CHAPTER 1

LOOKING AT DATA - DISTRIBUTIONS

SECTION 1.1

OVERVIEW

Section 1.1 introduces several methods for exploring data. These methods should only be applied after clearly understanding the background of the data collected. The choice of method depends to some extent upon the type of variable being measured. The two types of variables described in this section are

• **Categorical variables** - variables that record to what group or category an individual belongs. Hair color and gender are examples of categorical variables. Although we might count the number of people in the group with brown hair, we wouldn't think about computing an average hair color for the group, even if numbers were used to represent the hair color categories.

• **Quantitative variables** - variables that have numerical values and with which it makes sense to do arithmetic. Height, weight, and GPA are examples of quantitative variables. It makes sense to talk about the average height or GPA of a group of people.
To summarize the distribution of a variable, for categorical variables use bar charts or pie charts, while for numerical data use histograms or stemplots. Also, when numerical data are collected over time, in addition to a histogram or
stemplot, a **timeplot** can be used to look for interesting features of the data. When examining the data through graphs we should be on the alert for

- unusual values that do not follow the pattern of the rest of the data

- some sense of a central or typical value of the data

- some sense of how spread out or variable the data are

- some sense of the shape of the overall pattern

In addition, when drawing a timeplot be on the lookout for **trends** occurring over time. Although many of the graphs and plots may be drawn by computer, it is still up to you to recognize and interpret the important features of the plots and the information they contain.

**GUIDED SOLUTIONS**

**Exercise 1.3**

**KEY CONCEPTS - individuals and type of variables**

Identify the "individuals" or objects described, then the "variables" or characteristics being measured. Once the variables are identified, you need to determine if they are categorical (the variable just puts individuals into one of several groups) or quantitative (the variable takes meaningful numerical values for which arithmetic operations make sense).

The "individuals" in this problem are the funds. If we included share price, this would be a quantitative variable. If we had another variable, say a 1 if the year to date return was positive and 0 if it wasn't, this would still be a categorical variable even though we used numbers to represent the two categories. Now list the variables recorded and classify each as categorical or quantitative.

<table>
<thead>
<tr>
<th>Name of variable</th>
<th>Type of variable</th>
</tr>
</thead>
</table>

Exercise 1.7

KEY CONCEPTS - interpreting variables

To determine what a variable is telling us, we must know the purpose for which the variable is to be used. If the purpose is to make comparisons, as here, we need to consider whether the groups being compared differ only in the value of the variable or if they differ in other ways. If the groups differ in other ways, ask yourself if these differences could be partly responsible for differences in the value of the variable. For example, if two groups differ in size, then variables related to size (such as counts), are likely to differ even though the groups are identical in all other respects. In this exercise, consider ways in which the groups (different years) might differ and if these might explain the differences in cancer death rates even if cancer treatments are becoming more effective. Here are some scenarios that can be applied to different parts of the problem.

Suppose that cancer was detected earlier. What would that do to survival rates? Suppose that population size increases and 1% of the population dies of cancer each year. Suppose that progress is made on the fight against heart disease. What would be the effect on the death rates due to cancer?

Exercise 1.19

KEY CONCEPTS - interpreting a histogram

How would you describe this distribution? Which portion of the histogram do you think corresponds to the state schools? How about the more exclusive private schools? In general, how many groups of schools are there and what are the most important aspects of the distribution?

Exercise 1.26

KEY CONCEPTS - drawing histograms and stem-and-leaf plots, and interpreting their shapes
Looking at Data - Distributions

<table>
<thead>
<tr>
<th>DRP scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
</tr>
<tr>
<td>47</td>
</tr>
<tr>
<td>52</td>
</tr>
<tr>
<td>47</td>
</tr>
</tbody>
</table>
How to draw a histogram.

1. When drawing a histogram, choose class intervals that divide the data into classes of equal length. For this data set, the smallest DRP score is 14 and the largest is 54, so the class intervals need to cover this entire range. A simple set of class intervals would be 10 - < 20 (10 is included in the interval but not 20), 20 - < 30, 30 - < 40, and 40 - < 50 and 50 - < 60. Other sets of intervals are possible, although these have the advantage of using fairly simple numbers as endpoints.

2. Count the number of data values in each class interval. Using the class intervals above, complete the frequency table below. Remember, when counting the number in each interval be sure to include data values equal to the lower endpoint but not the upper endpoint.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
</table>

3. Draw the histogram, which is a picture of the frequency table. In addition to drawing the bars, this also requires labeling the axes. Using the frequency table you have computed, complete the histogram given. The x-axis has been labeled for you. Make sure to include an appropriate label for the y-axis, as well as numbers for the scale.
How to draw the stem-and-leaf plot - single stems and splitting stems

1. It is easiest although not necessary to first order the data. If the data has been ordered, the leaves on the stems they will be in increasing order. The DRP scores have been ordered for you below.

```
14 14 15 18 19 19 22 25 25 26 26 27 27 28 29 31 33 33 34
34 35 35 35 35 38 39 40 40 41 41 42 43 44 45 46 47 47
48 49 51 52 52 54
```

2. Using the stems below, complete the stem-and-leaf plot.

```
1
2
3
4
5
```

What are the similarities between this stem-and-leaf plot and the histogram you drew? Think about the relationship between the stems and the class intervals in this example, which helps to explain why the stem-and-leaf plot looks just like the histogram laid on its side. What class intervals do the stems correspond to? What is one important difference between the histogram and the stem-and-leaf plot?

3. Using the DRP scores, increase the number of stems by splitting each of the previous stems in two. Complete the split stem-and-leaf plot in the space given.
In this case, splitting the stems results in a plot similar to a histogram, which uses too many class intervals. (How would you construct a histogram that corresponds to this stem-and-leaf plot? What would the class intervals be?) The more regular features of the data set are becoming obscured by the extra details you are forced to look at with the extra stems. As a display of the data, would you prefer the histogram or the first stem-and-leaf plot? Why?
To finish up the example, think about the important features that describe a distribution. Does the distribution of the DRP scores have a single peak? Does it appear to be symmetric, or is it skewed to the right (tail with larger values is longer), or to the left? Are there any outliers that fall outside the overall pattern of the data?

Exercise 1.37

**KEY CONCEPTS** - drawing and interpreting a timeplot

a) Complete the timeplot on the graph below. The first three points are plotted for you

Over the 30 year period plotted, the interest rate shows several clear cycles. These cycles produce three well defined clear temporary peaks which involve more than a single year of increase in the rates. Identify these. There was an overall peak in interest rates in the early 1980's. Has there been a consistent trend downward trend in rates since that time?
Exercise 1.3

<table>
<thead>
<tr>
<th>Name of variable</th>
<th>Type of variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td>Categorical</td>
</tr>
<tr>
<td>Net Assets</td>
<td>Quantitative</td>
</tr>
<tr>
<td>Year to date return</td>
<td>Quantitative</td>
</tr>
<tr>
<td>Largest holding</td>
<td>Categorical</td>
</tr>
</tbody>
</table>

Exercise 1.7

a) As the population size increases and people also live longer, the number of people dying each year from cancer, as well as other causes will go up even if treatments are more effective, simply because there are more people. For example, suppose the number of people at risk for cancer (because of age) increases from 1000 to 1500. Suppose treatments have reduced the incidence of cancer from 5% to 4%. The number of cancer deaths will increase - from 50 to 60!

b) People die of something - if other death rates go down, the cancer death rate could still go up even if cancer treatments were more effective. People are more likely to survive long enough to contract and die from cancer than another disease.

c) Suppose cancer was detected earlier, but treatments were not more effective. People would appear to live longer just due to earlier diagnoses. For example, if breast cancer was being diagnosed one year earlier but life expectancy was not changed by treatment, patients would appear to be surviving one year longer (a woman with breast cancer dies at 55 - she was diagnosed at 52 instead of 53, so her "survival" time is 3 years instead of 2).

Exercise 1.19

There are three groups of schools. The state schools such as the University of Massachusetts have the lower tuitions, and form the group with tuitions of $6000 and below. The remaining private schools seem to be divided into two groups. There are 22 schools in the range $12000 - $18000 and include less expensive private schools such as Northeastern University. At the high end of the distribution (over $24000) are some of the most expensive private colleges which include, for example, Harvard University. This distribution provides an
example of a trimodal (three modes) distribution that has been created by including three distinct groups of schools in the distribution.
Exercise 1.23

<table>
<thead>
<tr>
<th>Interval</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 - &lt; 20</td>
<td>6</td>
<td>13.63%</td>
</tr>
<tr>
<td>20 - &lt; 0</td>
<td>9</td>
<td>20.45%</td>
</tr>
<tr>
<td>30 - &lt; 40</td>
<td>11</td>
<td>25.00%</td>
</tr>
<tr>
<td>40 - &lt; 50</td>
<td>14</td>
<td>31.82%</td>
</tr>
<tr>
<td>50 - &lt; 60</td>
<td>4</td>
<td>9.09%</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>99.99%</td>
</tr>
</tbody>
</table>

Stem-and-leaf plot of DRP score, with single stems

1 445899
2 25566789
3 1334455589
4 00112345667789
5 1224

Stem-and-leaf plot of DRP score, with split stems

1 44
1 5899
2 2
2 55667789
3 13344
3 55589
4 0011234
4 5667789
5 1224
In general, for this number of observations, the preference for a histogram over a stem-and-leaf plot is a personal preference. For smaller data sets, the stem-and-leaf plot is usually preferred, while for larger data sets most people prefer the histogram. In this particular example, the data set is probably starting to get a little large for the stem-and-leaf plot which, by keeping a record of every observation, is beginning to look a little cluttered. So our preference would be a histogram - it seems to make the general shape of the distribution a little more apparent.

Comments on the general shape: There are no obvious outliers that depart from the general pattern. The distribution is unimodal and skewed to the left. Test scores, which have an upper bound on the maximum score that students come close to, are often left-skewed. Although there is a lower bound of zero, the scores generally don't go quite that low, but instead slowly trail off on the lower end, giving the left-skewed appearance.

**Exercise 1.37**

b) The three well defined temporary peaks occurred in 1974, 1981 and 1989. While there is a "peak" in 1984, it is not part of a clear up-and-down movement, nor is the higher interest rate in 1995.
c) The overall peak in the interest rate over these years occurred in 1981. There was a general downward trend from 1981 until around 1993, but since 1993 the rates have come back up slightly and then leveled off.
SECTION 1.2

OVERVIEW

Although graphs give an overall sense of the data, numerical summaries of features of the data make more precise the notions of center and spread.

Two important measures of center are the mean and the median. If there are $n$ observations, $x_1, x_2, ..., x_n$, then the mean is

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n} \sum x_i$$

where $\sum$ means "add up all these numbers." Thus, the mean is just the total of all the observations divided by the number of observations.

While the median can be expressed by a formula, it is simpler to describe the rules for finding it.

How to find the median.

1. List all the observations from smallest to largest.

2. If the number of observations is odd, then the median is the middle observation. Count from the bottom of the list of ordered values up to the $(n + 1)/2$ largest observation. This observation is the median.

3. If the number of observations is even, then the median is the average of the two center observations.

The most important measures of spread are the quartiles, the standard deviation, and variance. For measures of spread, the quartiles are appropriate when the median is used as a measure of center. In general, the median and quartiles are more appropriate when outliers are present or when the data are skewed. In addition, the five-number summary, which reports the largest and smallest values of the data, the quartiles and the median, provides a compact description of the data that can be represented graphically by a boxplot. Computationally, the first quartile, $Q_1$, is the median of the lower half of the list of ordered observations and the third quartile, $Q_3$, is the median of the upper half of the list of ordered values.
If you use the mean as a measure of center, then the standard deviation and variance are the appropriate measures of spread. Remember that means and variances can be strongly affected by outliers and are harder to interpret for skewed data.
If we have \( n \) observations, \( x_1, x_2, \ldots, x_n \), with mean \( \bar{x} \), then the variance \( s^2 \) can be found using the formula

\[
\frac{s^2}{n-1} = \frac{1}{n-1} \sum (x_i - \bar{x})^2
\]

The standard deviation is the square root of the variance, i.e., \( s = \sqrt{s^2} \), and is a measure of spread in the same units as the original data. If the observations are in feet, then the standard deviation is in feet as well.

**GUIDED SOLUTIONS**

**Exercise 1.49**

**KEY CONCEPTS** - measures of center, five-number summary, drawing boxplots

a) Complete the back-to-back stemplot using the stems below. There are only a few observations over a fairly wide range, so the overall shapes of the distributions tend to be indistinct.

```
Women | Men
---|---
7 | 7
8 | 8
9 | 9
10 | 10
11 | 11
12 | 12
13 | 13
14 | 14
15 | 15
16 | 16
17 | 17
18 | 18
```

b) To find the means, find the sum of the scores, then divide by the number of scores. For the median, since the number of scores is even for both groups, the median is the average of the two middle scores. You can find this easily from the stemplot.
The feature that would suggest that $\bar{x} > M$ is not necessarily the same for both distributions, so look at them carefully.
c) The five-number summary consists of the median, the minimum and the maximum, and the first and third quartiles. Remember that the first quartile is the median of those observations below the median (for the women there are 9 observations below the median) and the third quartile is the median of those observations above the median. Fill in the table below. These numbers are used to draw the boxplot.

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>Q₁</td>
<td></td>
</tr>
<tr>
<td>Q₁</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>Q₃</td>
<td></td>
</tr>
<tr>
<td>Q₃</td>
<td>Maximum</td>
<td></td>
</tr>
</tbody>
</table>

To determine if the $1.5 \times \text{IQR}$ criterion flags the largest women's observation, first compute the IQR. It is $Q₃ - Q₁$.

\[
\text{IQR} = \quad 1.5 \times \text{IQR} = \]

Then add $1.5 \times \text{IQR}$ to the third quartile and see if the largest women's observation exceeds this value.

\[
Q₃ + (1.5 \times \text{IQR}) = \]

Is 200 larger than this number? If so, then it is flagged as an outlier. We have drawn the boxplot for the men. After making sure you understand the men's boxplot, add the women's boxplot to this picture.
The first thing is to draw the box, which goes from the first quartile to the third quartile. Next locate the median within the box. Finally, check for outliers. If there are no outliers, the lines from the box extend to the smallest and largest observations. If there are outliers, then the lines from the box extend to the smallest and largest observations which are not outliers. The outliers are then identified individually with a symbol (usually either a * or a dot).

Drawing one or two boxplots or other graphical display by hand is the best way to make sure you understand how to interpret the display. But after that, it is really best to leave the drawing of boxplots and most other graphical displays to a statistical computer package.

d) Use both the stem-and-leaf plot and boxplot to answer the questions. Which graphic makes it easier to answer the questions?

**Exercise 1.59**

**KEY CONCEPTS - measures of center**

When there are several observations at a single value, the key is to remember that the mean is the total of all the observations divided by the number of observations. When computing the total, remember to include a salary as many times as it appears. The same is true when ordering the observations to find the median - remember to include a salary as many times as it appears.

**Exercise 1.61**

**KEY CONCEPTS - measures of center, resistant measures**

The change of extremes affects the mean, but not the median. To compute the new mean you can figure out the new total (you don't need to add up all the numbers again - just think about how much it has gone up) and divide by the number of observations. Or else you can think about dividing up the salary increases by the number of observations and adding this value to the old mean.

**Exercise 1.65**
KEY CONCEPTS - standard deviation

There are two points to remember in getting to the answer - the first is that numbers "further apart" from each other tend to have higher variability than numbers closer together. The other is that repeats are allowed. There are several choices for the answer to (a) but only one for (b).
Exercise 1.73

KEY CONCEPTS - linear transformations

This is an exercise in recognizing when a new measurement can be expressed in terms of an old measurement by the equation $x_{new} = a + bx$. This form of transforming an old measurement to a new measurement is called a linear transformation. In part (a) you want to convert water temperature to a new measurement which is the difference between the water temperature and the "normal" body temperature. If the water temperature was 90 degrees, the difference between the water temperature and normal body temperature would be -8.6 degrees (Note that a negative sign would occur whenever the pool temperature was below normal body temperature). We obtained this result by taking $x - 98.6$, where $x$ is the water temperature. This corresponds to a linear transformation with $a = 1$ and $b = -98.6$.

Try and set up part (b) yourself. Linear transformations are important in statistics and will appear at several points in the book.

COMPLETE SOLUTIONS

Exercise 1.49

a)

<table>
<thead>
<tr>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 05</td>
<td>8 8</td>
</tr>
<tr>
<td>8 8</td>
<td>9 12</td>
</tr>
<tr>
<td>931 10</td>
<td>489 5 11</td>
</tr>
<tr>
<td>3455 966</td>
<td>12 6</td>
</tr>
<tr>
<td>77 13 2</td>
<td></td>
</tr>
<tr>
<td>80 14 06</td>
<td></td>
</tr>
<tr>
<td>442 15</td>
<td>1</td>
</tr>
<tr>
<td>55 16 9</td>
<td></td>
</tr>
<tr>
<td>8 17</td>
<td></td>
</tr>
<tr>
<td>18 07 19</td>
<td></td>
</tr>
<tr>
<td>0 20</td>
<td></td>
</tr>
</tbody>
</table>

b) For the men, the sum of the 20 scores is 2425 and the mean is $\bar{x} = 121.25$. For the women, the sum of the 18 scores is 2539 and the mean is $\bar{x} = 141.06$. 
Since the number of scores is even for both groups, the median is the average of the two middle scores. For the men $M = 114.5$ and for the women $M = 138.5$. For the women there is really little skewness - the mean exceeds the mean.
because of the outlier, and only slightly. For the men, the distribution is right-skewed, and the mean exceeds the median because of this.

c)  

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>70</td>
<td>101</td>
</tr>
<tr>
<td>Q1</td>
<td>98</td>
<td>126</td>
</tr>
<tr>
<td>M</td>
<td>114.5</td>
<td>138.5</td>
</tr>
<tr>
<td>Q3</td>
<td>143</td>
<td>154</td>
</tr>
<tr>
<td>Maximum</td>
<td>187</td>
<td>200</td>
</tr>
</tbody>
</table>

The IQR for women is 154 - 126 = 28, and 1.5 x IQR = 37. If we add 37 to the third quartile we get 154 + 37 = 191. Since 200 exceeds this, it is flagged as an outlier.

d) The boxplot makes the comparison of the plots easier. The symmetry of the women's scores, with the exception of the outlier is fairly obvious, as well as the right skewness of the men's distribution. It is also somewhat clearer from examining the boxplot that the women's scores tend to be higher than the men's, while the men's are more variable.

![Boxplot](image)

**Exercise 1.59**

The number of observations (individuals) is $5 + 2 + 1 = 8$ and the total of the salaries is
(5 x 25,000) + (2 x $60,000) + (1 x $255,000) = $500,000.
The mean is $500,000 / 8 = $62,500. Everyone earns less than the mean, except for the owner.

Since there are eight observations, the median is the average of the fourth and fifth smallest observations, which is $25,000. To see this, the eight ordered observations are $25,000, $25,000, $25,000, $25,000, $25,000, $60,000, $60,000, $255,000

**Exercise 1.61**

The owner has an increase in salary from $255,000 to $455,000, or an increase of $200,000. The total is increased by this amount, from $500,000 to $700,000 and the new mean is $700,000/8 = $87,500. Another way of thinking about it is that the $200,000 increase averaged among the 8 people is $25,000, so the mean must go up $25,000 to $62,500 + $25,000 = $87,500.

The fourth and fifth smallest observations are still the same, so the median is unaffected.

**Exercise 1.65**

a) The standard deviation is always greater than or equal to zero. The only way it can equal zero is if all the numbers in the data set are the same. Since repeats are allowed, just choose all four numbers the same to make the standard deviation equal to zero. Examples are 1, 1, 1, 1 or 2, 2, 2, 2.

b) To make the standard deviation large, numbers at the extremes should be selected. So you want to put the four numbers at zero or ten. The correct answer is 0, 0, 10, 10. You might have thought 0, 0, 0, 10 or 0, 10, 10, 10 would be just as good, but a computation of the standard deviation of these choices shows that two at either end is the best choice.

c) There are many choices for (a) but only one for (b).

**Exercise 1.73**

a) The difference between $x$ and 98.6 is $x - 98.6$ (positive numbers correspond to pool temperatures above body temperature and negative numbers to pool temperatures below body temperature). In general, $x_{new} = (x - 98.6)$. ($a = -98.6$ and $b = 1$)
b) A food with 120 milligrams corresponds to 100% of the RDA, 60 milligrams to 50% and 240 milligrams to 200%.

In general, \( \%\text{RDA} = 100 \left( \frac{\text{number of milligrams in food}}{120} \right) \quad (a = 0, \ b = 10/3) \)
SECTION 1.3

OVERVIEW

This section considers the use of mathematical models to describe the overall pattern of a distribution. A mathematical model is an idealized description of this overall pattern, often represented by a smooth curve. The name given to a mathematical model that summarizes the shape of a histogram is a density curve. The density curve is a kind of idealized histogram. The total area under a density curve is one and the area between two numbers represents the proportion of the data that lie between these two numbers. Like a histogram, it can be described by measures of center, such as the median (a point such that half the area under the density curve is to the left of the point) and the mean $\mu$ (the balance point of the density curve if the curve were made of solid material), and measures of spread, such as the quartiles and the standard deviation $\sigma$.

One of the most commonly used density curves in statistics is the normal curve and the distributions they describe are called normal distributions. Normal curves are symmetric and bell-shaped. The peak of the curve is located above the mean and median, which are equal since the density curve is symmetric. The standard deviation is the distance from the mean to the change-of-curvature points on either side. It measures how concentrated the area is around this peak. Normal curves follow the 68 - 95 - 99.7 rule, i.e., 68% of the area under a normal curve lies within one standard deviation of the mean (illustrated in the figure below), 95% within two standard deviations of the mean, and 99.7% within three standard deviations of the mean.

Areas under any normal curve can be found easily if quantities are first standardized by subtracting the mean from each value and dividing the result.
by the standard deviation. This standardized value is sometimes called the \( z \)-score. If data whose distribution can be described by a normal curve are standardized (all values replaced by their \( z \)-scores), the distribution of these
standardized values is called the **standard normal** distribution and they are described by the **standard normal curve**. Areas under standard normal curves are easily computed by using a standard normal table such as that found in Table A in the front inside cover of the text.

If we know the distribution of data is described by a normal curve, we can make statements about what values are likely and unlikely, without actually observing the individual values of the data. Although one can examine a histogram or stem-and-leaf plot to see if it is bell-shaped, the preferred method for determining if the distribution of data is described by a normal curve is a **normal quantile plot**. These are easily made using modern statistical computer software. If the distribution of data is described by a normal curve, the normal quantile plot should look like a straight line.

In general, density curves are useful for describing distributions. Many statistical procedures are based on assumptions about the nature of the density curve that describes the distribution of a set of data. **Density estimation** refers to techniques for finding a density curve that describes a given set of data.

**GUIDED SOLUTIONS**

**Exercise 1.79**

**KEY CONCEPTS** - density curves and area under a density curve

a) In this case, the density curve has unknown height $h$ between 0 and 2, and height 0 elsewhere. Thus, the density curve forms a rectangle with a base of length 2 and a height equal to $h$. Recall that the area of any rectangle is the product of the length of the base of the rectangle and the height. The area of this density curve is therefore (fill in the blanks)

$$\text{area} = \text{length of base} \times \text{height} = \underline{} \times \underline{}$$

The total area under a density curve must be 1. What must $h$ be so that this is the case?

Draw a graph of this density curve in the space provided.
b) The area of interest is the shaded region in the figure below.

Compute the area of the shaded region by filling in the blanks below.

area = length of base × height = ________ × ________ = __________

c) Try answering this part on your own, using the same reasoning as in b). First, shade in the area of interest in the figure below.

Now compute the area of your shaded region as in b).

Exercise 1.83

**KEY CONCEPTS** - the 68 - 95 - 99.7 rule for normal curves

Recall that the 68 - 95 - 99.7 rule says that for the normal distribution, approximately 68% of the observations fall between the mean minus one standard deviation and the mean plus one standard deviation, 95% of the observations fall between the mean minus two standard deviations and the mean plus two standard deviations, and approximately 99.7% of the observations fall between the mean minus three standard deviations and the mean plus three standard deviations. Also recall that the area under a density curve between two numbers corresponds to the proportion of the data that lies between these two numbers.
In this problem the mean is 336 days and the standard deviation is 3 days. From
the 68 - 95 - 99.7 rule we have, for example, that 68% of the lengths of all horse
pregnancies lie between $336 - (1 \times 3) = 333$ and $336 + (1 \times 3) = 339$ days.
a) The 68 - 95 - 99.7 rule says that approximately 99.7% of the observations fall between the mean minus three standard deviations and the mean plus three standard deviations. The mean minus three standard deviations is 327. This is the lower bound for the shaded region in the figure below. What is the upper bound for the shaded region below? Fill in the space provided in the figure.

b) We indicated above that the middle 68% of all horse pregnancies have lengths between 333 and 339 days. What percent are either less than 333 or longer than 339? What percent must therefore be longer than 339 (recall that the density curve is symmetric)?
Exercise 1.89

KEY CONCEPTS - computing relative frequencies for a standard normal distribution.

Recall that the proportion of observations from a standard normal distribution that are less than a given value \( z \) is equal to the area under the standard normal curve to the left of \( z \). Table A gives these areas. This is illustrated in the figure below.

In answering questions concerning the proportion of observations from a standard normal distribution that satisfy some relation, we find it helpful to first draw a picture of the area under a normal curve corresponding to the relation. We then try to visualize this area as a combination of areas of the form in the figure above, since such areas can be found in Table A. The entries in Table A are then combined to give the area corresponding to the relation of interest.

This approach is illustrated in the solutions that follow.

a) To get you started, we will work through a complete solution. A picture of the desired area is given on the next page.
This is exactly the type of area that is given in Table A. We simply find the row labeled -2.2 along the left margin of the table, locate the column labeled .05 across the top of the table, and read the entry in the intersection of this row and column. We find this entry is 0.0122. This is the proportion of observations from a standard normal distribution that satisfies $z < -2.25$.

b) Shade the desired area in the figure below.

Remembering that the area under the whole curve is 1, how would you modify your answer from part a)?

area = 
c) Try solving this part on your own. To begin, draw a picture of a normal curve and shade the region.

Now use the same line of reasoning as in part b) to determine the area of your shaded region. Remember, you want to try to visualize your shaded region as a combination of areas of the form in given in Table A.

d) To test yourself, try this part on your own. It is a bit more complicated than the previous parts, but the same approach will work. Draw a picture and then try and express the desired area as the difference of two regions for which the areas can be found directly in Table A.
Exercise 1.91

**KEY CONCEPTS** - finding the value $z$ (the quantile) corresponding to a given area under a standard normal curve

The strategy used to solve this type of problem is the "reverse" of that used to solve Exercise 1.89. We again begin by drawing a picture of what we know; we know the area, but not $z$. For areas corresponding to those given in Table A we have a situation like the following.

![Diagram showing a standard normal curve with shaded area and a z-score]

To determine $z$, we find the given value of the area in the body of Table A (or the entry in Table A closest to the given value of the area). We now look in the left margin of the table and across the top of the table to determine the value of $z$ that corresponds to this area.

If we are given a more complicated area, we draw a picture and then determine from properties of the normal curve the area to the left of $z$. We then determine $z$ as described above. The approach is illustrated in the solutions below.

a) A picture of what we know is given on the next page. Note that since the area given is larger than 0.5, we know $z$ must be to the right of 0 (recall that the area to the left of 0 under a standard normal curve is 0.5).
We now turn to Table A and find the entry closest to 0.80. This entry is 0.7995. Locating the $z$ values in the left margin and top column corresponding to this entry, we see that the $z$ that would give this area is 0.84.

b) Try this part on your own. Begin by sketching a normal curve and the area you are given on the curve. On which side of zero should $z$ be located? Thinking about which side of zero a point lies on is a good way to make sure your answer makes sense.

Exercise 1.93

**KEY CONCEPTS** - computing the area under an arbitrary normal curve

For these problems, we must first state the problem, then convert the question into one about a standard normal. This involves standardizing the numerical
conditions by subtracting the mean and dividing the result by the standard deviation. We then draw a picture of the desired area corresponding to these standardized conditions and compute the area as we did for the standard normal, using Table A. This approach is illustrated in the solutions below.
a) *State the problem.* Call the cholesterol level of a randomly chosen young woman $X$. The variable $X$ has the $N(185, 39)$ distribution. We want the percent of young women with $X > 240$.

**Standardize.** We need to standardize the condition $X > 240$. We replace $X$ by $Z$ (we use $Z$ to represent the standardized version of $X$) and standardize 240. Since we are told that the mean and standard deviation of cholesterol levels are 185 and 39, respectively, the standardized value ($z$-score) of 240 is (rounded to two decimal places)

$$z\text{-score of 240} = \frac{240 - 185}{39} = 1.41$$

In terms of a standard normal $Z$, the condition is $Z > 1.41$.

A picture of the desired area is

![Desired area](image)

**Use the table.** The desired area is not of the form given in Table A. However we note that the unshaded area to the left of 1.41 is of the form given in Table A and this area is 0.9207. Hence

$$\text{shaded area} = \text{total area under normal curve} - \text{unshaded area} = 1 - 0.9207 = 0.0793$$

Thus the percent of young women whose cholesterol level $X$ satisfies $X > 240$ is $0.0793 \times 100\% = 7.93\%$.

b) Try this part on your own. First *state the problem.*
Next *standardize*. To do so, compute $z$-scores to convert the problem to a statement involving standardized values. In terms of $z$-scores, the condition of interest is

**Standardized condition:**

Sketch the standard normal curve below and shade the desired region on your curve.

Now *use the table*. Use Table A to compute the desired area. This will be the answer to the question.

**Exercise 1.105**

**Key Concepts** - finding the value $x$ (the quantile) corresponding to a given area under an arbitrary normal curve

To solve this problem, we must use a reverse approach to that used in Exercise 1.93. First we *state the problem*. To make use of Table A, we need to state the problem in terms of areas to the left of some value. Next, we *use the table*. To do so, we think of having standardized the problem and we then find the value $z$ in the table for the standard normal distribution that satisfies the stated condition, i.e., has the desired area to the left of it. We next must *unstandardize* this $z$ value by multiplying by the standard deviation and then adding the mean to the result. This unstandardized value $x$ is the desired result. We illustrate this strategy in the solutions below.

*State the problem.* We are told in Exercise 1.104 that the WISC scores are normally distributed with $\mu = 100$ and $\sigma = 15$. We want to find the score $x$ that will place a child in the top 5% of the population. This means that 95% of the
population scores less than \( x \). We will need to find the corresponding value \( z \) for the standard normal. This is illustrated in the figure on the next page.
Use the table. The value $z$ must have the property that the area to the left of it is 0.95. Areas to the left are the types of areas reported in Table A. Find the entry in the body of Table A that has a value closest to 0.95. This entry is 0.9495. The value of $z$ that yields this area is seen, from Table A, to be 1.64.

Unstandardize. We now must unstandardize $z$. The unstandardized value is

$$x = (\text{standard deviation}) \times z + \text{mean} = 15z + 100 = 15 \times 1.64 + 100 = 124.6$$

Thus a child must score at least 124.6 to be in the top 5%. Assuming fractional scores are not possible, a child would have to score at least $x = 125$ to place in the top 5%.

Now see if you can determine the score needed for a child to place in the top 1%. Use the same line of reasoning as above.

State the problem. You may find it helpful to draw the region representing the $z$ value corresponding to the top 1%.

Use the table.
Unstandardize.
Exercise 1.111

**KEY CONCEPTS** - normal quantile plots and determining whether the distribution of a set of data can be described by a normal curve.

The data are given in Exercise 1.27. To make a normal quantile plot you should use statistical software. Consult the user manual for the procedure for your software packages. The first step is to enter the data values. If your software package uses a spreadsheet for data entry, enter the data in a single column. If you have access to an ASCII (text) file containing the data, you should import it into your software package. Then use the appropriate command for making a normal quantile plot. Below is such a plot. Yours should look similar.
To interpret the plot ask yourself the following questions:
• do the points appear to follow a straight line?
• if not, in what ways do they deviate? Are there outliers? Are there any unusual "bends" at either end of the plot? Is there evidence of skewness?

Refer to figures 1.31 to 1.34 in your text for some guidance in interpreting your plot. You might also make a histogram of the data to check your interpretation. Write down your interpretation and check your answer with the solution provided.
COMPLETE SOLUTIONS

Exercise 1.69

a) The area of the density curve in this case is

\[ \text{area} = \text{length of base} \times \text{height} = \frac{2}{\text{height}} \times h \]

In order for this area to be 1, h must be 0.5. A graph of the density curve is the following.

b) The area of interest is the shaded region indicated below.

This shaded rectangular region has area = length of base × height = 0.5 × 1 = 0.50.

c) The area of interest is the shaded region indicated below.
height = 0.5

This rectangular region has area = length of base $\times$ height $= 0.8 \times 0.5 = 0.4$. 
Exercise 1.83

a) The shaded region lies between 327 and 345 days.

b) Refer to the figure in the guided solution. If the shaded region gives the middle 68% of the area, then the two unshaded regions must account for the remaining 32%. Since the normal curve is symmetric, each of the two unshaded regions must have the same area and each must account for half of the 32%. Hence each of the unshaded regions accounts for 16% of the area. The rightmost of these regions accounts for the longest 16% of all pregnancies. We conclude that 16% of all horse pregnancies are longer than 339 days.

Exercise 1.89

a) A complete solution was provided in the guided solutions.

b) The desired area is indicated below.
This is not of the form for which Table A can be used directly. However, the unshaded area to the left of -2.25 is of the form needed for Table A. In fact, we found the area of the unshaded portion in part a). We notice that the shaded area can be visualized as what is left after deleting the unshaded area from the total area under the normal curve.

\[
\text{shaded area} = \text{total area under normal curve} - \text{area of unshaded portion} = 1 - 0.0122 = 0.9878.
\]

Thus the desired proportion is 0.9878.

c) The desired area is indicated in the figure below.
shaded region of interest
This is just like part b). The unshaded area to the left of 1.77 can be found in Table A and is 0.9616. Thus
\[
\text{shaded area} = \text{total area under normal curve} - \text{area of unshaded portion} \\
= 1 - 0.9616 = 0.0384.
\]

This is the desired proportion.

d) We begin with a picture of the desired area.

The shaded region is a bit more complicated than in the previous parts, however the same strategy still works. We note that the shaded region is obtained by removing the area to the left of -2.25 from all the area to the left of 1.77.
The area to the left of -2.25 is found in Table A to be 0.0122. The area to the left of 1.77 is found in Table A to be 0.9616. The shaded area is thus

\[
\text{shaded area} = \text{area to left of 1.77} - \text{area to left of -2.25} \\
= 0.9616 - 0.0122 \\
= 0.9494.
\]

This is the desired proportion.

**Exercise 1.91**

a) A complete solution was given in the guided solution.

b) A picture of what we know is given below. Note that since the area to the right of 0 under a standard normal curve is 0.5, we know that \( z \) must be located to the right of 0.

![Diagram of normal distribution with shaded area](image)

The shaded area is not of the form used in Table A. However, we note that the unshaded area to the left of \( z \) is of the correct form. Since the total area under a normal curve is 1, this unshaded area must be \( 1 - 0.35 = 0.65 \). Hence \( z \) has the property that the area to the left of \( z \) must be 0.65. We locate the entry in Table A closest to 0.65. This entry is 0.6517. The \( z \) corresponding to this entry is 0.39.

**Exercise 1.93**
a) A complete solution was given in the guided solution.

b) *State the problem.* The problem is to find the percent of young women with $200 < X < 240$. 
Standardize. We need to first standardize the condition $200 < X < 240$. We replace $X$ by $Z$ (we use $Z$ to represent the standardized version of $X$) and standardize 200 and 240. Since we are told that the mean and standard deviation are 185 and 39, respectively, the standardized values ($z$-scores) of 200 and 240 are (rounded to two decimal places)

$$z\text{-score of } 200 = \frac{(200 - 185)}{39} = 0.38$$

$$z\text{-score of } 240 = \frac{(240 - 185)}{39} = 1.41.$$  

Our condition in "standardized" form is $0.38 < Z < 1.41$. A picture of the desired area is

![Graph showing the standardized z-scores and the desired area](image)

Use the table. This area can be found by determining the area to the left of 1.41, the area of the left of 0.38, and then subtracting the area to the left of 0.38 from the area to the left of 1.41. From Table A

area to the left of 1.41 = 0.9207

area to the left of 0.38 = 0.6480

and the desired difference is $0.9207 - 0.6480 = 0.2727$. Thus, the percent of young women whose cholesterol level $X$ satisfies $200 < X < 240$ is $0.2727 \times 100\% = 27.27\%$. 

Exercise 1.105

The complete solution for the top 5% is given in the guided solution.

For the top 1% we proceed as follows.

State the problem. We want to find the score that will place a child in the top 1% of the population. We first find the corresponding value $z$ for the standard normal. This value $z$ must have the property that the area to the right of it under the standard normal curve is 0.01. This is illustrated in the figure below.

![Diagram](image)

From the figure we see that the area to the left of $z$ (the unshaded area) is

\[
\text{unshaded area} = \text{total area under normal curve} - \text{shaded area} = 1 - 0.01 = 0.99.
\]

Hence we must find the value $z$ such that the area to the left of it is 0.99.

Use the table. We find the entry in the body of Table A that has value closest to 0.99. This entry is 0.9901. The value of $z$ that yields this area is seen, from Table A, to be 2.33.

Unstandardize. We now must unstandardize $z$. The "unstandardized" value is

\[
x = (\text{standard deviation}) \times z + \text{mean} = 15z + 100 = 15 \times 2.33 + 100 = 134.95
\]

Thus, a child must score at least 134.95 to be in the top 1%. Assuming fractional scores are not possible, a child would have to score at least $x = 135$ to place in the top 1%. 
Exercise 1.111

If the data follow a normal distribution, the plot should look approximately like a straight line. In our plot, the three points in the lower-left corner lie below the line drawn through the remaining points. This suggests that the data may be slightly left skewed. Ignoring these three points, the rest of the plot is reasonably straight, suggesting that the remaining data are approximately normal. Making a histogram of the data, one sees the slight left skewness.