**KEY SKILLS:**

*Organize a data set into a frequency distribution.*
*Construct a histogram to summarize a data set.*
*Compute the percentile for a particular data value.*
*Compute quartiles and interquartile range and construct boxplots.*
*Compute the center and variability of a quantitative data set: mean/median, standard deviation/interquartile range.*
*Construct bar charts for a categorical data set.*

## 1. Constructing Frequency Distributions

When we encounter a **quantitative** data set, we can organize the data by constructing a **frequency distribution.** A frequency distribution is a simple chart that summarizes the data by **binning** into larger categories. For example, a data set of test scores from 0 to 100 might be binned into categories 0-5, 5-10, 10-15, 15-20, and so on, and we state the **count** or **frequency** of data in each bin. By our book's convention, a test score that lands on the boundary between two categories goes into the lower category (this convention is not universal).

### Ex. 1.1

In order to test the highway mileage claims made by a car manufacturer, a consumer advocacy group tests the gas mileage of 35 randomly selected cars by driving them 200 miles each. The raw gas mileage data is shown below in miles per gallon (mpg). Summarize the data in a frequency distribution using 10-20 bins.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 28.7 | 38.9 | 30.9 | 35.2 | 34.2 | 36.7 | 35.2 | 32.3 |
| 32.2 | 35.9 | 33.8 | 30.8 | 34.4 | 32.5 | 32.1 | 33.1 |
| 34.0 | 33.5 | 39.5 | 33.2 | 34.8 | 29.7 | 32.8 | 37.0 |
| 31.7 | 31.9 | 28.0 | 32.8 | 25.5 | 34.4 | 32.4 | 31.8 |
| 32.3 | 36.3 | 23.2 | | | | | |

The first step is to use technology to sort the data set. After sorting, we obtain the following list:

23.2, 25.5, 28.0, 28.7, 29.7, 30.8, 30.9, 31.7, 31.8, 31.9, 32.1, 32.2, 32.3, 32.3, 32.4, 32.5, 32.8, 32.8, 33.1, 33.2, 33.5, 33.8, 34.0, 34.2, 34.4, 34.4, 34.8, 35.2, 35.2, 35.9, 36.3, 36.7, 37.0, 38.9, 39.5

The minimum value is about 23, and the maximum value is about 40, so a bin size of 1.0 seems appropriate. We choose bins of 23-24, 24-25, . . . , 39-40, and we use the convention that boundary values go into the lower bin. We draw the frequency distribution and count how many data points lie in each bin:

| BIN | COUNT |
|---|---|
| 23-24 | 1 |
| 24-25 | 0 |
| 25-26 | 1 |
| 26-27 | 0 |
| 27-28 | 1 |
| 28-29 | 1 |
| 29-30 | 1 |
| 30-31 | 2 |
| 31-32 | 3 |
| 32-33 | 8 |
| 33-34 | 5 |
| 34-35 | 4 |
| 35-36 | 3 |
| 36-37 | 3 |
| 37-38 | 0 |
| 38-39 | 1 |
| 39-40 | 1 |

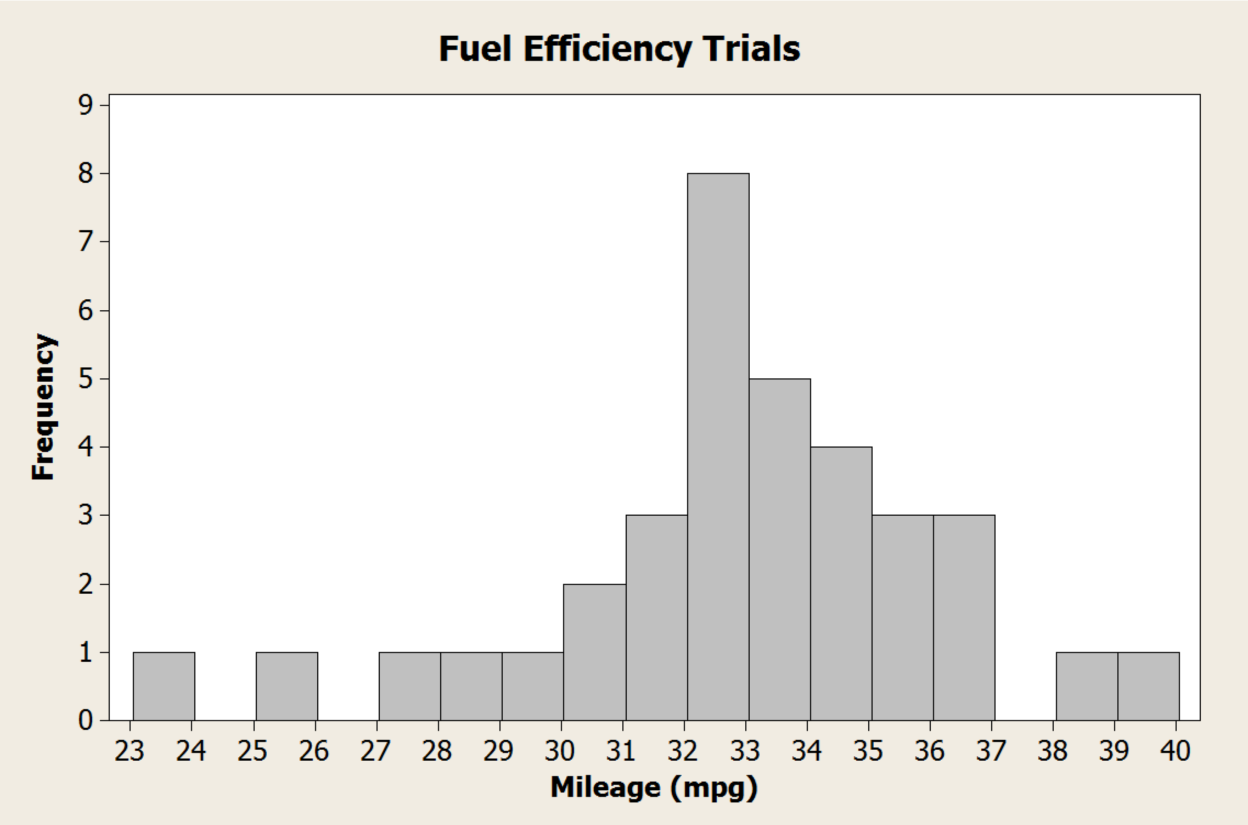
## **2.** Constructing Histograms

A **histogram** is a graphical representation of a frequency distribution. We label the **x-axis** with the cut-lines between bins, and we label the **y-axis** with the count in each bin. Histograms are a powerful tool for understanding the distribution of data: we can quickly estimate the center and spread of a data set with a quick glance at a well-constructed histogram.

### Ex. 2.1

Construct a histogram for the data in Ex. 1.1.

Here we use Minitab to create the histogram. Note that Minitab's binning does not follow our convention for data points on the boundary between two bins, so some modification of the graph is required.



**3.** Computing Percentiles

In a quantitative data set, the **percentile** of a particular data value tells us the percent of the data (strictly) less than that value. For example, a data point at the 90th percentile is greater than 90% of the data set. We generally use technology to rank the data set in ascending or descending order before computing percentiles.

**Ex. 3.1**

Using the data set in Ex. 1.1, compute percentiles for (a) 29.7 mpg and (b) 34.4 mpg. Round percentiles to the tenths place.

a. We refer to the sorted data set in Ex. 1.1 and count 4 data values less than 29.7. We compute the percentage of data less than 29.7 mpg: $\frac{4}{35} \approx .114 \text{ or } 11.4\%$. We say the car that got 29.7 mpg is at the 11.4th percentile.

b. There are two cars that got exactly 34.4 mpg, but we only count the number of cars that got *strictly* less than 34.4 mpg. There are 24 data points strictly less than 34.4, so the percentile is $\frac{24}{35} \approx .686 \text{ or } 68.6\%$. We say the cars that got 34.4 mpg are at the 68.6th percentile.

**4.** Median, Quartiles and the Interquartile Range.

A quantitative data set can be divided into four parts using **quartiles**. Roughly speaking, the first quartile (**Q1**) is the 25th percentile, the second quartile (**median**) is the 50th percentile and the third quartile (**Q3**) is the 75th percentile. For small data sets, the quartiles won't lie exactly at these percentiles, but they will be close.

To compute the **median** of a data set, we first sort the data using a calculator or computer. If the data set has an odd number of values, the median is the value in the middle. If the data set has an even number of values, the median is the average of the two values in the middle of the data set.

To compute the **first quartile**, we find the median of the lower half of the data. If the data set has an odd number of values, we *exclude the median value* from the lower half. Similarly, the **third quartile** is computed by finding the median of the upper half of the data set, excluding the middle value if the data set has an odd number of entries.

To measure the spread of a data set, we compute the **interquartile range** (IQR) by taking the difference Q3-Q1. Roughly speaking, the interquartile range tells us the range of values in the middle 50% of the data set.

**Ex. 4.1:**

Compute Q1, Median, Q3 and IQR for the data set in Ex. 1.1. What percentage of the data lies strictly between Q1 and Q3?

To compute the median, we sort the data set and pick the center value. Our data set has size 35, so the middle value has 17 values above and below. Carefully counting through the list, we see that the median value is **Med = 32.8 mpg**.

To compute Q1, we look at the lower half of the data set, excluding the median value of 32.8 mpg. The lower half of the data set is: 23.2, 25.5, 28.0, 28.7, 29.7, 30.8, 30.9, 31.7, 31.8, 31.9, 32.1, 32.2, 32.3, 32.3, 32.4, 32.5, 32.8. This list has an odd number of points, so the median is just the center value: 31.8. The first quartile is given by **Q1 = 31.8 mpg**.

To compute Q3, we look at the upper half of the data set, excluding the median value of 32.8 mpg. The upper half of the data set is: 33.1, 33.2, 33.5, 33.8, 34.0, 34.2, 34.4, 34.4, 34.8, 35.2, 35.2, 35.9, 36.3, 36.7, 37.0, 38.9, 39.5. The median value of this half is 34.8, so **Q3 = 34.8 mpg**.

The interquartile range is simply the difference Q3-Q1 = 34.8 – 31.8 = 3, so **IQR = 3 mpg**. Roughly speaking, the central 50% of the cars varied over only 3 mpg.

When we count the number of data points *strictly* between Q1 and Q3, we find that 17 data points lie in this region (about 49%).

**5.** Boxplots

We can summarize a quantitative data set in a chart called a boxplot. The book provides an adequate description of this process (pg. 35), but it doesn't provide a clear example. Note that I like to use the terms **upper fence** and **lower fence** where the book uses **max whisker reach** and **min whisker reach**.

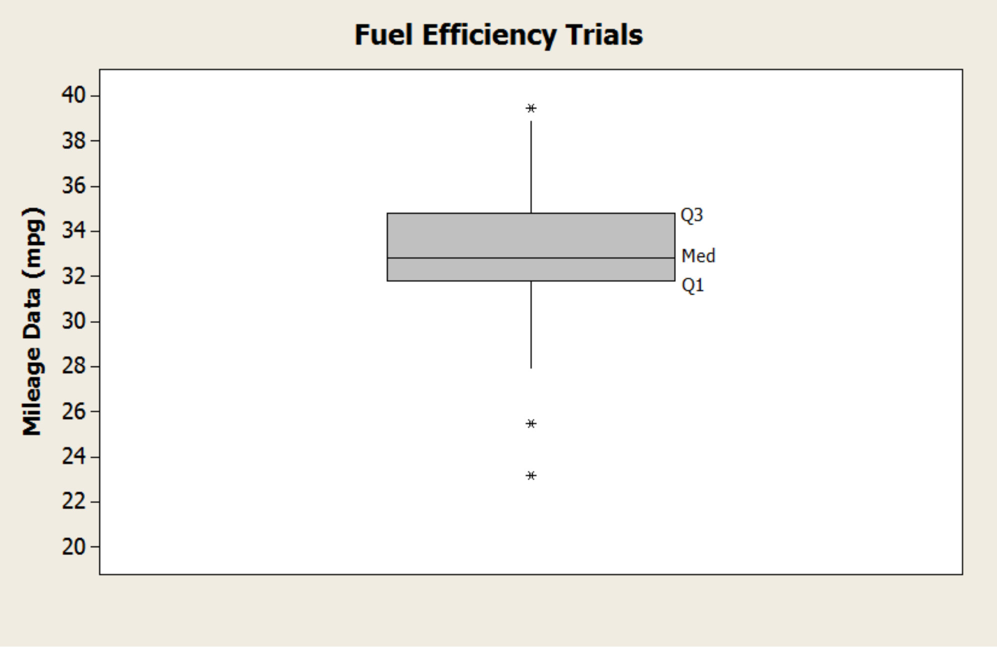
**Ex. 5.1:**

Construct a boxplot for the data set in Ex. 1.1.

We start with a vertical axis labeled from 20-40 to cover the full range of the original data set. We label Q1, Med and Q3 with horizontal line segments, then surround them with a box. We compute the upper and lower fences:

**Upper fence**: Q3+1.5(IQR) = 34.8+1.5(3) = **39.3 mpg**
**Lower fence**: Q1-1.5(IQR) = 31.8-1.5(3) = **27.3 mpg**

We classify any data point lying outside the fences as an **outlier** and mark it with an asterisk: 23.2, 25.5 and 39.5 are all outliers. Finally, the whiskers extend only to the largest and smallest data points that still lie *within* the fences.



**6.** Measures of center and variability.

The **center** and **variability** are two of the key features of a quantitative data set. We measure the center using the **mean** or the **median**, and we measure the variability using the **standard deviation** or the **interquartile range (IQR)**. The mean and standard deviation are given by precise mathematical formulas (see text), and we adopt the informal rule that roughly 70% of data lie within one standard deviation of the mean, provided that the data set is relatively well-behaved. For most statistical tests, we use the mean and standard deviation because they are more mathematically precise than median and IQR. However, some industries (e.g. real estate) use median and IQR because they are **robust** to outliers; that is, these statistics don't change very much even if we introduce extreme outliers to the data set.

### Ex. 6.1:

Use technology to compute the mean and standard deviation for the data set in Ex. 1.1. What percentage of the data lies within one standard deviation of the mean?

Using technology, we obtain a mean of $\bar{x} = 32.91 \text{ mpg}$ and a standard deviation of $s = 3.30 \text{ mpg}$.

We compute a lower bound on the 1-standard deviation interval about $\bar{x}$: $\bar{x} - s = 32.91 - 3.30 = 29.61 \text{ mpg}$.
Next, we compute the upper bound of the 1s interval: $\bar{x} + s = 32.91 + 3.30 = 36.21 \text{ mpg}$.

We count through the sorted data set and find that that 26 out of the 35 data points lie on the 1s interval. 26/35 = 74%.

### Ex. 6.2:

Suppose that one additional data point is added to the set from Ex. 1.1: a test drive in which a spark plug fails and the car only gets 4.1 mpg. Use technology to compute the new mean, standard deviation, median and IQR. Which statistics are the most robust to the outlier?

We add 4.1 to the data set, compute the new statistics and compare to the original values:

| **Center (mpg)** | | | | **Variability (mpg)** | | | |
|---|---|---|---|---|---|---|---|
| Original Mean: | 32.90 | New Mean: | 32.11 mpg | Original St. Dev.: | 3.30 | New St. Dev.: | 5.80 |
| Original Median: | 32.80 | New Median: | 32.80 mpg | Original IQR: | 3.00 | New IQR: | 2.85 |

We observe that the relative changes in the mean and standard deviation are very large compared to the relative changes in the median and IQR. In other words, the median and IQR are relatively *robust* to the outlier.

**7.** Constructing Bar Charts

We use **bar charts** to summarize categorical data graphically. Some bar charts use the height of a bar to show the count of individuals in each category, but we can also construct bar charts to show the *percentage* of individuals in each category. We can even construct side-by-side charts to compare multiple groups across the categories. Bar charts should always start with zero as the lowest value on the scale, otherwise the graph may exaggerate differences between categories.

**Ex. 7.1:**

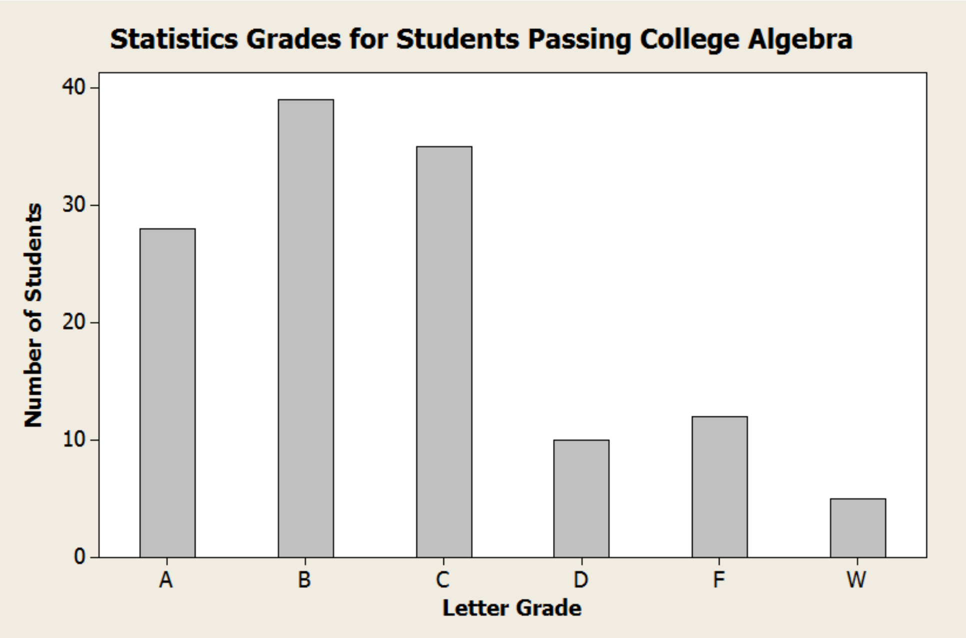
The table below shows the results of a survey taken in the math department at a small college. The results table shows the number of students receiving each letter grade in Introductory Statistics, and the results are divided up into two categories: students who used high school algebra as a prerequisite and students who used college algebra as a prerequisite.

| | A | B | C | D | F | W |
|---|---|---|---|---|---|---|
| HS Algebra | 25 | 40 | 32 | 15 | 17 | 22 |
| Coll Algebra | 28 | 39 | 35 | 10 | 12 | 5 |

a. Construct a bar chart for the letter grades of students who used college algebra as a prerequisite course for introductory statistics.
b. Construct a side-by-side bar chart for the *percent* of students in each grade category.

___________

a. We construct a carefully scaled bar chart, making sure to place the *x*-axis at zero (starting higher than zero can exaggerate differences between categories).

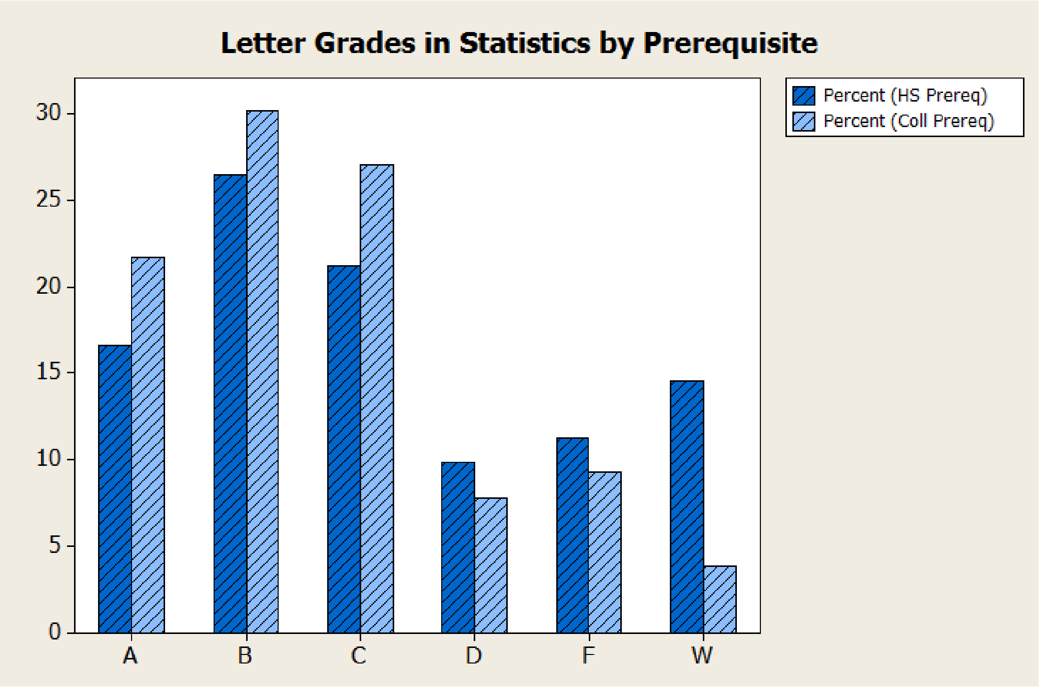


b. First, we need to find the total number of students in each prerequisite category. Adding up the rows, we obtain 151 “HS Algebra” students and 129 “Coll. Algebra” students.

Next, we compute the percent of students in each prerequisite category who received each grade. For example, the percent of “HS Algebra” students who received an A is 25/151 = 16.6%. A table of percentages is shown below:

| | A | B | C | D | F | W |
|---|---|---|---|---|---|---|
| HS Algebra | 16.6% | 26.5% | 21.2% | 9.9% | 11.3% | 14.6% |
| Coll Algebra | 21.7% | 30.2% | 27.1% | 7.8% | 9.3% | 3.9% |

Finally, we display the data in a side-by-side bar chart:



## UNIT I EXERCISES:

**Exercises 1-5:** The following (fictional) data set shows the adult smoking rate (%) in 58 different counties in California.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 17.7 | 4.7 | 18.8 | 17.3 | 17.2 | 12.6 | 13.5 | 15.2 | 18.2 | 15.1 | 13.3 | 17.0 | 13.6 | 17.4 |
| 5.8 | 14.6 | 17.3 | 14.1 | 13.5 | 16.3 | 19.5 | 13.9 | 13.8 | 18.9 | 8.3 | 12.5 | 29.9 | 11.9 |
| 18.5 | 15.5 | 13.0 | 15.2 | 14.0 | 7.8 | 16.6 | 10.1 | 12.8 | 13.9 | 17.9 | 15.7 | 9.3 | 18.8 |
| 16.2 | 13.5 | 17.1 | 13.4 | 24.0 | 23.7 | 19.7 | 13.6 | 12.0 | 14.5 | 14.9 | 16.9 | 17.2 | 12.6 |
| 19.1 | 13.0 | | | | | | | | | | | | |

**1.** Construct a frequency distribution and histogram for the data. Use a bin size of 2 starting from 4.0%.

**2.** Compute the percentile for
(a). a county with a 18.9% smoking rate.
(b). a county with a 12.6% smoking rate.

**3.** Compute Q1, Med, Q3, IQR and the fences for the data, then construct a modified boxplot. Make a list of the outliers.

**4.** Compute the mean and standard deviation for the data set using technology. What percent of the data lies within one standard deviation of the mean?

**5.** Suppose that one county was misreported: 23.7% was supposed to be 95%. Correct the data set and use technology to compute the mean/standard deviation and median/IQR again. Which of these statistics change the most?

-----

**6.** The (fictional) data set below shows the number of Americans who died in 2016 across the categories "Smoker" and "Non-Smoker".

**DEATHS IN THE U.S., 2016: SMOKING STATUS AND CAUSE OF DEATH**

| | **Heart Disease** | **Stroke** | **Cancer** | **Chronic Lung Dis.** |
|---|---|---|---|---|
| **Non-Smoker** | 400, 755 | 377,899 | 241,723 | 17,535 |
| **Smoker** | 185,100 | 124,320 | 150.208 | 55,254 |

a. Construct a bar chart of Percent of Deaths for Non-Smokers (calculate the percent of deceased non-smokers who died from heart disease, the percent who died from cancer and so on).

b. Use the table to construct a side-by-side bar chart of Number of Deaths for Non-Smokers and Smokers. How might this graph be misleading? What missing information would allow you to construct a more meaningful graph?

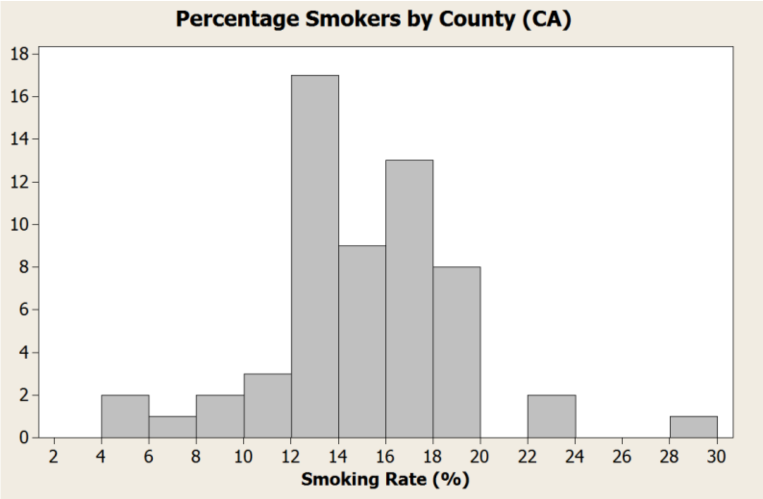
**Solutions to Exercises, Unit I Supplement**

**1-5:** After sorting the data on a calculator or computer, we obtain:

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4.7 | 5.8 | 7.8 | 8.3 | 9.3 | 10.1 | 11.9 | 12.0 | 12.5 | 12.6 | 12.6 | 12.8 | 13.0 | 13.0 | 13.3 |
| 13.4 | 13.5 | 13.5 | 13.5 | 13.6 | 13.6 | 13.8 | 13.9 | 13.9 | 14.0 | 14.1 | 14.5 | 14.6 | 14.9 | 15.1 |
| 15.2 | 15.2 | 15.5 | 15.7 | 16.2 | 16.3 | 16.6 | 16.9 | 17.0 | 17.1 | 17.2 | 17.2 | 17.3 | 17.3 | 17.4 |
| 17.7 | 17.9 | 18.2 | 18.5 | 18.8 | 18.8 | 18.9 | 19.1 | 19.5 | 19.7 | 23.7 | 24.0 | 29.9 | | |

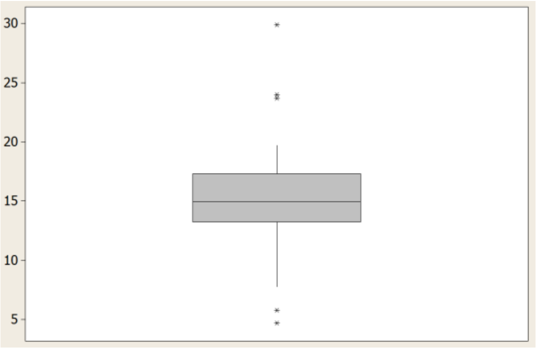
**1.** We bin the data starting from 4.0% with a bin size of 2.0%, remembering to place borderline cases in the lower bin (this is our textbook's convention).

| Bin | Count |
|---|---|
| 4-6 | 2 |
| 6-8 | 1 |
| 8-10 | 2 |
| 10-12 | 3 |
| 12-14 | 17 |
| 14-16 | 9 |
| 16-18 | 13 |
| 18-20 | 8 |
| 20-22 | 0 |
| 22-24 | 2 |
| 24-26 | 0 |
| 26-28 | 0 |
| 28-30 | 1 |



**2.** (a). There are 51 data points explicitly less than 18.9%, so the percentile is 51/58 ~.879 or 87.9th percentile.
(b). There are 9 data points explicitly less than 12.6%, so the percentile is 9/58 ~ .155 or 15.5th percentile.

**3.** Med: we average the middle two values (14.9+15.1)/2 = 15.0
Q1: we find the median of the lower half of the data. Since the data set is even, we just slice it in half to get the lower half (29 data points). The middle of this set is 13.3.
Q3: The median of the upper half of the data is 17.3.
IQR: Q3-Q1 = 4.0.
Upper fence: Q3+1.5*IQR = 23.3
Lower fence: Q1-1.5*IQR = 7.3
Outliers: anything outside the fences, so 4.7, 5.8, 23.7, 24.0, 29.9.
Boxplot:

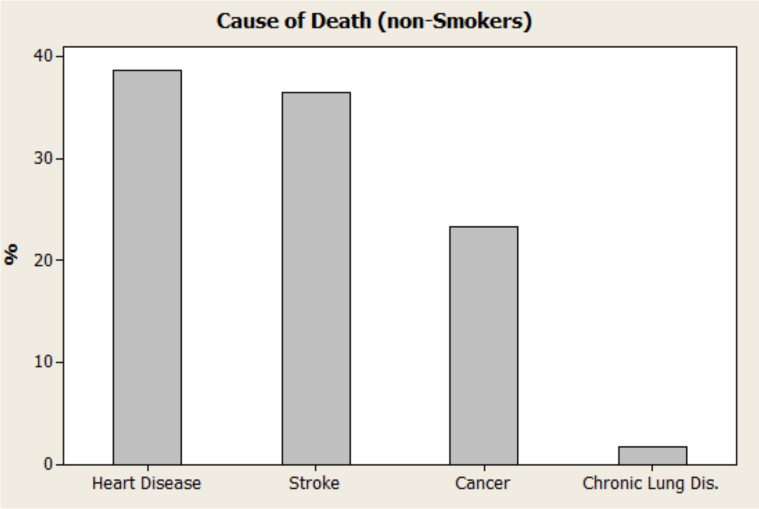


**4.** (use 1-Var Stats on a TI): Mean = 15.21 St. Dev. = 4.14
We look one st. dev. above and below the mean: 15.21-4.14=11.07 and 15.21+4.14=19.35
Count the number of data points between 11.07 and 19.35: I count 47 data points. 47/58 is 81%. This doesn't agree very well with the "70% rule", but that's not too surprising given that the data isn't very "normal"/bell-shaped.

**5.** Med: 15.0 Mean: 16.44 (median stayed the same, mean went up by ~1.2 points)
IQR: 4.0 St. Dev: 11.2 (IQR stayed the same, st.dev. went up by ~7 points)
Note: Med and IQR won't always stay *exactly* the same, but they are more robust to outliers than mean/st. dev.

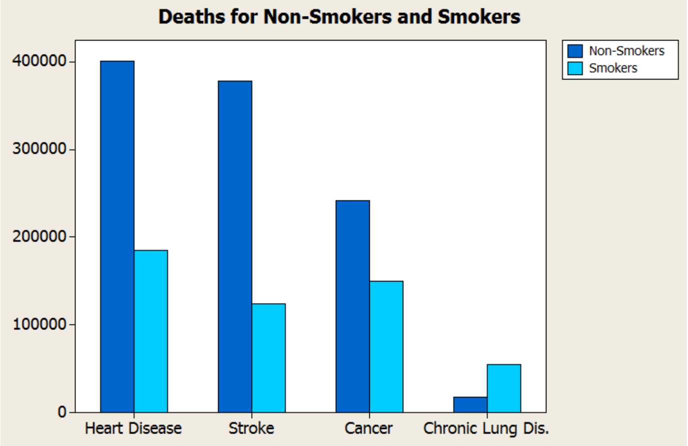
**6. (a).** we get a total for non-smoker deaths (1,037,912) then turn the raw numbers into percentages. Make sure the bar chart of percentages starts at zero.

| | |
|---|---|
| Heart Disease: | 38.6% |
| Stroke: | 36.4% |
| Cancer: | 23.3% |
| Chr. Lung Dis: | 1.7% |



Note: surely the data set should include "other" as a category, because these can't be the only causes of death. The percentages shown in the chart are the correct percentage of the shown data, but not the correct percentage of all the deaths from all causes.

**6. (b).** we construct the bar chart for the raw number of deaths in each group, side-by-side:



Because smokers are a minority of the population, our graph correctly shows fewer smokers dying in the first three categories, but the graph fails to communicate the higher risk of all these causes of death due to smoking. Maybe it should be re-done as the *percentage of all non-smokers* who die each year from each cause and the *percentage of all smokers* who die each year from each cause. In other words, we take into account the population size of smokers/non-smokers to produce a more meaningful graph.