Chapter 2
Summarizing Data With Graphs

This chapter considers bar graphs, box plots, dot plots, histograms, and stem plots. The distribution of a categorical variable lists counts (frequencies) or percents.

Definition 2.1. The *distribution* of a variable tells what values it takes and how often.

2.1 The Bar Graph for Categorical Data

Example 2.1. 2017 SIU student gender F: 47.3%, M: 52.7%. For decades, SIU has been one of the few universities where the female percentage is lower than the male percentage. In *R*, the bar graph is made with the barplot command. See Figure 2.1 and the *R* code below.

```
barplot(c(47.3,52.7),names.arg=c("F","M"))
```

Definition 2.2. A *bar graph* (or barplot or barchart or bar plot) is used to display categorical data. The vertical axis height = percent or count and the horizontal axis has categories. Separate the bars with a space. Bar widths are equal.

Example 2.2. The *R* data set `VADeaths` gives the death rates measured per 1000 population per year. They are cross-classified by age group (rows) and population group (columns). The age groups are: 50-54, 55-59, 60-64, 65-69, 70-74 and the population groups are Rural/Male, Rural/Female, Urban/Male and Urban/Female. Try the following command.

```
barplot(VADeaths)
```

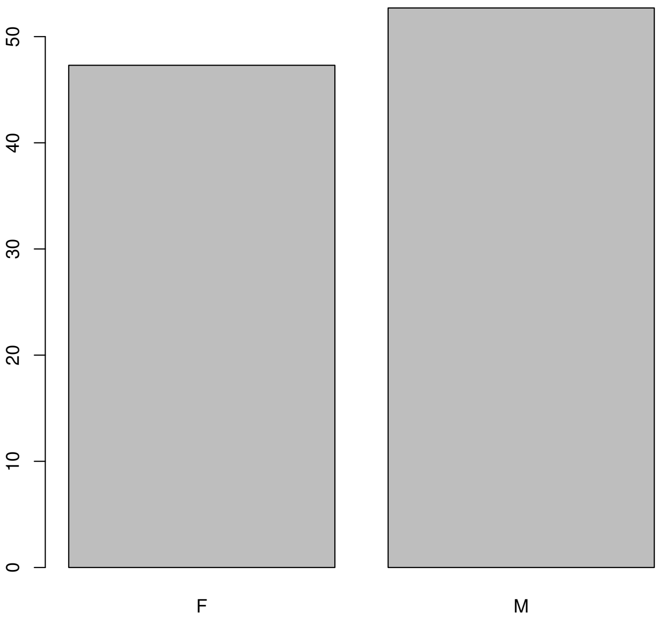

Fig. 2.1 Bar Graph for 2017 SIU Student Gender

2.2 Graphs for Quantitative Variables

We will use two data sets. The *R islands* data set gives the areas in thousands of square miles of the land masses which exceed 10,000 square miles. The Buxton (1920) data set has several variables that are measurements taken on 87 people. We will be interested in the *height* in mm in the variable `buxy`. Five heights were recorded near 19mm (about 0.7 inches) high. These five cases are outliers.

Definition 2.3. An *outlier* is a case that lies far away from the bulk of the data.

Definition 2.4. A *(frequency) histogram* is a graph that summarizes the distribution of a quantitative variable. Divide the range of the data into *classes* of *equal* width. Each observation should fall in *exactly one* class. Find the count of observations for each class (often use a tally) if a distribution table of classes and counts is not given. On the horizontal axis, mark the scale of the variable. Put the count (frequency) on the vertical axis. Bars have the same width. Labeling the top of each bar with the count can be useful.

Since the bars have equal width, the area of each bar is proportional to the percentage of observations in each class. Bar graphs have gaps between bars, histograms have no gaps unless a class has a count of 0.

Warning. Do not make breaks in the vertical axis for a bar graph or a histogram. If a break is made, then the area of the bars is no longer proportional to the percentage of observations in each class. Distances should be equally spaced. There can be one break on the horizontal axis for a histogram, but distances should be equally spaced. **This error is so common**, that the axes should be given for exams and quizzes for this class.

Example 2.3. The set of axes to the lower left has a break in the vertical axis and the distance from 0 to 300 is not equal to the distance between 300 and 600 or 600 and 900. The horizontal axis is wrong since the interval from 1100 to 1200 corresponding to a zero count has been omitted. The set of axes to the lower right is better although the distance from 0 to 300 appears to be less than the distance of 300 to 600. There is one break in the horizontal axis just before 1000, denoted by the two small vertical bars. The distance from the right edge of the graph to 1000 is not the same as the distances of 100 that separate the four numbers. The horizontal break can be useful to show detail that would be obscured if no break was used: 1000 to 1300 would be a small part of the graph if the horizontal axis used 0, 100, 200, ..., 1300, 1400. Note that an observation of 1200 goes with the 3rd class, not the 2nd class.

```
class              count
1000  <=  <  1100  600
1100  <=  <  1200   0
1200  <=  <  1300  900
      wrong                    better
900|        -----          |
   |        |    |      900|              -----
   |        |    |         |              |    |
600|   -----|    |      600|    -----     |    |
   |   |    |    |         |    |    |    |    |
   |   |    |    |      300|    |    |    |    |
300|_|||____|____|____     |_||_|____|____|____|__
       |    |    |              |    |    |    |
       1000 1100 1300           1000 1100 1200 1300
            1200
```

To interpret a histogram, look for an overall pattern and deviations from the pattern. Shape, center, and spread are important. The center is where the histogram is located (a typical or center value on the horizontal axis). There are three common shapes: left skewed, right skewed, and approximately symmetric. A graph is symmetric if graph is a mirror image about some midpoint. A graph is right skewed if it has a long right tail, e.g. income data, where most incomes are 50000 or less but a few incomes are very large. A graph is left skewed if it has a long left tail. The left tail is the leftmost part of the graph while the right tail is the rightmost part of the graph. If x is right skewed then $-x$ is left skewed. If x is left skewed then $-x$ is right skewed.

Example 2.4. Figure 2.2 has four histograms: the one in the upper left is approximately symmetric and the two tails are about the same length. The one in the upper right is right skewed with a long right tail. The one in the lower left is left skewed with a long left tail. The last histogram of $\log(x)$ is left skewed, but is less skewed than the histogram of x. Each artificial data set has 1000 cases. See the following *R* code.

```
par(mfrow=c(2,2)) #four graphs
hist(rnorm(1000))
x <- rexp(1000)
hist(x)
hist(-x)
hist(log(x))
par(mfrow=c(1,1))
```

Using $y = \log(x)$ can cause y to have less skew than x if x is skewed. This function is especially useful for reducing skew if $x > 0$ and $\max(x)/\min(x) \geq 10$ where $\max(x)$ is the largest (maximum) value in the data set, and $\min(x)$ is the smallest (minimum) value in the data set.

Definition 2.6. The logarithm function is the inverse function of exponentiation: $\log_b(b^y) = y$ where b is the base of the logarithm. Hence $\log_{10}(100) = \log_{10}(10^2) = 2$. We will usually use the natural logarithm with base $b = e \approx 2.72$, denoted by $y = \log(x)$. We need $x > 0$.

For the median and quartiles, see Chapter 3. The box plot is roughly symmetric if the line corresponding to the median is close to the middle of the plot, and the whiskers have about the same length. If the right whisker is longer than the left (or the circles extend further to the right), then the data is likely right skewed. If the left whisker is longer than the right (or the circles extend further to the left), then the data is likely left skewed.

Definition 2.7. The *five number summary* is the minimum, Q_1, the median, Q_3 and the maximum. The *box plot* or boxplot is a box from Q_1 to Q_3 with a line at the median $= Q_2$. If sketched by hand, whiskers extend from Q_1 to the minimum and from Q_3 to the maximum. *R* uses another rule to make the whiskers, say Q_1 to low and Q_3 to high, and puts circles past the whiskers to indicate possible outliers.

Warning. Computer output often gives several numbers besides the five number summary, such as the (sample) mean.

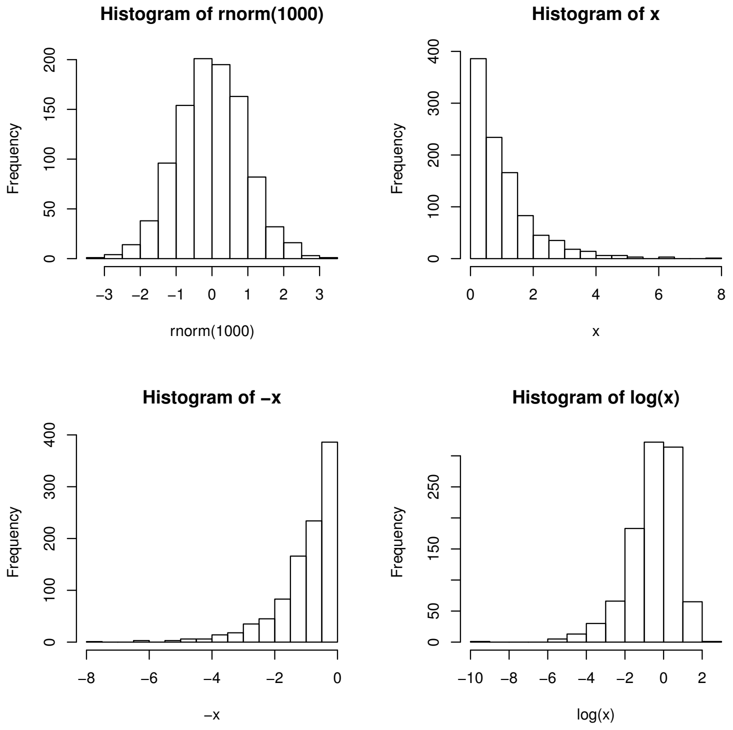

Fig. 2.2 Histograms for Example 2.4.

Example 2.5. Figure 2.3 shows histograms and boxplots for the islands data and the Buxton (1920) heights data. Note that the islands data is right skewed and the heights data has outliers (gaps in the plot). The box plot for the heights data looks roughly symmetric if the outliers are ignored.

Stem plots and dot plots are for small data sets.

Definition 2.8. To make a *stem plot*, divide the data into groups = stems that contain all but the final digit = leaf. Place the stems in order in a vertical column (e.g. smallest on top, largest on bottom). A vertical line separates the stems from the leaves. Each leaf is written to the right of its stem in increasing order. Write the stem and leaf units on the plot or tell where the decimal goes.

R tells where the decimal goes and truncates the data rather than rounding the data. Suppose the stem is 4 and the leaf is 5. If the decimal is j digits to the right of the | (stem), then the value is $4.5(10^0) = 4.5$ if $j = 0$ (the decimal is at the |), $4.5(10) = 4.5(10^1) = 45$ if $j = 1$, $4.5(100) = 4.5(10^2) = 450$ if

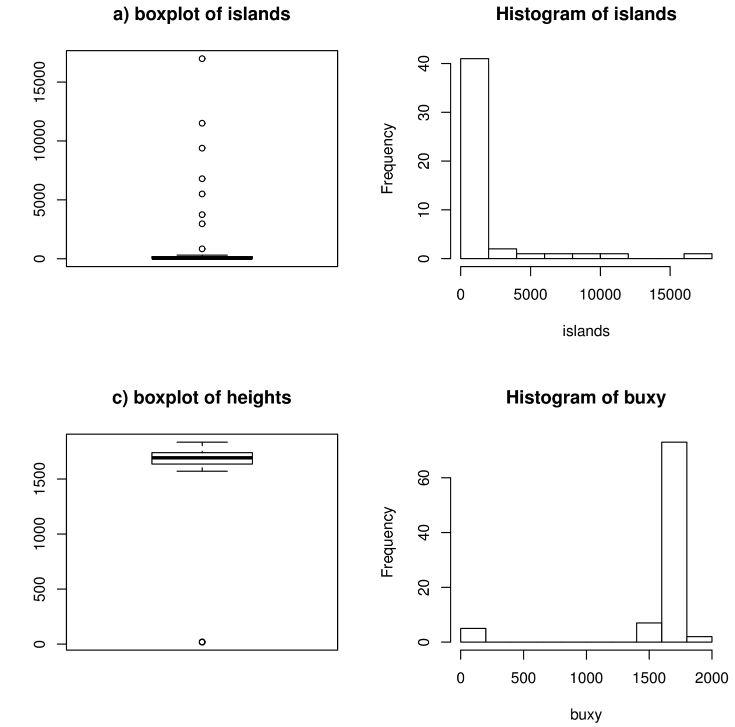

Fig. 2.3 Histograms and Box Plots for Example 2.5.

$j = 2$, and $4.5(10^3) = 4.5(1000) = 4500$ if $j = 3$. If the decimal was 1 digit to the left of the | (stem), then the value is 0.45. If the stem unit is ones and the leaf unit is tenths, then the value is $4(1) + 5(0.1) = 4.5$. If the stem unit is tens and the leaf unit is ones (very common), then the value is $4(10) + 5(1) = 45$.

Example 2.6. For the island data, the 4 and 5 correspond to $4500 = 4.5(1000)$. The 16|0 corresponds to $16000 = 16.0(1000)$. For log(islands), stem 9 with leaf 7 corresponds to 9.7 and $\log(16988) = 9.740$.

```
max(islands)  #largest value
[1] 16988
16000  #17 | 0 would have been better
> stem(islands)
```

```
  The decimal point is 3 digit(s) to the right of the |

   0 | 00000000000000000000000000000111111222338
   2 | 07
   4 | 5
   6 | 8
   8 | 4
  10 | 5
  12 |
  14 |
  16 | 0

> stem(log(islands))

  The decimal point is at the |

  2 | 566666778889
  3 | 0123444556778889
  4 | 134445
  5 | 22467
  6 | 7
  7 |
  8 | 0268
  9 | 147
```

```
par(mfrow=c(2,2)) #four graphs
boxplot(islands)
title("a) boxplot of islands")
hist(islands)
boxplot(buxy)
title("c) boxplot of heights")
hist(buxy)

stripchart(islands) #dot plot
title("a) dot plot for islands")
stripchart(log(islands))
title("b) dot plot for log(islands)")
stripchart(buxy)
title("c) dot plot for heights")
stripchart(buxy,method="jitter")
title("d) jittered dot plot for heights")

boxplot(log(islands))
?stripchart #same as help(stripchart)
hist(log(islands))
par(mfrow=c(1,1))
```

Definition 2.9. A *dot plot* consists of an axis and plotted points for each value of the data set.

Example 2.7. Figure 2.4 shows dot plots for the islands data and the Buxton (1920) heights data. Note that the islands data is right skewed and the heights data has outliers (gaps in the plot). Jitter adds some noise to the plotted points so it is easier to see the plotted points. The heights data has 5 outliers, and it is easier to see that there is more than one outlier in the jittered dot plot of Figure 2.4d) than in the dot plot of Figure 2.4c). See the above *R* commands.

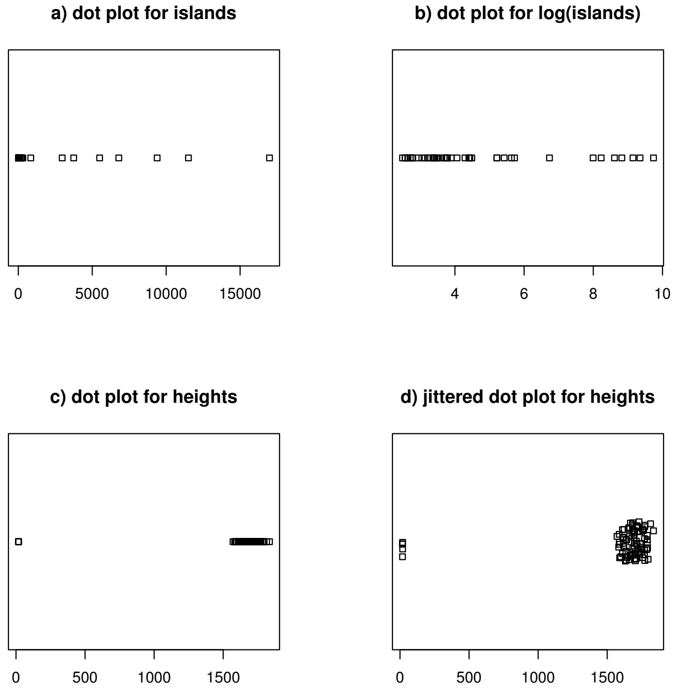

Fig. 2.4 Dot Plots for Example 2.7.

Example 2.8: Gender Intelligence. The following data is from the Nov. 16, 2011 *Chicago Tribune.* The data is the percentage of boys and girls meeting or exceeding the state standards at each grade level. Note that the math and science results are almost the same for grades 3 to 8, suggesting that gender intelligence is about the same. In 2001, more males than females took science and math in high school. According to a 2008 fall ABC news report, female 11th graders took as much science as males, and from about 2008, females have tended to score at least as high as males on standardized tests in science. Since each result has two categories: meeting or exceeding standards and failing to meet standards, the results could be displayed with a bar graph for each gender with two bars instead of four: just display the percent meeting or exceeding standards for each gender and for each grade discipline combination. (For grades 3 or 4 and 8 or 7, the first 3 categories were for the 1st grade, 3 or 8, while the last two categories science and social science were for grades 4 or 7.)

Table 2.1 Comparing Boys and Girls

grade	reading		writing		math		science		social science	
	B	G	B	G	B	G	B	G	B	G
3 or 4	60	65	53	63	74	74	66	65	62	60
5	57	60	63	77	61	62				
8 or 7	64	67	52	71	51	50	72	72	60	60
11	54	61	53	65	56	52	54	47	62	53

In some countries, gender differences are large (significant) for science and math. See Beaton et al. (1996). The differences were not likely due to gender differences in intelligence, but rather to the amount of classes taken by each gender, the percentage of female teachers in the sciences and mathematics, whether the teachers or students think that boys are better in science and math, gender opportunities for higher education, et cetera. For example, in the 1996 study, eighth grade girls disliked the sciences much more than boys, and 20% of the science teachers were female. In the US, 54% of the science teachers were female.

Example 2.9: Sexual Activity. If we use a Big Bang Theory term for sex, such as coitus, each time a man has sex with a woman, a woman has sex with a man, and vice versa. Hence the total number of times all women have sex in a year is equal to the total number of times all men have sex in a year. Men tend to claim to have more partners while women tend to claim to have fewer partners than the actual number. Table 2.2, taken from Student (1998), shows the estimated annual occasions of sex by age and gender in the USA. The numbers can be roughly explained by the numbers of men and women in each group. Also the male prison population is much higher than that of women. There are more young men than young women, so young women have the most sexual activity. Old men who are willing and able are greatly outnumbered by old women who are willing and able, and hence old men have much more sexual activity than old women.

Table 2.2 Estimated Annual Occasions of Sexual Activity

gender/age	18-24	25-34	35-40	40-54	55-64	65-74	75up
M	83	85	73	55	52	23	13
W	86	84	65	50	25	10	2

2.3 Summary

1) You should know how to make a bar graph for categorical data. Bar graph bars should have bases that have the **same length**.

2) **Do not make breaks in the vertical axis of bar graphs and histograms** because the area of the bars is proportional to the percentage of cases in each class.

3) Given a distribution table, make a histogram. Given a histogram and a rule for the endpoints (e.g. bar includes right endpoint but not the left endpoint), you should be able to make a distribution table.

4) Given a list of numbers, make a stemplot. **Include the stem units and leaf units on the plot.** For example the number 205 will have stem 20 and leaf 5 with stem units = tens and leaf units = ones. Note that $20(10)+1(5) = 205$.

5) Sometimes a list of numbers is presented as a stemplot, then you are asked to find the mean, median, and SD of the list. The stem and leaf units are used to determine what the list of numbers is.

6) Given a list of numbers or *R* output, find the five number summary: min, Q1, median, Q3, and max. Recall that the data is **sorted from smallest to largest.** The median is the "middle number", Q1 is the median of the sorted numbers to the left of the median, and Q3 is the median of the numbers to the right of the median. Use the five number summary to make a box plot.

7) Given a box plot, bar graph, stemplot, histogram, or dot plot, be able to give a short summary of what the plot tells you. For example are outliers present, is the histogram symmetric, right skewed or left skewed? How does the proportion of one category compare to the proportion of another category? *R* commands are `boxplot, barplot, stem, hist`, and `stripchart`.

8) From a story problem, you should be able to determine the individuals and the variables. You should know whether the variable is categorical or quantitative.

2.4 Complements

2.5 Problems

2.1. The stem-and-leaf display above is for 71 Stat 3011 final exam scores from around 1998. The lowest score was a 30 while the highest was a 92. The mean score was 69.2, the median score was 72, and the standard deviation of the scores was 15.8.

```
3 | 045678
4 | 266
5 | 01247899
6 | 01444555678889
7 | 011222334566888889        Stem: tens
8 | 1112222344456777888       Leaf: ones
9 | 122
```

If score of 59 or lower was a failing grade, what proportion of students failed this final?

a) 0.17 b) 0.24 c) 0.76 d) 0.024 e) 0.048

2.2. Twelve students took Math 580. From their quiz scores listed below, make a stemplot. Put the stem and leaf units to the right of the plot.

```
80  89  90  97  72  91  81  88  83  87  87  88
```

2.3. The weights, in ounces, of malignant tumors removed from 51 patients is displayed in the table below. Data is from Mendenhall and Beaver (1991, p. 19). Make a histogram or bar graph of the data and **state which plot you used**. Use labels like those shown below.

Class	Count
10 − < 20	5
20 − < 30	19
30 − < 40	10
40 − < 50	13
50 − < 60	4

```
|  |  |  |  |  |  |  |  |  |  |  |  |
0     10    20    30    40    50    60
```

2.4. Data for federal aid per capita for the 50 states in 1986 is summarized below. Data is from Mendenhall and Beaver (1991, p. 19). Make a histogram or bar graph of the data and **state which plot you used**Use labels like those shown below.

```
class           count
250  <=  <  450    26
450  <=  <  650    20
650  <=  <  850     1
850  <=  < 1050     1
1050 <=  < 1250     1
```

```
    __________________________________
    |  |  |  |  |  |  |  |  |  |  |  |
   250   450   650   850   1050  1250
```

2.5. Data is taken from the following newspaper article: Herndobler, K. (Aug. 27, 2002), "Illinois ACT Scores Bring Down the National Average," *Daily Egyptian*. The ACT scores during the 2001-2002 academic year were 22.8 for Carbondale, 20.8 for the USA, 20.1 for Illinois, and 16.5 for Chicago. (This was the first year that all Illinois high school juniors had to take the ACT, and the number taking the ACT jumped from 89000 in the previous year to 129000.) Display the ACT data with either a histogram or bar graph. Which plot did you use?

2.4. The lengths of reign of 13 rulers of England and Great Britain are listed below. Data is from Rossman and Chance (2011, p. 147).

```
21  13  35  19  35  10  17  56  35  20  50  22  13
```

Make a stem plot for the data. Do not forget to include the stem and leaf units.

2.5. The above data below is from Gould and Ryan p. 117. The numbers above are revenue (in millions) from the top ten Pixar animated movies as of June 2010. $Q_1 = 206, M = 245$, and $Q_3 = 261$. Draw a box plot for the movie data. Use labels like those shown below.

```
    192  163  246  256  340  261  244  206  224  293
```

```
                 ________________________________
                 |    |    |    |    |    |    |
                160  190  220  250  280  310  340
```

2.6. Suppose that the mean GRE math scores for the 50 states and the District of Columbia were entered into a computer. The computer gave the following descriptive statistics. From these statistics, draw a boxplot of the 51 GRE scores. Use labels like those shown below.

```
 N  MEAN   MEDIAN  STDEV   MIN    MAX    Q1     Q3
51  529.30 521.00  34.83   473.00 600.0  500.0  557.0
```

```
 _________________________________________________
  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
 450   475   500   525   550   575   600   625
```

2.7. The number of violent crimes (per 100000 people) in 2002 are summarized for the eight states below. Data is from Sullivan (2006, pp. 50, 87). From these statistics, draw a boxplot.

2.8. In the 1996 presidential election, about 50% of the voters voted Democrat and 40% voted Republican and about 10% of the voters voted "Other". Display this information with an appropriate plot and say what type of plot you used. Use labels like those shown below.

```
N  MEAN    MEDIAN  STDEV   MIN  MAX   Q1   Q3
8  479.5    497    112.60  311  599   388  578
```

```
  _____________________________________________
    |    |    |    |    |    |    |    |    |
   300  340  380  420  460  500  540  580  620
```

2.9. Fourteen students took Math 484. a) From their scores listed below, make a stemplot. Put the stem and leaf units to the right of the plot.

```
78 100 85 100 87 90 91 100 98 98 95 100 91 99
```

2.10. The proportion of US adults over 25 whose highest educational attainment (no high school degree, a high school degree, some college but no degree, BA or BS degree, or advanced degree) are given below. Data is from Sullivan (2014, p. 59). Make a histogram or bar graph of the data and **state which plot you used**. Use labels like those shown below.

noHS	0.1544
HS	0.3202
some college	0.1715
BA or BS	0.2612
adv. degree	0.0927

```
 _______________________________________________
 noHS      HS     somecollege   BAorBS   AdvDeg
```

2.11. A major newspaper reported the Budweiser's share of the beer market. Make a bar graph or histogram from the table below.

brand of beer	percent of market share
Budweiser	25%
Bud Light	22%
Other	53%

2.12. The percentage of intercity passengers travelling by bus, air, subway and Amtrack is listed below. Make a bar graph or histogram from the table below.

type	percent
Air	40.6 %

Amtrak	1.6%
Bus	29.2%
Subway	28.6%

2.13. Suppose that the grade distribution for Math 282 in 2010 was A - 12, B - 27, C - 8, D - 2, F - 1. (So 12 students get A's, 27 get B's, etc.) Make a histogram or bar chart of this data (one of these is inappropriate).

2.14. The November 16, 2001 *Chicago Tribune* reported the percentage of boys and girls meeting or exceeding state standards from achievement tests at several grade levels in Math, Science and Social Science. Make a histogram or bar graph of the data and **state which plot you used**. Use labels like those shown below. Does the plot suggest that males are
a) smarter, b) less smart **or** c) about as smart as females? (**Circle one.**)

group	label	% passing
Math 8th grade Female	(MF)	50
Math 8th grade Male	(MM)	51
Science 7th grade Female	(SF)	72
Science 7th grade Male	(SM)	72
Soc. Sci. 7th grade Female	(SSF)	60
Soc. Sci. 7th grade Male	(SSM)	60

MF MM SF SM SSF SSM