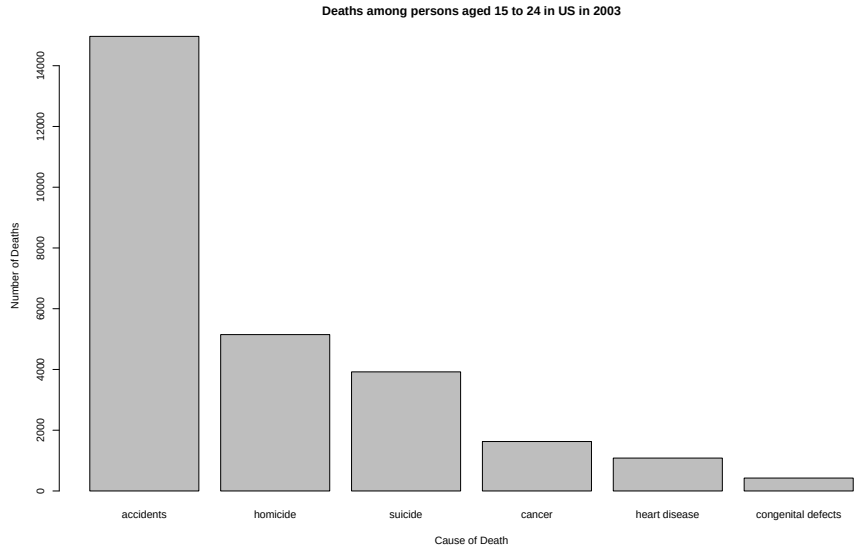
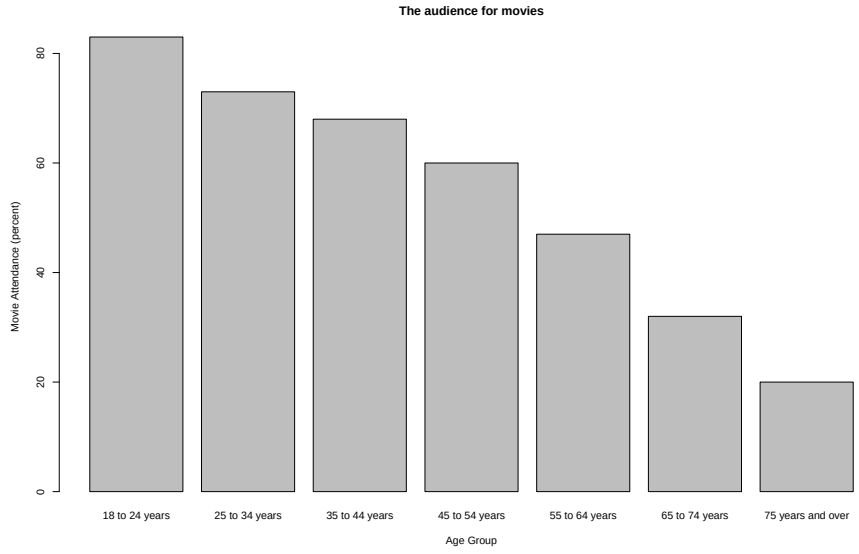
Stat 160 Fall 2008
Solutions to Assignment #1

#1.26 (a) A bar graph displaying the number of deaths among persons aged 15 to 24 years in the United States in 2003 due to accidents, homicide, suicide, cancer, heart disease, and congenital defects is shown below.



#1.26 (b) In order to make a pie chart, we would need to know the total number of deaths in this age group so that we could compute the number of deaths due to other causes.

#1.28 (a) A bar graph displaying the data on percentage of people aged 18 or older who attended a movie in the past 12 months is shown below.

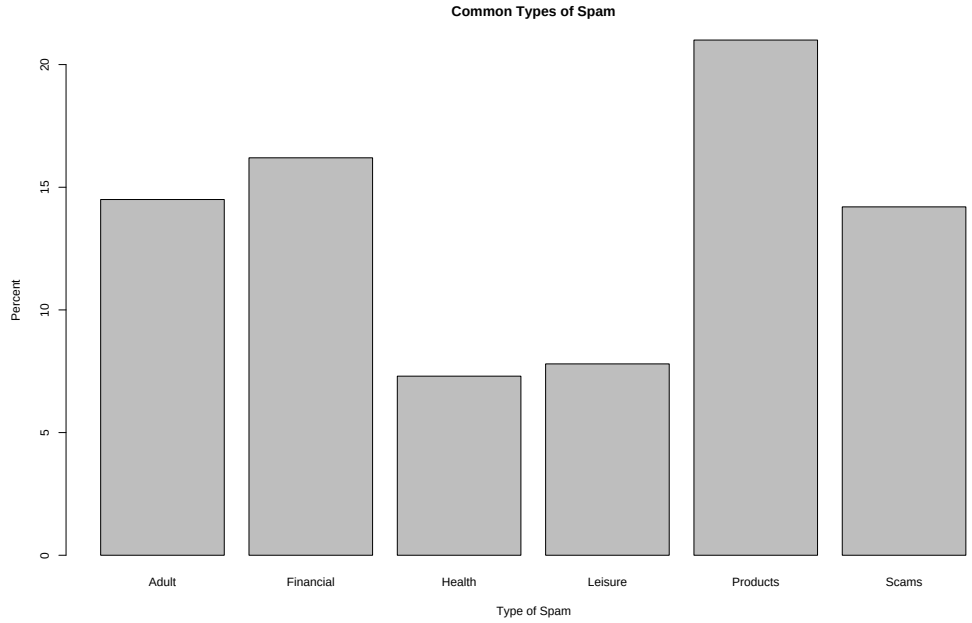


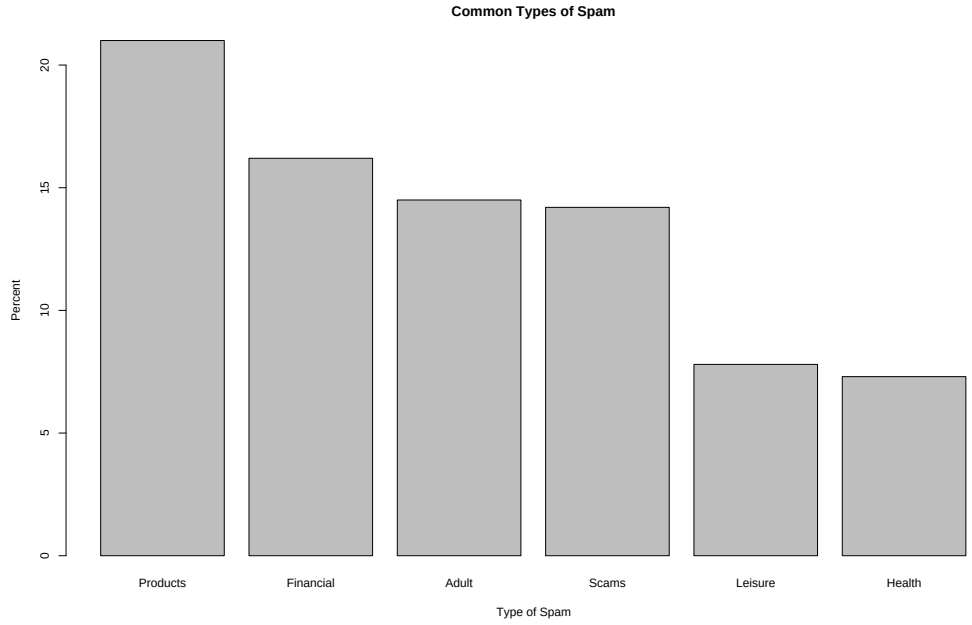
The main feature of this graph is that movie attendance decreases steadily with age.

#1.28 (b) It would not be correct to make a pie chart for these data. To create a pie chart, we would need to know what fraction of all moviegoers were in each age range, rather than what fraction of each age range goes to the movies.

#1.28 (c) In order to know what percent of the movie audience is 18 to 24 years old, we would need to know two things, namely the number of people who go to the movies, and the number of moviegoers in that age group. (We know neither of these things.)

#1.29 Two bar graphs displaying the most common types of spam are shown below. The first bar graph displays the bars ordered alphabetically, while the seond bar graph displays the bars ordered by height from tallest-to-shortest (left-to-right).





#1.32 Histogram (a) corresponds to variable (4). Minutes spent studying would likely be skewed to the right since many study for a short period of time, while only a few study longer.

Histogram (b) corresponds to variable (2), while histogram (c) corresponds to variable (1)—unless this was a particularly unusual class! We would expect that male/female counts should be somewhat close, while right-handed students should outnumber left-handed students substantially. (In fact, roughly 10% to 15% of the population as a whole is left-handed.)

Histogram (d) corresponds to variable (3). We would expect a fair amount of variation in student heights, but no particular skewness to such a distribution.

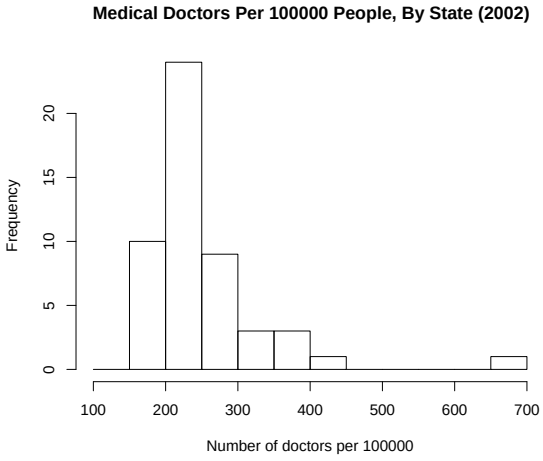
#1.34 (a) In a state with many people, more doctors are needed to serve the larger population. (For example, having 1000 doctors in Rhode Island would be very different from having 1000 doctors in California.) Thus, the raw numbers of doctors is not comparable across states, but normalizing the number of doctors per 100 000 people does allow for cross-state comparison.

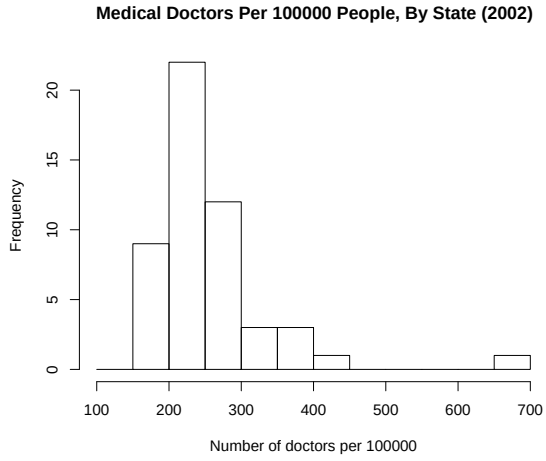
#1.34 (b) As noted in the textbook, it is extremely important to have well-defined histogram classes to ensure that no data points falls into two different classes. The choice you make is somewhat arbitrary, as is the width of each class. As noted on page 13:

> "There is no one right choice of classes in a histogram. Too few classes will give a *skyscraper* graph, with all values in a few classes with tall bars. Too many classes will produce a *pancake* graph, with most of the classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. You must use your judgment in choosing classes to display the shape."

Unfortunately, the text makes no comments about where to place the tick marks on the horizontal axis. This can be a problem since it is not obvious to which class a given tick mark belongs—the one to the left or the one to the right. As such, it is extremely important to include a clear definition of the classes that you choose.

To illustrate the previous comment, here are two possible histograms that both answer this question.





The class heights are all different because it is not clear to which classes the data points 200, 250, and 300 belong. This ambiguity is resolved only when the class definitions are included. For the histogram on the left, the classes are as follows.

class	count
150 <number of doctors per 100000≤ 200	10
200 <number of doctors per 100000≤ 250	24
250 <number of doctors per 100000≤ 300	9
300 <number of doctors per 100000≤ 350	3
350 <number of doctors per 100000≤ 400	3
400 <number of doctors per 100000≤ 450	1
450 <number of doctors per 100000≤ 500	0
500 <number of doctors per 100000≤ 550	0
550 <number of doctors per 100000≤ 600	0
600 <number of doctors per 100000≤ 650	0
650 <number of doctors per 100000≤ 700	1

For the histogram on the right, the classes are as follows.

class	count
150 ≤number of doctors per 100000< 200	9
200 ≤number of doctors per 100000< 250	22
250 ≤number of doctors per 100000< 300	12
300 ≤number of doctors per 100000< 350	3
350 ≤number of doctors per 100000< 400	3
400 ≤number of doctors per 100000< 450	1
450 ≤number of doctors per 100000< 500	0
500 ≤number of doctors per 100000< 550	0
550 ≤number of doctors per 100000< 600	0
600 ≤number of doctors per 100000< 650	0
650 ≤number of doctors per 100000< 700	1

Furthermore, it is worth mentioning that neither histogram is *more correct* than the other; both are correct displays of this data. It is, however, vital that the class definition be included with the histogram that is ultimately used.

The distribution is clearly skewed to the right, with the District of Columbia a high outlier. The states all have numbers between 161 and 427. The District of Columbia is different from the states in that it includes very little area that would be considered "rural," where we would expect the density of doctors would drop off considerably. (Observe that the states with large cities tend to have high numbers. The District of Columbia is an extreme case, because it consists mainly of a single large city, namely Washington, D.C.)

#1.36 There was a lot of confusion with the grading of this problem. Unfortunately, there appears to be an error in the solutions manual that the grader was following. The following correct solution matches the technique described in Example 1.10 on pages 19–21. The original data is given below.

173	1411	4700	505
234	1431	1702	304
616	1250	1119	425
344	2246	2407	214
515	1793	1049	385
576	1793	505	445
727	2809	998	676

The data rounded to the nearest hundred are as follows.

200	1400	4700	500
200	1400	1700	300
600	1300	1100	400
300	2200	2400	200
500	1800	1000	400
600	1800	500	400
700	2800	1000	700

We now drop the trailing zeroes to transform the data as follows.

2	14	47	5
2	14	17	3
6	13	11	4
3	22	24	2
5	18	10	4
6	18	5	4
7	28	10	7

In this last form, each data point is a positive whole number less than 50. Thus, we can use four stems, namely 0, 1, 2, 3, 4, each representing the tens-digit of the data in this last form. Thus, the corresponding stemplot is as follows.

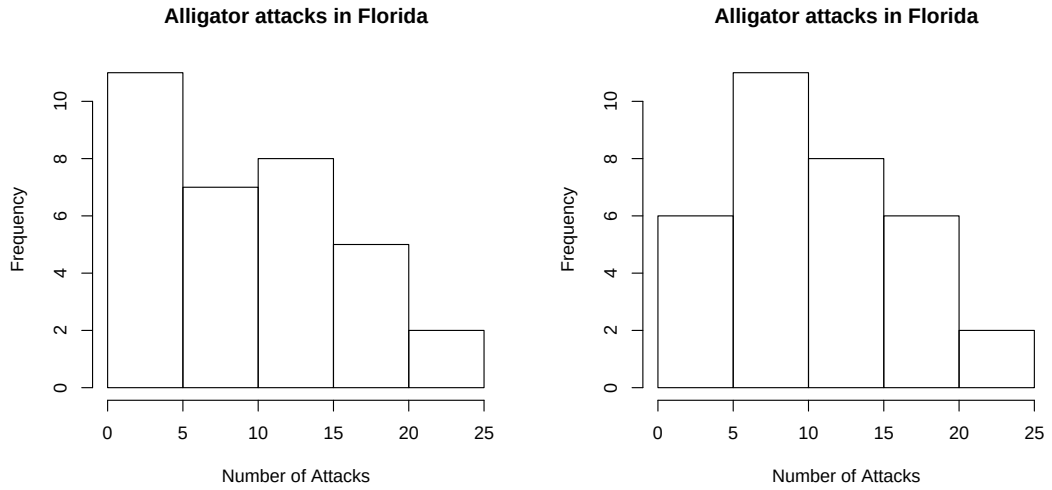
0	2	2	2	3	3	4	4	4	5	5	5	6	6	7	7
1	0	0	1	3	4	4	7	8	8						
2	2	4	8												
3															
4	7														

However, for this problem, we were specifically told to split the stems. The process for this is described in the second paragraph on page 21. The final answer is therefore as follows.

0	2	2	2	3	3	4	4	4
0	5	5	5	6	6	7	7	
1	0	0	1	3	4	4		
1	7	8	8					
2	2	4						
2	8							
3								
3								
4	7							

It is now clear from this stemplot that the distribution is skewed to the right with one apparent (high) outlier (namely, 4700 million sole from 1987). The centre of this data is around 700 million and the spread is from 173 million to 4700 million (or, with the outlier ignored, the bulk of the data is spread between 173 million and 2809 million).

#1.44 (a) Please refer to the comments about drawing histograms made in the solution to Exercise #1.34. To again illustrate the comments made in that solution, here are two possible histograms that both answer this question.



The class heights are all different because it is not clear to which classes the data points 5, 10, 15, and 20 belong. This ambiguity is resolved only when the class definitions are included. For the histogram on the left, the classes are as follows.

class	count
$0 <$attacks≤ 5	11
$5 <$attacks≤ 10	7
$10 <$attacks≤ 15	8
$15 <$attacks≤ 20	5
$20 <$attacks≤ 25	2

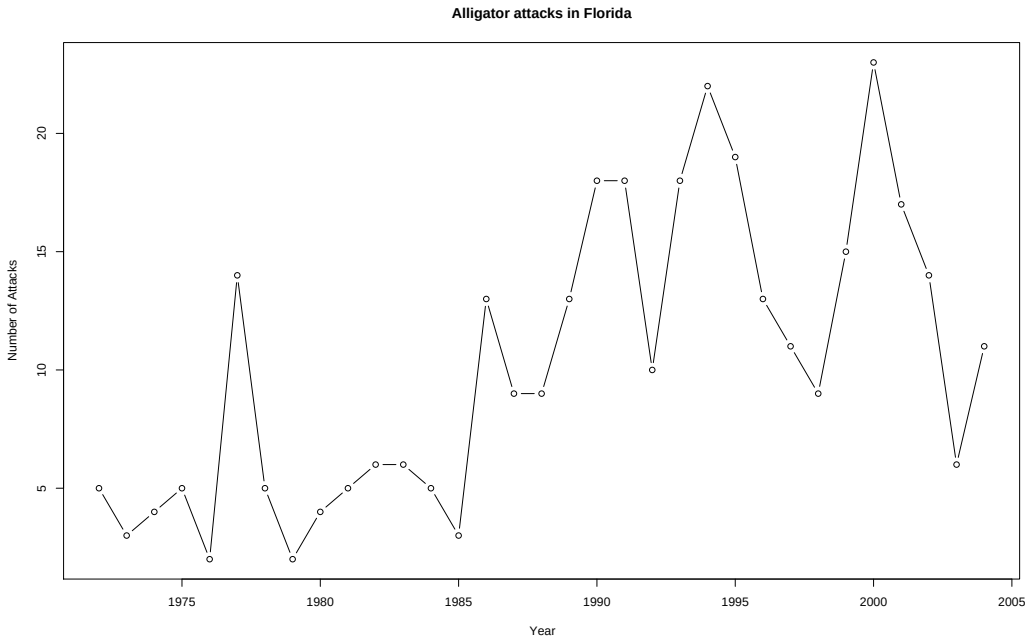
For the histogram on the right, the classes are as follows.

class	count
$0 \leq$attacks< 5	6
$5 \leq$attacks< 10	11
$10 \leq$attacks< 15	8
$15 \leq$attacks< 20	6
$20 \leq$attacks< 25	2

Furthermore, it is again worth mentioning that neither histogram is *more correct* than the other; both are correct displays of this data. It is, however, vital that the class definition be included with the histogram that is ultimately used.

Now, in either case, it appears that there is a slight skew to the right. (This is more pronounced in the histogram on the left, but also appears in the histogram on the right.) Furthermore, the middle number of attacks is 9 per year.

#1.44 (b) A time plot of the number of alligator attacks in Florida from 1972 through 2004 is shown below.



The time plot shows a lot of fluctuation, and so it is difficult to make any general conclusion. There has been, perhaps, a slight increase in recent years, although the two most recent years (2003, 2004) have both been lower than 1977. However, with the exception of 1977, all of the years after 1985 have had more attacks than any of the years through 1985. Thus, the typical number of attacks over the entire period is probably not a good indication of what to expect in the future; it appears, at least, to make future predictions that we should exclude data from years before 1985.

As noted in the text, the continuing increase in Florida's population also makes prediction difficult. Therefore, in order to predict future alligator attacks it might be more useful to consider the number of alligator attacks per 1 000 000 people. For those interested, population data is available online from Florida's Office of Economic & Demographic Research at (`http://edr.state.fl.us/population.htm`).