1 Introduction

The supplementary notes I hand out in this course are meant to fill in holes in the textbook, and are neither complete nor especially well organized. I will emphasize topics that I feel are important, and stress connections between different topics. You should use the textbook, these notes, notes from class, and notes on my homepage as sources for written information pertinent for this course. On assignments and tests, you are expected to synthesize the information from all these sources to come to an understanding of the underlying concepts.

Each section will contain a brief introduction, a list of topics discussed, several examples, example questions, and a check list of specific skills you should obtain. The road to mastering these skills is taken by working on homework problems.

2 Supplementary notes on exploratory data analysis

Exploratory data analysis is the process of examining and summarizing data to reveal features, patterns, and relationships in the variables of a data set. Data may be explored graphically and numerically, and we will use both methods heavily. When presented with a new data set, exploring it should be the first task. As a running example in this section, we will use the nutritional information from 73 brands of cold cereal. The source of this data set is DASL (data and story library), which is part of Statlib, a repository of data sets (and other goodies), maintained at the Department of Statistics at Carnegie Mellon University. (http://www.stat.cmu.edu for those interested in checking it out yourself. There's also a link from Bret Larget's homepage.)

2.1 Topics

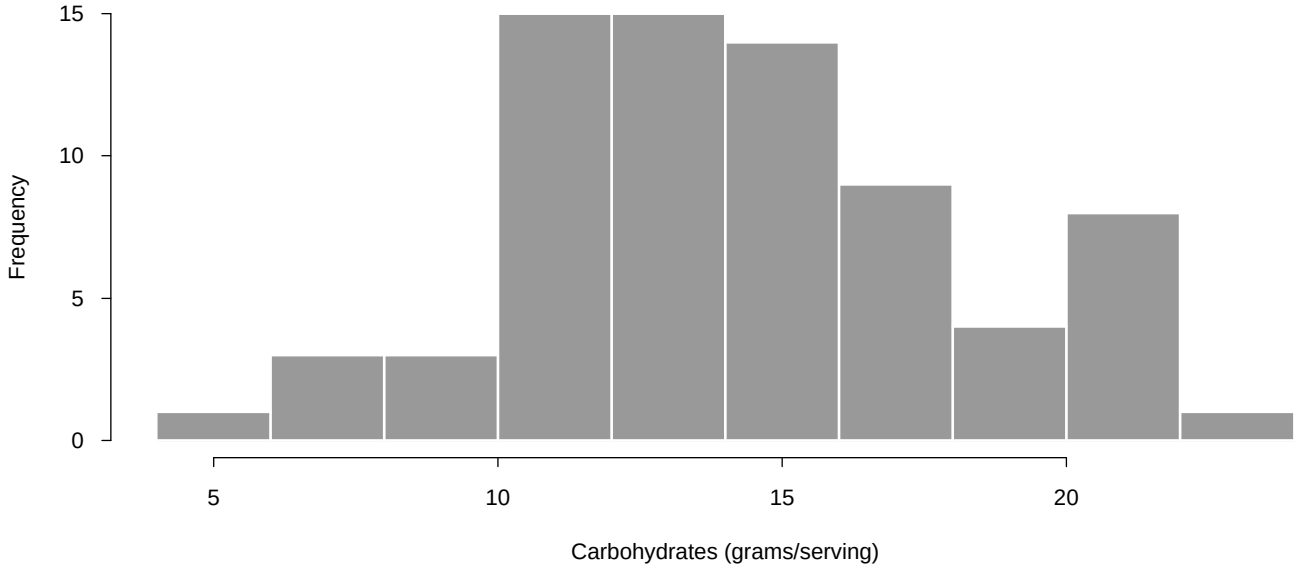
Each topics section will list the new topics and concepts presented. Remember, these notes are not complete. Topics which are covered satisfactorily in the textbook may be mentioned here, but not discussed in depth. Some topics in the textbook will not be mentioned here.

histogram	**stemplot**	**boxplot**	**scatter plot**
mean	**median**	**quartile**	**quantile**
standard deviation	**variance**	**range**	**interquartile range**

2.2 Histograms

A *histogram* is the most useful graph for observing a single quantitative variable. It is simply a graph of the count (or proportion) of observations that fall into each of several intervals or "bins", where the **area** of a shaded bar sitting over the entire bin represents the proportion of total observations in the bin. (Note that if the bins all have the same width, the **height** of the shaded bar may be used instead of the area.) If a bin is empty, there is no shaded box but the empty space remains to indicate the gap in the variable. Shaded bars in adjacent boxes should touch, and the x-axis should use the same scale throughout.

Example Here is a histogram of the variable carbo from the cereal example.



Notice that there are ten bins, and each bin width has 2. Roughly half the brands of cereal have less than 14 grams of cabohydrates per serving, which is evident because a vertical line at 14 would split the shaded area about in half. The smallest observation is no less than 4 and the highest observation is no more than 24. The shape of the distribution is roughly symmetric and bell-shaped with a mode at about 12, except for a second smaller mode near 21. A distribution is bell-shaped if most observtions are in the center and it tails off to both sides. It is roughly symmetric if flipping the shaded bars over from left to right would not greatly alter the picture. The distribution is not greatly skewed because the lower half of the variable (below 14) and the upper half of the variable (above 14) are spread out similar amounts.

The three characteristics of the variable summarized in the preceding paragraph are **center**, **spread**, and **shape**. We will discuss more formal numerical measures of these characteristics a little later on.

Further information may also be read from the histogram. For example, roughly 10% of the cereals have 10 g of carbohydrate per serving or less, because roughly 10% of the total shaded area is less than 10. There are 15 observations (out of 73) in the bin from 10 to 12. Observations which fall directly on boundaries are handled in different ways by different software packages. S-PLUS (which I used to produce this graph and which you will learn to use) takes the convention that observations on boundaries are grouped in the bin to the left. Hence the bin from 10 to 12 contains a count of all the brands of cereal where the carbohydrate measure was **greater** than 10 and **less than or equal to** 12. Details like this are usually not important for a general visual summary of the variable, but can be important if a particular question is of real interest.

A histogram is a summary of the variable, in that we cannot reproduce the variable precisely from the histogram. However, we have a better understanding of the variable than a quick perusal of the raw numerical values on a column of paper can provide. Furthermore, histograms can summarize arbitrarily large sets of data.

> **Question:** The bin with boundaries 6 and 8 has how many observations? Look at the complete data and determine which brands are counted in that bin.

How to make a histogram. We will always use computers to do the tedious work of binning data and drawing histograms. However, you ought to know how to do this by hand in principle. The general idea is to find an interval that spans the data, divide it into a reasonable number of bins of equal size (unless there is a compelling reason to choose unequal bin sizes), count the number of observations in each bin, and draw a shaded bar graph so neighboring bars touch.

Generally speaking, there should be around 5–20 bins. You will want to be on the high side for large data sets or if there are many "features" to show. A histogram should show all major modes (peaks) and any skewness in the distribution of the variable. Use an evenly scaled x-axis and leave bins empty if necessary.

Interpreting information in histograms. You should be able to do things such as:

- Identify if a variable has outliers.
- Get a rough idea of the center and the spread.

- Describe the shape (skewness and symmetry).
- Find the proportion of observations in intervals.
- Find intervals that contain a given proportion.

An *outlier* is an observation that sticks out of the overall pattern of the graph. There are no outliers in this example.

A histogram is approximately *symmetric* if the left and right halves are roughly mirror-images of one another. In a symmetric distribution, the spread in the two halves of the distribution is about the same

A histogram is *skewed to the right* if the right half is stretched out farther than the left half.

A histogram is *skewed to the left* if the left half is stretched out farther than the right half.

The histogram from this example is approximately symmetric and is not skewed.

2.3 Stemplots

A stemplot is a method to write down the data in a way that describes its shape. Stemplots are also called *stem and leaf diagrams*. A large vertical line divides the *stems* from the *leaves*.

The last significant digit of each observation is called the leaf. There is exactly one leaf per observation written on the right side of the display. For each row (stem) the leaves are ordered from smallest to largest. The leaves should line up exactly, and be flush with the vertical line, to give a visual description of how many observations are in each row.

The remaining portion of each observation is called the stem. Each stem is written only once, on the left side of the vertical line. Even if there are no observations for a stem, the stem should be written to show that there is a gap in the data.

Here is a stemplot of the same carbohydrate data.

stem	leaf
5	0
6	
7	0
8	00
9	0
10	0055
11	000005
12	0000000
13	000000005
14	000000
15	00000000
16	000000
17	000000
18	000
19	0
20	000
21	000000
22	00
23	0

`The break occurs at the decimal place.`

There are 73 observations. The minimum value is 5.0 and the maximum is 23.0. The stemplot also shows that there is a roughly symmetric bell-shaped distribution with most of the data in the middle, tapering off fairly evenly to both sides, except for the second mode at 21.0. Stemplots may be made fairly quickly by hand for moderately sized data sets. Creating a stemplot is a convenient way to sort a variable which is helpful for hand calculations of the median and quartiles described later in these notes.

When the data has too many significant digits, the summary is better if the data is **rounded** first. (Some software will **truncate** instead of round.) For example, the sodium variable has a minimum value of 0 and a maximal value of 320. If we used the ones place as the leaf, we would have 33 stems from 0 to 32, which would not be a very effective summary. Instead, we can round (or truncate) each value to the nearest 10, and then use the tens place as the leaf. If we do this, the smallest

value is 0 and the highest is 32, so we have stems of 0, 1, 2, and 3. This is a bit of an over summary. By **splitting** the stems, we can show more of the shape of the distribution. Use two rows for each stem with the first row for leaves 0–4 and the second row for leaves 5–9.

Here is a stemplot of the sodium data which has been truncated to the nearest ten and where the stems are split.

stem	leaf
0	0000000114
0	7799
1	223334444444
1	55567777788888999
2	00000001111222223344
2	556688999
3	2

`The break occurs two places to the left of the decimal place.`

Notes on Bret Larget's Web page give more examples. There is not always a single "best" decision as to whether a variable ought to be rounded or whether stems ought to be split. A good rule of thumb is that the total number of rows should be about the same as what you would choose for the number of bins in a histogram.

2.4 The mean and the median

The mean and the median are two different measures of center. The mean measures center by finding the balancing point. On a histogram, you can picture the mean as the point where the shaded bars would balance if they were made from some solid material. The median, on the other hand, is the middle value. half the observations ar at least as big and half are at least as small. For symmetric distributions, these two measures will be in the same place, but for skewed distributions, the mean is pulled in the direction of the skew.

The mean has the formula

$$\bar{x} = \frac{\sum x_i}{n}$$

where the notation $\sum$ means to add something up, in this case all of the individual observations, and n is the number of observations. The mean is simply the arithmetic average you are familiar with. The notation $\bar{x}$ is the conventional notation for the mean of a sample of n numbers. (In later chapters we will also use the notation μ to represent a mean.) The mean is the balancing point. The mean of the carbo data is 14.71. (It is a good rule of thumb to use one more place of accuracy than the original data when calculating a mean.) This seems to be a reasonable location for the balancing point of the associated histogram, by rough inspection.

The median is the middle number, after sorting from smallest to largest. If there are an odd number of observations, the median is uniquely defined. If there are an even number of observations, the median is taken to be the average of the two middle numbers.

Examples. Consider the sorted numbers 1, 2, 4, 5, 7. Here the middle number is 4. If the numbers had instead been 1, 2, 4, 5, 7, 10, both 4 and 5 would have been in the middle, and the median would be $(4+5)/2 = 4.5$.

For the carbohydrate data, there are 73 observations. If we could divide them evenly into two groups of equal size, we would have $73/2 = 36.5$ observations in each group. The 36 lowest values are in the low half and the 36 highest are in the high half. The 37th sorted number (from either the bottom or the top) is the unique median. Counting using the stemplot above, we see that the very last 14.0 is the 37th number, and so 14.0 is the median of the distribution.

On the histogram, a big vertical line at 14 would divide the shaded area about in half.

Technical definition of the median. A median of a distribution is any value m where at least half the observations are at least as large as m and at least half the observations are at least as small as m.

Comparisons between the mean and the median. The mean is greatly affected by outliers. If there are outliers present, the mean might not be a good representation of a "typical" value.

Example. The numbers 4, 5, and 6 have a mean of 5, a typical value in this sample, while the numbers 4, 5, and 600 have a mean of 203, which is not very typical of its sample.

The median is *robust* to outliers, and its value can most often be reasonably interpreted as typical. The median in both examples above is 5.

The mean and the median each are different measures of the center of a distribution. If the distribution is symmetric, then they will be in the same place. if the distribution is skewed to the right, then the mean will be larger than the median. If the distribution is skewed to the left, then the mean will be smaller than the median.

An advantage of the median over the mean, is that it is less susceptible to the effects of outliers, and is thus more likely to be close to a "typical" value for skewed distributions. We say the median is "robust" to outliers.

An advantage of the mean over the median, is that it is easier to compute, because it depends only on the sum of the data, not the entire set of data. With large sets of data, it is much faster to compute the mean than the median on a computer.

Also, the mean allows one to find the total, while the median does not contain this information.

Example. If the mean of ten numbers is 15.7, then the total of the numbers is 157.

If the median of ten numbers is 15.7, we cannot specify the total.

2.5 Quartiles and the interquartile range

The median divides a quantitative variable in half. The lower quartile cuts off the lower quarter of observations and the upper quartile cuts off the upper quarter. Together, these three statistics divide the distribution of the variable into four pieces of equal size.

Example. Sorted data: 1, 4, 4, 5, 6, 8, 10, 13, 13, 15, 17

There are 11 observations ($n = 11$). The median cuts them in half. $11/2 = 5.5$, so we'd like to have five and a half observations on each side. This means the sixth observation needs to be "cut in half" and shared. The sixth observation is the number 8, which sits precisely on the boundary between the upper and lower halves, and is the median. To divide the observations into four equal pieces, we want $11/4 = 2.75$ observations in each quartile. Thus, the numbers 1 and the first 4, are in the lowest piece, but we want to split the second 4. Because this 4 sits exactly on the border, it is the lower quartile. Working from the top by the same method, the boundary falls on the second 13 and 13 is the upper quartile.

Example. Sorted data: 1, 4, 4, 5, 6, 8, 10, 13, 13, 15, 17, 20

Now there are 12 observations. $12/2 = 6$ and we can divide the observations into two equal pieces of six observations each. The dividing line could be any point between 8 and 10. By convention, we choose 9. To find the quartiles, $12/4 = 3$, so there are exactly three observations in each piece. The lower quartile is between the second 4 and the 5. By convention we say it is 4.5. The upper quartile is between the second 13 and the 15, and we give it the value 14.

If $n/2$ is odd, the median will be the unique middle number. If $n/2$ is even, the median is the average of the two middle numbers. If $n/4$ has a fraction, we want to split numbers and the lower and upper quartiles are chosen exactly. (Just round up and count from the bottom and top.) If $n/4$ is an integer, the splitting points fall between observations and the lower and upper quartiles are averages of the two observations closest to the respective boundaries.

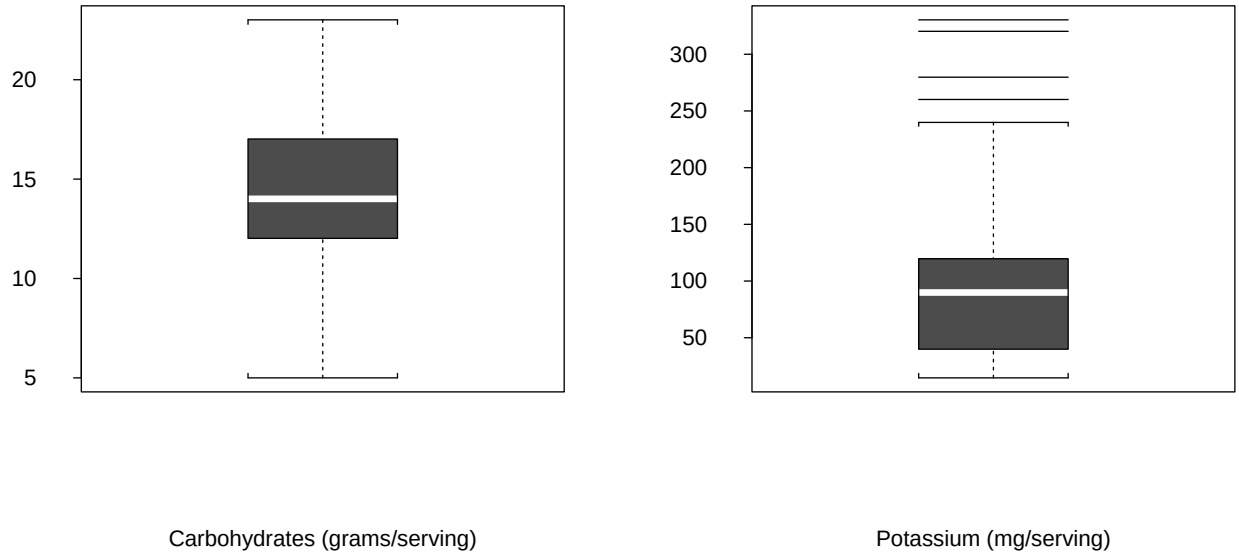
Example. For the carbohydrate variable, there should be $73/4 = 18.25$ observations in each quarter, so the 19th observation falls right on the boundary. The lower quartile is the 19th lowest observation, 12, and the upper quartile is 19th highest observation, 17.

The *interquartile range* or IQR is the difference between the upper quartile and the lower quartile. The IQR is frequently used in economics as a measure of the spread of a distribution. We will use it in constructing boxplots (below). For the two examples above, the interquartile ranges are $13 - 4 = 9$ and $14 - 4.5 = 9.5$. For the carbohydrate variable, the IQR is $17 - 12 = 5$.

2.6 Boxplots

A *boxplot* or *box and whisker plot* is a graphical display of this five number summary: minimum, lower quartile, median, upper quartile, maximum. Variants identify potential outliers. The "box" is formed with the top of the box at the upper quartile, the split in the box at the median, and the bottom of the box at the lower quartile. The box is the middle half of the data.

Any individual observations located a distance more than 1.5 IQR from the box are identified individually with a small point or line. A whisker is then drawn from the top of the box to the largest observation not identified as a potential outlier and a second whisker descends from the bottom of the box to the smallest value not identified as a potential outlier. The maximum and minimum value will be identified either as potential outliers or at the ends of the whiskers.



Example. The boxplot of carbohydrates does not contain any outliers. The lack of skewness is indicated by the similar sizes of the two whiskers, the absence of outliers, and the median being close to the middle of the box. In contrast, the boxplot of potassium indicates a strong skew to the right because the upper half of the data is spread so much more than the lower half. The quartiles of the potassium data are at 40 and 120, so the IQR is 80. Any individual observations more than $(1.5 \times 80 =)$ 120 from the box are marked.

> **Question:** Which four brands of cereal have unusually high potassium measurements? Is there something other than potassium they have in common? You cannot answer these questions from the boxplot alone, but will need to peruse the raw data.

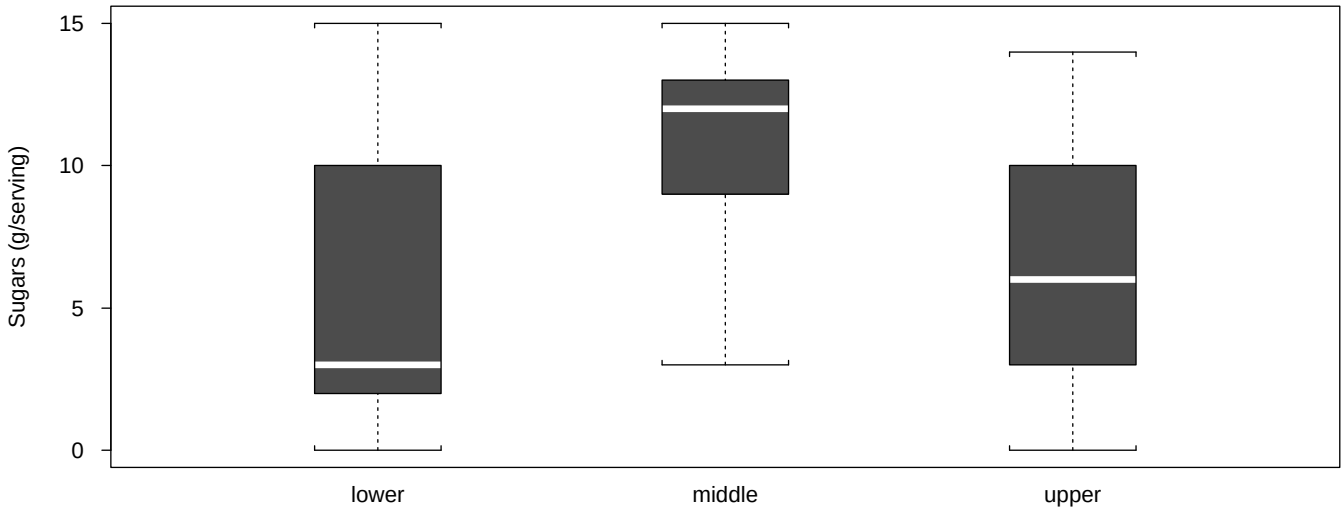
2.7 Comparing histograms, stemplots, and boxplots

Histograms are the most useful and flexible graphical summary of single quantitative variable. They are especially good at showing features of a distribution (if there is enough data). They are true summaries of the data, which cannot be precisely reconstructed.

Stemplots are quickest to construct by hand, and because they retain the sorted original data are a useful first step for computing medians and quartiles by hand. Stem plots are not useful when there are more than a few hundred observations.

A boxplot can indicate skewness and be useful for identifying outliers. Boxplots do not show features of a distribution such as the general shape or the location of modes. Histograms are generally more descriptive for summarizing a single variable, but drawing side by side boxplots is a nice way to simultaneously compare several variables measured on the same scale. These are especially useful when examining a single quantitative variable separately for each category of a categorical variable.

Example. Side-by-side boxplots of the quantitative variable sugar per serving versus the categorical variable shelf.



Question: These side-by-side boxplots substantial differences in the amount of sugar per serving for the three different shelf locations. The middle shelf is at the eye level of young children. Do you think this is coincidence?

2.8 The standard deviation and the variance

The crudest measure of spread is the *range*, the difference between the highest and lowest observations. The range of the carbohydrate variable is $23 - 5 = 18$. Because this measure uses only two of the values and ignores the rest, it is not very useful. A better measure we have already encountered is the interquartile range, or IQR. This measure is mainly used for descriptive purposes and is rarely employed in the health sciences.

The measure of spread which is most important for statistical inference and is ubiquitous in health science journal articles which use statistical methods is the *standard deviation*.

A natural way to measure spread is to quantify the overall distance between individual observations and some central measure. There are many possibilities, but the standard deviation is used most often, mainly because it (and the variance on which it is based) have nice mathematical properties.

The mean, $\bar{x}$, is the central measure we will use. Each observation x_i is some distance from the mean, $x_i - \bar{x}$. An observation's distance from the mean is called a deviation from the mean. This deviation is positive if the observation is above the mean and negative if it is below. While it is perfectly reasonable to consider the average absolute deviation from the mean as a measure of spread, we will do something different. Instead we will consider something that is "almost" the average squared deviation from the mean. This measure, the *variance*, has the formula

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

The numerator is the sum of the squared deviations from the mean and the denominator is one less than the number of observations instead of n, which is why the variance is only "almost" a mean. We will put off the reason for using $n-1$ instead of n until the time we begin discussing statistical inference.

The only difficulty with the variance is that the units are squared. If our variable is measured in grams, the variance is in grams squared. The standard deviation, s, is the square root of the variance, and has the same units as the original data lending itself to easier interpretation.

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

Your textbook goes through great pain to show how to calculate the standard deviation by hand with alternative formulas that give the same answer as the one above. Instead, I expect you to understand why the formula above measures spread and to be able to use technology to compute it. (I will ask you to do one calculation by hand to get a feel for the formula.) You will need a scientific calculator for this course which computes standard deviations.

Computing the standard deviation by hand Plugging into the equation amounts essentially to doing these steps.

1. Find the mean.
2. Subtract the mean from each individual observation.
3. Square each of these deviations.
4. Add the deviations up.
5. Divide by one less than the number of observations.
6. Take the square root.

Interpreting the standard deviation A standard deviation may be thought of as a typical deviation from the mean, in many instances. We will see that for most cases, it is very unusual for an observation to be more than 3 standard deviations from its mean in either direction. If a variable has a bell-shaped symmetric distribution, the mean and the standard deviation alone will give a very acceptable summary of the distribution.

For the carbohydrate data, the standard deviation is 3.91. A look at the histogram shows that 4 is indeed a typical deviation from the mean. Most observations are within 4 of the middle, but several are farther away. In fact, $51/73 = 70\%$ of the observations are within one standard deviation of the mean, $71/73 = 97\%$ are within two standard deviations and all are within three standard deviations. Next chapter we will see that these values are typical for symmetric bell-shaped variables.

2.9 Skills to master

After reading these notes and practicing on exercises I will provide, you should be able to do these things. This list is a list of technical skills you will need to master. Remember that mastering these skills is not the purpose of the course. We want to build toward *using these skills to enable critical thinking based on data* to make judgments on health science related problems. It is critical that you master these skills well enough so that they become tools you can use to approach problems on a deeper conceptual level without undo emphasis on simple calculation.

- Construct a histogram by hand from data (in principle).
- Answer questions like "about what proportion of brands of cereal have less than 16 grams of carbohydrate per serving?"
- Approximate the location of the median, mean, quartiles, or percentiles of a variable based on its histogram.
- Determine whether a distribution is skewed and, if so, the direction of the skew from histograms, stemplots, and boxplots.
- Determine from a graphical summary if the mean is much smaller than the median, about the same, or much greater without direct calculation.
- Estimate the standard deviation roughly from a histogram.
- Describe the shape of a distribution from a histogram or stemplot.
- Read statistics off of a boxplot.
- Compute statistics (mean, median, quartiles, standard deviation, IQR) from small data sets by hand.
- Find means and standard deviations using a calculator.