

AP Statistics Notes – Unit One: Exploring, Understanding and Analyzing Data

Syllabus Objectives: 1.1 – The student will construct graphical displays of distributions of univariate data. 1.2 – The student will interpret graphical displays of distributions of univariate data. 1.20 – The student will explore categorical data using frequency tables and bar charts.

Statistics is the science of learning from data. The two major branches of statistics are descriptive and inferential. **Descriptive statistics** is the branch of statistics that includes methods for organizing, displaying and summarizing data. **Inferential statistics** is the branch of statistics that involves generalizing from a sample to the population from which it was selected and assessing the reliability of such generalizations. The first semester of AP Statistics deals with descriptive statistics and the second semester concentrates on inferential. In this unit, we will be dealing with data. The first step in understanding the data is to hear what the data say. **Data analysis** is organizing, displaying and summarizing this data.

Individuals: Any set of data containing information about some group.

Variables: The characteristics we measure on each individual.

When you meet a new set of data, ask yourself the following key questions: **Who** are the individuals described by the data? **What** are the variables? **Why** were the data gathered? **When, where, how, and by whom** were the data produced?

The two major types of variables studied in this unit are categorical and quantitative variables.

A **categorical variable (qualitative data)** places an individual into one of several groups or categories. These values fall into separate, nonoverlapping categories.

Ex: marital status, eye color, birthplace, zip code, team standings

A **quantitative variable (numerical data)** takes numerical values for which arithmetic operations such as adding and averaging make sense. These measurements often have **units** like, dollars, degrees, inches, etc.

Ex: height, number of siblings, time taken on a test, salary, length of first name

The first rule of data analysis is to make a picture of the data. The **distribution** of the variable tells us what values the variable takes and how often it takes these values. Let's look at the different ways to graph our distributions.

Displaying and describing categorical data:

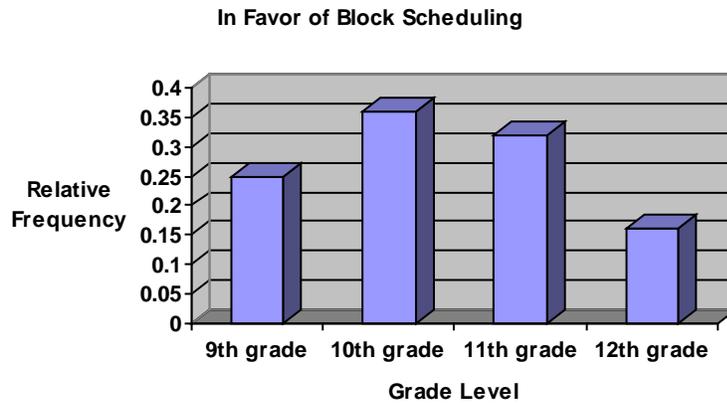
- Make and examine a table of counts. We can make a **frequency table** which organizes these counts. Here, students were asked "Do you favor a block schedule?". The number of yeses for each grade level is recorded in the table. A yes/no question is categorical.

○

Grade Level	Favoring block schedule (frequencies)	Relative frequency
9 th grade	5	0.25
10 th grade	9	0.36
11 th grade	8	0.32
12 th grade	4	0.16

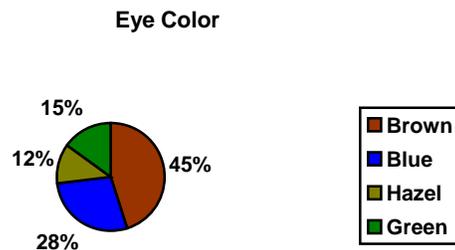
- A bar chart displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison. The bars should have spaces between them. The vertical axis of the bar chart can use the counts/frequencies or the percents, which are called the relative frequencies.

○



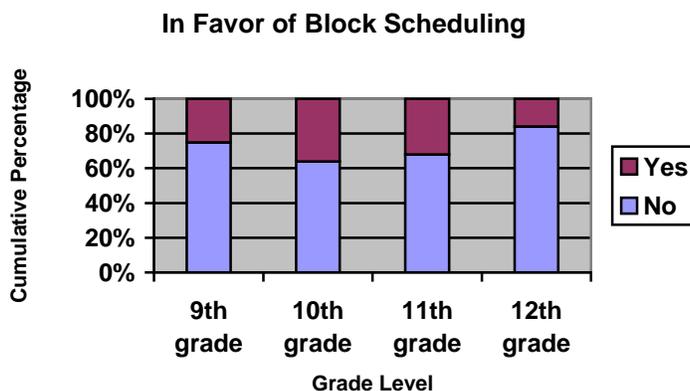
- A pie chart shows the whole group of cases as a circle. The piece has a size proportional to the fraction of the whole in each category. Here students were classified by their eye color.

○



- A segmented bar chart or stacked bar chart can represent data by showing the different categories as segments in one rectangle. The vertical axis of the segmented bar chart is the cumulative percentage - so each rectangle will add up to 100%.

○



Rules to remember when displaying categorical data:

1. Bar charts have spaces between each category of the variable.
2. The order of the categories is not important.
3. Either counts or proportions may be shown on the vertical axis.
4. Make sure that your data display has a descriptive title and that both of your axes are appropriately labeled.

When describing categorical distributions, make sure to describe them in the **CONTEXT** of the data. It is also not appropriate to describe the shape of the data. If making comparisons between two or more data sets, make sure to use comparative words like “larger” or “smaller”.

Displaying and describing numerical data:

There are two types of numerical data. **Discrete** numerical data almost always result from counting and are most often integers. Discrete examples would be number of siblings, length of first name, number of doctor visits in a year. **Continuous** numerical data are observations that are usually measurements and are numbers along an interval. Continuous examples are temperature, salary, height, weight, time to perform a task. Discrete data can be graphed on a dotplot, stem-and-leaf display, a bar chart, frequency table or boxplot. Continuous data can also be graphed using the displays mentioned previously, but instead of a bar chart, a histogram is used. The histogram has bars without spaces denoting the continuous nature of the data.

- A stem-and-leaf display contains all of the variables. The vertical column on the left is called the stem and the numbers to the right of the vertical line are called leaves. Leaves are ordered and you should have a title and a key on your display. Stem-and-leaf displays are good for small sets of data, as all values must be displayed. If the data is concentrated on just a few stems, repeated stems can be used to stretch the display. Divide the stems into two groups – “low” leaves (values 0 – 4) and “high” leaves (values 5 – 9). A comparative stem-and-leaf display can provide a better display for looking at two groups of data. The leaves for one group are to the right of the stem and the leaves for the second group are to the left of the stem.

- Stem-and-leaf display:

Number of Touchdown Passes for NFL (2000)

```
3 | 2337
2 | 001112223889
1 | 2244456888899
0 | 69
```

Key: 2 | 1 = 21

- Split or Repeated Stems Display:

Number of Touchdown Passes for NFL (2000)

```
3 | 7
3 | 233
2 | 889
2 | 001112223
1 | 56888899
1 | 22444
0 | 69
```

Key: 2 | 1 = 21

- Comparative stem-and-leaf display – comparing touchdown passes from two different years.

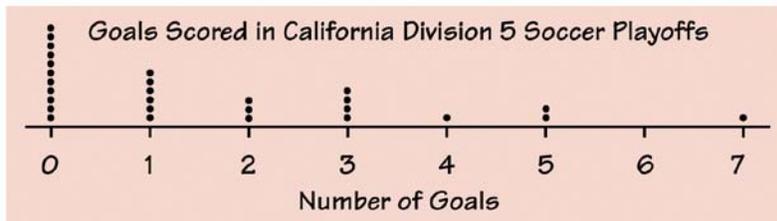
Number of Touchdown Passes for NFL

1998		2000
11	4	
	3	7
332	3	233
8865	2	889
443311110	2	001112223
98776665	1	56888899
321	1	22444
7	0	69

Key: 2 | 1 = 21

- A dotplot is a simple display. It places a dot along an axis for each case in the data. It is similar to a stem-and-leaf display, but with dots instead of digits for the leaves. The dotplot should have a scale, a variable label for the axis, and a title.

-



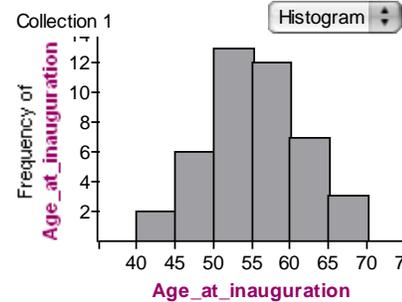
- When the data set is large, a dotplot and stem-and-leaf display are not the best graphs for displaying the data. A frequency distribution and a histogram are best suited for these types. When constructing a histogram, the data is divided into classes of equal widths and the heights of the bars are the frequencies or counts of the cases falling in each class. The vertical axis can be frequencies or relative frequencies. The relative frequency is just the frequency of the class divided by the total frequencies.

- Histogram tips:

- There is no one right choice of the classes in a histogram. Too few classes or too many classes will not give a good picture of the shape of the distribution.
- Five classes is a good minimum. Another rule of thumb is to take the square root of the number of values in the distribution and use this as your guide. For example, if there are 100 values in the data set, 10 bars would be used.
- Our eyes respond to the area of the bars in a histogram, so be sure to choose classes that are all the same width. Then, area is determined by height and classes are fairly represented.
- If you use a computer or graphing calculator, beware of letting the device choose the classes. Set your own window.

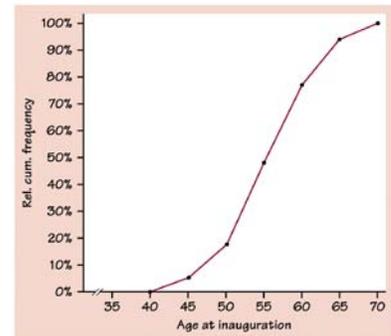
- Let's look at the ages of the presidents at inauguration. There are 43 values ranging from 42 to 69 years of age. Taking the square root of 43, we get approximately 6, so we will have 6 classes. To make division simple, start from age 40 and to reach 70, or 30 numbers and get 6 classes, the class width will be five: $(70 - 40) = 30 \div 6 = 5$. First, make the frequency table and then construct the histogram.

Class	Count	Rel. Freq.
40 - <45	2	.0465
45 - <50	6	.1395
50 - <55	13	.3023
55 - <60	12	.2790
60 - <65	7	.1628
65 - <70	3	.0698



- A histogram does a good job of displaying the distribution of values of a variable. But it tells us little about the relative standing of an individual observation. If we want this type of information, we construct a **relative cumulative frequency graph**, often called an **ogive**. Add the relative frequencies to get the cumulative frequencies and plot these as points that are then connected. The sum will add up to one – note .9999 in the chart due to round-off error.

Class	Count	Rel. Freq.	Cum. Rel. Freq.
40 - <45	2	.0465	.0465
45 - <50	6	.1395	.1860
50 - <55	13	.3023	.4883
55 - <60	12	.2790	.7673
60 - <65	7	.1628	.9301
65 - <70	3	.0698	.9999



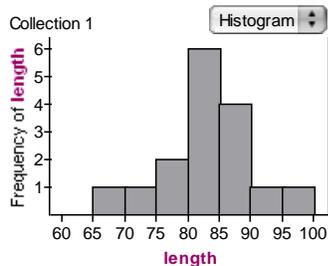
- To locate a value corresponding to a percentile, find the percentile; draw a horizontal line across the vertical axis until it meets the ogive. From there, draw a vertical line down to the vertical axis. For example, the 50%-tile for this distribution would be approximately 55.

Constructing a graph is just the first step. The next step is to interpret what you see. When you describe a NUMERICAL distribution, you should pay attention to the following features:

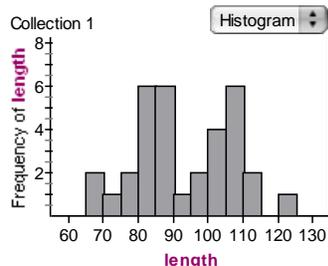
- In any graph of data, look for the **overall pattern** and for striking **deviations** from the pattern. Look for gaps and concentrations.
- You can describe the overall pattern of a distribution by its **shape, center and spread**.
- An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

To summarize, use your **SOCS**. **Shape, outliers, center and spread**.

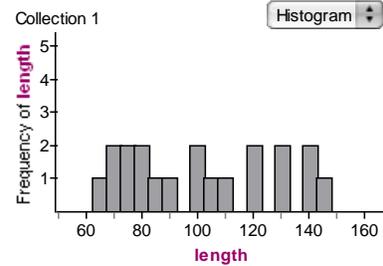
Before looking at center and spread, let's deal with shape and outliers. Identifying outliers is a matter of judgment. Look for points that are clearly apart from the body of the data, not just the most extreme observations in a distribution. It is not a good idea to just delete or ignore outliers. Shape is a bit more technical. Does the distribution have a single, central hump or several separated humps? These humps are called modes. A histogram with one peak is **unimodal**, two peaks is **bimodal** and for those with three or more, they are called **multimodal**. A histogram that has bars approximately the same height is called **uniform**.



Unimodal

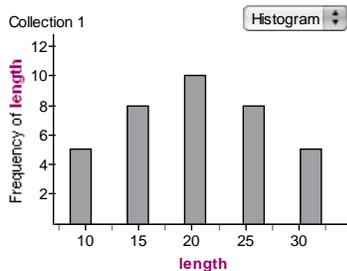


Bimodal

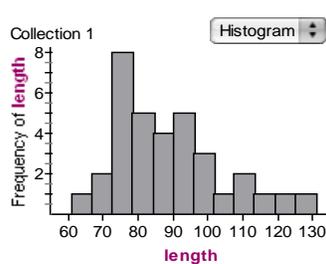


Uniform

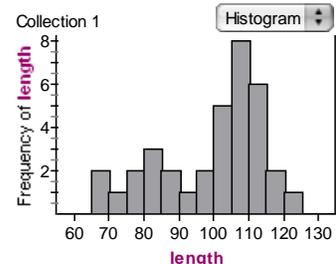
After looking at mode, ask, is the histogram **symmetric**? A distribution is symmetric if the values small and larger than its midpoint are mirror images of each other. It is **skewed** to the right if the right tail (larger values) is much longer than the left tail (smaller values). This is also known as positively skewed. If the left tail is longer it is negatively skewed or skewed to the left.



Symmetric



Skewed right or positively skewed



Skewed left or negatively skewed

Syllabus Objectives: 1.3 – The student will summarize distributions of data measuring the center using median and mean. 1.4 – The student will summarize distributions of data measuring the spread using range, interquartile range, and standard deviation. 1.5 – The student will summarize distributions of data measuring the position using quartiles and percentiles. 1.6 – The student will summarize distributions of data using box plots.

When describing numerical data, it is common to report a value that is representative of the observations. Such a number describes roughly where the data are located or “centered” along the number line. The two most popular measures of center are the **mean** and **median**.

The **mean** of a set of numerical observations is the familiar arithmetic average: the sum of the observations divided by the number of observations.

Definition: The **sample mean** of a sample of numerical observations,

$$x_1, x_2, \dots, x_n, \text{ denoted by } \bar{x} \text{ is } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x}{n}.$$

The **median** of a set of numerical observations is the middle value once the data values have been listed in order from smallest to largest. If the sample size n is odd then the median is the single middle value. If the sample size n is even, then there are two middle values and those values are averaged to obtain the sample median.

How do we decide when to use the mean or median? The mean and median of a symmetric distribution are close together. If the distribution is exactly symmetric, they are exactly the same. However, in a skewed distribution, the mean is farther out in the long tail than is the median. If the distribution is positively skewed, the mean will be larger than the median and in a negatively skewed distribution it is less than the median. Because the median considers only the order of the values, it is **resistant** to values that are extraordinarily large or small. Statistics that are not greatly affected by outliers are called **resistant**. The mean is **not** resistant to outliers because the formula takes all values (like the extremes) into account. So, to answer our question, first look at the data and make a judgment call. If the data is highly skewed or has many outliers, it is more appropriate to report the median. If the data is fairly symmetric with few outliers, report the mean.

Ex: Find the mean and median of Barry Bonds home runs from 1986 to 2005.

1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
16	25	24	19	33	25	34	46	37	33

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
42	40	37	34	49	73	46	45	45	5

By hand: Put the data in order:

5, 16, 19, 24, 25, 25, 33, 33, 34, 34, 37, 37, 40, 42, 45, 45, 46, 46, 49, 73

The middle number is between the 10th and 11th numbers, so we average 34 and 37 to get 35.5. Adding all of them up and dividing by 19, we get 35.4. Because these two values are so close together, we would expect to see a fairly symmetric distribution.

Notice the graphing calculator screens show the same answers for \bar{x} and Med.

```

3:000 CALC TESTS
1:Edit
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
    
```

```

EDIT 3:000 TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
    
```

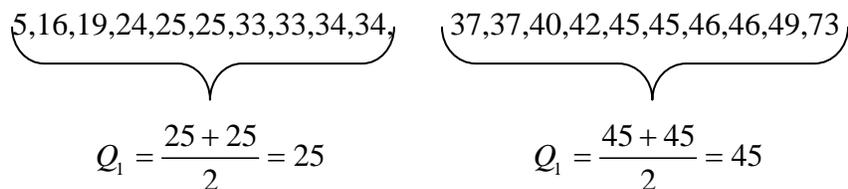
```

1-Var Stats
x̄=35.4
Σx=708
Σx²=29072
Sx=14.52547671
σx=14.15768343
↓n=20
    
```

```

1-Var Stats
↑n=20
minX=5
Q1=25
Med=35.5
Q3=45
maxX=73
    
```

Center is a very important measure, but a measure of center alone can be misleading. Spread is also important to report along with it. There are several measures of spread. The first and simplest is **range**, which is the difference between the largest and smallest observations. In the previous example, the range is 68 (73-5). It shows the full spread of the data, but it depends on only two observations and if there are outliers, the range can be misleading. The range, like the mean, is greatly affected by outliers and is not resistant. We can describe the spread or variability of a distribution by giving several percentiles. The **pth percentile** of a distribution is the value such that p percent of the observations fall at or below it. The median is the 50th percentile as half of the values fall above it and half below it. This leads us to a better way to describe the spread of a variable by ignoring the extremes and concentrating on the middle of the data. We could, for example, find the range of the middle half of the data. Divide the data in half at the median, now divide both halves in half again (the median is not included in either half), cutting the data into four quarters. We call these new dividing points **quartiles**. The first quartile, Q_1 , is the 25th percentile and the third quartile, Q_3 , is the 75th percentile. The difference between the quartiles tells us how much territory the middle half of the data covers and is called the **interquartile range**. It is most commonly abbreviated as the IQR. Using the example above, let's find the IQR. The data is arranged in order below and split into the first two halves. Since each half has an even amount of numbers (10), the middle two numbers must be averaged to find the quartiles.



The $IQR = 45 - 25 = 20$. The values of 25 and 45 are also seen on the graphing calculator screens above. Notice that the median, also known as the 50th percentile and Q_2 , is not used in the calculations of the first and third quartiles.

The **five-number summary** of a set of observations consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. These five numbers offer a reasonable complete description of center and spread. The five number summary for our example is: 5 25 35.5 45 73.

- A **boxplot** is a graph of the five-number summary.
 - A central box spans the quartiles Q_1 and Q_3 .
 - A line in the box marks the median M .
 - Lines extend from the box out to the smallest and largest observations.

```

5:Plot1...On
1:Plot1...On
  L1 1
2:Plot2...Off
  L1 L2
3:Plot3...Off
  L1 L2
4:PlotsOff

```

```

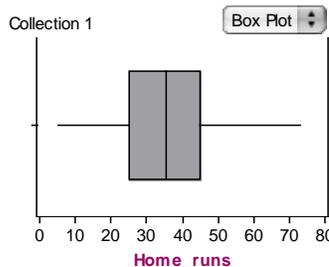
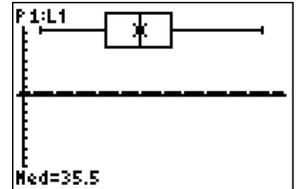
2:Plot2 Plot3
Off Off
Type:
Xlist:L1
Freq:1

```

```

MEMORY
3:Zoom Out
4:ZDecimal
5:ZSquare
6:ZStandard
7:ZTrig
8:ZInteger
ZoomStat

```



Note the scale and label on the x-axis. These should accompany all boxplots.

Up until this point, the term “outlier” has been used very loosely as a value that falls outside of the pattern of the distribution, but now we have a more formal definition. The **1.5 × IQR Rule for Outliers** states that an observation is an outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.

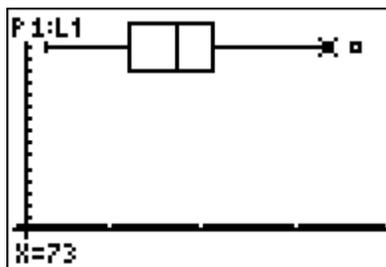
Ex: Applying this rule to the home run data,
 any values below $Q_1 - 1.5IQR = 25 - 1.5(20) = 25 - 30 = -5$
 or above $Q_3 + 1.5IQR = 45 + 1.5(20) = 45 + 30 = 75$.

In this example, there are no values below -5 or above 75 , therefore, there are no outliers. If outliers are present, a **modified boxplot** would then be constructed. In a modified boxplot, the lines extend out from the central box only to the smallest and largest observations that are NOT outliers. The outlier or outliers are then shown as points.

```

2:Plot2 Plot3
Off Off
Type:
Xlist:L1
Freq:1
Mark:

```



By adding the value 80 to the homerun data set, I can illustrate the outlier in a modified boxplot. The 80 is shown as a small box and the line extends out to the next highest value, 73.

More on measuring spread: The standard deviation

When a distribution is skewed or has outliers, the five-number summary is often the best way to describe the center and spread, but the most common numerical descriptions for center and spread belong to the mean to measure center and the standard deviation to measure spread. These measures are often reported when the distribution of the data is symmetric. The standard deviation measures spread by looking at how far the observations are from the mean. It is the average amount that the data varies from the mean.

Definition: The **sample variance**, denoted by s^2 , is the sum of squared deviations from the mean divided by $n - 1$. That is, $s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$.

Note that each value is subtracted from the mean to find its difference or *deviation* from the mean. To keep them from canceling out, the deviations are squared and then averaged. Since the units of the variance are in *squared* units, we then take the square root of this formula to find s , the standard deviation. The **sample standard deviation** is the positive square root of the sample variance and is denoted by s . A large amount of variability is indicated by a large standard deviation and a small standard deviation indicates a small amount of variability.

- Properties of the Standard Deviation
 - s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
 - $s = 0$, only when there is NO spread or variability. This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.
 - s , like the mean, is not resistant. A few outliers can make s very large.
 - s can be found on the graphing calculator using 1-Var Stats. The sample standard deviation is denoted by “Sx=” on the calculator.

Ex: Find the standard deviation of the data set:

Data: 2, 3, 5, 6, 9, 11

$$\bar{x} = \frac{36}{6} = 6$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
2	2-6 = -4	16
3	3-6 = -3	9
5	5-6 = -1	1
6	6-6 = 0	0
9	9-6 = 3	9
11	11-6 = 5	25

$$\sum(x - \bar{x})^2 = 60$$

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{60}{5} = 12$$

$$s = \sqrt{12} = 3.464$$

Note: $\sum(x - \bar{x}) = 0$

Syllabus Objectives: 1.7 – The student will analyze the effect of changing units on summary measures.

The same variable can be recorded in different units of measurement. It is easy to convert from one unit of measurement to another and this change in the measurement unit is a **linear transformation** of the measurements.

Definition: A **linear transformation** changes the original variable x into the new variable x_{new} given by an equation of the form: $x_{new} = a + bx$. Adding the constant a shifts all values of x upward or downward by the same amount. Multiplying by the positive constant b changes the size of the unit of measurement.

Ex: Original data set: 1,2,3,4,5 $\bar{x} = 3$ and $s = 1.58$

Add 1 to each value: 2,3,4,5,6 $\bar{x} = 4$ and $s = 1.58$

Ex: Now multiply each value by 10: 10,20,30,40,50 $\bar{x} = 30$ and $s = 15.8$

- Effects of a Linear Transformation
 - Adding the same number a (either positive, zero, or negative) to each observation adds a to measures of center and to quartiles but DOES NOT change measures of spread.
 - Multiplying each observation by a positive number b multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by b .

Ex: Maria measures the lengths of 5 cockroaches and these are the results in inches:

1.4	2.2	1.1	1.6	1.2
-----	-----	-----	-----	-----

The 1-Var Stats for the distribution is below.

```

1-Var Stats
x̄=1.5
Σx=7.5
Σx²=12.01
Sx=.4358898944
σx=.3898717738
↓n=5

```

```

1-Var Stats
↑n=5
minX=1.1
Q1=1.15
Med=1.4
Q3=1.9
maxX=2.2

```

Maria wants to change the units from inches to centimeters. (There are 2.54 cm in 1 inch). Below are the new summary statistics.

```

1-Var Stats
x̄=3.81
Σx=19.05
Σx²=77.483716
Sx=1.107160332
σx=.9902743054
↓n=5

```

```

1-Var Stats
↑n=5
minX=2.794
Q1=2.921
Med=3.556
Q3=4.826
maxX=5.588

```

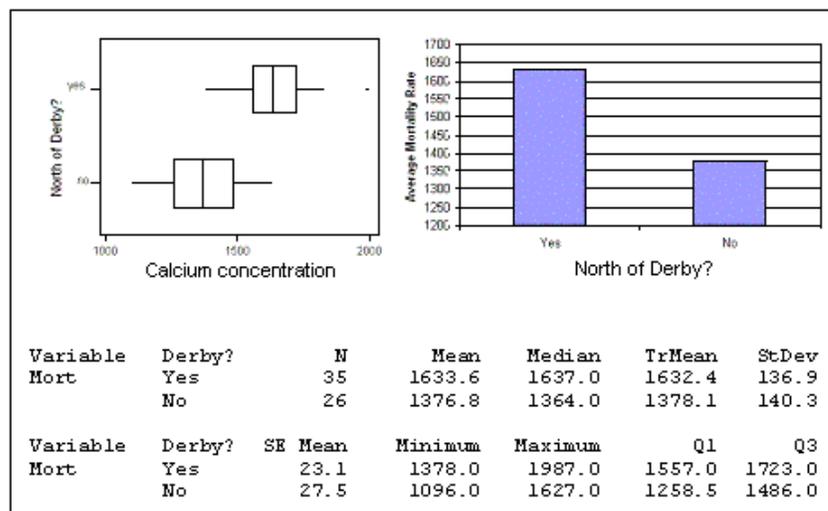
Note that all of the summary statistics have been multiplied by a factor of 2.54.

Syllabus Objectives: 1.8 – The student will compare distributions of data using dot plots, back-to-back stem-and-leaf plots, and parallel boxplots. 1.9 – The student will compare center, spread, and variation of data within a group and between groups. 1.10 – The student will compare distributions of data with respect to outliers and other unusual features. 1.11 – The student will compare distributions of data with respect to their shapes. 1.24 – The student will compare distributions of categorical data using bar charts.

Things to remember: Use double bar graphs to compare distributions of categorical data. Make back-to-back stemplots and parallel boxplots to compare distributions of quantitative variables. Always write narrative comparisons of the shape, center, spread, and outliers for two or more quantitative distributions. When asked to describe two data sets, make sure that you **COMPARE** the two data sets (reference similarities/differences in shape, center, and spread) and don't just describe each data set individually.

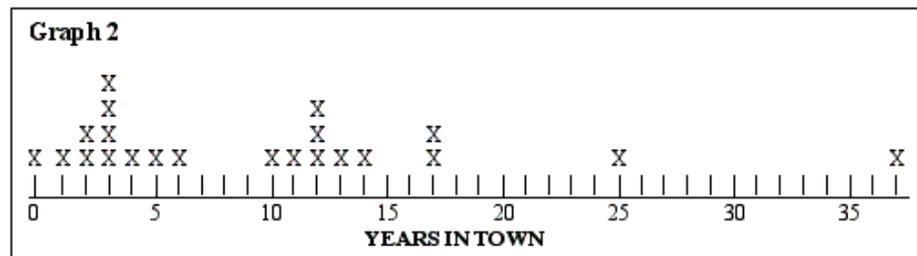
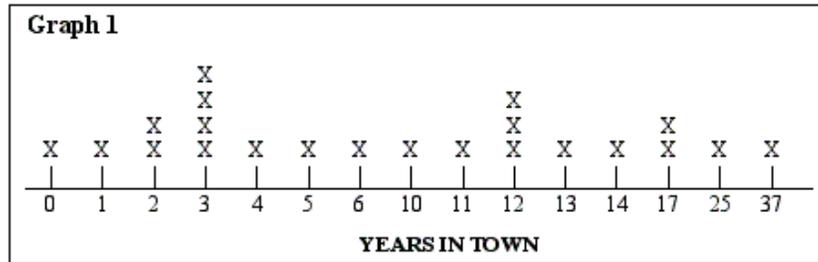
Ex: Data were collected in an investigation of environmental causes of disease. Variables include the annual mortality rate per 100,000 for males, averaged over the years 1958-1964 (think average number of deaths per year), and the calcium concentration (in parts per million) in the drinking water supply for a random sample of 61 large towns in England and Wales. (The higher the calcium concentration, the harder the water.) Another variable is whether or not the town is north of Derby (with towns at least as far north as Derby identified by a 1).

Below are graphical and numerical summaries of the calcium concentrations for towns north of Derby and towns not north of Derby. Summarize how the two groups compare.



Solution: The bar graph plots the categorical data which illustrates that the average mortality rate is much higher for towns north of Derby. The parallel boxplots show the calcium concentration in the drinking water. First, we see that the centers of the two distributions vary greatly. The median for towns north of Derby is much higher at 1637 versus the median not north of Derby at 1364. The shapes of the two distributions are fairly similar, both being quite symmetric. The spreads of the two distributions are also similar with standard deviations of 136.9 and 140.3. It appears that the distribution representing towns north of Derby has one outlier whereas the other distribution does not contain any outliers. (SOCS!)

Ex: A class of students recorded the number of years their families had lived in their town. Here are two graphs that students drew to summarize the data. Which graph gives a more accurate representation of the data? Then, use the appropriate display to describe the distribution.



Solution: Graph 2 gives a more accurate representation of the data. The scale on Graph 1 is not even. We can see the true shape of the distribution on the second graph. The distribution is skewed to the right (positively skewed) with possible outliers at 25 and 37. There are gaps between 6 and 9 and also at 15 and 16. The center (count to find the median) is at 8 and the spread (range) is 37. The first quartile is 3 and the third quartile is 13 and the *IQR* is 10.

Ex: These graphs were part of a newspaper story reporting on boating deaths in Tasmania. Write an interpretation of these graphs and what information they provide about boating deaths in Tasmania.

Solution: From the categorical bar graphs, we see that a majority of the accidents involved instances where the boaters were not wearing a life jacket and that alcohol was not a factor. The bar graph plotting the number of deaths per year shows us that the most amount of deaths occurred in 1999. Aside from that year, each previous year ranged from 0 to 6 boating deaths per year. This graph would be appropriate to show the large increase of deaths in 1999 when compared to previous years.

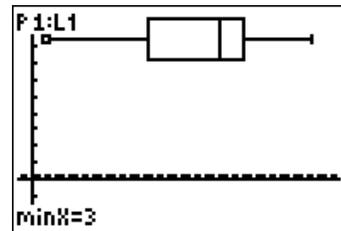


Ex: Ninety-eight (98) men and women were asked "How many hours did you work last week?" A stem-and-leaf plot of their responses is presented. Each stem represents 10s of hours. For example, a stem of 2 with a leaf of 8 represents 28 hours of work. Summary statistics are also reported below the stem-and-leaf plot.

0	3
0	7
1	3
1	5 5 6 6 8
2	0 0 0 0 0 0 0 0 0 1 1 2 4 4
2	5 5 5 5 5 6 6 8 9
3	0 0 0 0 0 4
3	5 5 5 6 7 7 8 8 9
4	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 2 3 3 4
4	5 5 5 5 5 5 8 8 8 8 8
5	0 0 0 0 0 1
5	5 5 5 6 7
6	0 0 0
6	

n = 98	Q ₁ = 25.0
\bar{x} = 35.51	Median = 40.0
sd = 12.77	Q ₃ = 44.5

Construct a boxplot of the data, showing outliers, if any. You must show ALL calculations needed to determine if there are outliers.



There are no outliers. The *IQR* is 19.5. Outliers would be below $25 - 1.5(19.5) = -4.5$ and above $44.5 + 1.5(19.5) = 73.75$.

Summarizing the distribution, the left side of the “box” is longer than the right side showing the data from the 25th percentile to the 50th percentile is more spread out than from the 50th to the 75th percentile, so the distribution is skewed to the left. This can also be noted by looking at the values for the mean and the median. The median is larger than the mean implying a negatively skewed (skewed to the left) distribution. To describe the center and spread of the distribution, we should use the median and *IQR* versus the mean and standard deviation due to the skewness of the boxplot. The distribution of work hours is centered at 40 and the spread of the middle half of the data is 19.5.

Ex: The goal of a nutritional study was to compare the caloric intake of adolescents living in rural areas of the United States with the caloric intake of adolescents living in urban areas of the United States. A random sample of ninth-grade students from one high school in a rural area was selected. Another random sample of ninth graders from one high school in an urban area was also selected. Each student in each sample kept records of all the food he or she consumed in one day.

The back-to-back stemplot below displays the number of calories of food consumed per kilogram of body weight for each student on that day.

<u>Urban</u>		<u>Rural</u>	
9 9 9 9 8 8 7 6	2		
4 4 3 1 0	3	2 3 3 4	
9 7 6 6 5	3	5 6 6 6 7	
2 0	4	0 2 2 2 4	
	4	5 6 8 8 9	Stem: tens
	5	1	Leaf: ones

(a) Write a few sentences comparing the distribution of the daily caloric intake of ninth-grade students in the rural high school with the distribution of the daily caloric intake of ninth-grade students in the urban high schools.

The mean (40.45 cal/kg) and median (41 cal/kg) daily caloric intake of ninth-grade students in the rural school are higher than the corresponding measures of center, mean (32.6 cal/kg) and median (32 cal/kg), for ninth-graders in the urban school. There is also more variability or spread in the daily caloric intake for students in the rural school (Range=19,SD=6.04,IQR=10) than in the daily caloric intake for students in the urban school (Range=16),SD=4.67,IQR=7). The shapes of the two distributions are also different. The distribution of daily caloric intake for rural students is more uniformly distributed (symmetric) between 32 cal/kg and 51 cal/kg while the distribution of daily caloric intake for urban students appears to be skewed toward the larger values (skewed right).

(b) Researchers who want to conduct a similar study are debating which of the following two plans to use.

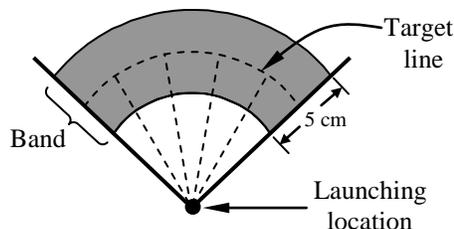
Plan I: Have each student in the study record all the food he or she consumed in one day. Then researchers would compute the number of calories of food consumed per kilogram of body weight for each student for that day.

Plan II: Have each student in the study record all the food he or she consumed over the same 7-day period. Then researchers would compute the average daily number of calories of food consumed per kilogram of body weight for each student during that 7-day period.

Assuming that the students keep accurate records, which plan, I or II, would better meet the goal of the study? Justify your answer.

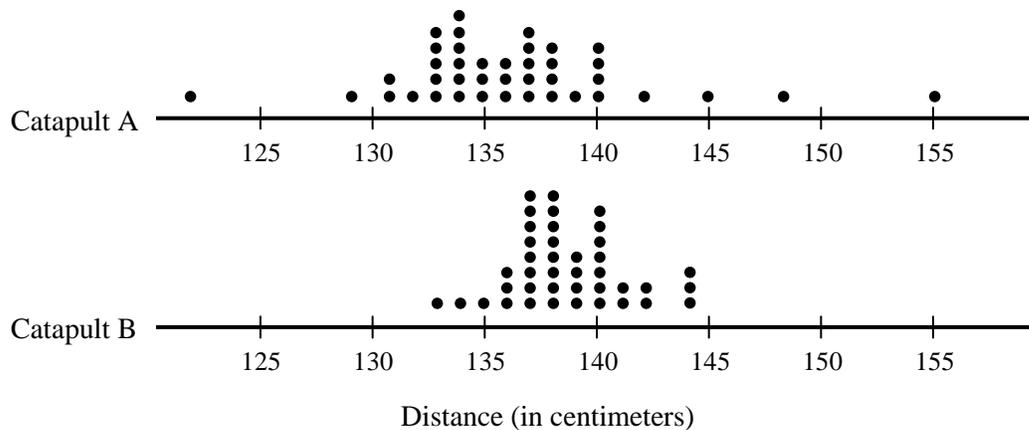
Since we are assuming that students keep accurate records, Plan II will do a better job of comparing the daily caloric intake of adolescents living in rural areas with the daily caloric intake of adolescents living in urban areas. Both plans take body weight into account by converting to food consumed per kilogram of body weight. Plan II includes a 7-day period (possibly days in school and days at home on the weekend), and there are differences in caloric intake among days. It would therefore be better to average over the 7-day period rather than considering only the food consumed in one day, as is the case with Plan I. Plan II would provide a more precise estimate of the average daily intake.

Ex: Two parents have each built a toy catapult for use in a game at an elementary school fair. To play the game, students will attempt to launch Ping-Pong balls from the catapults so that the balls land within a 5-centimeter band. A target line will be drawn through the middle of the band, as shown in the figure below. All points on the target line are equidistant from the launching location.



If a ball lands within the shaded band, the student will win a prize.

The parents have constructed the two catapults according to slightly different plans. They want to test these catapults before building additional ones. Under identical conditions, the parents launch 40 Ping-Pong balls from each catapult and measure the distance that the ball travels before landing. Distances to the nearest centimeter are graphed in the dotplots below.



(a) Comment on any similarities and any differences in the two distributions of differences traveled by balls launched from catapult A and catapult B.

Both distributions of distances are roughly symmetric and somewhat mound-shaped. The center of the distances for catapult A (median A = 136 cm) is slightly lower than the center of the distances for catapult B (median B = 138 cm). There is more variability in the distances traveled by the Ping-Pong balls launched with catapult A. There are distances that are extreme enough to be called potential outliers in the catapult A distribution, but there are no outliers among the catapult B distances.

(b) If parents want to maximize the probability of having the Ping-Pong balls land within the band, which one of the two catapults, A or B, would be better to use than the other? Justify your choice.

Catapult B would be best because the distances vary less about the center of the distribution for catapult B. If catapult B is properly placed, the balls launched will have a higher probability of landing in the narrow (only 5 cm wide) target band.

(c) Using the catapult that you choose in part (b), how many centimeters from the target line should this catapult be placed? Explain why you chose this distance.

The catapult should be placed 138 cm from the target line. Since the distribution of distances for catapult B seems to be fairly symmetric and somewhat mound-shaped, the median (138 cm) is a good representation of the center of the distribution. Placing catapult B at this location would have resulted in a high proportion ($30/40=0.75$) of Ping-Pong balls from this sample of launches landing in the target band.