed in the units of the original observations, while the variance is in squared units, which are difficult to interpret.

Let us now consider a linear transformation of x_i defined as $y_i = ax_i + b$ for $i = 1, \ldots, n$. Using some algebra, one can check that the sample variance of the transformed data is equal to $s_y^2 = a^2 s_x^2$ and the sample standard deviation is $s_y = |a|s_x$ (see Problem 2.3). This means that both statistics are not impacted by a shift in data, and scaling of data by a positive constant results in the same scaling of the sample standard deviation.

Another measure of variability is the *interquartile range* (IQR), defined as the difference between the third and first quartiles, which is the range covering the middle 50% of the data.

2.4 DATA VISUALIZATION

We all know that a picture is worth a thousand words. In the statistical context, it means that valuable information can be extracted from graphs representing data—information that might be difficult to notice and convey when reporting only numbers. For an efficient graphical presentation, it is important that the maximum amount of information is conveyed with the minimum amount of ink. This allows representations of large data sets and at the same time keeps the graphs clear and easy to interpret. This concept has been popularized by Tufte (2001), who used the information-to-ink ratio as a measure of graph efficiency. In those terms, bar charts and pie charts are very inefficient, and indeed they are of very little value in data analysis.

2.4.1 Dot Plots

One of the simplest graphs is a dot plot, where one dot represents one observation, and one axis (such as the horizontal axis as in Figure 2.3) is devoted to showing the range

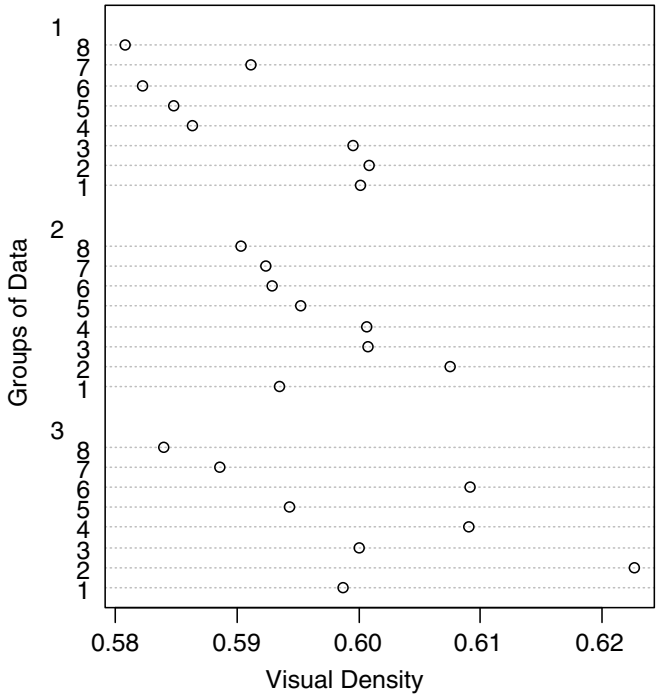

Figure 2.3 Dot plot for Visual Density of eight patches of cyan printed on three different pages (groups).

of values. The second axis may not be used at all (with all dots lined up along a horizontal line), or it can be used to show additional information such as grouping of observations, or their order. One advantage of a dot plot is that it can be created in any software program capable of plotting dots in a system of coordinates.

Example 2.2 As part of a printing experiment described in Appendix B, three pages were printed with an identical pattern of color patches, such as the one shown in Figure 1.3 in the context of Example 1.2. On each page, there were eight patches of cyan (at maximum gradation, or amount, of the cyan ink). For each patch, Visual Density was measured as a quality control metric. Figure 2.3 shows a dot plot of Visual Density for the three pages as three groups. The horizontal lines within each group represent eight patches. The three groups of data (as pages) seem to be somewhat different, but it is unclear if the differences could have happened by chance or if they manifest a real difference. No real difference would be good news because it would mean consistent printing from page to page. This question would need to be addressed by statistical inference discussed in Chapters 3 and 4.

In Figure 2.3, we may have an impression of a slanted shape of points within each group, where the patches with a higher identification number tend to give lower densities. This suggests a possible pattern from patch to patch. In order to test this hypothesis, we can group data into eight groups (for eight patches) of three observations each and create a dot plot with patches as groups. In that case, the number of groups is fairly large, and it makes sense to use a different version of a dot plot, where each group is plotted along one horizontal line as in Figure 2.4. We can

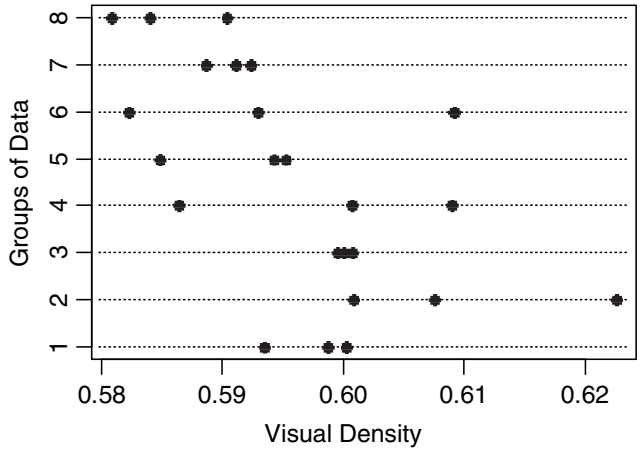

Figure 2.4 Dot plot for Visual Density of eight patches of cyan (as groups) printed on three different pages.

now see that Patches 5, 7, and 8 tend to have lower Visual Density values than some other patches, especially Patch 2. Since we have only three observations per patch, it is unclear if this effect is incidental, or if there is a real systematic difference among patches. Again, this question needs to be answered with some formal statistical methods that will be discussed in Chapter 3. □

2.4.2 Histograms

Dot plots are convenient for small to medium-sized data sets. For large data sets, we start getting significant overlap of dots, which can be dealt with by stacking the points, but this requires extra programming or a specialized function. Also, it becomes difficult to assess the shape of the distribution with too many points. In those cases, we can use a *histogram*, which resembles a bar chart, except that the bars represent adjacent bins or subintervals of equal length defined within the range of given data. For example, the histogram in Figure 2.5 uses bins of width 0.05. The tallest bar represents the bin from 0 to 0.05, the next bin to the right is from 0.05 to 0.1, and so on. The height of the bar shows the number of points (frequency) in the bins. In this

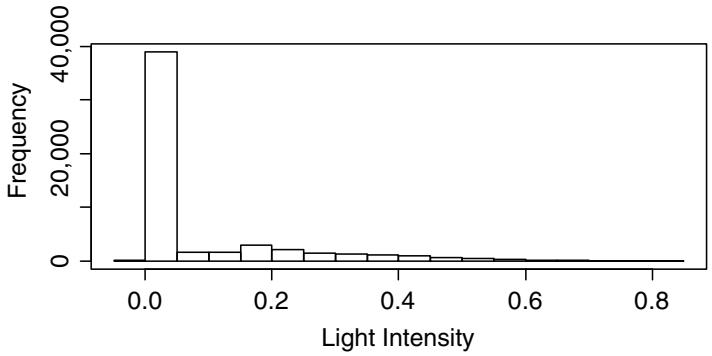

Figure 2.5 A histogram of the Light Intensity values from an image of a fish as used in Example 2.3.

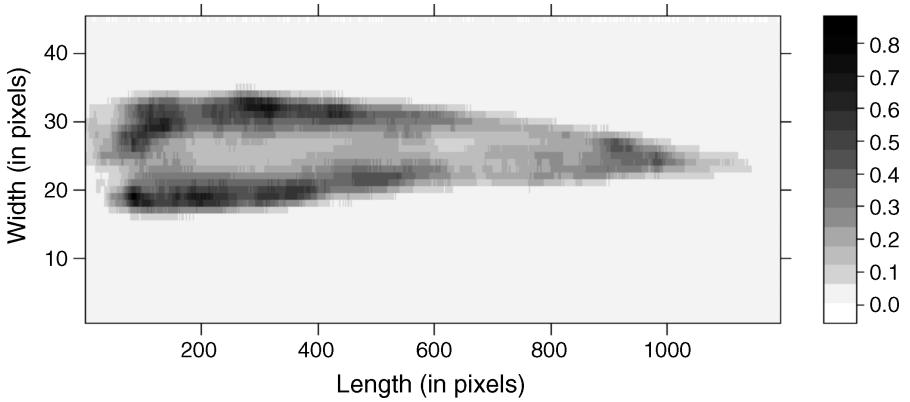

Figure 2.6 Light Intensity values from an image of a fish as used in Example 2.3.

example, there are almost 40,000 observations in the bin from 0 to 0.05. The bins in a histogram are adjacent with no gaps between them. Consequently, there are usually no gaps between the bars. If there is a gap in the bars, it means that the respective bin had zero frequency and was not plotted (or had zero height). In very large data sets, the height of a bar might be larger than zero but still be so small (in relation to the vertical scale of frequencies) that the bar is not visible.

Example 2.3 Consider Fish Image data set representing an image of a fish on a conveyer belt as explained in Appendix B. The average transflected Light Intensity over 15 image channels was calculated for each image pixel and plotted in Figure 2.6. We use a convention that higher values are shown in darker colors. This produces better displays in most cases than the traditional approach in imaging to use white for the highest values. Using white for largest values may seem logical from the point of view of color management, but it usually produces poor quality displays.

There are 45 pixels along the width of the conveyer belt and 1194 pixels along its length, for a total of 53,730 pixels. In a paper by Wold et al. (2006), a threshold on the Light Intensity was used to distinguish between the fish and non-fish pixels, but no details were provided as to the process of selecting the threshold. In order to determine the threshold, it is helpful to perform exploratory analysis of the data. To this end, we can create a histogram of all 53,730 Light Intensity values as shown in Figure 2.5, so that we can look for a natural cutoff point between the two sets of pixels. Unfortunately, that histogram is not very useful because the majority of observations fall into one bin, and then not much can be seen in the remaining bins. This is partially because of the scaling of the vertical axis being dictated by the very high frequency for that one bin. It turns out that the largest Light Intensity is above 0.82, and as many as 33 values are above 0.7. Yet, one cannot see any frequency bars above 0.7. The reason has been discussed earlier. The resulting height of the bar is too small to be seen. It also turns out that 182 values are exactly zero, and they were included in the first (tiny) bar on the left.

One way to improve the histogram in Figure 2.5 is to use a logarithmic scale. To this end, we calculated a logarithm to base 10 of all positive values and created a histogram shown in Figure 2.7. A larger number of bins were used, so that finer details of the

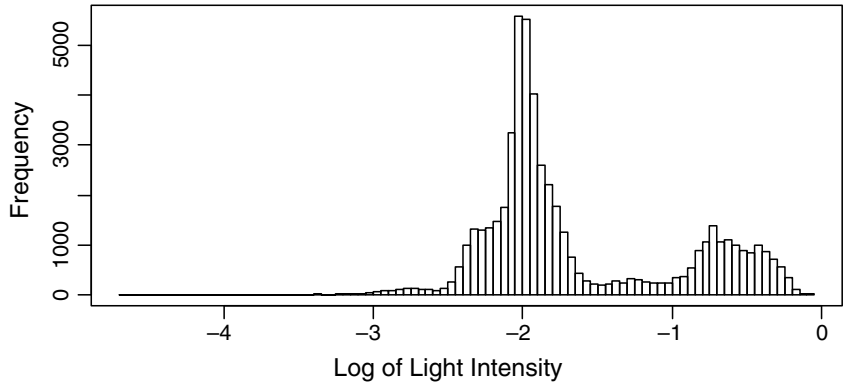

Figure 2.7 A histogram of a base 10 logarithm of the Light Intensity values from an image of a fish as used in Example 2.3.

distribution could be seen. The computer software for creating histograms usually has a built-in algorithm for a default number of bins, but users often have an option to specify their own preference. Some experimentation may be needed to find a suitable number of bins.

Based on the data in Figure 2.6, we know that there are more pixels representing the conveyer belt than those representing the fish. We also know that the higher values represent the fish. This information, together with Figure 2.7, suggests the threshold value identifying the fish pixels to be somewhere between -1.5 and -1 for $\log_{10}(\text{Light Intensity})$, which corresponds to $0.0316 < \text{Light Intensity} < 0.1$. However, it is unclear which exact value would be best. In order to find a good threshold value, we can look at spatial patterns of pixels identified as fish. Since each image pixel represents an area within the viewing scene, it is often represented as a rectangle, like those in Figure 2.8. We could require that the set of selected pixels forms a connected set because the image represents a fish in one piece. In the context of a pixilated image, we define a set A of pixels as a *connected set*, if for any pair of pixels from A, one can find a path connecting the pixels. The path can directly connect two pixels only when they are neighbors touching at the sides (but not if they only touch at

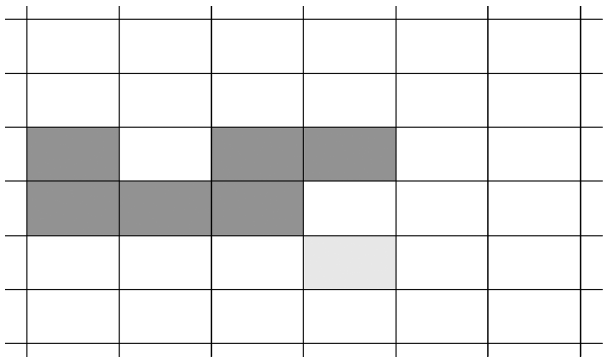

Figure 2.8 The darker shaded area is a connected set, but when the lighter shaded pixel is added, the set of pixels is not connected.

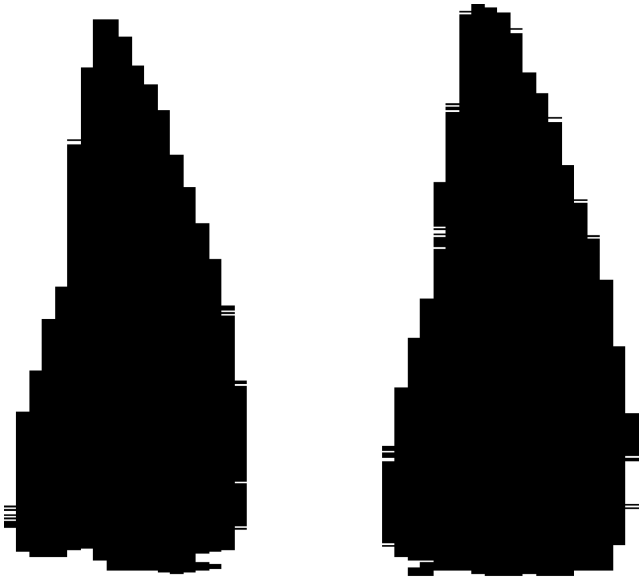

Figure 2.9 Dark areas show connected sets of pixels with Light Intensity above 0.08104 (a) and above 0.03809 (b), based on Fish data from Example 2.3.

corners). The darker shaded area in Figure 2.8 is a connected set, but when the lighter shaded pixel is added, the set of pixels is not connected.

When selecting all pixels with Light Intensity above 0.08104, one obtains a connected set of pixels shown as the black area in Figure 2.9a. Reducing the threshold below 0.08104 adds additional pixels that are not connected with the main connected set. An algorithm was used, where the threshold value was lowered, and the number of

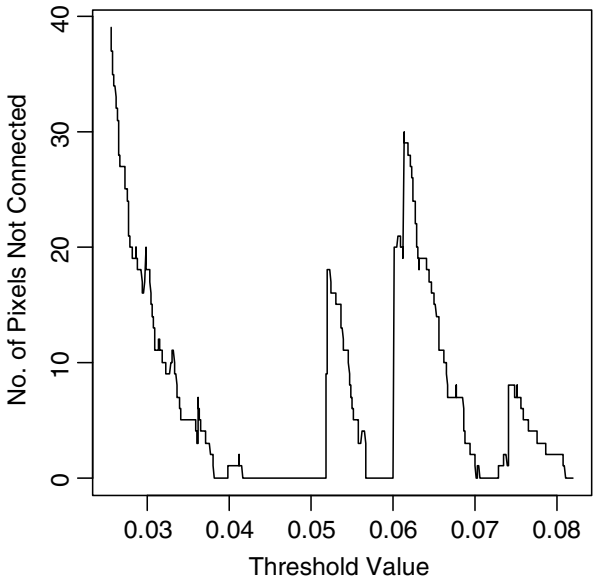

Figure 2.10 The number of pixels not connected to the main connected set shown as a function of the threshold value (for Fish data from Example 2.3).

pixels not connected to the main connected set was recorded and shown in Figure 2.10 as a function of the threshold value. We can see that for thresholds slightly above 0.07, the selected pixels again form a connected set (because the number of pixels not connected equals zero). This happens again at several ranges of the smaller threshold value until the smallest such value at 0.03809 (the place most to the left in Figure 2.10 where the function value is still zero). Below that value, the number of pixels not connected goes to very high values (beyond the range shown in Figure 2.10). Clearly, a good choice for the threshold value would be the one for which the number of pixels not connected is zero. However, Figure 2.10 still leaves us with a number of possible choices. Further investigation could be performed by looking at the type of graphs shown in Figure 2.9 and assessing the smoothness of the boundary lines. □

2.4.3 Box Plots

Another useful graph for showing the distribution of data is a *box plot* (sometimes called a box-and-whisker plot). An example of a box plot is shown in Figure 2.11, where a vertical axis is used for showing the numerical values. The box is plotted so that its top edge is at the level of the third quartile, and the bottom edge is at the level of the first quartile. A horizontal line inside the box is drawn at the level of the median. In the simplest version of a box plot, vertical lines (called whiskers) extend from the box to the minimum and maximum values. Some box plots may show outliers with special symbols (stars, here), and the whiskers extending only to the highest and lowest values that are not outliers (called upper and lower adjacent values). Clearly, this requires an automated decision as to which observations are outliers. Computer software often uses some simplified rules based on the interquartile range. For example, an observation might be considered an outlier when it is above the third quartile or below the first quartile by more than $1.5 \cdot \text{IQR}$. However such rules are potentially misleading because any serious treatment of outliers should also take into account the sample size. We discuss outliers and their detection in Section 3.6.

Example 2.4 In Example 2.2, we discussed the Visual Density of cyan patches on three pages printed immediately after the printer calibration. In the experiment described in Appendix B, the printer was then idle for 14 h, and a set of 30 pages

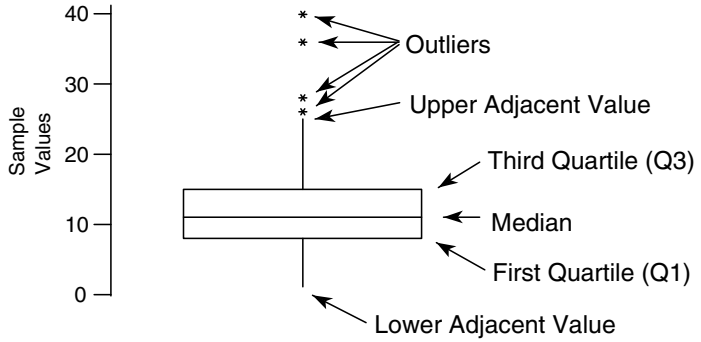

Figure 2.11 An example of a box plot.

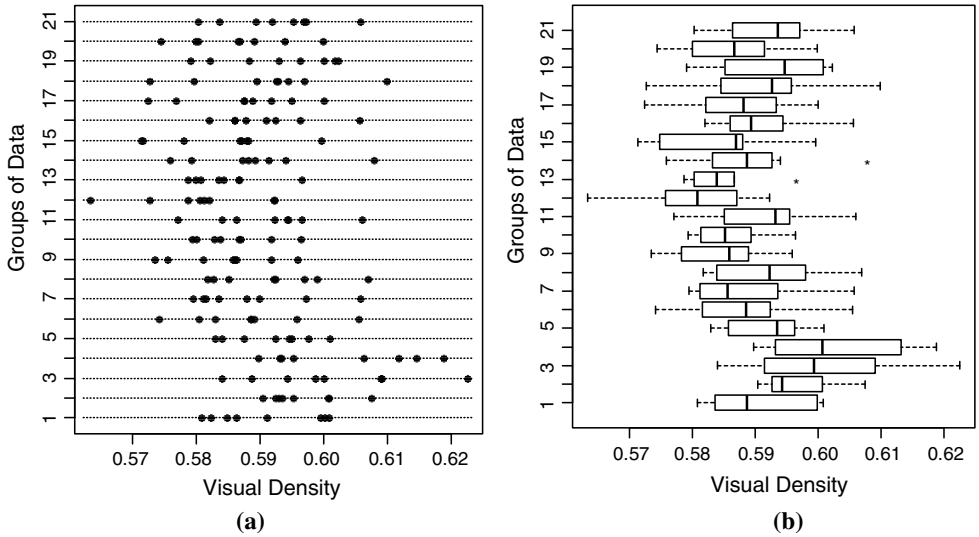

Figure 2.12 Visual Density of cyan printed on 21 pages shown as groups in the dot plots (a) and the box plots (b).

was printed, of which 18 pages were measured by a scanning spectrophotometer. This gives us a total of 21 pages with eight measurements of cyan patches in each page. Figure 2.12 shows the data in 21 groups using the dot plots (panel (a)) and the box plots (panel (b)). The box plots are somewhat easier to interpret, and this advantage increases with the increased number of groups and observations per group.

In Figure 2.12, we cannot see any specific patterns in Visual Density changes from page to page, which means that the idle time and subsequent printing of 30 pages had no significant impact on the quality of print as measured by the Visual Density of cyan patches.

2.4.4 Scatter Plots

When two characteristics, or variables, are recorded for each observation, or row, in the statistical database, we can create a two-dimensional *scatter plot* (as shown in Figure 2.13), where each observation is represented as a point with the two coordinates equal to the values of the two variables. A specific application of a scatter plot is best illustrated by the following example.

Example 2.5 This is a follow-up on Example 1.1, where you can find some background information about eye tracking. Here we want to consider an RGB image obtained in an Eye Tracking experiment as explained in Appendix B. This is a 128 by 128 pixel image (shown in Figure 2.14). The image consists of 16,384 pixels, which are treated as observations here. For each pixel, we have the intensity values (ranging from 0 to 1) for the three colors: Red, Green, and Blue, which can be regarded as three variables.

Figure 2.13 shows a scatter plot of Red versus Green values for that image. The pixels (observations) are represented as very small dots, so that thousands of them can

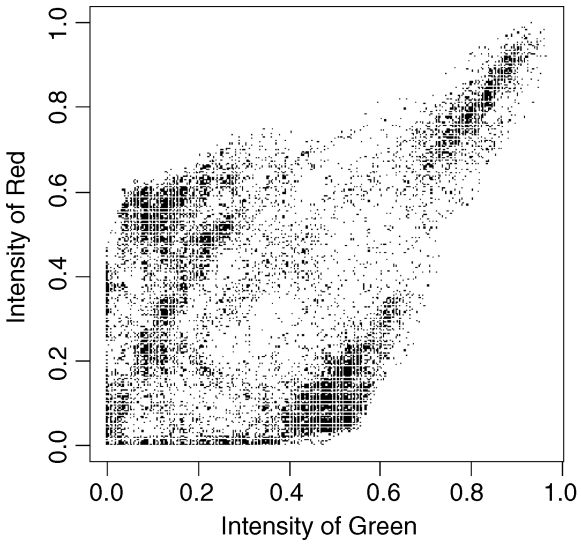

Figure 2.13 A scatter plot of intensities from the Eye Tracking image discussed in Example 2.5 and shown in Figure 2.14.

be seen as separate dots in the graph. A scatter plot is intended for continuous variables, and a primary color intensity is a continuous variable in principle. However, the three colors in the RGB image were recorded using 8 bits, which means that there are only 256 gradations of each color. This causes some discreteness of values, which can be seen as a pattern of dots lining up horizontally and vertically in Figure 2.13. It also turns out that there are many pixels in this image with exactly the same combination of gradations for the two colors. That is, some dots in the scatter plot represent more than one pixel. In order to deal with this issue, a technique of random jitter can be used, which amounts to adding a small random number to each point coordinate, before the points are plotted. This way, the dots do not print on the top of

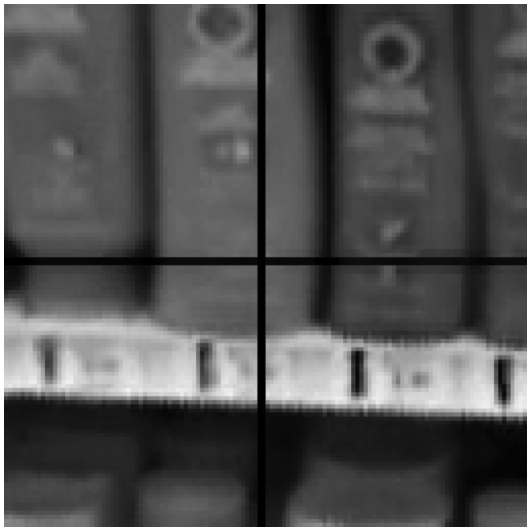

Figure 2.14 An RGB image from the Eye Tracking data set.

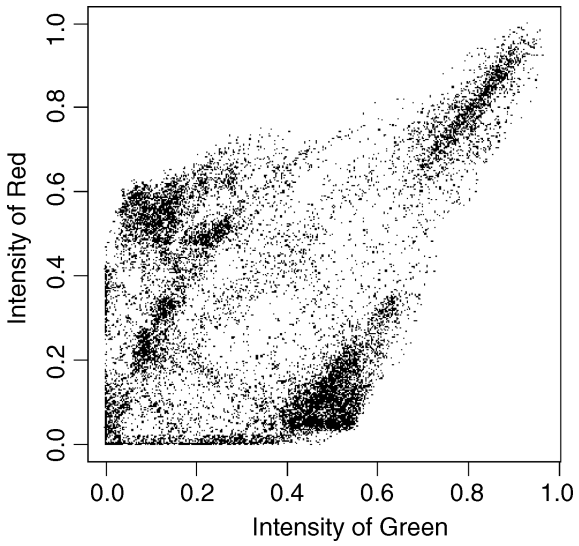

Figure 2.15 A scatter plot of color intensities from the Eye Tracking image shown in Figure 2.14. A small amount of random jitter was added to each dot.

each other. In Figure 2.15, a jitter in the amount equal to $(U-0.5)/256$ was used, where U is a random variable with the uniform distribution on the interval $(0, 1)$. The jitter improved the image, which no longer exhibits granulation, and we can better see where the larger concentrations of dots are. The use of jitter becomes even more important for highly discrete data.

The scatter plot shown in Figure 2.15 tells us that many pixels have high values both in Red and in Green. There is also a large group of pixels with approximately 50% of red and a small amount of green and then another group of pixels with approximately 50% of green and a small amount of red. There are no pixels with a very large value in one color and a low value in the other color, which is why the top left corner and the bottom right corner are both empty. □

2.5 PROBABILITY AND PROBABILITY DISTRIBUTIONS

2.5.1 Probability and Its Properties

In statistics, we typically assume that there is some randomness in the process we are trying to describe. For example, when tossing a coin, the outcome is considered random, and one would expect to obtain heads or tails with the same probability of 0.5. On the other hand, a physicist may say that there is nothing random about tossing a coin. Assuming full knowledge about the force applied to the coin, one should be able to calculate the coin trajectory as well as its spin, and ultimately predict heads or tails. However, it is usually not practical to collect that type of detailed information about the coin toss, and the assumption of 50–50 chances for heads or tails is regarded as sufficient, given lack of additional information. In general, one can say that randomness is a way of dealing with insufficient information. This would explain why, for a

given process, one can build many models depending on the available information. Also, the more information we have, the more likely we are to reduce the randomness in our model.

In order to calculate a probability of an event, we need to assume a certain probabilistic model, which involves a description of basic random events we are dealing with and a specification of their probabilities. For example, when assuming 50–50 chances for heads or tails, we are saying that each of the two events, heads and tails, has the same probability of 0.5. We can call this simple model a fair-coin model. Assuming this model, one can then calculate the probability of getting 45 tails and 55 heads in 100 tosses of the coin.

In statistics, we use this information in order to deal with an inverse problem. That is, let's assume we observe 45 tails and 55 heads in 100 tosses of a coin, but we do not know if the coin is fair with the same chances of heads or tails. Statistics would tell us, with certain confidence, what the probabilities are for heads or tails in one toss. It would also tell us if it is reasonable to assume the same probability of 0.5 for both events. If you think we can safely conclude, based on these 100 tosses, that the coin is fair, you are correct. What would be your answer if you observed 450 tails in 1000 tosses? If you are not sure, you can continue reading about the tools that will allow you to do the calculations needed to answer this question.

Before we introduce a formal definition of probability, we need to define a sample space as follows.

Definition 2.5. A *sample space* is the set of all possible outcomes of interest in a given situation under consideration.

The outcomes in a sample space are mutually exclusive, that is, only one outcome can occur in a given situation under consideration. For example, when a coin is tossed three times, the outcome is a three-element sequence of heads and tails. When we take 10 measurements, the outcome is a sequence of 10 numbers.

Definition 2.6. An *event* is a subset of a sample space.

When a coin is tossed three times, observing heads in the first toss is an event consisting of four outcomes: (H, H, H), (H, H, T), (H, T, H), and (H, T, T), where H stands for heads and T stands for tails. In a different example, when we take 10 measurements on a continuous scale, we can define an event that all of those measurements are between 20 and 25 units.

Definition 2.7. *Probability* is a function assigning a number between 0 and 1 to all events in a sample space such that these two conditions are fulfilled:

1. The probability of the whole sample space is always 1, which acknowledges the fact that one of the outcomes always has to happen.
2. For a set of mutually exclusive events A_i, we have $P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i)$, where k is the number of events, which may also be infinity.

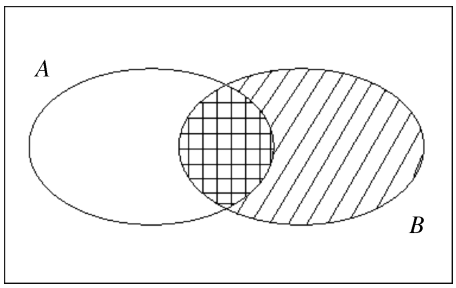

Figure 2.16 A Venn diagram showing two intersecting events. The probability $P(A|B)$ equals $P(A \cap B)$ as a fraction of $P(B)$.

We can say that probability behaves like the area of a geometric object on a plane. The sample space can be thought of as a rectangle with an area equal to 1, and all events as subsets of that square. Many properties of probability can be better understood through such geometric representation. Figure 2.16 discussed below shows an example of such representation called a Venn diagram.

When the sample space is finite, we often try to construct it so that all outcomes are equally likely. In this way, the calculation of probability is reduced to the task of counting the number of cases, such as permutations, combinations, and other combinatorial calculations. More on these rudimentary topics in probability can be found in most books on the fundamentals of statistics such as Devore (2004) or Mendenhall et al. (2006).

Definition 2.8. For any two events A and B, where $P(B) > 0$, the *conditional probability of A given that B has occurred* is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{2.6}$$

Without any information about B, we would use the unconditional probability $P(A)$ as a description of the probability of A. However, once we find out that B has happened, we should use the conditional probability $P(A|B)$ to describe the probability of A. One can think of the conditional probability as probability defined on the subset B as the whole sample space, and consequently, we consider only that part of A that also belongs to B as shown in Figure 2.16.

If A and B are disjoint events, then $P(A|B) = 0$, which means that A cannot happen if B has already occurred. A different concept is that of independence of events, which can be defined as follows.

Definition 2.9. Two events A and B are *independent* if and only if $P(A \cap B) = P(A) \cdot P(B)$.

When $P(B) > 0$, the events A and B are independent if and only if $P(A|B) = P(A)$, which means that the probability of A does not change once we find out that B has occurred. Some people confuse independent events with disjoint events, but the two

concepts are very different. If the events A and B are both independent and disjoint, then $0 = P(A|B) = P(A)$, which means that this can happen only for an uninteresting case when one of the sets has probability zero.

The event that B has not occurred is denoted as a complement set $B^c = \mathbb{S} \setminus B$, where $\mathbb{S}$ is the whole sample space. When $P(B^c) > 0$, the events A and B are independent if and only if $P(A|B^c) = P(A)$, which means that knowing that B has not occurred also does not change the probability of A happening. We can say that knowing whether B has occurred or not is not helpful in predicting A. The following theorem is often useful for calculating conditional probabilities.

Theorem 2.1 (**Bayes' Theorem**). Let $A_1, \ldots, A_k$ be a set of mutually exclusive events such that $P(A_i) > 0$ for $i = 1, \ldots, k$ and $\bigcup_{i=1}^{k} A_i$ is equal to the whole sample space. For any event B such that $P(B) > 0$, we have

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^{k} P(B|A_j) \cdot P(A_j)} \quad \text{for } i = 1, \ldots, k. \tag{2.7}$$

This theorem is often used to calculate the probabilities $P(A_i|B)$, when we know the conditional probabilities $P(B|A_i)$. The following example illustrates such an application.

Example 2.6 Medical imaging is often used to diagnose a disease. Consider a diagnostic method based on magnetic resonance imaging (MRI), which was tested on a large sample of patients having a particular disease. This method confirmed the disease in 99% of cases of the disease. Consider a randomly chosen person from the general population, and define A as the event that the person has the disease and B as the event that the person tested positive. Based on the above testing, we say that the probability $P(B|A)$ can be estimated as 0.99. This probability is called the sensitivity of the diagnostic method. The high sensitivity may seem like a proof of the test's good performance. However, we also need to know how the test would perform on people without the disease. So, the MRI diagnostic method was also tested on a large sample of people not having the disease. Based on the results, the probability $P(B^c|A^c)$ of testing negative for a healthy person was estimated as 0.9. This probability is called the specificity of the diagnostic method. Again, this may seem like a well performing method.

In practice, when using MRI on a patient, we do not know if the patient has the disease, so we are interested in calculating the probability $P(A|B)$ that a person testing positive has the disease. In order to apply Bayes' theorem, we also need to know $P(A)$, that is, the prevalence of the disease in the general population. In our example, it turns out that approximately 0.1% of the population has the disease, that is, $P(A) = 0.001$. Under these assumptions, the probability $P(A|B)$ can be calculated as 0.0098, which is surprisingly low (see Problem 2.4). The key to understanding why this happens is to consider all people not having the disease. They constitute 99.9% of the general population, and about 10% of them may test positive. On the other hand, only 0.1% of

Table 2.2 Examples of Probabilities of Disease if Tested Positive as a Function of Sensitivity, Specificity, and Disease Prevalence

Disease Prevalence $P(A)$	Sensitivity $P(B\|A)$	Specificity $P(B^c\|A^c)$	Probability of Disease if Tested Positive $P(A\|B)$
0.5	0.9	0.9	0.9
0.01	0.99	0.9	0.0909
0.001	0.99	0.9	0.0098
0.001	0.99	0.99	0.0902
0.001	0.99	0.999	0.4977
0.001	0.99	0.9999	0.9083
0.001	0.99	0.99999	0.9900

all people have the disease, which is a very small fraction (approximately 1%) of all people testing positive. This explains why most people testing positive do not have the disease. Table 2.2 shows some other interesting scenarios on how the probability $P(A|B)$ that a person with positive test result has the disease depends on sensitivity, specificity, and disease prevalence. □

2.5.2 Probability Distributions

We can now precisely define a random variable as a function assigning a number to each outcome in the sample space $\mathbb{S}$, that is, $X : \mathbb{S} \to \mathbb{R}$, where $\mathbb{R}$ is the set of real numbers. A value of the random variable is called a realization of X. For example, when a coin is tossed three times, define X as the number of times we observe heads. For each possible outcome, that is, a three-element sequence of heads and tails, we can count the number of heads. This will be the value of the random variable X.

Each random variable defines a probability measure on the set of real numbers $\mathbb{R}$. For each subset $A \subset \mathbb{R}$, we define $P(A) = P_{\mathbb{S}}(X^{-1}(A))$, where $P_{\mathbb{S}}(\,\cdot\,)$ is the probability defined on the sample space $\mathbb{S}$ and $X^{-1}(A)$ is the set of those outcomes in $\mathbb{S}$ that are assigned a value belonging to the set A (note that $X^{-1}(\,\cdot\,)$ is the inverse function). This probability measure is called the probability distribution of X. Continuing our example with X being the number of heads in three tosses, and taking A consisting of one number, say $A = \{2\}$, we obtain $P(A) = P_{\mathbb{S}}((H,H,T),(H,T,H),(T,H,H))$, which is equal to 3/8. This probability is more conveniently denoted by $P(X = 2)$.

In scientific applications, it is often impractical to list all possible events leading to a given value of X. For example, let X be the reflectance of a ceramic tile as measured in the spectral wavelength band between 400 and 410 μm. The random variable X will be subject to variability due to many factors such as the condition of the instrument, the process followed by the instrument operator, and so on. It would be difficult to describe all possible events that can happen during such measurements. For all practical purposes, it is sufficient to deal with the probability distribution of X without explicitly defining sample space and probability on it.

We now need to introduce some mathematical tools in order to describe probability distributions. It is convenient to distinguish two types of distributions: discrete distributions for discrete random variables, and continuous distributions for continuous random variables.

Definition 2.10. A random variable is *discrete* when all of its possible values can be counted using whole numbers.

Definition 2.11. A random variable is *continuous* when all of its possible values consist of an interval or a union of intervals on the real line $\mathbb{R}$.

A discrete probability distribution is described by a *probability mass function* $p(x) = P(X = x)$ defined for each possible value x of the random variable X. For example, if X is the number of heads in three tosses,

$$p(x) = \begin{cases} 1/8 & \text{for } x = 0 \text{ or } 3, \\ 3/8 & \text{for } x = 1 \text{ or } 2. \end{cases} \tag{2.8}$$

Property 2.1 A function defined on a discrete set D is a probability mass function of a certain distribution if and only if $p(x) \geq 0$ and $\sum_{x \in D} p(x) = 1$.

The set D in the above definition is the set of all possible values of X. Examples of some useful discrete distributions are shown in Appendix A.

A continuous probability distribution is described by *a probability density function* $f(x)$ such that for any two numbers a and b with $a \leq b$

$$P(a \leq X \leq b) = \int_a^b f(x)dx. \tag{2.9}$$

An example of a probability density function is plotted in Figure 2.17 as a bold bell-shaped curve. This is a density function of a *normal* distribution that approximates the distribution of data from two different samples. Each sample was generated randomly from the normal distribution. For the sample size of $n = 40$ in the left panel, the sampling variability is fairly large, and the histogram is not very well approximated by the density function. For the large sample size of $n = 400$, the approximation is much better, and it gets even better with larger samples. One can think of a density function as an idealized histogram for a very large or infinite sample size.

Property 2.2 A function $f : \mathbb{R} \to \mathbb{R}$ is a probability density function of a certain distribution if and only if $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$.

Examples of some useful continuous distributions and their density functions are shown in Appendix A. For continuous random variables, $P(X = x)$ is always equal

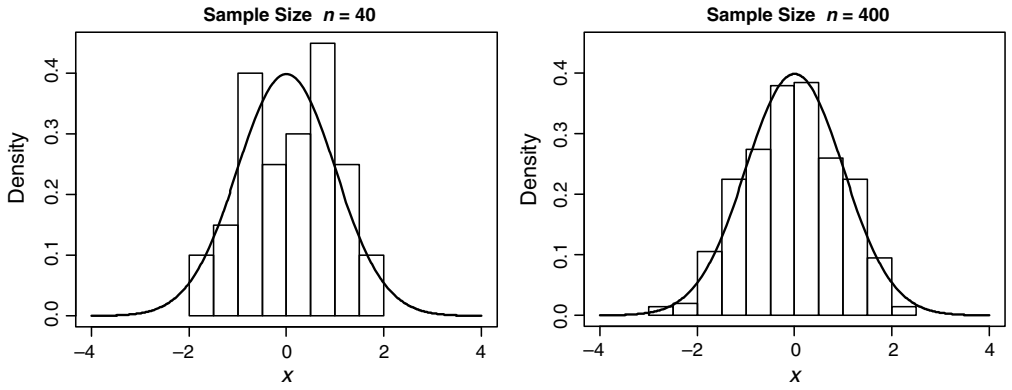

Figure 2.17 Histograms of two samples approximated by the normal density function describing the model from which the data were generated.

to zero, so a probability mass function would not be useful for describing such distributions.

Another way to describe any distribution (including a discrete or continuous one) is to use a cumulative distribution function (CDF) defined as

$$F(x) = P(X \leq x). \tag{2.10}$$

For any continuous distribution, the derivative of the CDF is equal to the density function, that is, $F'(x) = f(x)$ for any point x such that the derivative $F'(x)$ exists.

We can calculate probabilities of events associated with a given random variable X with the help of the CDF. Often, we also need to solve a reverse problem, that is, to find x such that $F(x)$ is equal to a given probability.

Definition 2.12. Let p be a number between 0 and 1. The $(100p)$th *percentile* of the distribution defined by $F(x)$ is a number η_p such that $p = F(\eta_p)$.

Often, it is convenient to define the upper percentile as follows.

Definition 2.13. Let p be a number between 0 and 1. The $(100p)$th *upper percentile* of the distribution defined by $F(x)$ is a number τ_p such that $p = 1-F(\tau_p)$.

It is easy to see that the $(100p)$th percentile η_p is equal to the $(100(1-p))$th upper percentile τ_{1-p} of the same distribution. For continuous distributions, the percentile η_p exists for any value $p \in (0, 1)$. Tables of percentiles for some important statistical distributions can usually be found in statistical textbooks. These days, one can often obtain percentiles from computer software, but we still provide some percentile values in Appendix A for added convenience. Appendix A shows the notation used throughout this book for percentiles of a wide range of distributions.

Even though a distribution is precisely defined by its cumulative distribution function or by a density or mass function (for continuous or discrete distributions, respectively), it is often beneficial to characterize distributions by using single numbers or parameters. Some important characteristics are the first and the third

quartile (the 25th and 75th percentiles, respectively) and the median (the 50th percentile). Other characteristics of distributions are defined in the next section.

2.5.3 Expected Value and Moments

The expected or mean value of a random variable X is defined as

$$E(X) = \begin{cases} \int_{-\infty}^{\infty} x \cdot f(x)dx & \text{if } X \text{ is continuous,} \\ \sum_{x \in D} x \cdot p(x) & \text{if } X \text{ is discrete.} \end{cases} \tag{2.11}$$

The expected value describes an average outcome based on a theoretical distribution. It is different from the sample mean $\overline{x}$ calculated from data. If data are generated from the distribution of X, the sample mean $\overline{x}$ should be close to $E(X)$ and it will get closer, on average, as the sample size increases. The expected value $E(X)$ is often denoted by μ, but a subtlety here is that μ should be considered as a parameter, while $E(X)$ is an operation on the distribution of X that produces a number.

Based on the linear property of integrals and summations, one can show (see Problem 2.5) that for any constants a and b

$$E(aX + bY) = aE(X) + bE(Y). \tag{2.12}$$

For any natural number k, the kth moment of X is defined as the expectation of X^k

$$E(X^k) = \begin{cases} \int_{-\infty}^{\infty} x^k \cdot f(x)dx & \text{if } X \text{ is continuous,} \\ \sum_{x \in D} x^k \cdot p(x) & \text{if } X \text{ is discrete,} \end{cases} \tag{2.13}$$

and the central moments are defined as moments centered around the mean, that is, $E\left[(X-E(X))^k\right]$. The mean value is interpreted as a position parameter or a "central" point, because it is an average of possible values of X weighted by their probabilities or by density. The second central moment, called variance, is denoted by

$$\text{Var}(X) = E\left[(X-E(X))^2\right] = E(X^2) - [E(X)]^2. \tag{2.14}$$

The variance measures variability around the mean value, while the noncentral moment $E(X^2)$ measures variability of X around zero. By using property (2.12), one can show that for any constants a and b

$$\text{Var}(aX + b) = a^2 \,\text{Var}(X), \tag{2.15}$$

which means that the variance is not affected by a shift (adding a constant). This makes sense because a simple shift does not impact variability. Since the variance is

expressed in the squared units of X, it is convenient to introduce the concept of standard deviation defined as the square root of variance and denoted by

$$\text{StDev}(X) = \sqrt{\text{Var}(X)}. \tag{2.16}$$

The standard deviation is a measure of variability expressed in the units of X, and its interpretation is further explained in Section 2.5. From equation (2.15), we obtain

$$\text{StDev}(aX + b) = |a| \cdot \text{StDev}(X), \tag{2.17}$$

which means that multiplying X by a positive constant results in the same multiplication of the standard deviation. The standard deviation is often denoted by σ, but again we have a subtlety here, where σ should be thought of as a parameter, while $\text{StDev}(X)$ is an operation on the distribution of X that produces a number.

The standard deviation as a parameter is often considered a scale parameter. We can use the standard deviation σ and the mean (expected value) $\mu = E(X)$ to standardize X, that is, we define the standardized variable $Z = (X-\mu)/\sigma$. It is easy to see that $E(Z) = 0$ and $\text{Var}(Z) = 1$. Since $(X-\mu)$ and σ are in the same units, the variable Z has no units.

2.5.4 Joint Distributions and Independence

Consider two random variables X and Y. We can study their relationship by considering a random vector (X, Y). This random vector can also be treated as a random point (X, Y) on the plane $\mathbb{R}^2$. Assume that we observe a large number of values, or realizations, of (X, Y). Each realization or data point can be plotted in the system of x and y coordinates as a point. Figure 2.18a shows a scatter plot of such points as an example. The relationship between X and Y is fully described by the joint

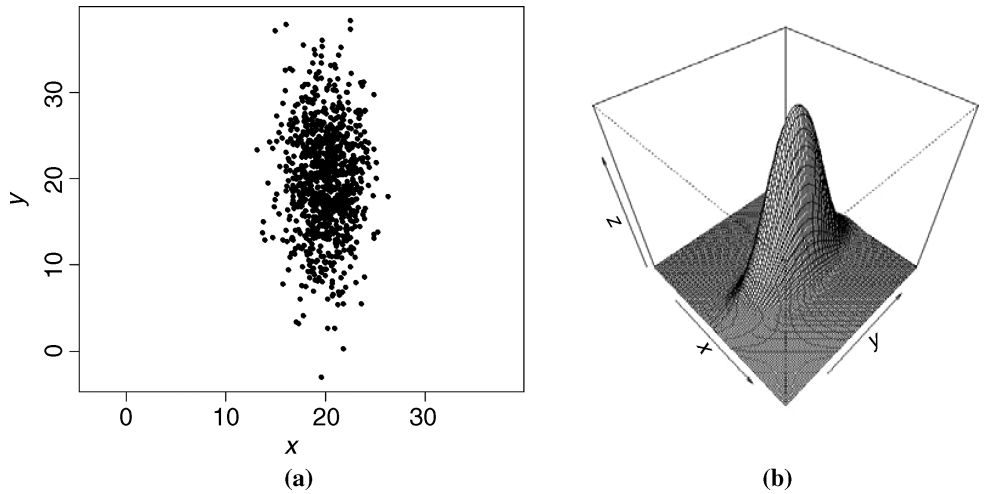

Figure 2.18 Panel (a) shows a scatter plot of (x, y) values generated as realizations of a random vector (X, Y) with the joint density function shown in panel (b).

distribution of these variables on the plane $\mathbb{R}^2$. The joint distribution, in turn, can be fully described by the cumulative bivariate distribution function defined as

$$F(x, y) = P(X \leq x \text{ and } Y \leq y). \tag{2.18}$$

For a continuous bivariate distribution, there exists a bivariate density function $f(\cdot, \cdot)$ such that

$$P((X, Y) \in A) = \iint_A f(s, t) ds\, dt \quad \text{for any } A \subset \mathbb{R}^2. \tag{2.19}$$

In particular, we have this property of the cumulative distribution function

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(s, t) ds\, dt. \tag{2.20}$$

Figure 2.18b shows a density function of the form $f(s, t) = f_1(s) f_2(t)$, where f_1 is the density function of the normal distribution $N(20, 2)$ with the mean of 20 and standard deviation of 2, and f_2 is the density function of $N(20, 6)$ (see Appendix A for the specific formula). The points in Figure 2.18a are values or realizations generated from the distribution defined by $f(s, t)$. The higher concentration of points around the center $(20, 20)$ corresponds to the higher value of the joint density function shown in Figure 2.18b. The range of x coordinates is smaller than the one for the y coordinates because of the smaller standard deviation in the x direction as seen in the elongated shape of the density in panel (b).

In the context of the bivariate distribution of (X, Y), the distributions of X and Y are called marginal distributions. For a continuous bivariate distribution of (X, Y), the marginal density function of one of the variables, let's say X, can be calculated by "summing up" the probabilities associated with the other variable, say Y, that is,

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy. \tag{2.21}$$

As another example, define a bivariate density function

$$f_0(x, y) = \begin{cases} 0.5 & \text{if } -1 \leq y + x \leq 1 \text{ and } -1 \leq y - x \leq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{2.22}$$

which is positive inside of a rotated square shown in Figure 2.19b.

The marginal distributions are obtained by "projecting" the bivariate density on the x or y axes, respectively. This is best understood by projecting the points in Figure 2.19a on one of the axes. Figure 2.20a shows a histogram of projections of

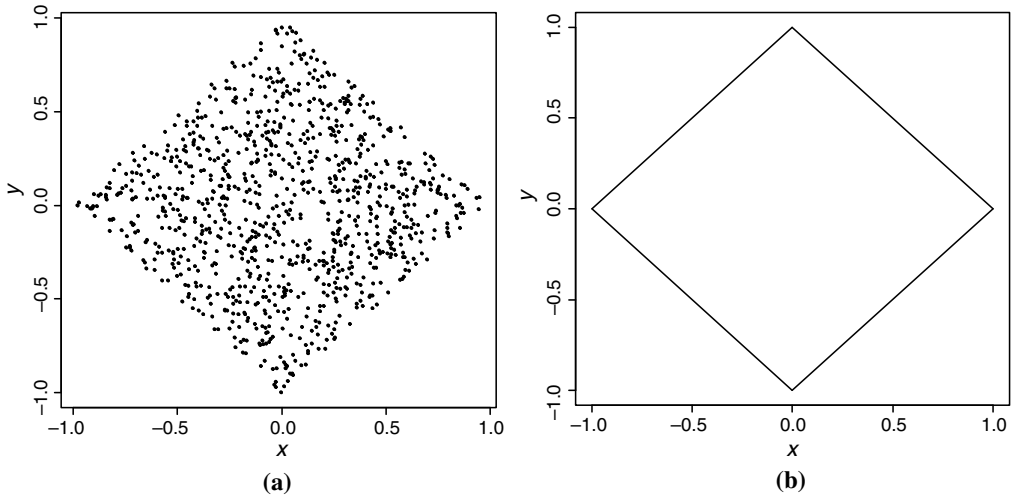

Figure 2.19 Panel (a) shows a scatter plot of (x, y) values generated as realizations of a random vector (X, Y) with the joint density function equal to 0.5 inside of the rotated square shown in panel (b) and zero outside of the square.

those points onto the x-axis. Figure 2.20b shows a theoretical distribution of X derived from (2.21) and (2.22) and given by the formula

$$f_X(x) = \begin{cases} 1-|x| & \text{if } |x| \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.23}$$

When dealing with a bivariate distribution of (X, Y), we might be interested in knowing the distribution of Y given an observed value of $X = x$, which represents a new piece of information. That distribution is called a *conditional distribution* of Y

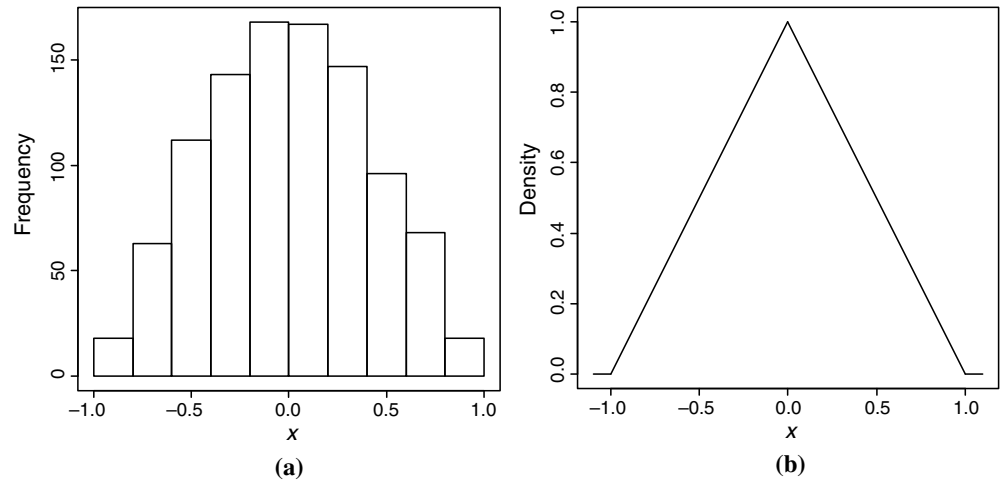

Figure 2.20 Panel (a) shows a histogram of projections of points from Figure 2.19a onto the x-axis. Panel (b) shows a theoretical distribution of the projection on the x-axis, that is, the marginal distribution of X.

given $X = x$. While values of the random vector (X, Y) lie on the plane $\mathbb{R}^2$, the conditional distribution of Y given $X = x$ is concentrated on the subset of the plane, namely, vertical line crossing x-axis at x.

For a continuous bivariate distribution of (X, Y), the conditional probability density function of Y given $X = x$ is defined as

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} \quad \text{for } -\infty < y < \infty, \tag{2.24}$$

where f_X is the marginal density function defined in (2.21) and x is any value such that $f_X(x) > 0$.

Intuitively, the conditional distribution is obtained by taking a cross section of the distribution of (X, Y), such as the one depicted in Figure 2.19b, along the line $X = x$. Dividing by $f_X(x)$ in formula (2.24) reflects the fact that the conditional distribution is a probability measure defined on a smaller space determined by $X = x$. This also makes the resulting function a density function, but it does not change the shape of the function.

Consider the bivariate density function $f_0(x, y)$ defined by (2.22). From Figure 2.19b, we can see that that for any $|x| \le 1$, $f_0(x, y)$ as a function of y is positive and constant on the interval $[-(1-|x|),\ 1-|x|]$ and zero outside this interval. Therefore, for any $|x| \le 1$, the conditional distribution of Y given $X = x$ is the uniform distribution concentrated on the interval $[-(1-|x|),\ 1-|x|]$. Since the length of this interval is $2(1-|x|)$, the conditional density function is given by the formula

$$f_{Y|X}(y|x) = \begin{cases} 1/(2(1-|x|)) & \text{if } -(1-|x|) \le y \le 1-|x|, \\ 0 & \text{otherwise.} \end{cases} \tag{2.25}$$

This formula can also be derived directly from definition (2.24) and formulas (2.22) and (2.23). Notice that in this example, the conditional distribution of Y depends on the observed value $X = x$. This information changes the range of possible values of Y from the general range $[-1,\ 1]$, without any knowledge of X, to the narrower range $[-(1-|x|),\ 1-|x|]$ when the value of $X = x$ is already known. This means that Y is dependent on X.

We now extend the definition of independence from random events to random variables.

Definition 2.14. The random variables X and Y are called *independent* when the events associated with those variables are independent, that is, for any sets $A, B \subset \mathbb{R}$

$$P(X \in A \text{ and } Y \in B) = P(X \in A) \cdot P(Y \in B). \tag{2.26}$$

For a continuous bivariate distribution of (X, Y), X and Y are independent if and only if

$$f(x,y) = f_X(x) f_Y(y) \quad \text{for all pairs of } x \text{ and } y \text{ values.} \tag{2.27}$$

As expected, independence of random variables is closely related to the conditional distributions. For a continuous bivariate distribution of (X, Y), X and Y are independent if and only if

$$f_{Y|X}(y|x) = f_Y(y) \quad \text{for all pairs of } x \text{ and } y \text{ values such that } f_X(x) > 0. \tag{2.28}$$

We can say that X and Y are independent if and only if information contained in X is not helpful in predicting Y. For example, the random variables X and Y with the joint distribution shown in Figure 2.19b are not independent because the conditional distribution shown in formula (2.25) depends on x, as we discussed previously. In Figure 2.18b, we depicted the joint distribution of two independent variables X and Y. We can imagine that although the cross sections of the surface taken at various values of x are different, smaller for x farther from the mean value of 20, they have the same bell shape and after normalizing by the marginal density $f_1(x)$, they all are identical to $f_2(y)$, the marginal density function of Y.

2.5.5 Covariance and Correlation

In order to capture an important property of the joint distribution, it is useful to define *covariance* of the random variables X and Y as

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]. \tag{2.29}$$

With some algebra, one can show that

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y). \tag{2.30}$$

Using equation (2.12), one can show that for any constants a and b

$$\text{Cov}(aX + bY, Z) = a\,\text{Cov}(X, Z) + b\,\text{Cov}(Y, Z) \tag{2.31}$$

for any random variable Z, which means that the covariance is linear with respect to its first argument. From symmetry, the same property holds for the second argument of the covariance. Since $\text{Var}(X) = \text{Cov}(X, X)$, we obtain

$$\text{Var}(aX + bY) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X, Y). \tag{2.32}$$

Let us now take $Y \equiv 1$, that is, a random variable equal to a constant 1. Then $Y = E(Y)$ and $\text{Var}(Y) = 0$. We can also see from (2.29) that the covariance of a constant variable Y with an arbitrary random variable Z is zero, that is, $\text{Cov}(Y, Z) = 0$. We can now write formula (2.31) as

$$\text{Cov}(aX + b, Z) = a\,\text{Cov}(X, Z), \tag{2.33}$$

which means that the covariance is not affected by a shift of X (adding a constant), but it is affected by the scale, that is, when X is multiplied by a constant. In the same way, we could obtain property (2.15) as a special case of (2.32).

The covariance $\text{Cov}(X, Y)$ measures a degree of linear association between X and Y. Unfortunately, that measure is distorted by the impact of a scale change. To make the measure scale independent, we introduce the *correlation coefficient* defined by the covariance scaled by the standard deviations of the variables as follows:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{StDev}(X)\text{StDev}(Y)}, \tag{2.34}$$

where $\text{StDev}(X) > 0$ and $\text{StDev}(Y) > 0$. It can be proven that $|\text{Corr}(X, Y)| \leq 1$, and the equality holds if and only if there exist constants $a \neq 0$ and b such that $Y = aX + b$ with probability 1, which means that X and Y are perfectly collinear. The correlation coefficient $\text{Corr}(X, Y)$ is often denoted by $\rho_{X,Y}$ or simply ρ.

Definition 2.15. The random variables X and Y are called *uncorrelated* when $\text{Corr}(X, Y) = 0$.

From (2.32), we conclude that the random variables X and Y are uncorrelated, if and only if

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y). \tag{2.35}$$

From (2.30), we conclude that the random variables X and Y are uncorrelated, if and only if

$$E(XY) = E(X)E(Y). \tag{2.36}$$

When two random variables are independent, they are also uncorrelated. However, in general, the reverse implication is not true. We have already discussed the random variables X and Y with the joint distribution shown in Figure 2.19 as being dependent. One can show that they are also uncorrelated. Another example of uncorrelated dependent variables is shown in Figure 2.21, where Y clearly depends on X, but not in a

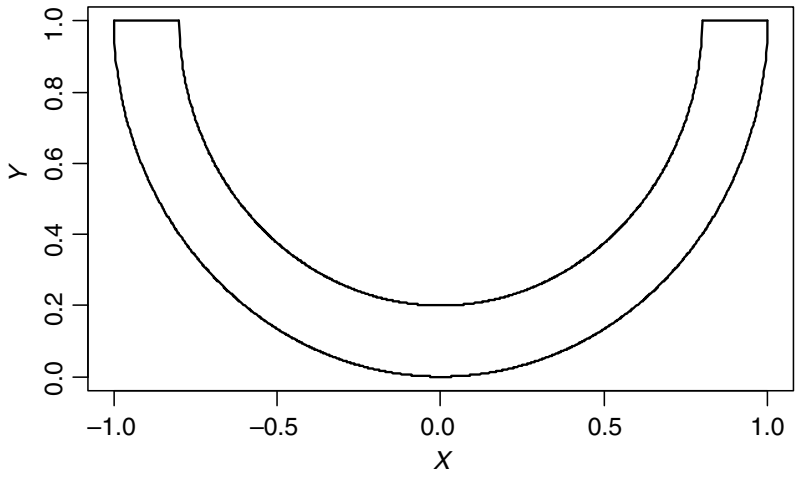

Figure 2.21 An example of uncorrelated and dependent random variables X and Y. The joint density function of the random variables X and Y is equal to a positive constant within the U-shaped area shown here and is equal to zero outside of that area.

linear fashion. In both cases, the lack of correlation can be concluded from the following property (see Problem 2.10 for an outline of the proof).

Property 2.3 If the joint distribution of X and Y is symmetric with respect to a vertical or horizontal straight line, and the correlation $\text{Corr}(X, Y)$ exists, then $\text{Corr}(X, Y) = 0$.

2.6 RULES OF TWO AND THREE SIGMA

In Section 2.2, we introduced the sample standard deviation as a measure of sample variability. In Section 2.4, we discussed the population standard deviation σ as a measure of variability in a random variable, say X, where $\sigma = \text{StDev}(X) = \sqrt{\text{Var}(X)}$. We now want to provide interpretation of the standard deviation σ by describing what the knowledge of σ can tell us about the variability in X. We will start by assuming that X follows the normal (Gaussian) distribution $N(\mu, \sigma)$ defined in Appendix A. The normal distribution is the most important distribution in probability and statistics, because data distributions and some theoretical distributions are often well approximated by the normal distribution. The reasons for that will be discussed in Section 2.7.

Property 2.4 If X follows the normal (Gaussian) distribution, then for any constants $a \neq 0$ and b, the variable $aX + b$ also follows the normal distribution. (See Problem 2.11 for a hint on the proof.)

We standardize X by defining $Z = (X-\mu)/\sigma$. It is easy to see (from (2.12) and (2.15)) that $E(Z) = 0$ and $\text{Var}(Z) = 1$. From Property 2.4, the standardized variable Z has the normal distribution $N(0, 1)$, which is called the *standard normal distribution*.

$$P(|X-\mu| \leq k\sigma) = P(\mu-k\sigma \leq X \leq \mu + k\sigma) = P(-k \leq Z \leq k)$$

$$= \Phi(k)-\Phi(-k) = 2\Phi(k)-1, \tag{2.37}$$

where $k > 0$ and Φ is the CDF of the standard normal distribution. For some specific values of k, we get

$$P(\mu-\sigma \leq X \leq \mu + \sigma) = P(-1 \leq Z \leq 1) \approx 0.68, \tag{2.38}$$

$$P(\mu-2\sigma \leq X \leq \mu + 2\sigma) = P(-2 \leq Z \leq 2) \approx 0.95, \tag{2.39}$$

$$P(\mu-3\sigma \leq X \leq \mu + 3\sigma) = P(-3 \leq Z \leq 3) \approx 0.997. \tag{2.40}$$

The properties (2.38), (2.39), and (2.40) are called the one-, two-, and three-sigma rules, respectively, and are illustrated in Figure 2.22. Since many distributions are well approximated by the normal distribution, these rules are widely used, especially for a

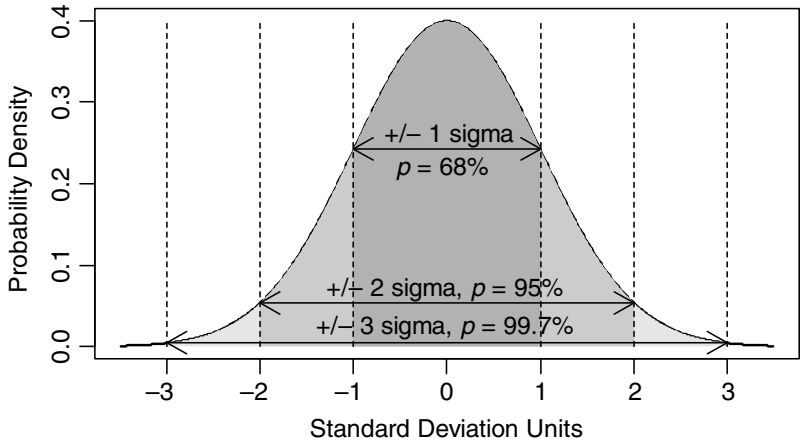

Figure 2.22 One-, two-, and three-sigma rules shown as areas under the normal density curve.

quick and intuitive understanding of the amount of variability associated with a given value of the standard deviation σ. For example, the two-sigma rule tells us that approximately 95% of the distribution lies within two standard deviations from the mean.

Even though the approximation by the normal distribution works quite well in many contexts, it would be good to know the significance of σ in other types of distributions. The following theorem addresses this issue in a general context.

Theorem 2.2 (Chebyshev's Inequality). For a random variable X with a finite mean μ and standard deviation σ, we have

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}, \tag{2.41}$$

where $k > 0$ is an arbitrary constant.

The proof can be found in Ross (2002). When applying (2.41) with $k = 2$, we can see that at least 75% of the distribution lies within two standard deviations from the mean, compared to the 95% based on normality. With $k = 3$, we obtain $8/9$ or at least 88.8% of the distribution being within three standard deviations from the mean, compared to the 99.7% based on normality.

2.7 SAMPLING DISTRIBUTIONS AND THE LAWS OF LARGE NUMBERS

In Section 2.2, we discussed a sample $x_1, x_2, \ldots, x_n$ of n measurements or observations as a set of specific numbers. However, before the observations are collected, there is uncertainty about their values. Also, if another set of observations were collected from the same unchanged process, the values would be somewhat different due to natural variability. This is why we often treat observations as random variables, so that we can study their properties in repeated sampling. For example, if we want to measure reflectance of a given surface as a single number

(let's say, in a narrow spectral band), it is convenient to consider this measurement as a random variable, say X. Each time the measurement is taken, we may get a somewhat different number, which will be regarded as a (random) value of that variable. If we measure that surface three times, we can introduce three random variables $X_1, X_2,$ and X_3 representing the three measurements. Each time we repeat the experiment, we will obtain three numbers as values of those three variables. It often makes sense to assume that the measurements are independent, that is, a measurement does not change the process under investigation, and the subsequent measurements are not impacted by the previous ones.

Definition 2.16. The random variables $X_1, X_2, \ldots, X_n$ are said to form a *(simple) random sample*, if they are independent, and each has the same distribution. They are called i.i.d. (independent, identically distributed) random variables.

For each sample, we can calculate a statistic, such as the sample mean, which can be treated as a random variable $\overline{X}$, since its values will vary in repeated samples. The distribution of $\overline{X}$ is called its sampling distribution in order to emphasize the fact that it describes the behavior of $\overline{X}$ over repeated samples.

Consider a random sample $X_1, X_2, \ldots, X_n$ from an arbitrary distribution G with a finite mean μ. The law of large numbers tells us that $\overline{X}$ approaches μ as n tends to infinity. Technical details about this convergence can be found in Ross (2002) and Bickel and Doksum (2001). The convergence means that we can draw conclusions about the population (represented by the distribution G) based on the sample $X_1, X_2, \ldots, X_n$, and there is a benefit from having larger samples. For very large n, the mean $\overline{X}$ will be very close to μ. Another far-reaching consequence can be concluded from the following construction. Let A be an arbitrary probabilistic event with a certain probability $P(A)$. The event A could be "obtaining heads in a single toss of a coin." Consider repeated independent trials (coin tosses), where the event A can happen with probability $P(A)$. For the ith trial, define Y_i as equal to 1 when A happens and 0 otherwise. Note that $P(Y_i = 1) = P(A) = E(Y_i) = \mu$. The sample mean $\overline{Y}$ is the relative frequency of the event A in n trials (fraction of *heads* in n tosses). The law of large numbers tells us that the fraction of trials when A happens (fraction of heads) in n trials approaches the probability $P(A)$ of the event (heads) as n tends to infinity. This may seem intuitively obvious, but it is good to have a confirmation of this fact as a basis for this interpretation of probability.

The law of large numbers tells us that $\overline{X}$ approaches μ as n tends to infinity, but it does not tell us how fast it is approaching. This information would be very useful from a practical point of view, so that we know the consequences of using a specific sample size n. From properties (2.15) and (2.35), one can show that

$$\mathrm{StDev}\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}, \tag{2.42}$$

where $\sigma = \mathrm{StDev}(X_i)$, $i = 1, \ldots, n$. This means that we can standardize $\overline{X}$ by defining $Z_n = \left(\overline{X} - \mu\right)/(\sigma/\sqrt{n}) = \sqrt{n}\left(\overline{X} - \mu\right)/\sigma$, such that $\mathrm{Var}(Z_n) = 1$. We know that $\left(\overline{X} - \mu\right)$ converges to 0. When it is multiplied by $\sqrt{n}$, it no longer converges

to 0, nor does it go to infinity (since $\mathrm{Var}(Z_n) = 1$). We could say that $\sqrt{n}$ is just the right multiplier to make Z_n "stable." For example, if we used the multiplier n^a, then $n^a(\overline{X}-\mu)/\sigma$ would approach infinity for $a > 0.5$, and it would approach 0 for $a < 0.5$.

If the random sample $X_1, X_2, \ldots, X_n$ comes from the normal distribution $N(\mu, \sigma)$, the distribution of Z is standard normal $N(0, 1)$ (see Property 2.4). This allows us to tell how close $\overline{X}$ is to μ with the probability given by

$$P\left(|\overline{X}-\mu| < k\frac{\sigma}{\sqrt{n}}\right) = P(|Z| < k) = 2\Phi(k)-1. \tag{2.43}$$

When the distribution G of the sample is not normal, the distribution of Z often is not easy to calculate, and it also depends on n. Fortunately, the following theorem allows an approximation of the distribution of Z for large n.

Theorem 2.3 (The Central Limit Theorem, CLT). Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables, each having a finite mean μ and standard deviation σ. Then the distribution of Z approaches the standard normal distribution as n tends to infinity, that is,

$$\lim_{n\to\infty} P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < k\right) = \Phi(k). \tag{2.44}$$

The proof can be found in Ross (2002).

The CLT allows us to use equation (2.43) as an approximation in cases of samples from non-normal distributions. Various sources give some rules of thumb (e.g., $n \geq 30$) as to how large n is needed for the normal approximation. This could be potentially misleading. The precision of the normal approximation depends on the shape of the X_i's distribution. For example, the convergence is generally slower for nonsymmetric distributions. Figure 2.23 shows an example of the density functions of Z, when the distribution of X_i's is chi-squared with one degree of freedom and n is equal to 3, 10, and 30, respectively. The density of the standard normal distribution is also shown for comparison. The CLT approximation using (2.44) can be better assessed based on Figure 2.24, where the CDFs of the same distributions are shown. Precision of the normal approximation is further discussed in Chapter 3.

The CLT explains why real data often follow the normal distribution (approximately). Many characteristics are sums of a large number of small independent factors. For example, height in a large population depends on influences of particular genes, elements in the diet, and other factors. Hence, the height distribution is typically well approximated by the normal distribution. Another example is when we take multiple measurements of the same object. The measurement error usually depends on many independent small factors (environmental conditions, gauge conditions, operators' impact, etc.) that add up to the final result. Again, the measurement error is typically well approximated by the normal distribution.

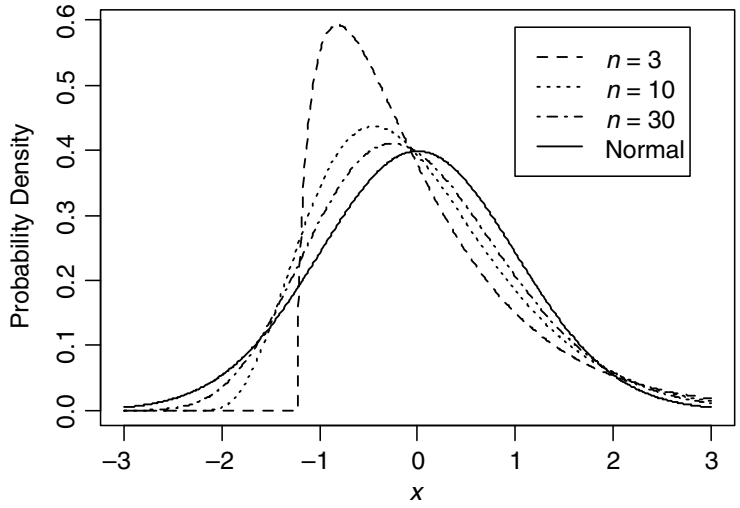

Figure 2.23 The density functions of Z, when the distribution of X_i's is chi-squared with one degree of freedom and n is equal to 3, 10, and 30, respectively. The solid line is the density of the standard normal distribution intended as the approximation.

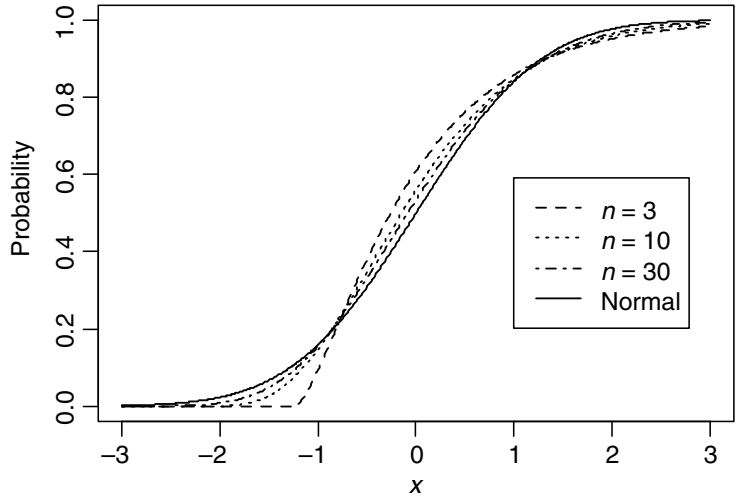

Figure 2.24 The CDF of Z, when the distribution of X_i's is chi-squared with one degree of freedom and n is equal to 3, 10, and 30, respectively. The solid line is the density of the standard normal distribution intended as the approximation.

2.8 SKEWNESS AND KURTOSIS[1]

The first two moments characterize the location and variability in a distribution. In order to characterize the shape of a distribution, it is convenient to consider the standardized variable $Z = (X-\mu)/\sigma$. Since the first two moments of Z are already

[1] This section is more technical and is not needed in the remaining part of this book.

determined ($E(Z) = 0$ and $\mathrm{Var}(Z) = 1$), we will use higher order moments in order to elicit the information about the distribution shape.

The lack of symmetry around the mean value in a distribution, that is, skewness, is measured by the *coefficient of skewness* defined as

$$\gamma_1 = E\left(Z^3\right) = \frac{E\left[\left(X - E(X)\right)^3\right]}{\left[\mathrm{Var}(X)\right]^{3/2}}. \tag{2.45}$$

For any symmetric distribution, $\gamma_1 = 0$. The fourth moment of Z is defined as *kurtosis* of X, that is,

$$\mathrm{Kurt}(X) = E\left(Z^4\right) = \frac{E\left[\left(X - E(X)\right)^4\right]}{\left[\mathrm{Var}(X)\right]^2}. \tag{2.46}$$

For a normal distribution, $\mathrm{Kurt}(X) = 3$, which is why an *excess kurtosis* is often defined as $\gamma_2 = \mathrm{Kurt}(X) - 3$. Since for a normal (Gaussian) distribution, $\gamma_1 = 0$ and $\gamma_2 = 0$, the skewness and kurtosis are sometimes used for checking normality. This approach is utilized in independent component analysis, an advanced multivariate method. A lack of symmetry in a distribution is fairly easy to recognize, but the interpretation of kurtosis is much less obvious. This is why we will detail more information about kurtosis and some related terminology.

A positive value of γ_2 indicates a super-Gaussian distribution (also called leptokurtic), which is often characterized by "fat tails," that is, the density function decreases slowly for large x values. A negative value of γ_2 indicates a sub-Gaussian distribution (also called platykurtic), which is often characterized by "thin tails," that is, the density function decreases rapidly for large x values. The kurtosis is also described as a measure of "peakedness" of a distribution at the center $E(X)$. These interpretations are true only to some extent. We will now discuss a different interpretation that clarifies the matter.

Since Z is a standardized variable, we have $\mathrm{Var}(Z) = E(Z^2) = 1$, and it might be of interest to know how far Z^2 is from 1. This can be measured by the mean square $E(Z^2 - 1)^2$, which is equal to $E(Z^4) - 1 = \mathrm{Kurt}(X) - 1$. We can also write

$$\mathrm{Kurt}(X) = E\left(Z^2 - 1\right)^2 + 1. \tag{2.47}$$

If Z^2 is close to 1 (i.e., X is concentrated around $\mu - \sigma$ or $\mu + \sigma$), then $\mathrm{Kurt}(X)$ is small. If $Z^2 \equiv 1$ (which happens for the Bernoulli distribution with only two possible values, each with the same probability $p = 0.5$), then $\mathrm{Kurt}(X) = 1$, which is its smallest possible value (i.e., the excess kurtosis γ_2 is always at least -2). If Z^2 is far from 1, then $\mathrm{Kurt}(X)$ is large. The variable Z^2 can be far from 1 when it is concentrated around 0 (high "peakedness") or concentrated on very large values ("fat tails" in the sense of large probabilities for X much larger than $\mu + \sigma$ in units of σ). Hence, in general, one of these conditions is sufficient to produce large kurtosis, but both of them

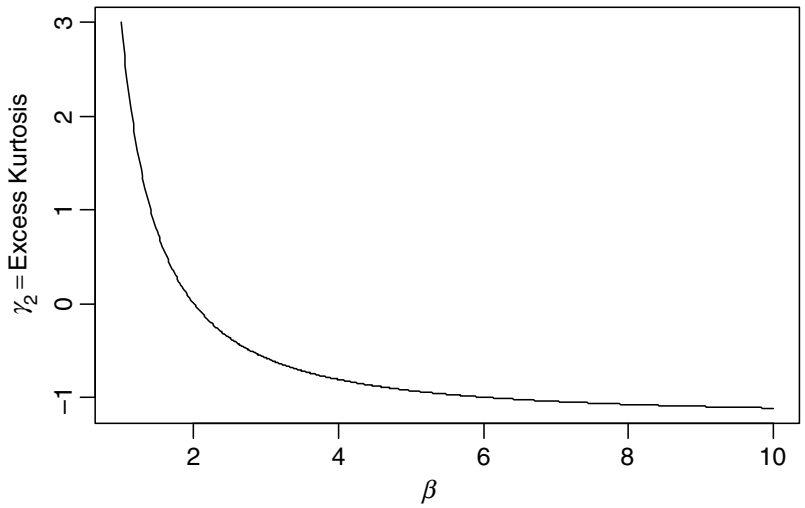

Figure 2.25 Excess kurtosis γ_2 as a function of the shape parameter $1 \leq \alpha \leq 10$ for the exponential power distribution.

give an even larger kurtosis. At the same time, the "peakedness" has small direct impact on the kurtosis because values close to 0 can never be really far from 1. The impact of "peakedness" is indirect. Since we always have $E(Z^2) = 1$, values of Z close to 0 allow some other Z values to be very large (and create "fat tails" in the sense described above).

Example 2.7 Consider the family of exponential power distributions defined in Appendix A. Its excess kurtosis is given by the formula

$$\gamma_2 = \frac{\Gamma(5/\alpha)\Gamma(1/\alpha)}{\Gamma(3/\alpha)^2} - 3, \tag{2.48}$$

where $\alpha > 0$ is the shape parameter. Figure 2.25 shows γ_2 as a function of α on the interval $[1, 10]$. The value $\alpha = 1$ corresponds to the Laplace distribution with $\gamma_2 = 3$, the value $\alpha = 2$ corresponds to the normal distribution with $\gamma_2 = 0$, and with α approaching infinity, the exponential power distribution approaches the uniform distribution having $\gamma_2 = -1.2$. When α approaches 0, the excess kurtosis γ_2 approaches infinity. The exponential power distributions with $\alpha < 2$ are super-Gaussian with "fat tails," while those with $\alpha > 2$ are sub-Gaussian with "thin tails." We illustrate this point in Figure 2.26, where we show densities of the exponential power distributions with $\alpha = 1$ (Laplace), $\alpha = 2$ (normal), and $\alpha = 10$.

Example 2.8 The interpretation of kurtosis as an indication of "fat" versus "thin" tails is not always as clear-cut as shown in Example 2.7. As discussed earlier, a sub-Gaussian distribution is often associated with "thin tails," but here we construct a sub-Gaussian distribution with "fat tails." Consider X following a chi-squared random variable with n degrees of freedom. The excess kurtosis of X is equal to $12/n$, so it is

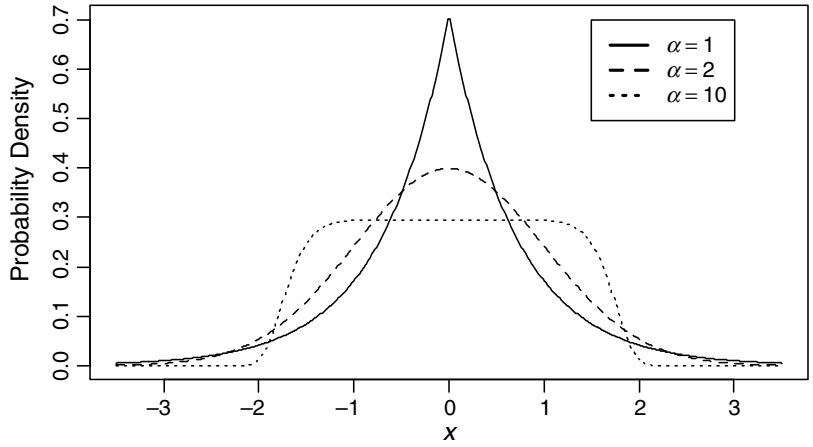

Figure 2.26 The densities of the exponential power distributions with $\alpha = 1$ (Laplace), $\alpha = 2$ (normal), and $\alpha = 10$, chosen so that the mean is 0 and the variance is 1.

considered super-Gaussian with "fat tails." We define a new variable $Y = D(X + a)$, where a is a positive constant and D is a random variable independent of X such that $P(D = 1) = P(D = -1) = 0.5$. The density of the Y variable is symmetric with respect to zero, and it consists of two symmetric shapes of the density of the chi-squared distribution with a gap in between (zero density on the interval $(-a, a)$). We call this distribution *double chi-squared*. Figure 2.27 shows an example of such density for $n = 4$ and $a = 0.41$. If we move the two pieces of the density function farther apart (by increasing a), its general shape does not change. This means that the density of Y has tails that are "fatter" than those of the normal density. However, for $a = 0.41$, one can calculate that $\gamma_2 = 0$. This tells us that the double chi-squared

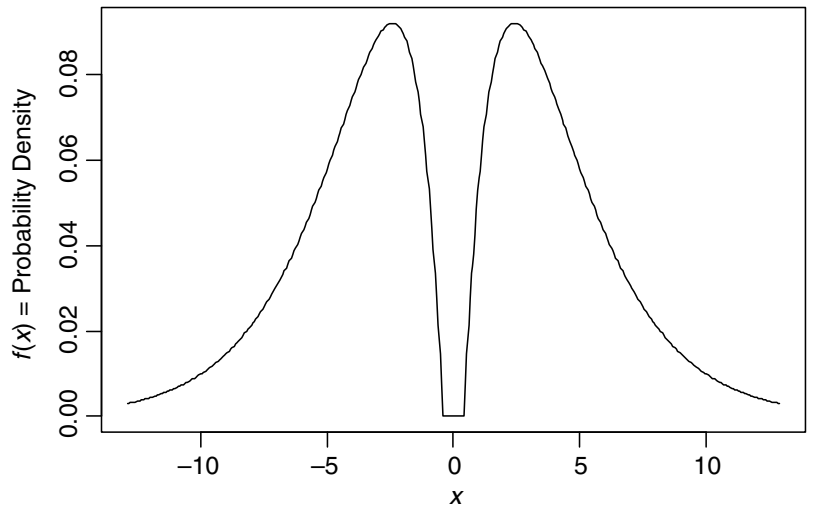

Figure 2.27 Density function of the random variable $Y = D(X + a)$ with zero excess kurtosis ($a \approx 0.41$), where X is a chi-squared random variable with four degrees of freedom and D is a random variable independent of X such that $P(D = 1) = P(D = -1) = 0.5$.

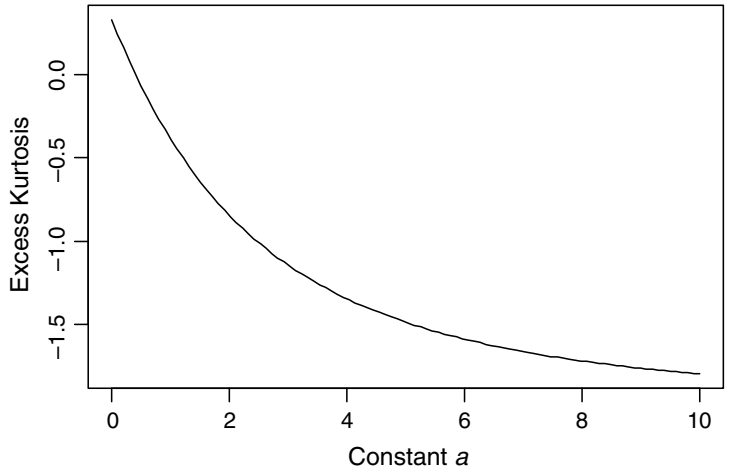

Figure 2.28 The excess kurtosis γ_2 of $Y = D(X + a)$ as a function of a, where X is a chi-squared random variable with four degrees of freedom and D is a random variable independent of X such that $P(D = 1) = P(D = -1) = 0.5$.

distribution shown in Figure 2.27 is an example of a "fat-tailed" distribution with the kurtosis equal to that of the Gaussian (normal) distribution.

Figure 2.28 shows how the excess kurtosis γ_2 of Y depends on a for $n = 4$. Clearly, the distribution becomes highly sub-Gaussian for large a. See Problem 2.12 for directions on how to perform the calculations for this example. □

The excess kurtosis is sometimes used to measure how far a distribution is from the normal distribution. This can be potentially misleading. The double chi-squared distribution introduced in the above example shows an example of a distribution with $\gamma_2 = 0$ (for $a \approx 0.41$), which is far from normal. The density function of that distribution is shown in Figure 2.27.

The construction used in Example 2.8 is more general and can be applied to many other distributions. For example, we can take any symmetric distribution and modify it to become a highly sub-Gaussian distribution. Notice that any random variable W with a distribution symmetric around zero can be represented as DX, where D is a random variable independent of X such that $P(D = 1) = P(D = -1) = 0.5$ and X describes the distribution of W on the positive numbers. Specifically, we can take $X = |W|$. We now define a new variable

$$Y = D(X + a), \tag{2.49}$$

where $a > 0$. The distribution of Y consists of two symmetric halves with the distribution of the same shape as that of X. With increasing a, the two halves move farther apart. As a tends to infinity, the distribution of Y becomes highly sub-Gaussian, approaching the most extreme case of $\gamma_2 = -2$ as shown by the following property (see Problem 2.13 for a sketch of the proof).

Property 2.5 For the excess kurtosis γ_2 of Y defined by (2.49), we have $\lim_{a \to \infty} \gamma_2 = -2$.

PROBLEMS

2.1. For the five observations of the Output Power variable in Example 2.1, find the 90th percentile calculated by a linear extrapolation using formula (2.4).

2.2. Prove that the deviations from the mean, defined as $d_i = x_i - \overline{x}$, have the property that they sum up to zero, that is, $\sum_{i=1}^{n} d_i = 0$.

2.3. Consider n observations x_i, $i = 1, \ldots, n$, and their linear transformations defined as $y_i = ax_i + b$ for $i = 1, \ldots, n$. Prove that $s_y^2 = a^2 s_x^2$ and $s_y = |a| s_x$.

2.4. In the context of Example 2.6, develop a formula for the probability of having the disease if testing positive as a function of sensitivity, specificity, and disease prevalence. Verify the numbers shown in Table 2.2.

2.5. Prove that for any constants a and b and random variables X and Y, we have (formula (2.12))

$$E(aX + bY) = aE(X) + bE(Y).$$

2.6. Prove formula (2.15).

2.7. Prove that for any constants a and b and random variables X, Y, and Z, we have (formula (2.31))

$$\mathrm{Cov}(aX + bY, Z) = a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z).$$

2.8. Prove that for any constants a and b and random variables X and Y, we have (formula (2.32))

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y).$$

2.9. Prove formula (2.33).

2.10. * Prove Property 2.3. *Hint*: Assume that the joint distribution is symmetric with respect to the line $x = \mu$. Define

$$d_+(x, y) = \begin{cases} 1 & \text{for } x > \mu, \\ 0 & \text{otherwise,} \end{cases} \qquad d_-(x, y) = \begin{cases} 1 & \text{for } x < \mu, \\ 0 & \text{otherwise.} \end{cases}$$

From the symmetry assumption, we have

$$E[d_+(X, Y)(X - \mu)(Y - E(Y))] = -E[d_-(X, Y)(X - \mu)(Y - E(Y))].$$

Since

$$\mathrm{Cov}(X,Y) = E[d_+(X,Y)(X-\mu)(Y-E(Y))] + E[d_-(X,Y)(X-\mu)(Y-E(Y))],$$

we have $\mathrm{Cov}(X,Y) = 0$.

2.11. Let X be a random variable following the normal (Gaussian) distribution $N(\mu,\sigma)$ defined in Appendix A. Show that for any constants $a \neq 0$ and b, the variable $aX + b$ also follows the normal distribution (and the distribution is $N(a\mu + b, |a|\sigma)$). *Hint*: Find the CDF of $aX + b$ from definition and perform integration by substitution.

2.12. * Consider the random variable X following a chi-squared distribution with n degrees of freedom. As in Section 2.8, define $Y = D(X + a)$, where a is a positive constant and D is a random variable independent of X such that $P(D = 1) = P(D = -1) = 0.5$. Find the formula for the kurtosis γ_2 of Y as it depends on a and n. Confirm that $\gamma_2 = 0$ for $n = 4$ and $a \approx 0.41$. Confirm the plots obtained in Figures 2.27 and 2.28. *Hint*: $E(Y^k) = E(D^k)E\left[(X-a)^k\right] = E\left[(X-a)^k\right]$ for k even and $E(Y^k) = E(D^k)E\left[(X-a)^k\right] = 0$ for k odd because $E(D^k) = 1$ for k even and $E(D^k) = 0$ for k odd. The formula for the moments of the chi-squared distribution can be found in Appendix A.

2.13. * Prove Property 2.5. *Hint*: Clearly, $E(Y) = 0$. Show that $E(Y^k) = E\left[(X + a)^k\right]$ for k even. Then show that $E\left[(X + a)^4\right]$ and $\left\{E\left[(X + a)^2\right]\right\}^2$ are four-degree polynomials with respect to a, having the coefficient 1 by the term a^4. This leads to $\lim_{a\to\infty} \mathrm{Kurt}(X) = 1$.