Chapter 2
Descriptive Statistics I: Tabular and Graphical Summary

2.1 Generalities

Consider the problem of using data to learn something about the characteristics of the group of units which comprise the sample. Recall that the distribution of a variable is the way in which the possible values of the variable are distributed among the units in the group of interest. A variable is chosen to measure some characteristic of the units in the group of interest; therefore, the distribution of a variable contains all of the available information about the characteristic (as measured by that variable) for the group of interest. Other variables, either alone or in conjunction with the primary variable, may also contain information about the characteristic of interest. A meaningful summary of the distribution of a variable provides an indication of the overall pattern of the distribution and serves to highlight possible unusual or particularly interesting aspects of the distribution. In this chapter we will discuss tabular and graphical methods for summarizing the distribution of a variable and in the following chapter we will discuss numerical summary methods.

Generally speaking, it is hard to tell much about the distribution of a variable by examining the data in raw form. For example, scanning the Stat 214 data in Table 1 of Chapter 1 it is fairly easy to see that the majority of these students are female; but, it is hard to get a good feel for the distributions of the variables which have more than two possible values. Therefore, the first step in summarizing the distribution of a variable is to tabulate the frequencies with which the possible values of the variable appear in the sample. A **frequency distribution** is a table listing the possible values of the variable and their frequencies (counts of the number of times each value occurs). A frequency distribution provides a decomposition of the total number of observations (the sample size) into frequencies for each possible value. In general, especially when comparing two distributions based on different sample sizes, it is preferable to provide a decomposition in terms of relative frequencies. A **relative frequency distribution** is a table listing the possible values of the variable along with their relative frequencies (proportions). A relative frequency distribution provides a decomposition of the total relative frequency of one (100%) into proportions or relative frequencies (percentages) for each possible value.

Many aspects of the distribution of a variable are most easily communicated by a graphical representation of the distribution. The basic idea of a graphical representation of a distribution is to use area to represent relative frequency. The total area of the graphical representation is taken to be one (100%) and sections with area equal to the relative frequency (percentage) of occurrence of a value are used to represent each possible value of the variable.

2.2 Describing qualitative data

In this section we consider tabular and graphical summary of the distribution of a qualitative variable. Assuming that there are not too many distinct possible values for the variable, we can summarize the distribution using a table of possible values along with the frequencies and relative frequencies with which these values occur in the sample. Recall that a frequency distribution provides a decomposition of the total number of observations (the sample size) into frequencies for each possible value; and, a relative frequency distribution provides a decomposition of the total relative frequency of one (100%) into proportions or relative frequencies (percentages) for each possible value. In most applications, and especially for comparisons of distributions, it is better to use relative frequencies rather than raw frequencies. When forming a relative frequency distribution for a nominal qualitative variable we can list the possible values of the variable in any convenient order. On the other hand, the possible values of an ordinal qualitative variable should always be listed in proper order to avoid possible confusion when reading the table.

Table 1 contains the frequency distributions and relative frequency distributions of the two qualitative variables sex and classification for the Stat 214 example. Notice that the possible values of the ordinal variable, classification of the student, are listed in proper order to avoid possible confusion when reading the table.

Table 1. Relative frequency distributions for the sex and classification distributions in the Stat 214 example.

Sex distribution.

sex	frequency	relative frequency
female	51	.761
male	16	.239
total	67	1.000

Classification distribution.

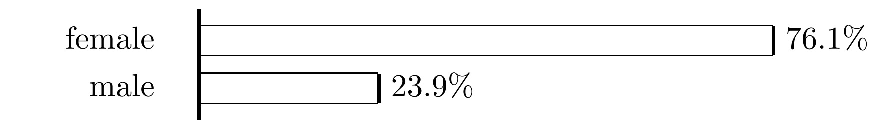
classification	frequency	relative frequency
freshman	27	.403
sophomore	16	.239
junior	16	.239
senior	8	.119
total	67	1.000

Bar graphs summarizing the sex and classification distributions for the Stat 214 example are given in Figure 1. Again, to avoid confusion, the possible classification values are presented in proper order. A **bar graph** consists of a collection of bars (rectangles) such that the combined area of all the bars is one (100%) and the area of a particular bar is the relative frequency of the corresponding value of the variable. Two other common forms for such a graphical representation are segmented bar graphs and pie graphs. A

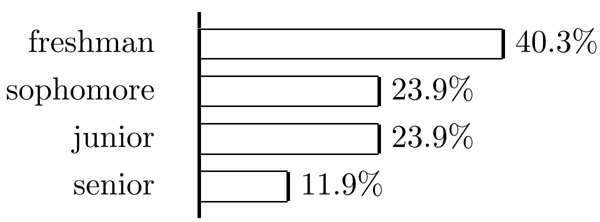
segmented bar graph consists of a single bar of area one (100%) that is divided into segments with a segment of the appropriate area for each observed value of the variable. A segmented bar graph can be obtained by joining the separate bars of a bar graph. If the bar of the segmented bar graph is replaced by a circle, the result is a pie graph or pie chart. In a **pie graph** or pie chart the interior of a circle (the pie) is used to represent the total area of one (100%); and the pie is divided into slices of the appropriate area or relative frequency, with one slice for each observed value of the variable.

Figure 1. Bar graphs for the sex and classification distributions in the Stat 214 example.

Sex distribution.



Classification distribution.



For these two sections of Stat 214 it is clear that a large majority (76.1%) of the students are female. There are two simple explanations for the predominance of females in this sample: The proportion of females among all undergraduate students at this university is roughly 65%; and, the majors which require this particular course traditionally attract more females than than males. Turning to the classification distribution, it is clear that relatively few (11.9%) of the students in these sections are seniors. This aspect of the classification distribution is not surprising, since Stat 214 is a 200 level (nominally sophomore) course. It is somewhat surprising, for a 200 level course, to find that the most common classification is freshman (40.3%). We might wonder whether these characteristics of the sex and classification distributions are applicable to both sections of Stat 214. The section variable can be used as an indicator variable (a qualitative explanatory variable used for grouping observations) to divide the sample of 67 students into the group of 36 students in section 1 and the group of 31 students in section 2. The sex and classification distributions for these two sections are summarized in Tables 2 and 3 and Figures 2 and 3. Notice that, because of rounding of the relative frequencies, the sum of the relative frequencies is not exactly one in Table 3.

Table 2. Relative frequency distributions for sex, by section.

section 1

sex	frequency	relative frequency
female	28	.778
male	8	.222
total	36	1.000

section 2

sex	frequency	relative frequency
female	23	.742
male	8	.258
total	31	1.000

Figure 2. Bar graphs for sex, by section.

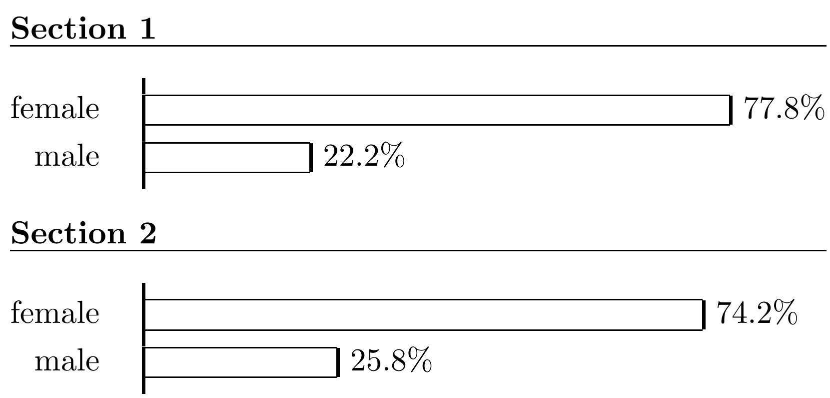


Table 3. Relative frequency distributions for classification, by section.

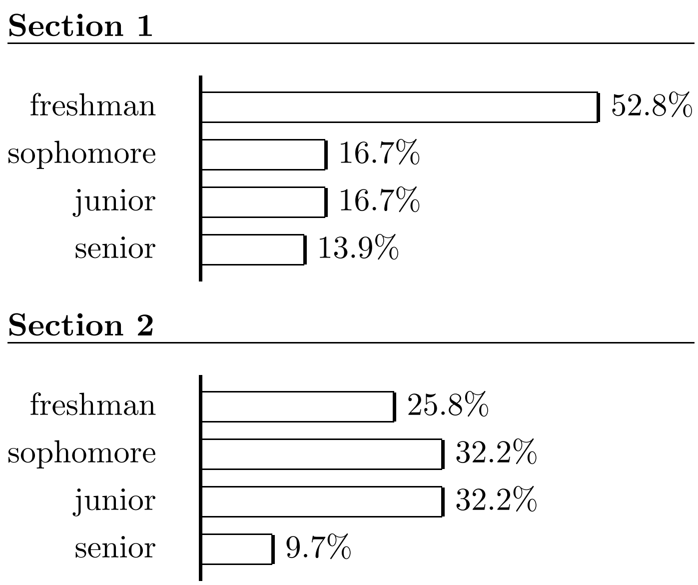
section 1

classification	frequency	relative frequency
freshman	19	.528
sophomore	6	.167
junior	6	.167
senior	5	.139
total	36	1.001

section 2

classification	frequency	relative frequency
freshman	8	.258
sophomore	10	.322
junior	10	.322
senior	3	.097
total	31	.999

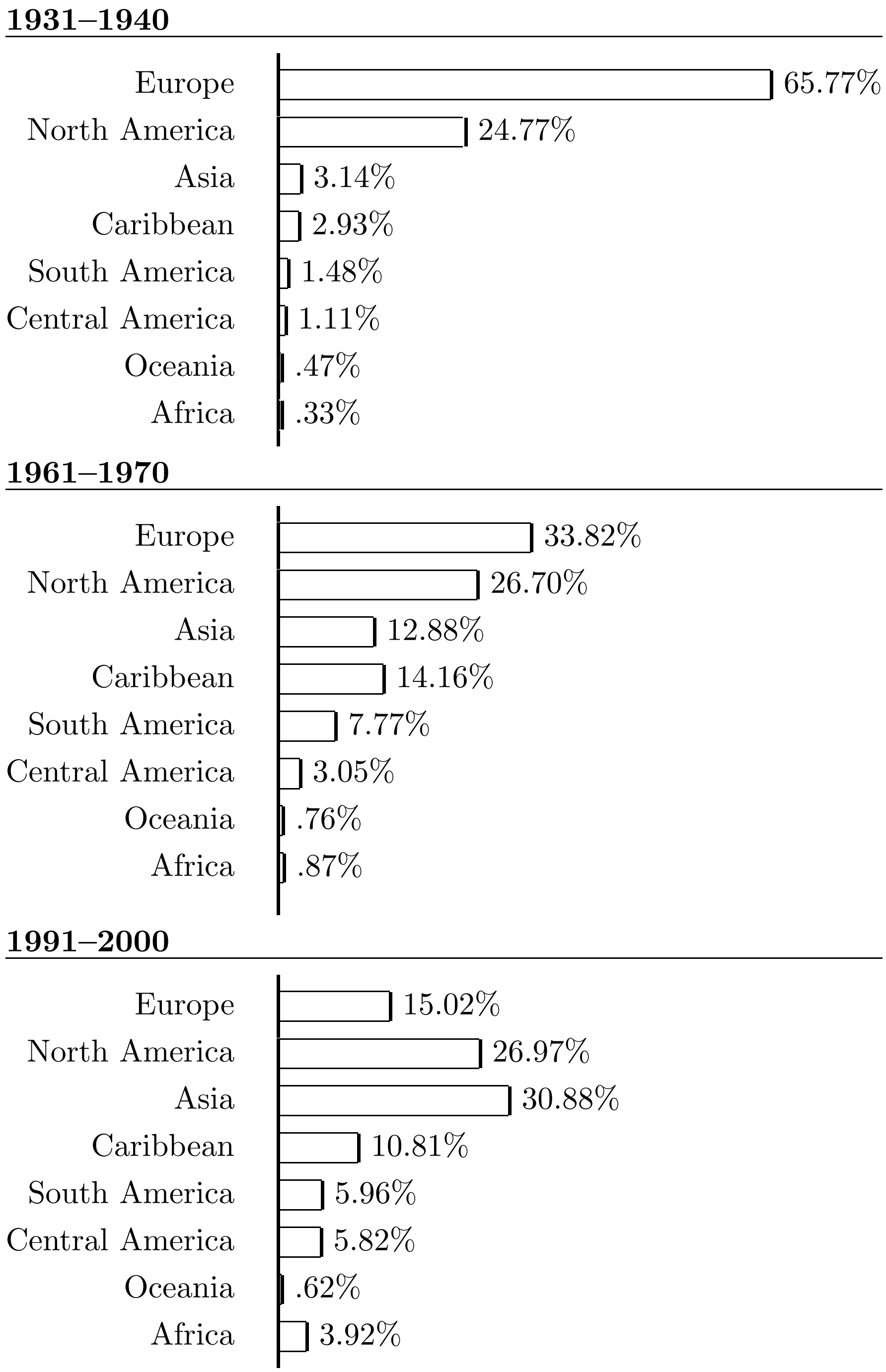
Figure 3. Bar graphs for classification, by section.



The sex distributions for the two sections are essentially the same; in both sections approximately 75% of the students are female. On the other hand, there is a clear difference between the two classification distributions. The section 1 classification distribution is similar to the combined classification distribution with a very large proportion of freshmen. In fact, more than half (52.8%) of the students in section 1 are freshmen. This predominance of freshmen does not happen in section 2 where there is not a single dominant classification value. The most common classifications for section 2 are sophomore and junior, with each of these classifications accounting for 32.2% of the students. In summary, we find that for section 1 the majority of students (52.8%) are freshmen but that for section 2 the majority (64.4%) of the students are sophomores (32.2%) or juniors (32.2%). It is interesting to notice that in both sections the proportions of sophomores and juniors are equal.

Example. Immigrants to the United States. The data concerning immigrants admitted to the United States summarized by decade as raw frequency distributions in Section 1.2 were taken from the *2002 Yearbook of Immigration Statistics*, USCIS, (`www.uscis.gov`). Immigrants for whom the country of last residence was unknown are omitted. For this example a unit is an individual immigrant and these data correspond to a census of the entire population of immigrants, for whom the country of last residence was known, for these decades. Because the region of last residence of an immigrant is a nominal variable and its values do not have an inherent ordering, the values in the bar graphs (and relative frequency distributions) in Figure 4 have been arranged so that the percentages for the 1931–1940 decade are in decreasing order.

Figure 4. Region of last residence for immigrants to USA, by decade.



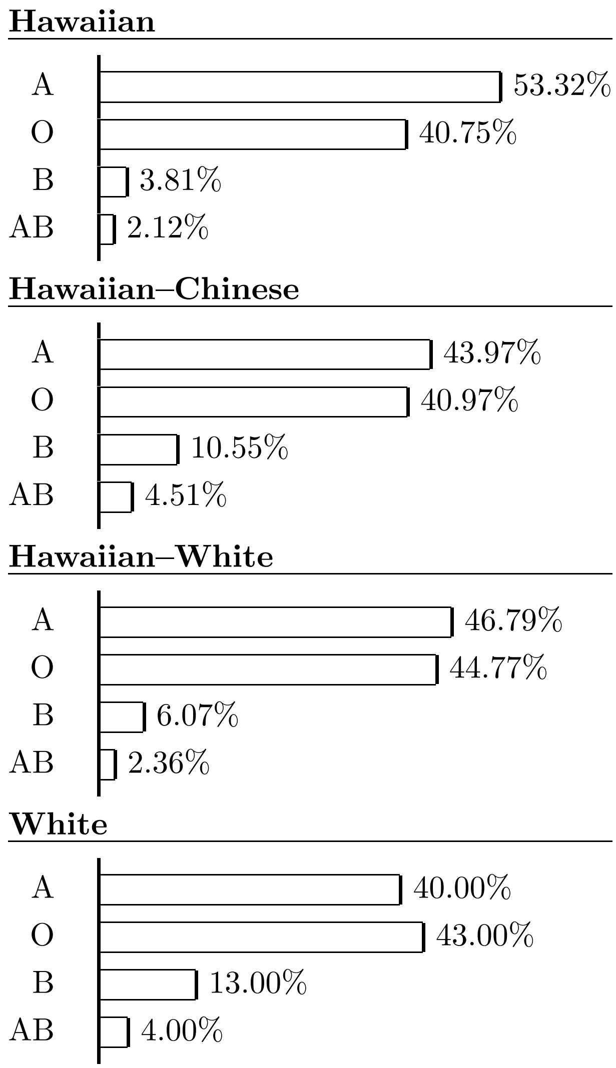
Two aspects of the distributions of region of origin of immigrants which are apparent in these bar graphs are: The decrease in the proportion of immigrants from Europe; and, the increase in the proportion of immigrants from Asia. In 1931–1940 a large majority (65.77%) of the immigrants were from Europe but for the later decades this proportion steadily decreases. On the other hand, the proportion of Asians (only 3.14% in 1931–1940) steadily increases to 30.88% in 1991–2000. Also note that the proportion of immigrants from North America is reasonably constant for these three decades. The patterns we observe in these distributions may be attributable to several causes. Political, social, and economic pressures in the region of origin of these people will clearly have an impact on their desire to immigrate to the US. Furthermore, political pressures within the US have effects on immigration quotas and the availability of visas.

Example. Hawaiian blood types. This example is based on the description in Moore and McCabe, *Introduction to the Practice of Statistics*, Freeman, (1993) of a study discussed in A.E. Mourant, *et al.*, *The Distribution of Blood Groups and Other Polymorphisms*, Oxford University Press, London, 1976. The Blood Bank of Hawaii cross–classified 145,057 individuals according to their blood type (A, AB, B, O) and their ethnic group (Hawaiian, Hawaiian–Chinese, Hawaiian–White, White). The frequencies for each of the 16 combinations of the 4 levels of these two qualitative variables are given in Table 4. This sample of individuals is most likely a convenience sample of blood donors. We will use the classification of an individual by ethnic group as an explanatory (indicator) variable and consider the (conditional) distributions of the nominal qualitative variable blood type for the four ethnic groups. The four columns corresponding to the ethnic groups in Table 4 provide the conditional (frequency) distributions for these groups. Because there is no inherent ordering among the blood types the arrangement of the blood types in Figure 5 is such that the percentages for the Hawaiian group are in decreasing order. The conditional (relative frequency) distributions of blood type for the ethnic groups summarized in Figure 5 clarify the differences in the distributions of blood type for these four ethnic groups.

Table 4. Blood type and ethnic group observed frequencies.

	ethnic group				
blood type	Hawaiian	Hawaiian–Chinese	Hawaiian–White	White	total
A	2490	2368	4671	50008	59537
O	1903	2206	4469	53759	62337
B	178	568	606	16252	17604
AB	99	243	236	5001	5579
total	4670	5385	9982	125020	145057

Figure 5. Conditional distributions of blood type by ethnic group.



2.3 Describing discrete quantitative data

The tabular representations used to summarize the distribution of a discrete quantitative variable, *i.e.,* the frequency and relative frequency distributions, are defined the same as they were for qualitative data. Since the values of a quantitative variable can be viewed as points on the number line, we need to indicate this structure in a tabular representation. In the frequency or relative frequency distribution the values of the variable are listed in order and all possible values within the range of the data are listed even if they do not appear in the data.

First consider the distribution of the number of siblings for the Stat 214 example. The relative frequency distribution for the number of siblings is given in Table 5.

Table 5. Relative frequency distribution for number of siblings.

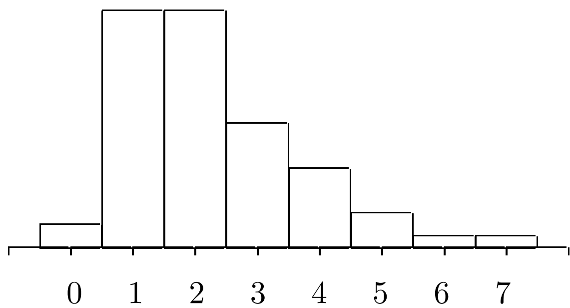
number of siblings	frequency	relative frequency
0	2	.030
1	21	.313
2	21	.313
3	11	.164
4	7	.104
5	3	.045
6	1	.015
7	1	.015
total	67	.999

We will use a graphical representation called a histogram to summarize the distribution of a discrete quantitative variable. Like the bar graph we used to represent the distribution of a qualitative variable, the histogram provides a representation of the distribution of a quantitative variable using area to represent relative frequency. A **histogram** is basically a bar graph modified to indicate the location of the observed values of the variable on the number line. For ease of discussion we will describe histograms for situations where the possible values of the discrete quantitative variable are equally spaced (the distance between any two adjacent possible values is always the same).

Consider the histogram for the number of siblings for the Stat 214 example given in Figure 6. This histogram is made up of rectangles of equal width, centered at the observed values of the variable. The heights of these rectangles are chosen so that the area of a rectangle is the relative frequency of the corresponding value of the variable. There is not a gap between two adjacent rectangles in the histogram unless there is an unobserved possible value of the variable between the corresponding adjacent observed values. For this example there are no gaps; but, there is a gap in the histogram of Figure 8.

In this histogram we are using an interval of values on the number line to indicate a single value of the variable. For example, the rectangle centered over 1 in the histogram of Figure 6 represents the relative frequency of a student having 1 sibling; but its base extends from .5 to 1.5 on the number line. Because it is impossible for the number of siblings to be strictly between 0 and 1 or strictly between 1 and 2, we are identifying the entire interval from .5 to 1.5 on the number line with the actual value of 1. This identification of an interval of values with the possible value at the center of the interval eliminates gaps in the histogram that would incorrectly suggest the presence of unobserved, possible values.

Figure 6. Histogram for number of siblings.

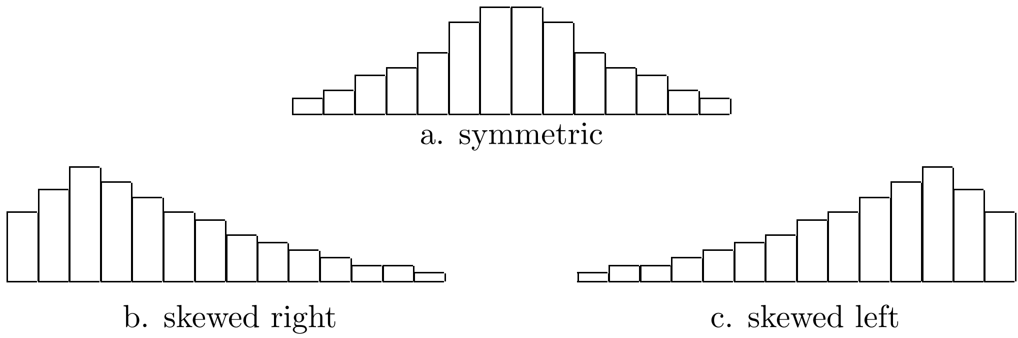


The histogram for the distribution of the number of siblings for the Stat 214 example in Figure 6 has a mound shaped appearance with a single peak over the values 1 and 2, indicating that the most common number of siblings for a student in this group is either 1 or 2. In fact, 31.3% of the students in this group have one sibling and 31.3% have two siblings. It is relatively unusual for a student in this group to be an only child (3%) or to have 5 or more siblings (7.5%).

The histogram of Figure 6, or the associated distribution, is not symmetric. That is, the histogram (distribution) is not the same on the left side (smaller values) of the peak over the values 1 and 2 as it is on the right side (larger values). This histogram or distribution is said to be skewed to the right. The concept of a distribution being skewed to the right is often explained by saying that the right "tail" of the distribution is "longer" than the left "tail". That is, the area in the histogram is more spread out along the number line on the right than it is on the left. For this example, the smallest 25% of the observed values are zeros and ones while the largest 25% of the observed values include values ranging from three to seven. In the present example we might say that there is essentially no left tail in the distribution.

The number of siblings histogram and the histograms for the next three examples discussed below are examples of a very common type of histogram (distribution) which is mound shaped and has a single peak. This type of distribution arises when there is a single value (or a few adjacent values) which occurs with highest relative frequency, causing the histogram to have a single peak at this location, and when the relative frequencies of the other values taper off (decrease) as we move away from the location of the peak. Three examples of common mound shaped distributions with a single peak are provided in Figure 7. The **symmetric** distribution is such that the histogram has two mirror image halves. The **skewed** distributions are more spread out along the number line on one side (the direction of the skewness) than they are on the other side.

Figure 7. Mound shaped histograms with a single peak.



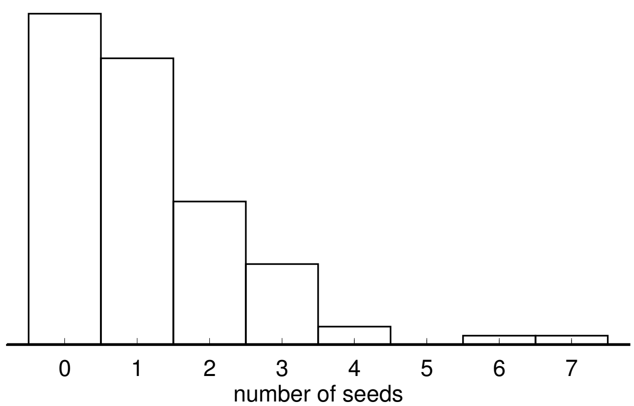
A distribution with a single peak is said to be unimodal, indicating that it has a single mode. The formal definition of a **mode** is a value which occurs with highest frequency. In practice, if two adjacent values are both modes, as are 1 and 2 in the number of siblings example, then we would still say that the distribution is unimodal. Some distributions are bimodal (or multimodal) in the sense of having two distinct modes which are separated by an interval of values with lower relative frequencies. The degree of cloudiness example below provides an example of an extreme version of a bimodal distribution. A more common situation when a bimodal distribution might arise is when the sample under study is a mixture of two subgroups (say males and females) with distinct and well separated modes.

Example. Weed seeds. C. W. Leggatt counted the number of seeds of the weed *potentilla* found in 98 quarter–ounce batches of the grass *Phleum praetense*. This example is taken from Snedecor and Cochran, *Statistical Methods*, Iowa State, (1980), 198; the original source is C. W. Leggatt, *Comptes rendus de l'association international d'essais de semences*, **5** (1935), 27. The 98 observed numbers of weed seeds, which varied from 0 to 7, are summarized in the relative frequency distribution of Table 6 and the histogram of Figure 8. In this example a unit is a batch of grass and the number of seeds in a batch is a discrete quantitative variable with possible values of $0, 1, 2, \ldots$. The distribution of the number of weed seeds is mound shaped with a single peak at zero and it is skewed to the right. The majority of these batches of grass have a small number of weed seeds; but, there are a few batches with relatively high numbers of weed seeds.

Table 6. Weed seed relative frequency distribution.

number of seeds	frequency	relative frequency
0	37	.3776
1	32	.3265
2	16	.1633
3	9	.0918
4	2	.0204
5	0	.0000
6	1	.0102
7	1	.0102
total	98	1.0000

Figure 8. Histogram for number of weed seeds.

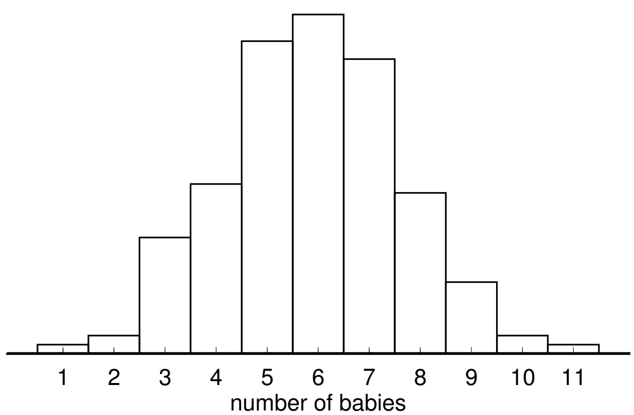


Example. Vole reproduction. An investigation was conducted to study reproduction in laboratory colonies of voles. This example is taken from Devore and Peck, *Statistics*, (1997), 33; the original reference is the article "Reproduction in laboratory colonies of voles", *Oikos*, (1983), 184. The data summarized in Table 7 and Figure 9 are the numbers of babies in 170 litters born to voles in a particular laboratory. In this example a unit is a litter of voles and the number of babies in a vole litter is a discrete quantitative variable with possible values of $1, 2, 3, \ldots$. In this example we see that the distribution of the number of vole babies is mound shaped with a single peak at 6 and it is reasonably symmetric. For these vole litters the majority of the litters have around 6 babies. There are a few litters with relatively small numbers of babies and there are a few with relatively large numbers of babies.

Table 7. Vole baby relative frequency distribution.

number of babies	frequency	relative frequency
1	1	.0059
2	2	.0118
3	13	.0765
4	19	.1118
5	35	.2059
6	38	.2235
7	33	.1941
8	18	.1059
9	8	.0471
10	2	.0118
11	1	.0059
total	170	1.0002

Figure 9. Histogram for number of vole babies.

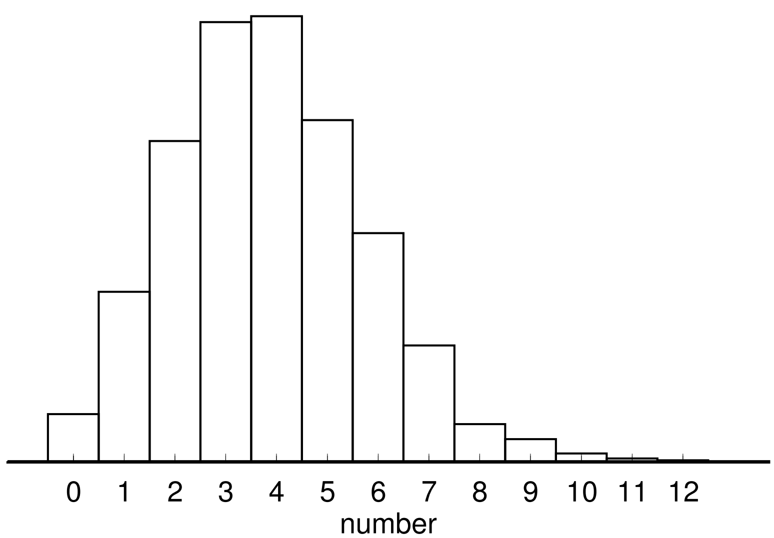


Example. Radioactive disintegrations. This example is taken from Feller, *An Introduction to Probability Theory and its Applications, vol.1*, Wiley, (1957), 149 and Cramér, *Mathematical Methods of Statistics*, Princeton, (1945). In a famous experiment by Rutherford, Chadwick, and Ellis (*Radiations from Radioactive Substances*, Cambridge, 1920) a radioactive substance was observed during 2608 consecutive time intervals of length 7.5 seconds each. In this example a unit is a 7.5 second time interval and the number of particles reaching a counter during the time period is a discrete quantitative variable with possible values of $0, 1, 2, \ldots$. The distribution of the number of radioactive disintegrations is summarized in Table 8 and Figure 10. In this example we see that the distribution of the number of particles per time interval is mound shaped with a single peak around 3 and 4. This distribution is reasonably symmetric but there is some skewness to the right.

Table 8. Radioactive disintegrations relative frequency distribution.

number	frequency	relative frequency
0	57	.0219
1	203	.0778
2	383	.1469
3	525	.2013
4	532	.2040
5	408	.1564
6	273	.1047
7	139	.0533
8	45	.0173
9	27	.0104
10	10	.0038
11	4	.0015
12	2	.0008
total	2608	1.0001

Figure 10. Histogram for radioactive disintegrations.

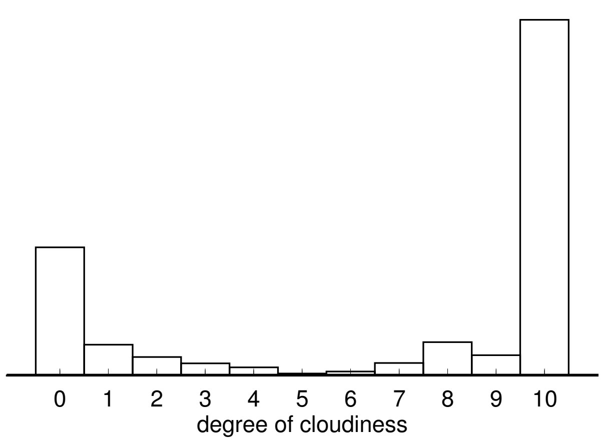


Example. Degree of cloudiness at Breslau. This example is taken from P.R. Rider (1927), *J. Amer. Statist. Assoc.* **22**, 202–208. The estimated degree of cloudiness at Breslau for days during the decade 1876–1885 is summarized in Table 9 and Figure 11. Zero degrees of cloudiness corresponds to an entirely clear day and 10 degrees of cloudiness corresponds to an entirely overcast day. This measurement of degree of cloudiness is essentially a ranking on a scale from 0 to 10 and this variable is properly viewed as being an ordinal qualitative variable. However, as long as we are careful with our interpretation of its numerical values it is reasonable to treat the degree of cloudiness as a discrete quantitative variable. The distribution of the degree of cloudiness is “U”–shaped with a peak at each of the extremes and relatively low relative frequencies in the middle of the range. More properly, we might say that there is a primary peak (mode) at 10 and a smaller secondary peak (mode) at 0. This “U”–shape indicates that for most of these days it was either entirely clear or nearly clear or it was entirely overcast or nearly overcast. There were relatively few days when the degree of cloudiness was in the middle of the range. The most common value was 10, entirely overcast (57.19%), and the second most common value was 0, entirely clear (20.56%).

Table 9. Degree of cloudiness at Breslau relative frequency distribution.

degree of cloudiness	frequency	relative frequency
0	751	.2056
1	179	.0490
2	107	.0293
3	69	.0189
4	46	.0126
5	9	.0025
6	21	.0057
7	71	.0194
8	194	.0531
9	117	.0320
10	2089	.5719
total	3653	1.0000

Figure 11. Histogram for degree of cloudiness at Breslau.



2.4 Describing continuous quantitative data

There is a fundamental difference between summarizing and describing the distribution of a discrete quantitative variable and summarizing and describing the distribution of a continuous quantitative variable. Since a continuous quantitative variable has an infinite number of possible values, it is not possible to list all of these values. Therefore, some changes to the tabular and graphical summaries used for discrete variables are required.

In practice, the observed values of a continuous quantitative variable are discretized, *i.e.,* the values are rounded so that they can be written down. Therefore, there is really no difference between summarizing the distribution of a continuous variable and summarizing the distribution of a discrete variable with a large number of possible values. In either case, it may be impossible or undesirable to actually list all of the possible values of the variable within the range of the observed data. Thus, when summarizing the distribution of a continuous variable, we will group the possible values into intervals.

Figure 12. Stem and leaf histogram for weight.

In this stem and leaf histogram the stem represents tens and the leaf represents ones. (pounds)

stem	leaf
9	56
10	355
11	0000000255567
12	000555
13	00003455555
14	000000055
15	00556
16	0255
17	000005
18	0055
19	000
20	5

To make this discussion more concrete, consider the weights of the students in the Stat 214 example. We can group the possible weights into the intervals: 90–100, 100–110, . . ., 200–210. We will need to adopt an endpoint convention so that each possible weight belongs to only one of these intervals. We will adopt the endpoint convention of including the left (lower) endpoint and excluding the right (upper) endpoint. Under this convention the interval 90–100 includes 90 but excludes 100. A stem and leaf histogram of the weights of the Stat 214 students is given in Figure 12. The **stem and leaf histogram** is an easily constructed version of a **frequency histogram** (a histogram based on frequencies instead of relative frequencies). The stem and leaf histogram uses the numbers themselves to form the rectangles of the histogram. The stem indicates the interval of values while the leaves provide the "rectangle." For the weight data the actual weight of a student is decomposed into tens (the stem) and ones (the leaf). For example, the first weight is 95 pounds which is decomposed as $95 = 9$ tens $+ 5$ ones. Therefore, the weight 95 appears as a 5 leaf in the leaves of the 9 stem.

Notice that the leaves in the stem and leaf histogram of Figure 12 are arranged in increasing order within each stem. Having ordered leaves in the stem and leaf histogram makes certain subsequent tasks easier. For example, having ordered leaves makes it easier to change the stems (change the intervals used to group the values) if this is necessary to obtain a more informative stem and leaf histogram. Furthermore, we can easily determine certain summary statistics directly from a stem and leaf histogram with ordered leaves. (We will discuss summary statistics in Chapter 3.) When constructing a stem and leaf histogram from unordered data, the best way to get ordered leaves is to first form a preliminary stem and leaf histogram with unordered leaves and then revise it to get ordered leaves.

Once the stem and leaf histogram is formed it is easy to construct a frequency distribution, a relative frequency distribution, and a formal relative frequency histogram, if these are desired. By counting the numbers of leaves corresponding to each stem in the stem and leaf histogram we can easily form the corresponding frequency and relative frequency distributions and the formal (relative frequency) histogram.

The weight distribution histogram of Figure 12 has an asymmetric mound shape peaking in the 110–120 pound range and showing skewness to the right. There is a lot of variability in the weights of these students. The majority of the students have weights in the 110–150 pound range; but, weights in the 150–190 pound range are also fairly common.

The appearance of two peaks, one in the 11 (110 pound) stem and one in the 13 (130 pound) stem is probably due to the way these students rounded their weights; therefore, it seems reasonable to say that this distribution has a single peak. Notice that, in general, the appearance of a stem and leaf histogram or a formal histogram for a continuous variable depends on the choice of the intervals used in its construction and minor features such as multiple local modes which are not very far apart might disappear if the intervals were shifted slightly.

Since this weight distribution corresponds to a group consisting of both females and males, we might expect to see two separate peaks; one located at the center of the female weight distribution and another located at the center of the male weight distribution. However, there does not appear to be much evidence of this. Separate stem and leaf histograms for the weight distributions of the females and the males are given in Figure 13. Some care is required in comparing these two stem and leaf (frequency) histograms due to the disparate sample sizes. There are 51 female weights but only 16 male weights.

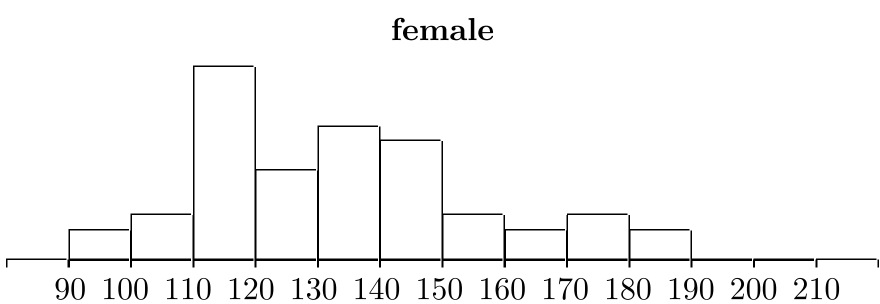
The peak in the female weight distribution stem and leaf histogram appears to be much more pronounced than the peak in the male weight distribution stem and leaf histogram. However, the formal (relative frequency) histograms exhibited in Figure 14 show that this difference in peakedness is not so large. The female weight distribution is skewed to the right with one peak in the 110–150 pound range. The male weight distribution is much more uniform without strong evidence of skewness.

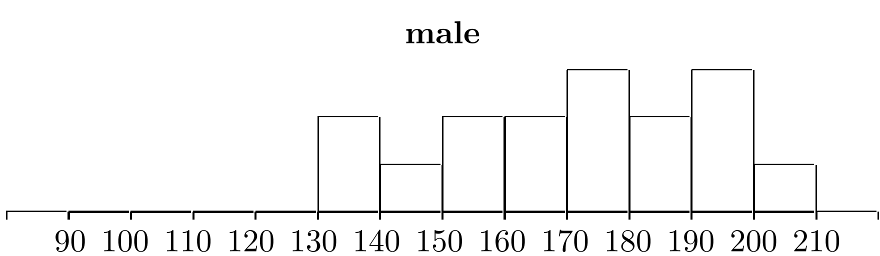
Figure 13. Stem and leaf histograms for weight, by sex.

In these stem and leaf histograms the stem represents tens and the leaf represents ones. (pounds)

Female		**Male**	
9	56	9	
10	355	10	
11	0000000255567	11	
12	000555	12	
13	000345555	13	05
14	00000055	14	0
15	005	15	56
16	05	16	25
17	000	17	005
18	05	18	05
19		19	000
20		20	5

Figure 14. Histograms for weight, by sex.





The stem and leaf histograms and formal histograms we formed for the weight distribution were based on intervals of length 10 (10 pounds), *e.g.*, 90–100, 100–110, *etc.* Notice that we chose this interval length when we chose to use the last digit of the weight of a student as the leaf and the remaining digits of the weight as the stem in the stem and leaf histogram. In some situations using the last digit of the variable value as the leaf may yield inappropriate intervals. Consider the stem and leaf histogram for the height distribution for the Stat 214 example given in Figure 15. Clearly the majority of the heights are in the 60's, but the shape of the distribution is not clear from this stem and leaf histogram. The intervals used here are too long causing the stem and leaf histogram to be too compressed along the number line to give a useful indication of the shape of the distribution.

Figure 15. Stem and leaf histogram for height.

In this stem and leaf histogram the stem represents tens and the leaf represents ones. (inches)

5	9
6	001111222222333334444444455556666666666677778888888999
7	000112222245

We can refine the stem and leaf histogram by changing the lengths of the intervals into which the data are grouped. This refinement can be viewed as a splitting of the stems of the stem and leaf histogram. To avoid distortion we need to subdivide the intervals (split the stems) so that each of the resulting intervals is of the same length. We can easily do this by either splitting the stems once, yielding 2 intervals of length 5 for each stem instead of 1 interval of length 10, or by splitting the stems five times, yielding 5 intervals of length 2. To demonstrate this splitting of stems stem and leaf histograms of the height distribution with stems split in these fashions are provided in Figures 16 and 17, respectively. In this particular case the stem and leaf histogram of Figure 17 (stems split into five) seems to provide the most informative display of the shape of the height distribution. The height distribution is reasonably symmetric with a single peak at the 66–67 interval.

Figure 16. Stem and leaf histogram for height with stems split into two.

In this stem and leaf histogram the stem represents tens and the leaf represents ones. (inches)

Low stem leaves: 0,1,2,3,4
High stem leaves: 5,6,7,8,9

5	
5	9
6	0011112222223333344444444
6	55556666666666677778888888999
7	00011222224
7	5

Figure 17. Stem and leaf histogram for height with stems split into five.

In this stem and leaf histogram the stem represents tens and the leaf represents ones. (inches)

First stem leaves: 0,1
Second stem leaves: 2,3
Third stem leaves: 4,5
Fourth stem leaves: 6,7
Fifth stem leaves: 8,9

5	9
6	001111
6	22222233333
6	444444445555
6	666666666667777
6	8888888999
7	00011
7	22222
7	45

To complete our discussion of stem and leaf histograms consider the hypothetical example, with values between -3.9 and 3.9, of Figure 18. The first thing you should notice is that there is a -0 stem and a 0 stem. The negative 0 stem corresponds to the interval from -1 to 0 (not including -1) and the positive 0 stem corresponds to the interval from 0 to 1 (not including 1). If there were any zero observations we could place half of them with each of the zero stems. Notice also that the leaves for the negative stems decrease from left to right so that as we read through the histogram (going from left to right) the values increase from the minimum -3.9 to the maximum 3.9.

Figure 18. A stem and leaf histogram for hypothetical data with negative values and positive values.

-3	98664
-2	8773210
-1	7664432110
-0	9776555442211
0	11222344467
1	1222345578
2	34466789
3	445569

2.5 Summary

In this chapter we discussed tabular and graphical methods for summarizing the distribution of a variable X, *i.e.*, methods for summarizing the way in which the possible values of X are distributed among the units in the sample. The basic idea underlying these summaries is that of using relative frequencies (proportions or percentages) to show how the total relative frequency of one (100%) is partitioned into relative frequencies for each of the possible values of X.

A relative frequency distribution is a table listing the possible values of X and the associated relative frequencies with which these values occurred in the sample. For a qualitative variable or a discrete quantitative variable it is usually possible to tabulate all of the possible values and their relative frequencies. For a discrete quantitative variable with many possible values or a continuous quantitative variable, there are generally too many possible values to list each individually and it is necessary to group the possible values into intervals and then tabulate the relative frequencies for each of these intervals of values.

A graphical representation of the distribution of X is based on the identification of area with relative frequency. Thus, a graphical representation provides a decomposition of a region of area one, representing the total relative frequency of one (100%), into subregions of area equal to the relative frequencies of each of the possible values (or intervals of values) of X. For qualitative variables we emphasized the bar graph with rectangular regions for each value of X. For quantitative variables we used a histogram which is basically a bar graph with the bars suitably arranged along the number line to indicate the relative locations of the values of X. We also discussed stem and leaf histograms which are easily constructed raw frequency histograms; and we noted that stem and leaf histograms should be converted to proper relative frequency histograms before making comparisons of two or more distributions.

The representation of the distribution of a quantitative variable via a histogram allows us to discuss the shape of the distribution. We discussed some basic shapes with most of our emphasis on the distinction between skewed and symmetric mound shaped distributions. For skewed distributions we defined the terms skewed left and skewed right.

2.6 Exercises

For each of the examples in Section 1.2 (excluding those already treated in this chapter): construct suitable tabular and graphical summaries of the distribution(s) and discuss the distribution of the variable(s).

Notes:

For the examples with two or more groups (DiMaggio and Mantle, Guatemalan cholesterol, gear tooth strength), compare and contrast the distributions of the variable for the two (or more) groups.

For the paired data examples (wooly–bear cocoons, homophone confusions) find the differences for each pair of data values and describe the distribution of the differences.