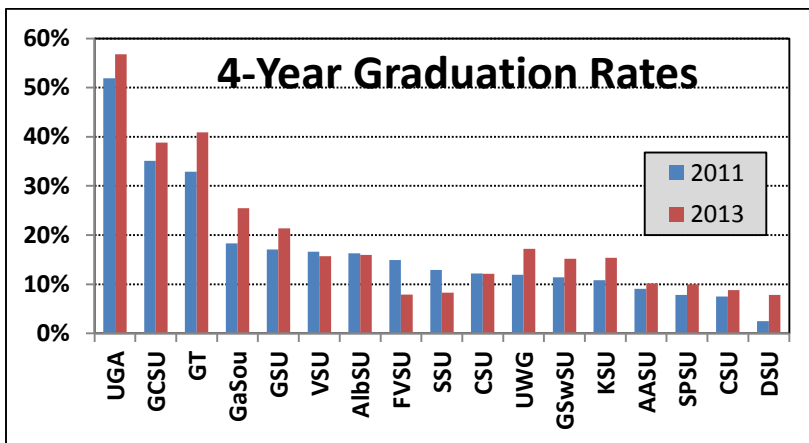


Chapter 1 – Introduction to Statistics

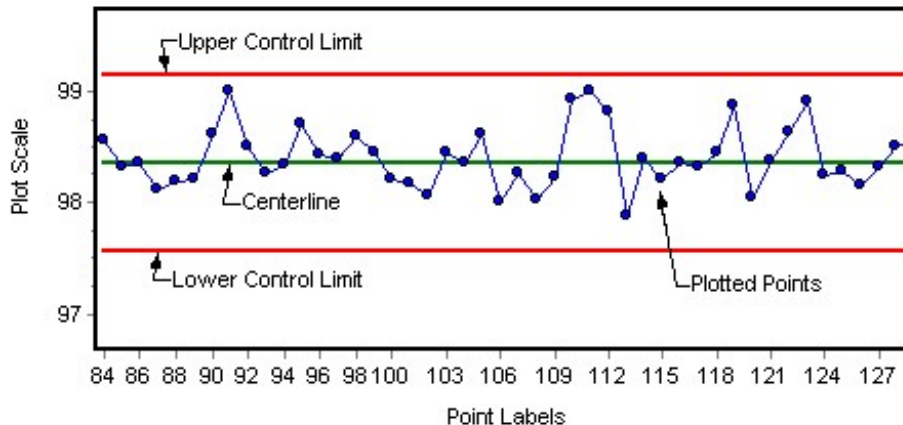
This chapter introduces the basic ideas and terminology of statistics.

1.1 – Definitions 1

1. These are some basic definitions. We will refine them later.
 - a. **Data** – A collection of information about something we are interested in.
 - b. **Statistics** – A summary of data.
2. **Why do we use statistics?**
 - a. To learn more about things we are interested in: [Music](#), [Google](#), [US](#), [Sports](#), [Statista.com](#) (scroll down page), [worldometers](#), [BrainOfBrian](#)
 - b. Provide evidence that a problem exists



- c. Gain insight into the causes of a problem.
- d. To measure the effectiveness of solutions to a problem.
- e. To ensure the things conform to specifications.



Source: <http://www.acqnotes.com/Images/Control%20Chart.png>

- f. To help us make informed decisions about problems we face.
- Education: How can we help students graduate faster? Learn better?
 - Societal: How can we reduce the rate of teenage pregnancies?
 - Business: How can a business be more profitable? Have better customer service?
 - [Sports](#): How can a golfer improve his swing?
 - Science: How do changes in the ozone layer affect glaciers?
 - Engineering: How can we produce more energy efficient cars?
 - [Psychology](#)

3. **Population** – We define a *population* to be the set of *all* things that have a particular characteristic of interest, the totality of the things we are interested in. The population is determined by what the researcher wants learn about. Example: How can we help students graduate faster?

<u>Who is Interested?</u>	<u>Possible Population</u>
<ul style="list-style-type: none">• US Department of Education	All students enrolled in 4-year college or university in the US
<ul style="list-style-type: none">• USG Board of Regents	All students enrolled in a USG 4-year college or university
<ul style="list-style-type: none">• Valdosta State University	All students enrolled in Valdosta State University

4. **Data** – Data, loosely defined is simply a collection of information that is measured, observed, or self-reported about items in a population (or sample). We often use the terms *data*, *data set*, and even *sample* interchangeably.

5. **Variable** – A *variable* is a characteristic that we measure or observe about items in a population in order to obtain data. In most realistic situations there are many variables that are of interest for a particular situation.

a. Example 1:

- Question: How long does it take students to graduate?
- Population: All VSU students who graduated Spring 2014
- Variables: sex, number of semesters to graduate, GPA, number of hours taken each semester, SAT score, *etc.*

b. Example 2:

- Question: How long does it take a customer at a drive-through from the time they enter the line to the time they drive away?
- Population: All customers who use the drive-through during lunch hours (11:30-12:30) on weekdays.
- Variables: Total time, time to order, waiting time, service time, number of people in vehicle, total price of items.

6. Some variables are:

- a. Measured – service time at a drive through, weight of a tomato, height of a plant, speed of a CPU, breaking strength of a steel beam, *etc.*

- b. Observed – sex of a driver, color of a car, using a cell phone or not, presence of a defect on a product or not, etc.
- c. Self-Reported – We may be interested in the amount of money people spend on clothing in a year. We could certainly measure this but it would be very time consuming. In such a situation we might ask randomly selected people to go back over their credit card statements and estimate how much they spent on clothing in the last year. Self-reported data is usually not as accurate as data that is directly measured or observed.

7. **Census** – A *census* is when we measure (or observe) every item in the population. In theory, a census provides exact information about a population. Some problems are amenable to using a census.

- In the case of studying graduation rates in the USG, we can obtain *exact* information on a number of variables due to records being kept in databases and the relatively low cost of querying the databases.

In other cases, it is too time consuming and/or expensive to do a census.

- In the case of studying the total volume of lumber contained in a 500-acre tract of land, it would be very time consuming and expensive to measure the height and diameter of every tree on the tract.

In other cases, conducting a census is fraught with error.

- The US conducts a census of the US population every 10 years to determine how many people there are, where they live, etc. Currently, 38 states have filed lawsuits related to redistricting as a result of the 2010 census. There many other lawsuits by other groups and for other reasons relating to the 2010 census.¹

Why are there errors in conducting a census of the US population?

- If we did try to do a census of the height and diameter of every tree on a 500-acre tract of land, what errors might result?

In summary, there are three factors to consider: time to conduct a census, cost to conduct a census, whether the results will contain errors. In the case where errors may result, the problem is that we have no way of knowing how much error there is!

For instance, in the case of the US census, we know there are errors, but what where are the errors (which district, town, county, *etc*) and how much error is there (10 people, 20,000 people, *etc.*)?

Finally, in the case where a census is time or cost prohibitive, or where resulting errors are not quantifiable we should use a *sample* to obtain information about a population.

8. **Sample** – When it is not feasible to conduct a census, we use a *sample* which is a subset of the population that is (usually) chosen so that it is *representative* of the population in some way.

- **Question:** How do VSU students rate the extra-curricular opportunities offered?
Population: All full-time VSU students enrolled in the Spring 2014 semester.
Sample: Send the survey to all students. The ones that respond are the sample.

¹ http://ballotpedia.org/wiki/index.php/Redistricting_lawsuits_relating_to_the_2010_Census

- **Question:** What is the public’s sentiment about a proposed bill before congress
Population: All registered voters
Sample: Select registered voters at random until 2000 people have answered the survey. (Almost all national opinion polls survey 1000-2000 people chosen randomly.)
- **Question:** Do manufactured products conform to required standards
Population: all products coming off an assembly line
Sample: Every 50th product is pulled for testing to see if it conforms to standards.

9. **Representative Sample** – When we use a sample, we are not using every item in the population, so we want the sample to be *representative* of the population so that we can have confidence in conclusions we draw about the population.

- A sample of the first 20 products that came off an assembly line during a day’s production would probably not be representative; there may be a warm-up affect occurring as the machines get up to speed.
- A survey 100 residents of mid-town Atlanta to learn about purchasing habits would not be representative of the entire population of Atlanta.
- A survey of students in evening classes might not be representative of the entire student population.

A *biased* sample doesn’t represent the population as a whole and thus leads to biased conclusions about a population. Later, we will consider this further.

Note: there is a LOT more to choosing samples that we will not consider.

10. The population should be carefully and accurately defined.

- We don’t want to define the population as, “all 18-24 year-olds in the US” if we can only obtain data from students at VSU.

11. When we use a sample, we know there is error in our results for the simple reason that we did not measure every element in the population. However, as we will learn, statistical inference allows us to quantify the error in a sample.

- A survey of 2000 voters revealed that 63% support a particular candidate with a margin of error of 3%.

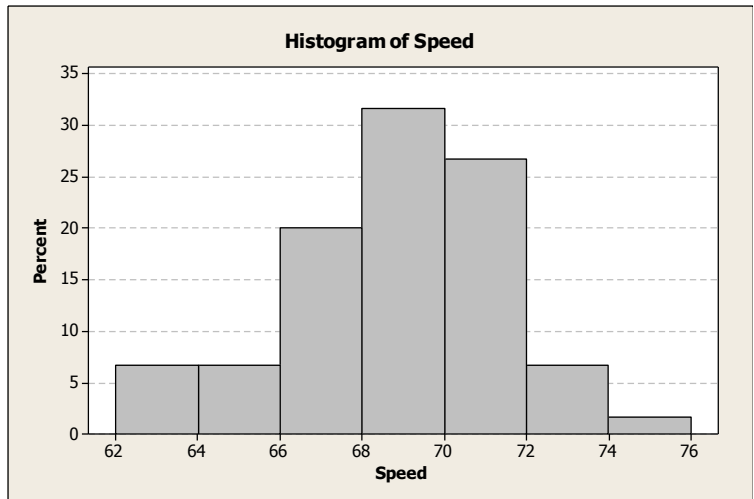
12. **Statistic** – A *statistic* is simply a number, graph, or table that summarizes data.

a. Example 1 – A sample of speeds (*mph*) of cars on I75, at exit 18:

68.2	70.7	65.3	70.8	71.6	69.9	68.0	68.0	67.0	72.4
72.0	64.4	68.5	66.9	71.2	66.2	66.5	70.7	70.1	69.2
72.5	69.2	74.9	66.2	70.3	63.8	69.5	66.1	66.0	68.5
71.4	66.4	63.0	67.0	71.1	70.5	68.0	67.9	69.1	71.5
65.9	69.2	71.1	70.8	68.1	69.2	63.3	69.1	70.9	72.5
63.9	70.9	68.8	71.3	68.6	69.9	69.1	65.1	69.5	67.5

What does this data tell us? These individual data values don't really tell us much. However, these *statistics* tell us more:

- The average speed is 68.8 *mph*.
- The *minimum* is 63.0 *mph*.
- The *maximum* is 74.9 *mph*.
- The *standard deviation* is 2.6 *mph*.
- The histogram on the right tells us about the distribution of speeds:
- The chart below also tells about the distribution of speeds:



Speed	Count	Percent
60-65	5	10%
65-70	34	60%
70-75	21	40%

- a. Example 2 – Consider a baseball player over the course of a season. The following values are all statistics:

Number of at-bats	384
Number of hits	129
Batting Average	$129/384 = 0.336$

- b. The *percentage* of people in class on a given day is a statistic. For example, there are 33 students present in a class of 36. Thus, $\frac{33}{36} * 100 = 91.7\%$ of students are present.

13. **Science of Statistics** – The *science of statistics* is the study of how we collect, analyze, interpret and present data.

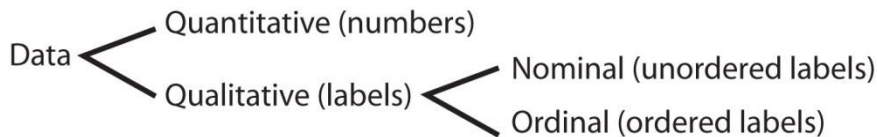
14. Types of Populations

- a. A *concrete* population is one where we can (in theory) directly identify (measure) every item in the population. Examples:
- Pine trees in Georgia. Variables: height, volume, age, *etc.*
 - People 18-24 years old that reside in the United States (about 27 million). Variables: height, weight, SAT, salary, *etc.*
 - 4-year non-profit colleges and universities in the US (about 2170). Variables: tuition, enrollment, *etc.*
 - Cell phones in the world (about 6 billion), land-lines in the world (about 1.2 billion). Variables: length of phone calls.
 - Ponds in Lowndes County. Variables: volume, surface area, *etc.*

- b. A *conceptual* population is one where the data (items) are part of an ongoing process.
- Customers arrive at the drive-through, hour after hour, day after day. Variables: length of time to place an order, length of time to receive order.
 - The bottles of soda that come off a filling line. Variable: amount of liquid in bottle.
 - People that get their hair cut at a salon. Variable: length of time to cut hair.
 - The amount of CFC's in the atmosphere.

1.2 – Types of Data

1. These are the types of data we will consider. (There is a fourth type, ratio, that is related to quantitative which we will not consider).



2. **Quantitative data** – Quantitative data are measures of something. The data values are numeric. The arithmetic difference between two data values measures how much more or less there is when the two values are compared. Examples of quantitative variables:

- height of a pine tree
- weight of a tomato
- volume of soda in a can
- length of a phone call
- speed of a computer processor
- distance between two molecules
- time that it takes to a drug to absorb into the bloodstream.

3. **Qualitative Data** – Some data are not actually measured, they are simply observed and we usually give labels to these observations. For example if I am observing the sex of people who shop at a store, I might record the data in any of these forms which all convey the same meaning:

- {Male, Female, Female, Male, Male, Male, Female}
- {M, F, F, M, M, M, F}
- {Yes, No, No, Yes, Yes, Yes, No}

Thus, with qualitative data, the data values themselves are simply *labels*. We can *count* qualitative data values, but we can't do arithmetic operations on the data values themselves. Qualitative data is sometimes called *categorical data*. Qualitative data can be broken down into two types:

- a. **Nominal Data** – The data values cannot be ranked (ordered). Examples:

- sex – { Male, Female }
- occupation – { Engineer, Manufacturing, Service, Other }
- color of car – { Red, Blue, Tan, Black, White, Other }
- presence of a defect in a product – { Present, Not Present }
- wearing sun glasses – { Wearing, Not Wearing }

b. **Ordinal Data** – The data values can be ranked. Examples:

- grades for a course {A, B, C, D, F}
- student rank in class { “first”, “second”, “third” }
- SOI (Student Opinion of Instruction), e.g. course ratings – each question is answered with a response between 1 and 5 corresponding to: 5 (strongly agree), 4 (agree), 3 (neutral), 2 (disagree), and 1 (strongly disagree).
- Survey card in a restaurant, “How was your service today? poor, fair, ok, good, great.”

4. **Example** – Consider the data set shown on the right.

Student	Sex	Final Avg.	Grade	Rank
Anna	F	83.7	B	4
Jean	F	97.9	A	1
Walt	M	73.3	C	6
Cindy	F	86.2	B	3
Sam	M	97.3	A	2
Dave	M	79.6	B	5

- Sex is a nominal variable (qualitative),
- Final Avg is quantitative
- Grade and Rank are both ordinal variables (qualitative). You may be tempted to say that Rank is quantitative because you can calculate differences between the values (as required for quantitative data); however, the differences don’t have a consistent meaning. For example, Anna and Sam have a difference in rank of 1 as do Sam and Cindy. However, their respective differences in Final Averages is not consistent: 0.6 (97.9-97.3) for Anna and Sam while it is 11.1 (97.3-86.2) for Sam and Cindy. Thus, this variable is ordinal, not quantitative.

5. **Example** – Consider a restaurant survey as shown on the right. You are asked a to rate variables on a scale from 0 to 5. Above each number, we usually see descriptions like, “Not at all important”, ..., “Very Important”.

Using a scale of 0= Not at all important to 5=Very Important, please rate the following aspects of our service in the restaurant?

	Not at all important 0	1	2	3	4	Very Important 5	No Opinion
Speed of Service	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Friendliness of Staff	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Helpfulness of Staff	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Value for Money	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Thus, the data are labels and can be ranked so the data is ordinal.

6. **Scales of Measurement** – The amount of information contained in data can be classified in terms of its strength. The stronger data is, the more information that is available, and the more potential to learn from it. Quantitative data has the most information, followed by ordinal, and then nominal. In other words, you can gain more insight into a subject when you have quantitative data as compared to qualitative. Similarly, ordinal data has more information than nominal data.

- For example, suppose a small part in a toy is required to have a width of 10 mm with a tolerance of 0.5 mm. We learn more about the process that produces the parts by measuring the width of a sample of parts as opposed to simply classifying each part in the sample as defective or not defective.

1.3 – Sources of Data

Where do we get data?

1. **Historical** – This is sometimes referred to as *past data*. It is data that already exists in databases on computers, or sheets of paper in a filing cabinet, *etc.* On the internet you can find historical data on almost any subject you can imagine.
2. **Observational** – We *observe* the subject that we are interested in to collect data in the present. In this situation we do not manipulate the subjects in any way, we simply *observe*. For instance, we may time how long a shopper spends in a store.
3. **Experimental** – This type of data results from designed experiments and computer simulation. This type of data promotes advances in science, medicine, and engineering. In designed experiments, we usually manipulate the subjects in some way. Examples:
 - In the study of a drug's effectiveness, we might gather a random sample of people that have a symptom(s) that the drug might cure. Half the subjects are given the drug and the other half is given a placebo. However, all subjects are lead to believe they are taking a drug that will cure their symptom. Often, doctors who monitor the subjects during the experiment don't know which patients have received the placebo.
 - If we are studying a plant fertilizer, we might alter the amount of fertilizer that we give to groups of plants in order to see what the optimal amount is to produce growth.
 - Computer simulation involves using a computer to simulate some physical phenomena. For instance, computer simulation is used to help decide how to arrange products in a warehouse, how many people are needed to pick orders, and how to batch orders to give to an order picker. It would be very expensive and time consuming to physically experiment with a real warehouse so computer simulation is used to evaluate different configurations of product arrangement, number of order pickers, and strategies for batching orders.

In this course, we will mostly consider historical and observational data. More advanced statistical techniques are used to evaluate experimental data.

1.4 – Basic Definitions 3

1. **Simple Random Sample (SRS)** – In statistics there are lots of ways to obtain a sample. The simplest type of sample, and the one we will use almost always in this class is called a *simple random sample (SRS)*. An SRS is chosen by making sure that the following three conditions are met:
 - a. Each item selected comes from the same population
 - b. Each item is selected independently of the others.
 - c. Each item in the population has an equal chance to be in the sample.

Let's consider these conditions in more detail:

a. *Each item selected comes from the same population*

This sounds simple, and often is. But things can get contaminated sometimes. Sometimes, a data value will be unusually large or small. Such a value can drastically influence the statistics we compute from the data. When we find such a value, we might decide upon careful analysis that it doesn't belong to the population and should be excluded from the sample.

- Suppose that you are studying how long it takes a person to check their bags and obtain their ticket upon arrival to the Delta ticket counter at the Jacksonville airport. Suppose that you time peoples' waits and almost all are between 3 and 7 minutes. However, one person doesn't speak English and a translator must be paged. This person's time ends up being 45 minutes. You might argue that this person doesn't belong to the population you are studying. If so, technically, we should modify the description of the population to say that we are only studying people speaking English.

You also want to be proactive to prevent this type of contamination.

- For instance if you were studying how much individuals spent on goods at Sams for their personal consumption, then you would want to have a sampling plan that made sure you did not take data from someone buying products for a business.

b. *Each item is selected independently of the others.*

What this means is that each data value does not influence the value of any other data value.

- Suppose you want to study how fast people walk. You decide to measure and mark a distance of 20 feet along a sidewalk. Then, you time random people as they walk this distance. However, you would need to be more specific in your definition of your population. Are you timing individual people or groups of people? If you were interested in individual people, then you would not time separately, two people who were walking together as their times would be dependent.
- Suppose you want to study how long people have to wait in line to checkout at a grocery store. You would not want to collect data from two people who are in the line one after the other because the second person's wait time *depends* on the first person's wait time. These two people are not independent. One way you could do this would be to randomly pick a line and time the next person who gets in that line. Once the person has finished, then you randomly pick another line and start the process over again.

However, if we were studying the *service time*, the time it takes the cashier to scan the groceries, bag them, and accept payment, then we would not run into this problem. Why?

c. *Each item in the population has an equal chance to be in the sample.*

Conceptually, this is very simple to do. Give every item in your population a unique number. Then put all the numbers on pieces of paper, put them in a hat, shake it up, and then choose say 50 numbers from the hat. Then, the items corresponding to the 50 chosen numbers comprise the sample. Such a sampling plan ensures that each item in the population has an equal chance to be in the sample. In practice, this is harder to do.

- Suppose we want to survey what adults in America think about a bill pending before Congress. Suppose we decide to use a telephone survey. Not all adults have a telephone. But telephone surveys are done and are often very accurate. How are phone numbers picked? Telephone surveys typically utilize computer programs to create random phone numbers and dial them. This ensures that people with unlisted phone numbers will be sampled as well as people with listed numbers. However, such a sample doesn't perfectly represent the population.
 - Suppose you had a cage with 100 rats and you want to choose a sample of 10. What is wrong with using the first 10 rats you can catch?
2. **Outlier** - An outlier unusually large or small data value. We will learn several techniques to detect outliers. When we detect an outlier, we must ask if it belongs to the population. If the answer is "yes", we usually keep the outlier. If the outlier is determined to be from some other population, we discard it. Sometimes an outlier is a part of the population, but we have a strong reason to exclude it. In either case, we always document the removal of the outlier.
 - a. Many companies practice quality improvement programs where a group of employees will study a problem area in the company. One tenet of such programs is that to change something, the first step is to measure it. When outliers are detected the group will study them to find the cause(s) of the outlier. If the outlier is not desired, then the company can consider steps to eliminate (reduce) the conditions that caused the outlier.
 3. **Descriptive Statistics** – Descriptive statistics summarize a sample. We use tables, graphs, numbers, percentages, averages, etc. to summarize the information in a sample and all of these are descriptive statistics.
 4. **Statistical Inference** – Statistical inference refers to the process of using the information in a sample (descriptive statistics) to make a statement about a population or evaluate a claim about a population. This statement is usually accompanied by a margin of error. We will discuss what a margin of error is later (Ch. 6).
 5. **Casual Statistical Inference** – Inferences made on a population without a margin of error or accompanying basis for the statement.
 6. **Examples** – The following are examples of descriptive statistics, statistical inference, or casual inference. Notice that when we make a descriptive statement we refer to a specific sample, study, survey, etc. In contrast, an inferential statement refers either implicitly (casual inference) or explicitly (statistical inference) to a population. The point is to be aware of the language that is being used.
 - a. Salary Statistics:
 - **Descriptive:** A survey of 100 chemical engineers during January 2015 revealed that the average salary was \$102,046.
 - **Casual Inference:** The average salary of chemical engineers is \$102,046.
 - **Statistical Inference:** The average salary of salary of chemical engineers is \$102,046 with a margin of error of \$5,200.
 - b. Unemployment Statistics:

- **Descriptive:** A sample of 200 shopping at Wal-mart found that 22 (11%) were unemployed.
- **Casual Inference:** 11% of shoppers at Wal-mart are unemployed.
- **Statistical Inference:** 11% ($\pm 4.3\%$) of shoppers at Wal-mart are unemployed.

c. Arithmetic Skills Statistics:

- **Descriptive:** A nationwide study of 500 first graders found that the average amount of time it takes a student to do a simple addition problem is 3.4 seconds.
- **Casual Inference:** On average, first graders require 3.4 seconds to solve simple addition problems.
- **Statistical Inference:** First graders require 3-3.8 seconds on average to complete a simple arithmetic problem.

Homework 1.1

1. Define all the terms in this chapter: data, statistic, science of statistics, variable, population, census, sample, practice of statistics, quantitative data, qualitative data, nominal data, ordinal data, scales of measurement, sources of data, simple random sample, outlier, descriptive statistics, statistical inference, casual statistical inference, sampling plan
2. Write several statements that use descriptive statistics (make up examples that relate to your major).
3. Write several statements that use statistical inference.
4. When you read a sentence that contains statistics, how can you tell if it is descriptive statistics or statistical inference?

Chapter 2 – Descriptive Statistics

This chapter discusses how to compute and interpret descriptive statistics for qualitative and quantitative data. Comparing statistics using percentage differences is a useful when presenting statistical results. Percentage differences are reviewed in the Chapter 2 – Appendix.

2.1 – Overview of Descriptive Statistics

This is an outline of the topics we will consider in this chapter:

- A. Qualitative Data
 - 1. Frequency Distribution
 - 2. Relative Frequency Distribution
 - 3. Bar Chart
 - 4. Pie Chart
 - 5. Mode

- B. Quantitative Data
 - 1. Numerical Descriptive Statistics
 - a. Measures of Central Tendency – Average, Median
 - b. Measures of Spread – Range, Variance, Standard Deviation, Interquartile Range
 - c. Miscellaneous – Sample Size, Minimum, Maximum, Percentiles, Quartiles, 5-number Summary

 - 2. Graphical Descriptive Statistics
 - a. Dot Plot, Stem and Leaf Plot
 - b. Histogram
 - c. Ogive
 - d. Boxplot

 - 3. Other Topics:
 - a. Empirical Rule
 - b. z-scores

2.2 – Descriptive Statistics for Qualitative Data

This section considers descriptive statistics for qualitative data: frequency distribution, relative frequency distribution, bar chart, mode.

1. Definitions:

- a. **Sample Size** – The number of data values in the sample
- b. **Frequency Distribution** – Count the number of data values that fall into each *category*.
- c. **Relative Frequency Distribution** – Divide each frequency by the sample size. Many times we multiply the relative frequencies by 100 to obtain *percents*.
- d. **Bar Chart** – The *Bar Chart* is a graph composed of bars that correspond to categories. The height of each bar is either the frequency (count) or relative frequency.

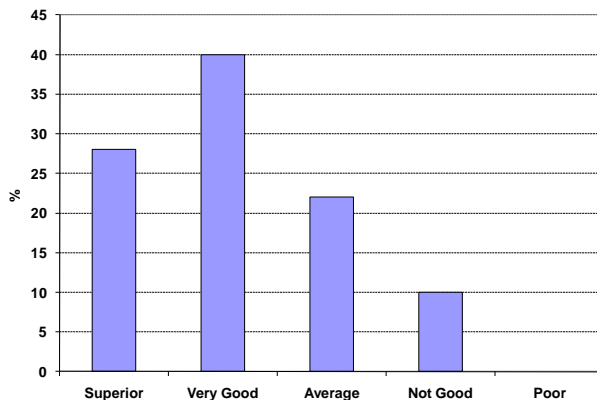
2. **Example** – Restaurant patrons are given a survey about the quality of their overall experience on that particular visit. One of the questions asks, "Which response most accurately reflects your overall experience in our restaurant tonight: Superior(S), Very Good(V), Average(A), Not Good(N), or Poor(P)." The data is shown below:

Class	Frequency	Relative Frequency
Superior	17	$17/60 = 0.283 = 28\%$
Very Good	24	$24/60 = 0.400 = 40\%$
Average	13	$13/60 = 0.217 = 22\%$
Not Good	6	$6/60 = 0.100 = 10\%$
Poor	0	$0/60 = 0.000 = 0\%$
	n = 60	sum = 1 = 100%

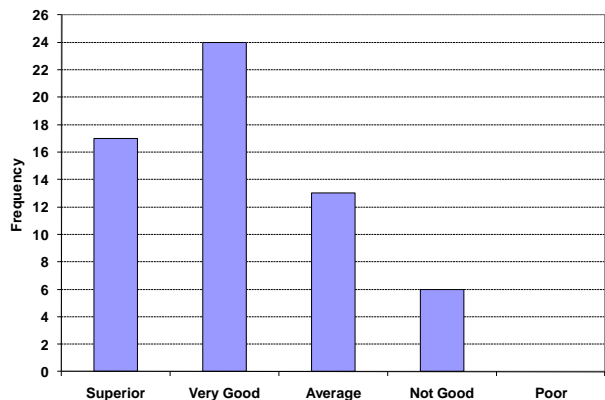
N	N	V	A	A	A	A	V	N	V	V	V	A	A	V	N	V	N	N	A	S	A	V	S	S	S	A	A	S	V
S	V	V	A	V	S	V	V	S	S	S	A	S	V	S	V	V	V	V	S	V	V	S	S	A	V	S	S	V	V

A summary of the data is shown in the table above on the right and in the charts below.

Relative Frequency (%) Bar Chart



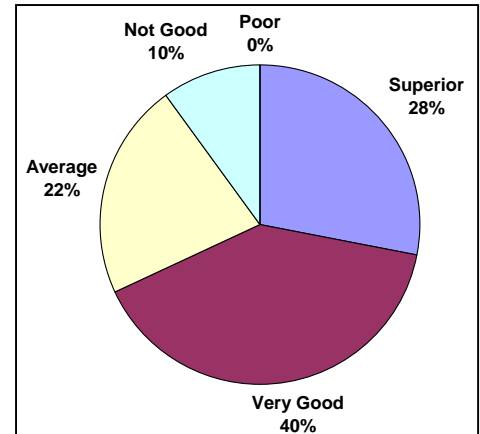
Frequency Bar Chart



Note that visually the two graphs are the same. Although you see frequency bar charts, and they can be useful, especially with a small data set, I recommend you use the *relative frequency bar chart*. The reason is that it is usually more informative. For instance, when we look at the frequency bar chart above (right), it indicates that 17 people had a *Superior* experience. But what exactly does this mean? Is it 17 out of 20 people or 17 out of 100, *etc.*? However, if we look at the relative frequency bar chart, it shows that about 28% of people had a *Superior* experience, or we could say that *almost 3 out of 10* people had a superior experience, or more than a quarter had a superior experience. This seems more informative than dealing with the raw numbers.

Note that we can compare categories by directly comparing the heights of the bars visually. For instance, we can see that people tend to rate their experience *Very Good* at about twice as often as they rate their experience *Average* because the *Very Good* bar is about twice the height of the *Average* bar.

- Pie Chart** – This is a very common type of representation of the relative frequencies. The Pie Chart shown at right represents the data from the example above. The angle of each wedge of the pie (category) is simply the relative frequency multiplied by 360 degrees.



The Pie Chart is recommend for use with six or fewer classes and when the values aren't too close in value. It is easier to interpret differences in heights of bars than determining differences in the angles represented in a pie chart. Thus, I recommend the bar chart.

- Mode** – The *mode* is the most frequently occurring data value(s). For the example above, the Mode is *Very Good*. The mode is not necessarily unique. For instance, suppose that a survey revealed that 15 people have black cars, 15 white, 10 blue, and 8 red. The mode is *Black* and *White*.

Homework 2.1

- A restaurant is analyzing the popularity of the different sandwiches it sells. A simple random sample of data is shown. (a) Find the relative frequency distribution. (b) Make a bar graph. (c) Find the Mode.

Vegetarian	Turkey	Ham	Ham	Ham	Turkey	Ham	Turkey	Turkey
Vegetarian	Turkey	Corned Beef	Ham	Vegetarian	Turkey	Turkey	Ham	Vegetarian
Vegetarian	Turkey	Vegetarian	Ham	Corned Beef	Vegetarian	Ham	Ham	Vegetarian
Turkey	Turkey	Vegetarian	Turkey	Vegetarian	Turkey	Corned Beef	Ham	Vegetarian
Ham	Vegetarian	Turkey	Turkey	Ham	Corned Beef	Vegetarian	Vegetarian	
Turkey	Turkey	Turkey	Turkey	Ham	Ham	Turkey	Turkey	
Ham	Vegetarian	Corned Beef	Turkey	Turkey	Vegetarian	Vegetarian	Vegetarian	

2.3 – Overview of Exploring Quantitative Data

Now, we will consider descriptive statistics for quantitative data. Before we begin with the techniques, let's discuss some things we look for when analyzing quantitative data:

- The *center* of the data. What is the average of the data? The median?
- The *spread* of the data. How spread out is the data? Or how tightly grouped is the data? Does the data range from 2.5 to 4.3 seconds? Or does it range from 2.5 seconds to 56.3 seconds?
- The *shape* of the distribution and whether the data is *symmetric* or *skewed*. How is the data distributed? Are there a lot of small values and a few large values? Or a few small values and a lot of large values? Or are there approximately an equal number of small and large values?

We'll learn more about these throughout this chapter. We will also see how we can use the various descriptive statistics to answer these questions.

2.4 – Measures of Central Tendency

1. **Sample Size** – We use the symbol n to denote the number of items in a sample and refer to this as the *sample size*.
2. **Data Labels** – As shown below, each data value is assigned a label. Many times the label is not shown and just understood to be there. It is convenient to introduce the idea of data labels so that we can understand the formulas that follow. In the example below, $x_1 = 4$ refers to the first data value. Similarly, $x_2 = 9$ refers to the second value.

Sample (Data)	4	9	3	6	5
Data Labels	x_1	x_2	x_3	x_4	x_5

3. **Sample Average** – We use the symbol, \bar{x} to denote the sample average. The *sample average* is also called the: *average, sample mean, mean, x-bar*. It is the numerical average of the data values obtained by adding up all data values and dividing by the sample size. The sample average is calculated with this formula: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Note: The capital Greek letter Σ (sigma) in mathematics means, “*add up the values to the right of it*”. For instance, $\sum_{i=1}^n x_i$ means “*add up all the x's from the first one (i=1) to the last one (n)*.”

Example (using data from table above): $\bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i = \frac{1}{5} (4 + 9 + 3 + 6 + 5) = \frac{1}{5} (27) = 5.4$

4. **Median, M** - is the *middle* data value which means that approximately 50% of the data have values *less* than the median and 50% have values *greater* than the median. Steps to find the median of a data set:

Steps to calculate median:

1. Sort data, smallest to largest.
2. Re-label sorted data.
3. Determine whether sample size (n) is *odd* or *even*.
 - a. If n is odd, then median is middle observation,
Calculate $i = \frac{n}{2}$, then *round i up to the next integer*. Finally, $M = x_i$
 - b. Else if n is even, then median is the average of two middle observations.
Calculate $i = \frac{n}{2}$, then. Finally, $M = \frac{x_i + x_{i+1}}{2}$

5. Examples

a. Find the median of the following data:

2	8	12	9	5	6	4
---	---	----	---	---	---	---

- Sort and
- Relabel data:

i	1	2	3	4	5	6	7
x_i	2	4	5	6	8	9	12

- Case *a*. Since the sample size is *odd*, $n = 7$, calculate $i = \frac{n}{2} = \frac{7}{2} = 3.5$, Then *round i up* so that $i = 4$. Finally, the median is $M = x_i = x_4 = 6$.

b. Find the median of the following data:

2.4	8.3	12.1	9.6	5.2	6.8	4.7	11.2	10.4	4.9
-----	-----	------	-----	-----	-----	-----	------	------	-----

- Sort and
- Relabel data:

i	1	2	3	4	5	6	7	8	9	10
x_i	2.4	4.7	4.9	5.2	6.8	8.3	9.6	10.4	11.2	12.1

- Case *b*. Since the sample size is *even*, $n = 10$, calculate $i = \frac{n}{2} = \frac{10}{2} = 5$. Thus, the median is the

$$\text{average of the 5}^{\text{th}} \text{ and 6}^{\text{th}} \text{ data values: } M = \frac{x_i + x_{i+1}}{2} = \frac{x_5 + x_6}{2} = \frac{6.8 + 8.3}{2} = 7.55.$$

c. Mean and Median – A 5-Word Word Search Problem. Men and Women VSU students were timed as described in Chapter 1. The statistics shown below were generated by Minitab, the statistical software you will use for your projects.

As shown in the table below, women completed the puzzle on average about 28 seconds quicker than the men (94 seconds for women and 122 seconds for men). This shows that the women were about 23% faster on average than the men. However, the median times for women and men differ by only 9 seconds (82 seconds and 91 seconds, respectively), revealing only a 9% difference.

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Women	50	93.49	82.09	90.38	60.35	8.54
Men	50	121.6	90.6	117.3	76.7	10.9

Variable	Minimum	Maximum	Q1	Q3
Women	20.23	221.39	40.13	133.25
Men	25.6	300.9	55.4	187.1

6. **Outliers** – An *outlier* is defined to be an extremely large or small data value. The median is said to be *resistant to outliers*. What this means is that the median is not as influenced by an outlier(s) compared to the sample mean. We also say that the sample mean is *not resistant to outliers*. Later, we will see two methods to detect possible outliers: boxplots and z-scores.

7. Examples

- a. Consider the random sample of annual salaries in Valdosta shown below.

44,705 35,301 30,788 38,994 24,724 40,020 45,452 37,715 41,963 225,000

As shown in the table below, the sample average is $\bar{x} = \$56,466$ while the median is $M = \$39,507$. Which value more accurately measures the *center* of this data set? It hardly seems the average is a good measure of the center when we see that all the data is less than the average except the one salary of \$225,000. Since the median represents the middle of the data it is a better measure of the center in this case.

The preceding example was rather extreme. We should be leery of small data sets with possible outliers. Next, we remove the outlier and re-compute the mean and median as shown in the table below. We see that if we remove the outlier, the median changes very little.

Mean

With Outlier	\$56,466
Outlier Removed	\$37,740
Difference	\$18,726

Median

With Outlier	\$39,507
Outlier Removed	\$38,944
Difference	\$563

When salaries (and house prices) are reported in the media, the median is almost always stated, not the average. Why is this so?

- b. 5-Word Word Search Problem. The largest data value for the men is 300.9 seconds and more than a minute longer than any of the women. Removing this data value results in the table shown below. Again, we see that the Median is less influenced by outliers. However, we stress that we should have a valid reason to exclude any data value from analysis.

Descriptive Statistics

Variable	N	Mean	Median
Men-Orig	50	121.6	90.6
Men-Outlier Removed	49	117.9	89.9

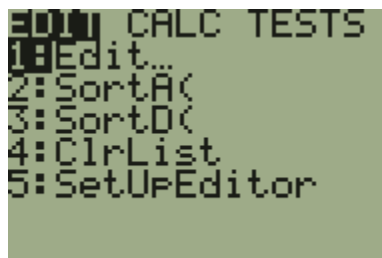
Difference		3.7	0.7

2.5 – Numerical Descriptive Statistics on TI-83 Calculator

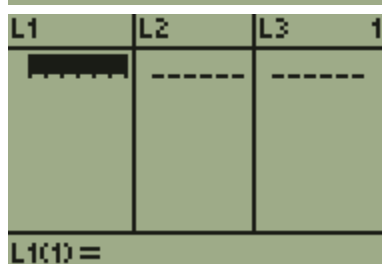
This section provides instructions on how to use the TI-83/84 calculator to obtain various numerical descriptive statistics. The calculator uses *Lists* to store data. Lists have names like "L1", "L2", etc and are arranged in columns.

1. Display the List window:

a. Press the *Stat* key

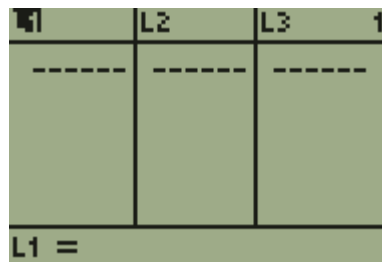


b. Chose *Edit* by pressing the *1* key



c. If there is data in L1, clear the list:

- Press the *Up Arrow* key which highlights L1.



- Press the Clear key and Enter

2. Enter data:

Type a number and press *Enter*, repeat. Enter these values:

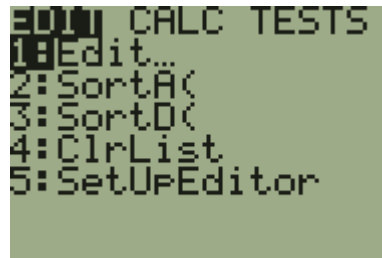
7, 39, 13, 9, 25, 8, 22, 0, 2, 18, 2, 30, 7,
35, 12, 15, 8, 6, 5, 29, 0, 11, 39, 16, 15.



L1	L2	L3	1
29			
0			
11			
39			
16			
15			
L1(26) =			

3. You can sort the data if you want to.

a. Press the *Stat* key:



STAT TESTS
1: Edit...
2: SortA(
3: SortD(
4: ClrList
5: SetUpEditor

b. Choose *SortA* by pressing the 2 key.



SortA(

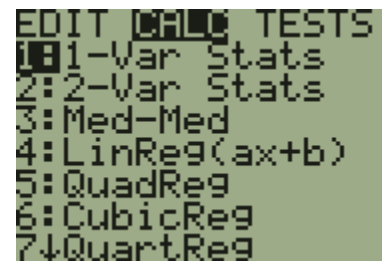
c. Supply the list you want to sort. To sort L1, press the 2nd key and then the 1 key and the press *Enter*.

d. Press the *Stat* key and then the 1 key to see the data sorted.

4. Display descriptive statistics:

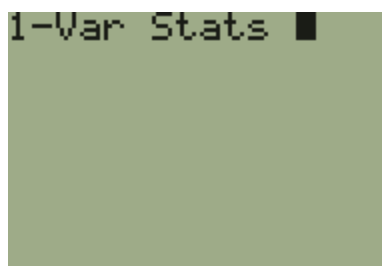
a. Press the *Stat* key

b. Use the right-arrow key to move over *Calc*.

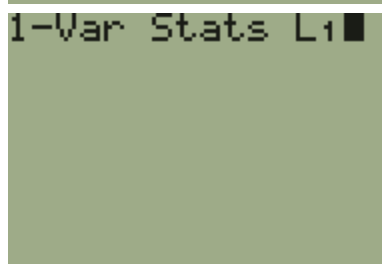


EDIT CALC TESTS
1: 1-Var Stats
2: 2-Var Stats
3: Med-Med
4: LinReg(ax+b)
5: QuadReg
6: CubicReg
7: QuartReg

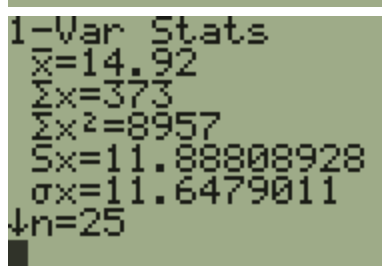
c. Press the 1 key for 1-Var Stats



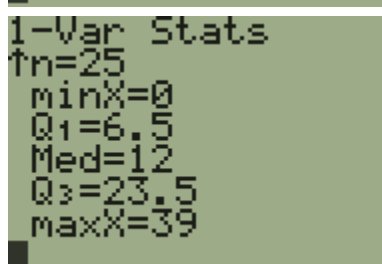
d. Supply the list you want to use. To use L1, press the 2nd key and then the 1 key



e. Press Enter



f. Use the down-arrow key several times to see more of the display.



Your window will show these descriptive statistics. We will learn about the ones in red.

Symbol	Value	Comment
\bar{x}	14.92	Sample Average
$\sum x$	373	Not considered
$\sum x^2$	8957	Not considered
s_x	11.888	Sample standard deviation
σ_x	11.648	Not considered
n	25	Sample size
$\min x$	0	Minimum
Q_1	6.5	First Quartile
Med	12	Median
Q_3	23.5	Third Quartile
$\max x$	39	Maximum

Homework 2.2

- The data below shows the number of hours per week that a sample of men and women studied while taking a full load of classes at a university. The times are recorded in hours and include the time spent in the classroom as well as studying outside of class. (Use Calculator) (a) Compare the averages for men and women. (b) Compare the medians for men and women.

Men	43.2	28.7	44.9	29.3	31.7	29.8	54.3	30.1	30.9	31
	29.7	31.1	32.1	51.7	32.2	34.3	34.4	34.9	35.3	30.4
	36.2	36.4	56.6	37.4	39.2	60	41.4	41.9	44.3	41.2
	35.8	41.5	47.9	28.2	49.3	29.5	55.2	38.5	58.1	47.5
Women	48.1	48.7	47.5	33.8	40.3	46.1	54.2	55.2	41.5	46.7
	61.2	45.4	50.4	43.1	51.2	34	54.1	47.9	58.3	41.3
	53.8	54.2	44.6	48.3	33.9	43	44.4	57.2	27.5	45.1
	43.6	48	40	36.6	47.9	43.8	34.8	30.5	44.2	46.3

2.6 – Measures of Location

- p^{th} Percentile** – The p^{th} percentile is the number such that approximately p percent of the data have values *smaller* than it and approximately $(100 - p)\%$ of the data have values *larger* than it. Thus, if the 90th percentile on a test is 86, this means that approximately 90% of people scored less than 86 and 10% scored more than 86. The percentile is not usually one of the actual data values.
- Steps to Calculate a Percentile** – Suppose you have n data values and you want to calculate the p^{th} percentile.
 - Sort data, smallest to largest.
 - Label sorted data: x_1, x_2, \dots, x_n
 - Calculate the *index*, $i = p \cdot n$, where p is the percentile you want to calculate (expressed in decimal form).
 - Determine whether the *index*, i is an integer or a decimal number.
 - If i is an integer then the p^{th} percentile is the average of the data values in positions i and $i + 1$:
$$p^{th} \text{ percentile} = \frac{x_i + x_{i+1}}{2}$$
 - If i is a decimal then *round i up* to the next integer value and then the p^{th} percentile is the i^{th} data value: $p^{th} \text{ percentile} = x_i$

3. **Percentile Examples** – Consider the data shown in the table below (Original Data) which represents salaries (in \$1000's) for recent college graduates in the computer science field. Find the following percentiles: (a) 10th (b) 25th (c) 75th (d) 80th (e) 87th. First, sort and relabel the data.

Data		Sorted	
<i>i</i>	x_i	<i>i</i>	x_i
1	34.4	1	24.7
2	31.2	2	28.4
3	32.1	3	29.6
4	31.3	4	30.2
5	35.6	5	30.5
6	38.7	6	31.1
7	38.0	7	31.2
8	40.3	8	31.3
9	35.3	9	32.1
10	41.4	10	32.8
11	35.6	11	33.2
12	32.8	12	33.2
13	24.7	13	33.4
14	44.1	14	34.4
15	29.	15	35.1
16	38.5	16	35.3
17	33.2	17	35.5
18	28.4	18	35.6
19	33.4	19	35.6
20	42.9	20	37.4
21	33.2	21	38.0
22	30.5	22	38.1
23	37.4	23	38.5
24	35.5	24	38.7
25	31.1	25	40.3
26	35.1	26	41.4
27	42.1	27	42.1
28	30.2	28	42.1
29	38.1	29	42.9
30	42.1	30	44.1

a. 10th percentile: $i = p * n = 0.1(30) = 3$

Since i is an integer, the 10th percentile is:

$$\frac{x_i + x_{i+1}}{2} = \frac{x_3 + x_4}{2} = \frac{29.6 + 30.2}{2} = 29.9$$

b. 25th percentile: $i = p * n = 0.25(30) = 7.5$

Since i is not an integer, we *round i up* so that $i = 8$. Finally, the 25th percentile is: $x_i = x_8 = 31.3$

c. 75th percentile: $i = p * n = 0.75(30) = 22.5$

Since i is not an integer, we *round i up* so that $i = 23$. Finally, the 75th percentile is: $x_i = x_{23} = 38.5$

d. 80th percentile: $i = p * n = 0.8(30) = 24$

Since i is an integer, the 80th percentile is:

$$\frac{x_i + x_{i+1}}{2} = \frac{x_{24} + x_{25}}{2} = \frac{38.7 + 40.3}{2} = 39.5$$

e. 87th percentile: $i = p * n = 0.87(30) = 26.1$

Since i is not an integer, we *round i up* so that $i = 27$. Finally, the 87th percentile is: $x_i = x_{27} = 42.1$

4. **Quartiles** – Several percentiles have special names:

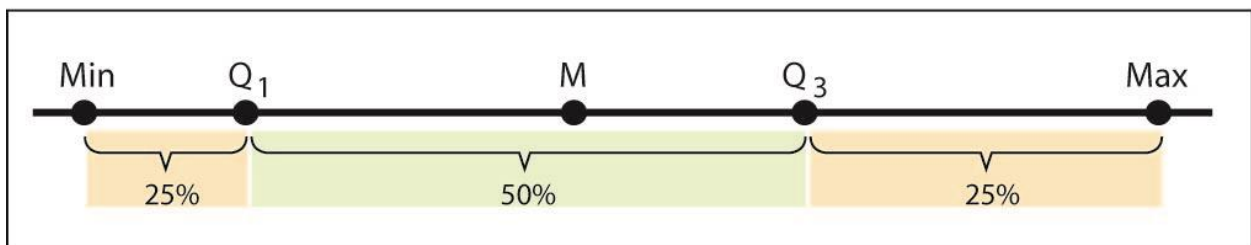
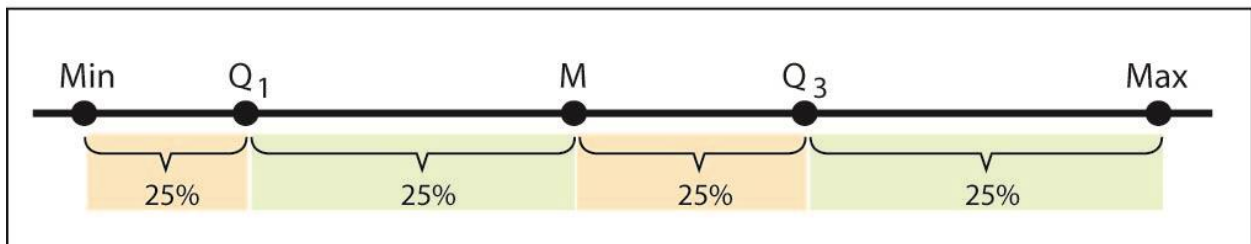
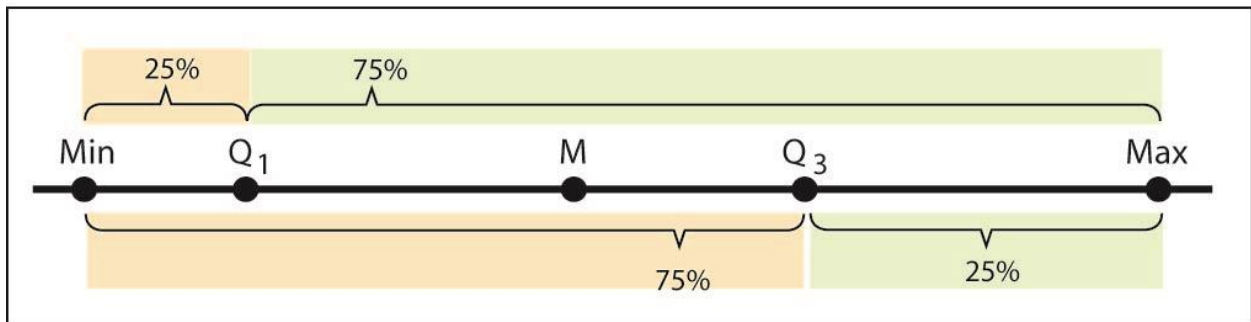
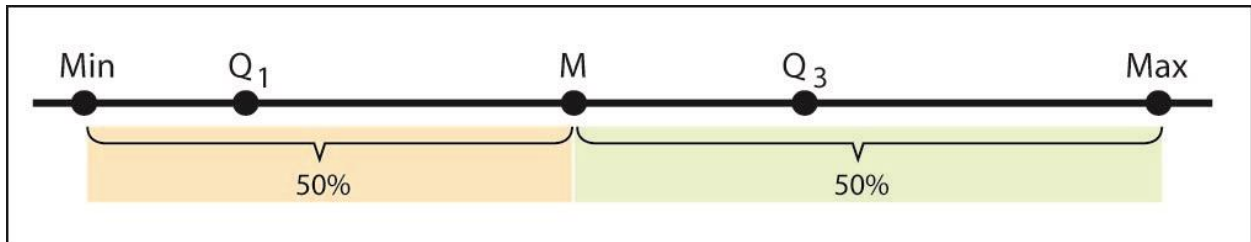
- A. First Quartile is $Q_1 = 25^{th}$ percentile
- B. Second Quartile is $Q_2 = 50^{th}$ percentile (the Median)
- C. Third Quartile is $Q_3 = 75^{th}$ percentile

Example – The first quartile for the data above is $Q_1 = 31.3$ and the third quartile is $Q_3 = 38.5$

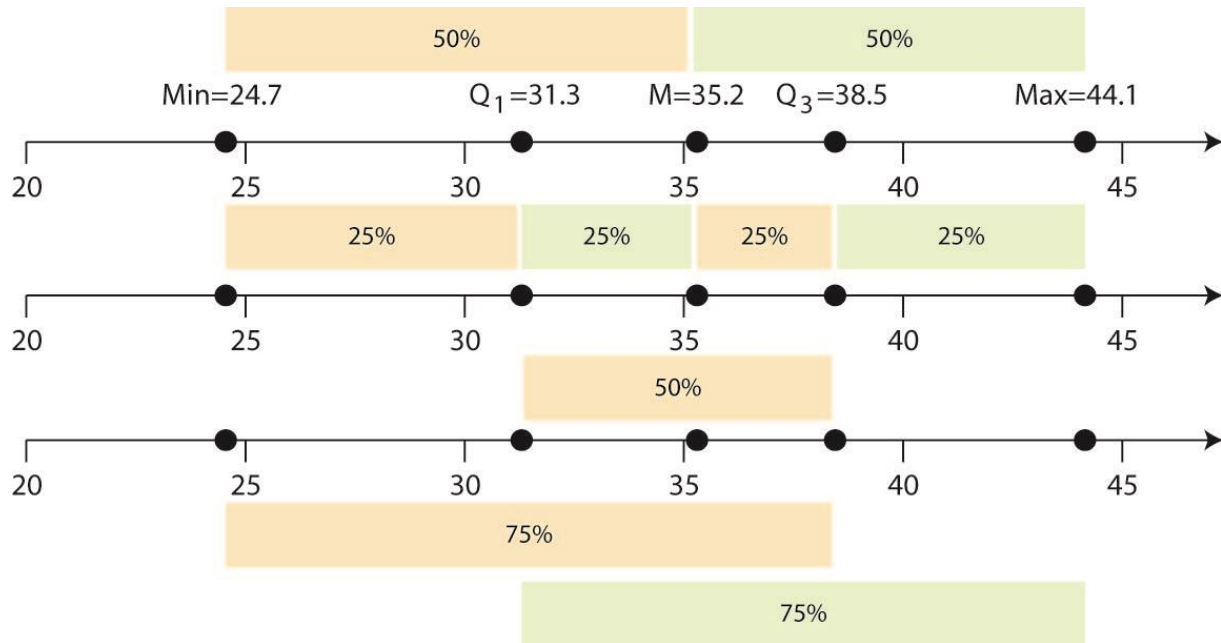
5. **5-Number Summary** – The 5-Number Summary for a data set is these five numbers:

Minimum, Q_1 , M , Q_3 , Maximum

From the 5-number summary, we can see the distribution of the data in a number of ways. For instance, the third figure shows the data broken into 4 regions where each region contains the same amount of data (25%). Note that the width of these regions is unequal in general. This is useful to show us where the data may be clumped up, spread out, or evenly distributed. The figure below summarizes the 5-Number Summary.



6. **Example**, continued from data above:

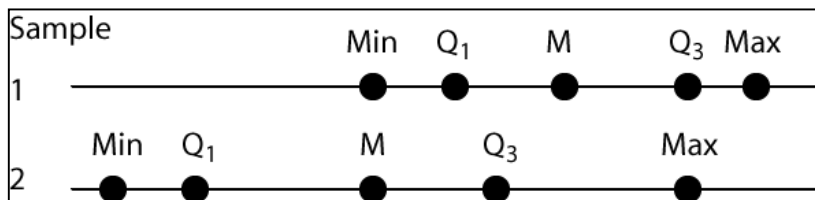


Some things we can observe from this 5-number summary above:

- about 50% of salaries fall between \$31,300 and \$38,500,
- 25% of salaries are less than \$31,300,
- 25% of salaries are greater than \$38,500,
- Half the salaries are above \$35,200.
- The third quartile is more tightly grouped than fourth quartile.
- The center most 50% of the data is more tightly grouped than outer values.

Homework 2.3

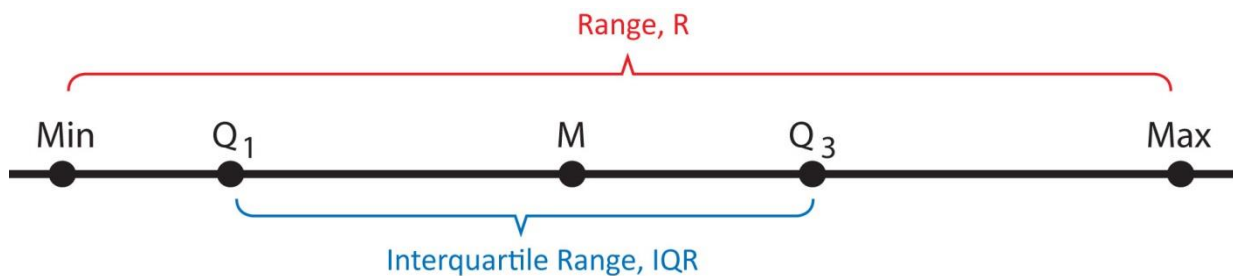
- Consider the data from Homework 2.2. (a) Make a chart of the following values for men and women: first quartile, third quartile, 63rd percentile, and 80th percentile. (b) Study the chart and write some sentences that summarize what you observe. Can you make some comparisons between the men's and women's values?
- Compare the minimum and maximums for the data from Homework 2.3.
- Consider the 5-number summaries shown below for two samples of data. Make some statements that compare these two samples.



2.7 – Measures of Spread (Variation)

When we collect data, say the time it takes to get an order at a drive through, all the values are different. Each order takes a different amount of time to fill. So, a natural question is, how much do the values vary? For instance, the high temperatures in July in Valdosta may vary from 89-100 degrees whereas the high temperatures in January may vary from 28-55 degrees. Which month has more variability in temperatures. In this section we consider several numerical measures of how much *variability* there is in the data. By doing this, we can determine how spread out the data is or how tightly grouped it is.

1. **Range** – The *range* is the *difference* between the largest and smallest data values, $R = \text{Maximum} - \text{Minimum}$
2. **Interquartile Range** – This is the range of the centermost 50% of the data, $IQR = Q_3 - Q_1$. This value is more resistant to outliers than the range.



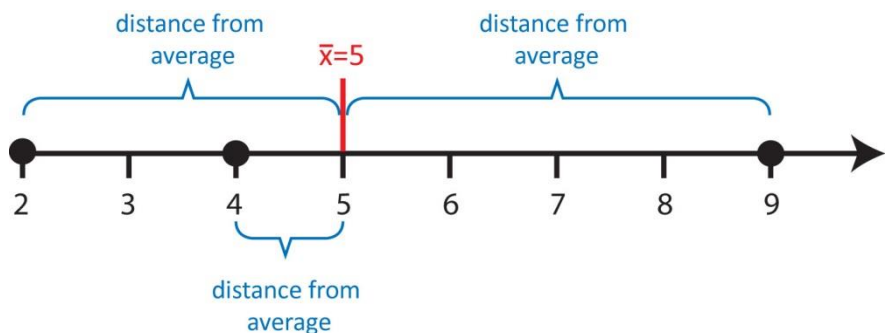
3. **Sample Variance and Standard Deviation** – We use the sample variance and sample standard deviation as measures of the variability in data.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{Sample Variance}$$

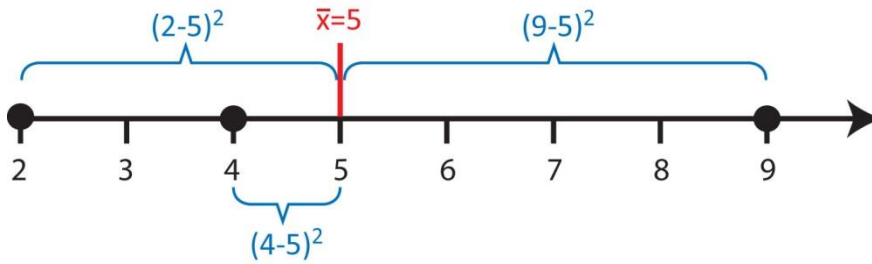
$$s = \sqrt{s^2} \quad \text{Sample Standard Deviation}$$

Let's take a closer look at the formula for variance. Suppose we have the dataset: 2, 4, 9. What is the variance and standard deviation? The sample average of this data is 5. The formula for the standard deviation says:

- a. See how far each data value is from the average: $x_i - \bar{x}$



- b. Square each of the distances: $(x_i - \bar{x})^2$



- c. Add up the squared distances: $\sum_{i=1}^n (x_i - \bar{x})^2 = 3^2 + 1^2 + 4^2 = 26 \text{ sec}^2$

- d. "Average" the squared distances to obtain the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{3-1} 26 = 13 \text{ sec}^2$$

Thus, in some sense, sample variance is the average squared distance of data values from their average. We can see from the formula that the larger the sample variance (or standard deviation), the more spread out the data is.

- e. Take square-root to obtain sample standard deviation: $s = \sqrt{13 \text{ sec}^2} = 3.6 \text{ sec}$

The *units* of standard deviation are units of the data, for instance, *sec* in the example above. In statistics, we almost always use the sample standard deviation as opposed to the sample variance.

Thus, in some sense, sample standard deviation is the average distance of data values from their average.

Previously, we showed the numerical descriptive statistics display on the TI-83 calculator. There, the symbol, s_x is used to represent the standard deviation. Most statistical software and your TI-83/84 calculator only report the standard deviation. The sample variance, of course can be calculated by squaring the standard deviation.

4. **Examples** – A study was done to see how long people spend in Applebees and Ryans. The Range, IQR, and standard deviation are shown for each data set in the table below (all data is in minutes):

	Range	IQR	Standard Deviation
Ryans	129	25.8	24.5
Applebees	42	10.5	10.3

Thus, we see that the variability at Ryans is more than twice the value at Applebees as shown by any of these measures of variability. Can you think of reasons why there might be more variability at Ryans compared to Applebees? (Ryans is a buffet restaurant and Applebees is a restaurant where you order from your table). Can you think of why this information might be useful to someone opening a restaurant?

Homework 2.4

1. Consider the data from Homework 2.2. (a) Make a chart of the following values for men and women: Range, standard deviation, and inter-quartile range. (b) Study the chart and write some sentences that compare the spread for men and women.

2.8 – Making a Histogram

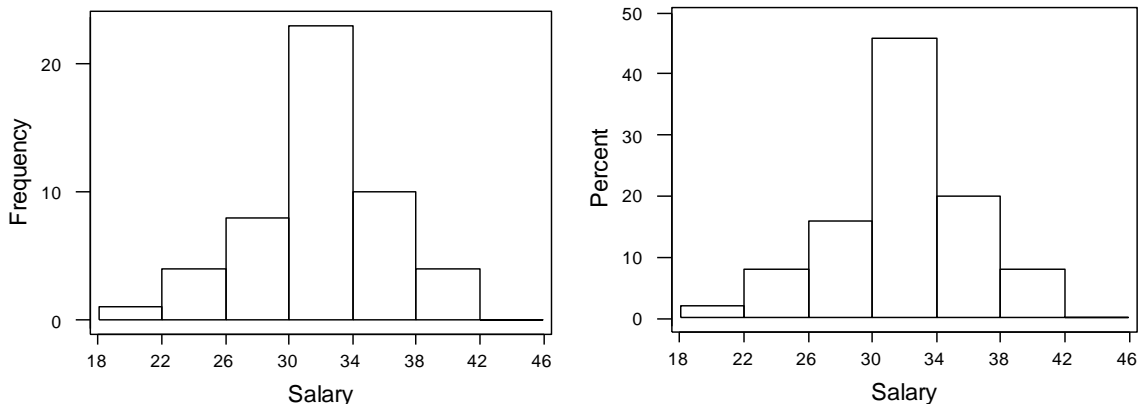
1. **Histogram** – A *histogram* is a graphical representation of the data that is similar to the bar chart we considered with qualitative data: it shows the distribution of data into different groups. The general idea is to:
 1. Divide the data into groups (classes),
 2. Count the data values in each group (class),
 3. Graph the counts (or relative frequencies) using bars.
2. **Example** – Consider the salary data shown below which is a sample of size $n = 50$, showing the starting salary (in thousands of dollars) of recent Computer Science graduates:

19.5	26.8	29.1	30.5	31.5	32.2	33.2	33.9	35.4	37.4
24.2	27.4	29.1	30.9	31.9	32.4	33.3	34.8	36.1	39.2
24.8	27.8	29.6	30.9	32.1	32.5	33.3	34.9	36.1	39.7
25.3	27.8	30.1	31.1	32.1	32.8	33.4	35.0	36.3	40.2
25.7	28.6	30.3	31.4	32.1	33.1	33.4	35.3	36.9	41.0

If we decide to define the classes as shown in the table below, we see that there is one salary in the range $\$18,000 < salary < \$22,000$ and 4 salaries in the range, $\$22,000 \leq salary < \$26,000$, etc. We can also calculate the relative frequencies by dividing each frequency (count) by the sample size.

	Frequency Distribution	Relative Frequency Distribution
Class	Frequency (count)	Relative Frequency
(18,22)	1	$1/50 = 0.02 = 2\%$
[22, 26)	4	$4/50 = 0.08 = 8\%$
[26, 30)	8	$8/50 = 0.16 = 16\%$
[30, 34)	23	$23/50 = 0.46 = 46\%$
[34, 38)	10	$10/50 = 0.20 = 20\%$
[38, 42)	4	$4/50 = 0.08 = 8\%$
	n = 50	sum = 100%

The graph on the left is the *frequency histogram* where the height of each bar is the count of the number of data values that fall in that class. The graph on the right is a *relative frequency histogram* and shows the percentage (or fraction) of data that falls in each class.



We notice that no matter whether we use the frequencies or the relative frequencies to draw the histogram, the *shape* is the same. In this case, we notice that the histogram is symmetric and bell-shaped (mound-shaped). *Symmetric* means that we can draw a line down the middle of the graph and the two halves are *approximately* mirror images of one another. When we say that day has a *bell-shaped* distribution, this implies that it is symmetric but it also conveys to us that relatively, most of the data is clustered around the center, but there are a few relatively small and large values.

3. Guidelines for making a histogram:

- a. Choose number of bars (classes, groups, intervals). Choose between 5 and 20. A histogram is sensitive to the number of classes, so you may want to try several values in practice. A rule-of-thumb is to use about \sqrt{n} classes for a histogram.
 - A sample size of 30 means $\sqrt{30} = 5.477$, thus we might use 5 or 6 classes in the histogram.
 - A sample size of 250 means $\sqrt{250} = 15.81$ suggesting we might want anywhere from 14, 15, or 16 classes.
- b. Determine class width, beginning point of histogram and ending point. Keep in mind that:
 - i. Classes have no overlap.
 - ii. The *lower* limit of the *smallest* class is usually *less* than the smallest data value. The upper limit of the *largest* class is usually *greater* than the largest data value. However, if the data cannot be negative, it doesn't make sense to choose a starting point less than 0.

There is no single correct value for class width. A good place to start is to use this formula:

$$\text{approximate class width} = \frac{\text{Maximum} - \text{Minimum}}{\text{number of bars}} = \frac{R}{\text{number of bars}}$$

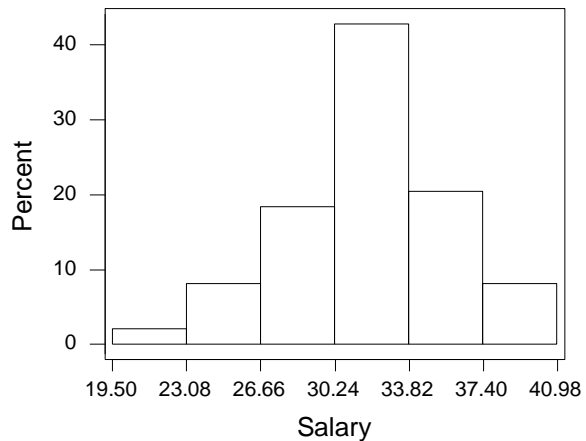
- For instance, considering the salary data above, there are 30 data values so we will begin with 6 bars. Thus, *approximate class width* = $\frac{41-19.5}{6} = 3.58$ so that each bar represents \$3,580. However, if we adhere to rule *b.ii* above, we need to choose a value a little larger than 41 for the ending point and a little smaller than 19.5 as the beginning point. If we choose 42 and 18, respectively, we see that the class width is $\frac{42-18}{6} = 4$. Thus, the first class will go from 18 to 18+4, *i.e.* (18, 22). The second class will go from 22 to 22+4, *i.e.* [22, 26) as shown in the table above.

- Construct the frequency and relative frequency distributions.
- Graph the relative frequency distribution.

4. Choosing Class Limits

Note that your calculator and the software we will use on a computer will *automatically* compute a histogram for us and thus, automatically compute class widths and starting and ending points. This is a good place to begin, but then we usually alter these values a bit to make the class widths (and/or starting and ending points) more understandable.

For instance, in the example above, software may compute the class widths to be 3.58 and result in a histogram as shown at right.



Although the shape of this histogram is similar to the one above with a class width of 4, the interpretation of the lower and upper limits for each class is more cumbersome. For instance, the second bar in this histogram indicates that about 9% of people had salaries between \$23,080 and \$26,660. Compared to the earlier histogram (easier to interpret) which showed that about 5% of people had salaries between \$22,000 and \$26,000. Decimals of course cannot always be avoided when choosing a class width; however, a judicious choice should be made so that class widths are easier to interpret.

Homework 2.5

- Use the data from Homework 2.2, to (by hand) (a) build a frequency distribution for each data set, (b) construct histograms for both datasets. Label both axes appropriately. Use any number of classes that is reasonable.

2.9 – Making a Histogram on TI-83 Calculator

This section provides instructions on how to use the TI-83/84 calculator to make a histogram

1. Enter this data in L1 (or some other list):

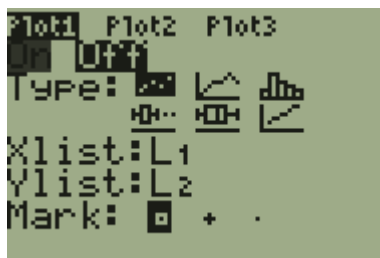
7, 39, 13, 9, 25, 8, 22, 0, 2, 18, 2, 30, 7, 35, 12, 15, 8, 6, 5, 29, 0, 11, 39, 16, 15

2. Configure the calculator to make a histogram.

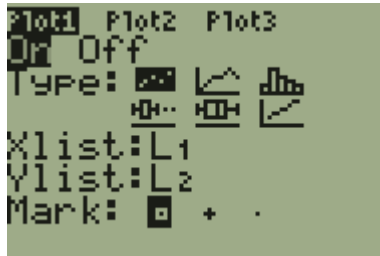
- a. Choose: *Stat Plot* by pressing the 2nd key and then they *Y=* (upper left, top row) key.



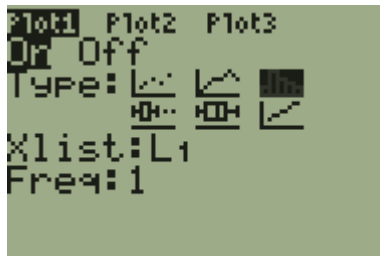
- b. Press the *1* key and then *Enter*.



- c. Use the left-arrow key to highlight *On*



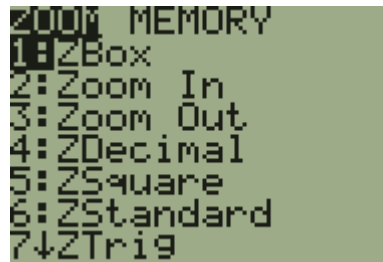
- d. Use the down-arrow key to move to *Type*. Then use the right-arrow key to choose the histogram icon (last item on top row). Press *Enter*.



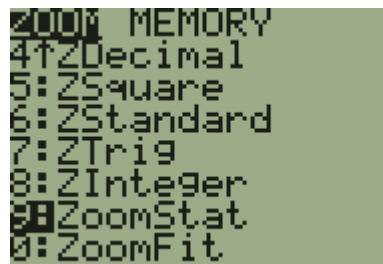
- e. Use the down-arrow key to move to *Xlist*. Enter the correct list if it is not already shown.
- f. Make sure *Freq* is set to 1

3. Make a default histogram

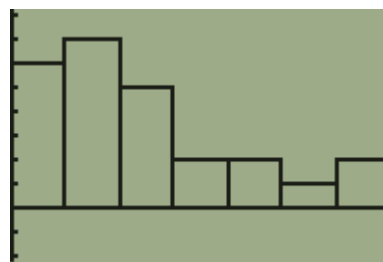
a. Press the *Zoom* key (top row, middle).



b. Use the down-arrow key to scroll down to 9 (ZoomStat).

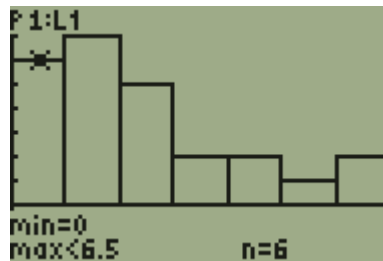


c. Press *Enter*. A default histogram is displayed.



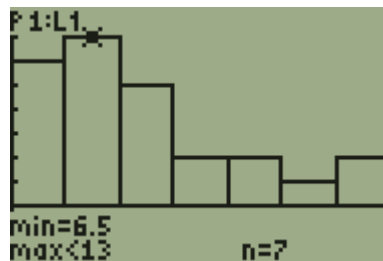
4. Investigate the histogram

a. Press the *Trace* key (top row). This will place an icon at the top of the first bar.



In this case, the text below tells us that the bar begins at 0 and ends at 6.5 (less than) and that the bar is 6 units high. Thus, there are 6 data values in $0 \leq x < 6.5$

b. Use the right-arrow key to scroll to the next bar.



In this case, the text below tells us that the bar begins at 6.5 and ends at 13 (less than) and that the bar is 7 units high. Thus, there are 7 data values in $6.5 \leq x < 13$

c. Continue scrolling to the right to investigate each of the bars.

5. Customize the histogram. Suppose we want a histogram with 5 bars. We can use the descriptive statistics to see that the minimum value is 0 and the maximum value is 39. Thus, we might want the histogram to span 0 to 40 with 5 bars. Thus, we would have a bar width of $\frac{40-0}{5}=8$.

a. Press the *Window* key (top row).

```
WINDOW
Xmin=0
Xmax=45.5
Xscl=6.5
Ymin=-2.10483
Ymax=8.19
Yscl=1
Xres=1
```

b. Use the arrow keys and numbers to change these values:

Xmin = 0
 Xmax = 40
 Xscl = 8

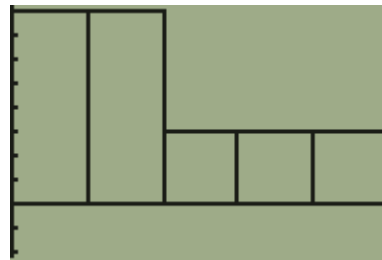
```
WINDOW
Xmin=0
Xmax=40
Xscl=8
Ymin=-2.10483
Ymax=8.19
Yscl=1
Xres=1
```

Note that *Xscl* is the bar width.

Do not change any of the other values.

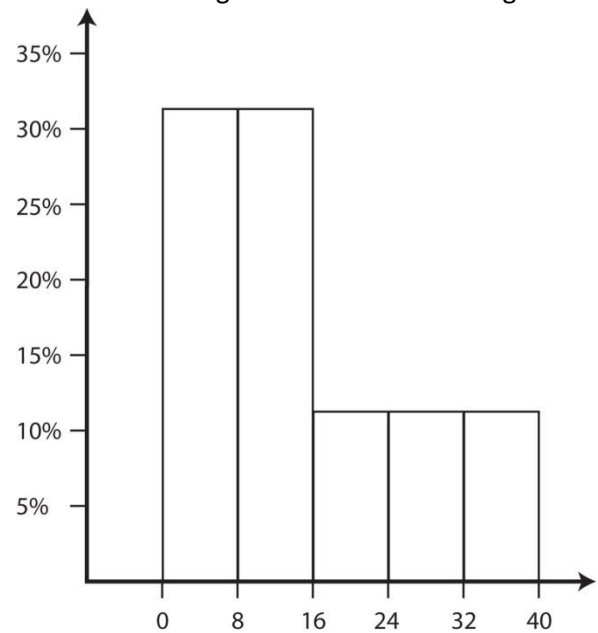
c. Press the *Graph* key (top row, far right) to display the customized histogram.

Use the *Trace* key to investigate.



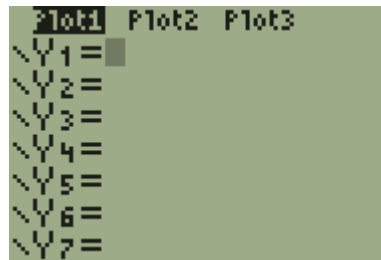
6. Suppose I ask you to draw a *relative frequency histogram* with the axes properly labeled. Tracing the histogram we can build this chart shown below on the left. Then, we can draw the histogram as shown on the right.

Class	Frequency	Relative Freq.
(0,8)	8	$8/25 = 0.32 = 32\%$
[8, 16)	8	$8/25 = 0.32 = 32\%$
[16, 24)	3	$3/25 = 0.12 = 12\%$
[24, 32)	3	$3/25 = 0.12 = 12\%$
[32, 40)	3	$3/25 = 0.12 = 12\%$
	n = 25	sum = 100%



7. After you have modified the window settings, if you then go back and press the *Zoom* and *9* keys (displaying the default histogram), then this will reset the window settings.
8. If you graph does not display correctly, try these things to fix it:

- a. Press the *Y=* key (top row). There should be nothing to the right of any of the equals signs. If there is, delete it.



- b. Choose *Stat Plot* by pressing the *2nd* key and then the *Y=* key. Make sure Plot 1 is *On* and all other Plots are *Off*.



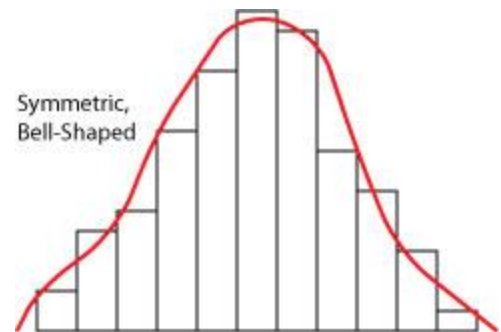
Homework 2.6

1. Use your calculator to draw a histogram from the data in Homework 2.2 using 5, 6, and 7 classes for both men's and women's data. Sketch each histogram on paper with properly labeled axes.

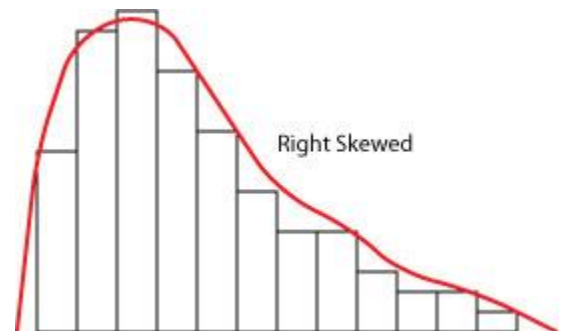
2.10 – Histogram Shapes and Skew

1. **Common Shapes of Distributions** – To determine the shape of a distribution, we sketch a curve over the histogram and analyze the curve.

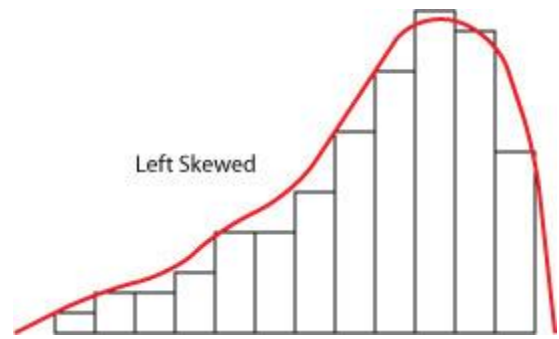
Symmetric, Bell-Shaped – Most of the data is towards the center but curve falls off rapidly in the larger and smaller directions.



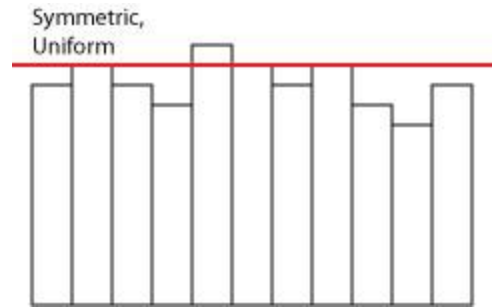
Right Skewed – Most of the data have *small values* but there are a few large data values



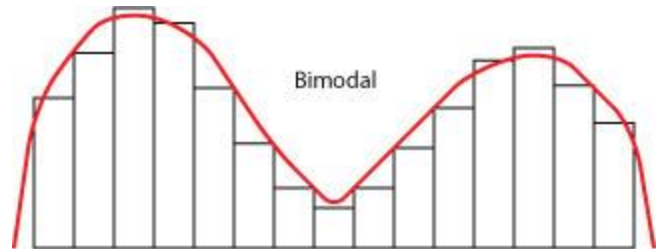
Left Skewed – Most of the data have *large* values but there are a few small values.



Symmetric, Uniform – Small, medium, and large data values occur with similar frequencies.



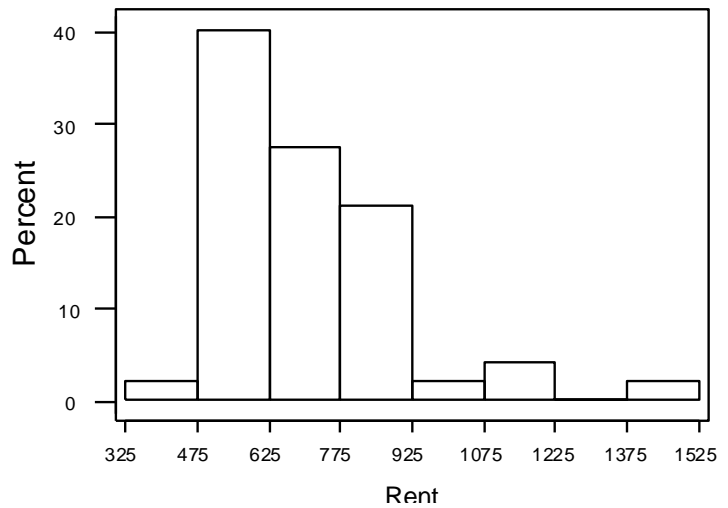
Bimodal – There are two distinct peaks in the data. This is not very common. Sometimes when we have bimodal data, we have inadvertently sampled from two populations.



2. **Examples**

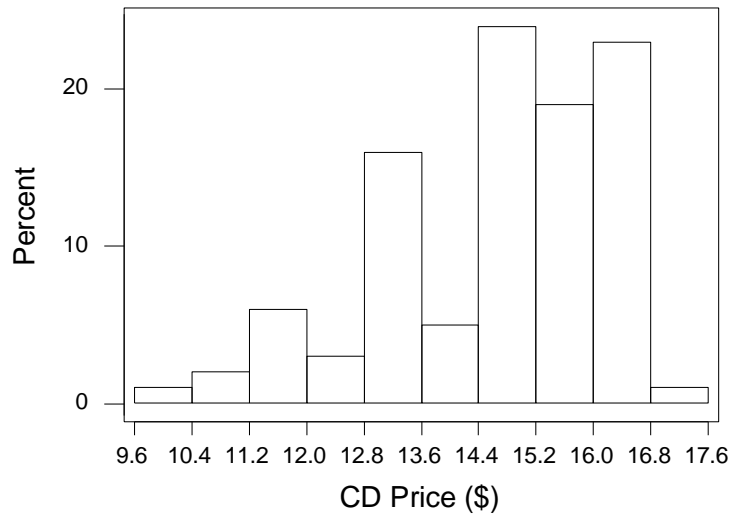
Example - 2 Bedroom Apartment Prices in Atlanta, Ga. This data was collected in April, 1995 from the Creative Loafing, a free, weekly newspaper in Atlanta. All 2 bedroom advertisements were included. Most of the apartments (about 90%) are between \$475 and \$925; however, there are a few with much higher prices.

Skewed Right



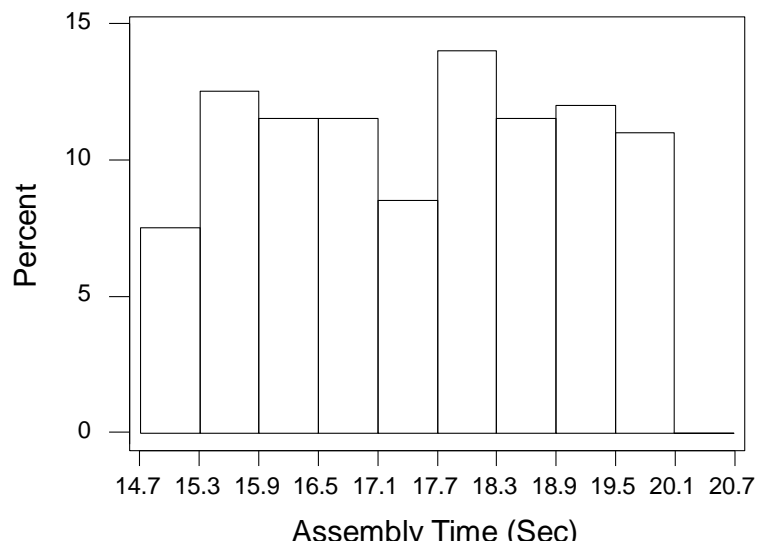
Example - New release CD prices. Most CD's (about 65%) cost between \$14.40 and \$16.80; however there are a few with lower prices. This could be due to large department stores using new release CD's as *loss leaders*.

Skewed Left



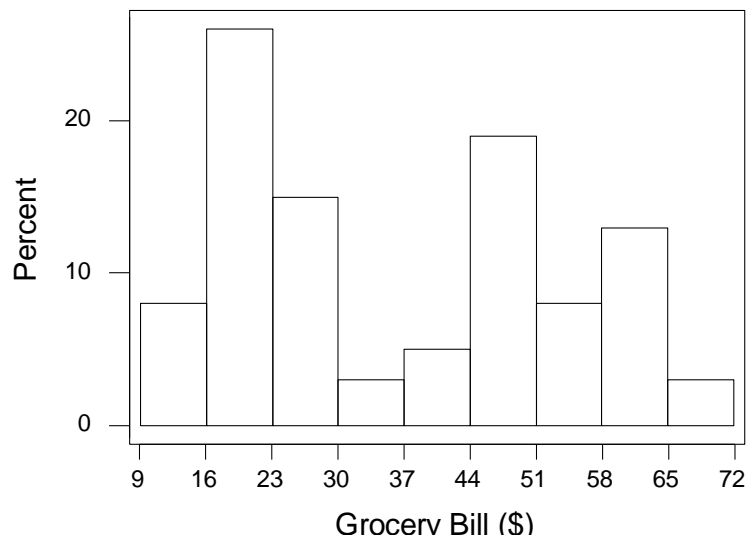
Example - Time (seconds) it takes a worker to assemble a part on an assembly line. The time to assemble a part is fairly uniform over the range 15 seconds to 20 seconds.

Symmetric, Uniform



Example - Total bill (\$) at a small grocery store. Perhaps there are a number of people who just stop in to buy a few items (lower total bill) and a similar number of people who shop for a week (higher grocery bill). Thus, perhaps there are two populations represented here: quick stop shoppers and weekly shoppers.

Bimodal



Homework 2.7

- Describe the distribution of the men's and women's data using the histograms from Homework 2.6.

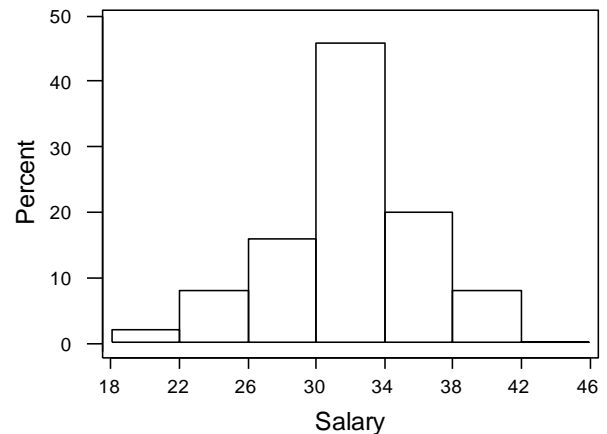
2.11 – Making an Ogive

- Cumulative Relative Frequency** – The percentage of data that have values below the upper class limit for a class.
- Ogive** - A graph of the *cumulative relative frequencies*. This graph shows how the classes/bars add up.
- Example** – Consider the salary data previously considered:

Class	Frequency (count)	Relative Frequency	Cumulative Relative Frequency	
(18,22)	1	$1/50 = 0.02 = 2\%$	2%	of salaries are less than \$22,000
[22, 26)	4	$4/50 = 0.08 = 8\%$	$2\% + 8\% = 10\%$	of salaries are less than \$26,000
[26, 30)	8	$8/50 = 0.16 = 16\%$	$10\% + 16\% = 26\%$	of salaries are less than \$30,000
[30, 34)	23	$23/50 = 0.46 = 46\%$	$26\% + 46\% = 72\%$	of salaries are less than \$34,000
[34, 38)	10	$10/50 = 0.20 = 20\%$	$72\% + 20\% = 92\%$	of salaries are less than \$38,000
[38, 42)	4	$4/50 = 0.08 = 8\%$	$92\% + 8\% = 100\%$	of salaries are less than \$42,000
	n = 50	sum = 100%		

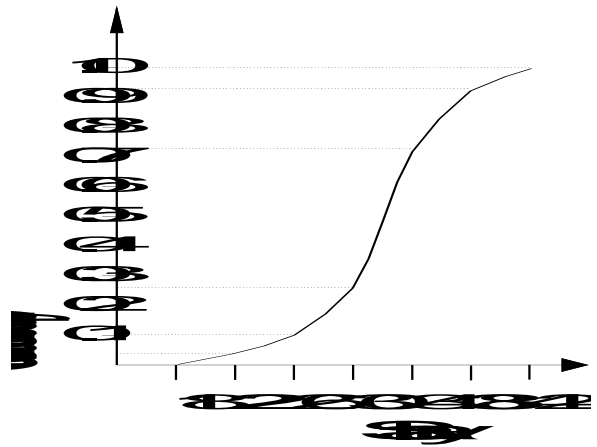
To compute the cumulative relative frequencies shown in the table above, we simply ask ourselves a series of questions: what percentage of salaries are:

- less than 22?* Answer: the height of the first bar in the histogram (shown at right), 2%.
- less than 26?* Answer: the height of the first two bars in the histogram, $2\% + 8\% = 10\%$.
- less than 30?* Answer: the height of the first three bars in the histogram, $2\% + 8\% + 16\% = 26\%$.
etc.



Finally, we graph the cumulative relative frequencies to obtain the *Ogive* as shown in the figure below:

- a. We graph the upper class limit with the corresponding cumulative relative frequency for each class. For instance, we graph, (22,2), (26,10), (30, 26), (34, 72), (38, 92), (42, 100).
- b. We go back and plot the lower class limit for the *first* class with the value 0. For instance, we plot (18,0).
- c. Connect the dots with a smooth curve.

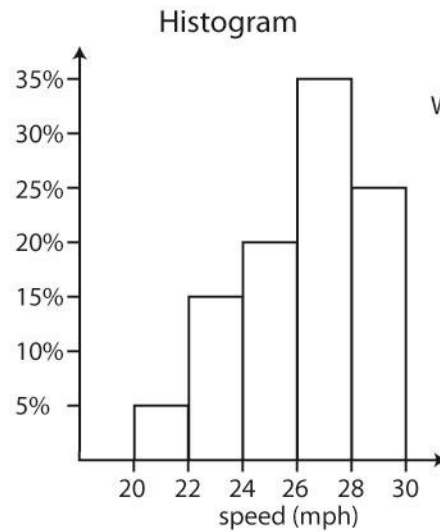


Homework 2.8

1. Choose a men's and women's histogram from Homework 2.6 and draw an ogive for each.

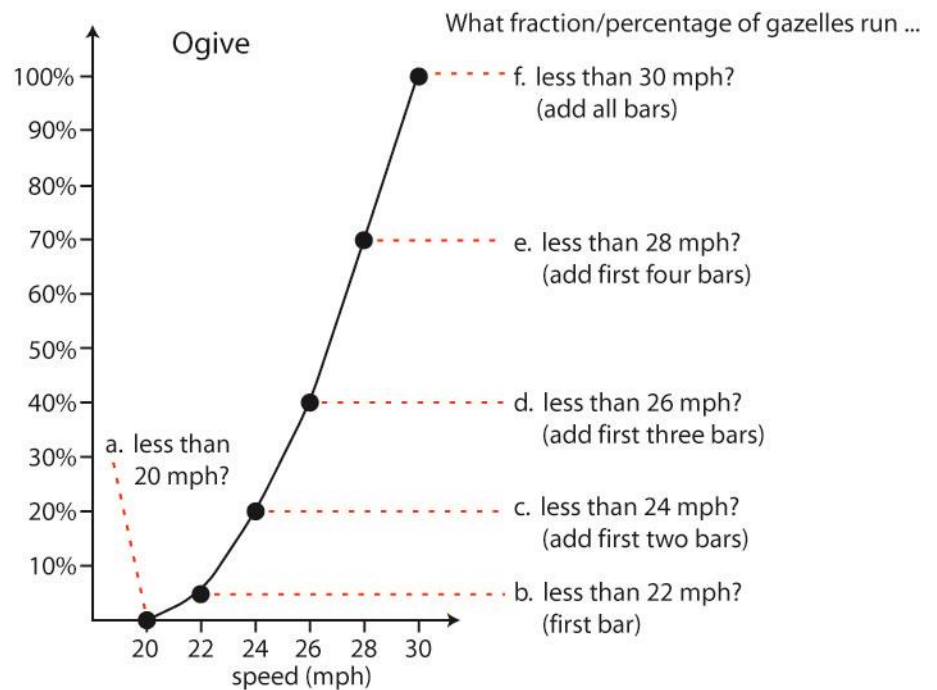
2.12 – Interpreting a Histogram and Ogive

1. The *height* of a bar in a histogram tells the percentage of data that falls *between* the lower and upper limits of the bar.
2. The *height* at some point along the curve of an ogive represents the percentage of data that is less than the value on the x-axis that is directly below the point on the curve.

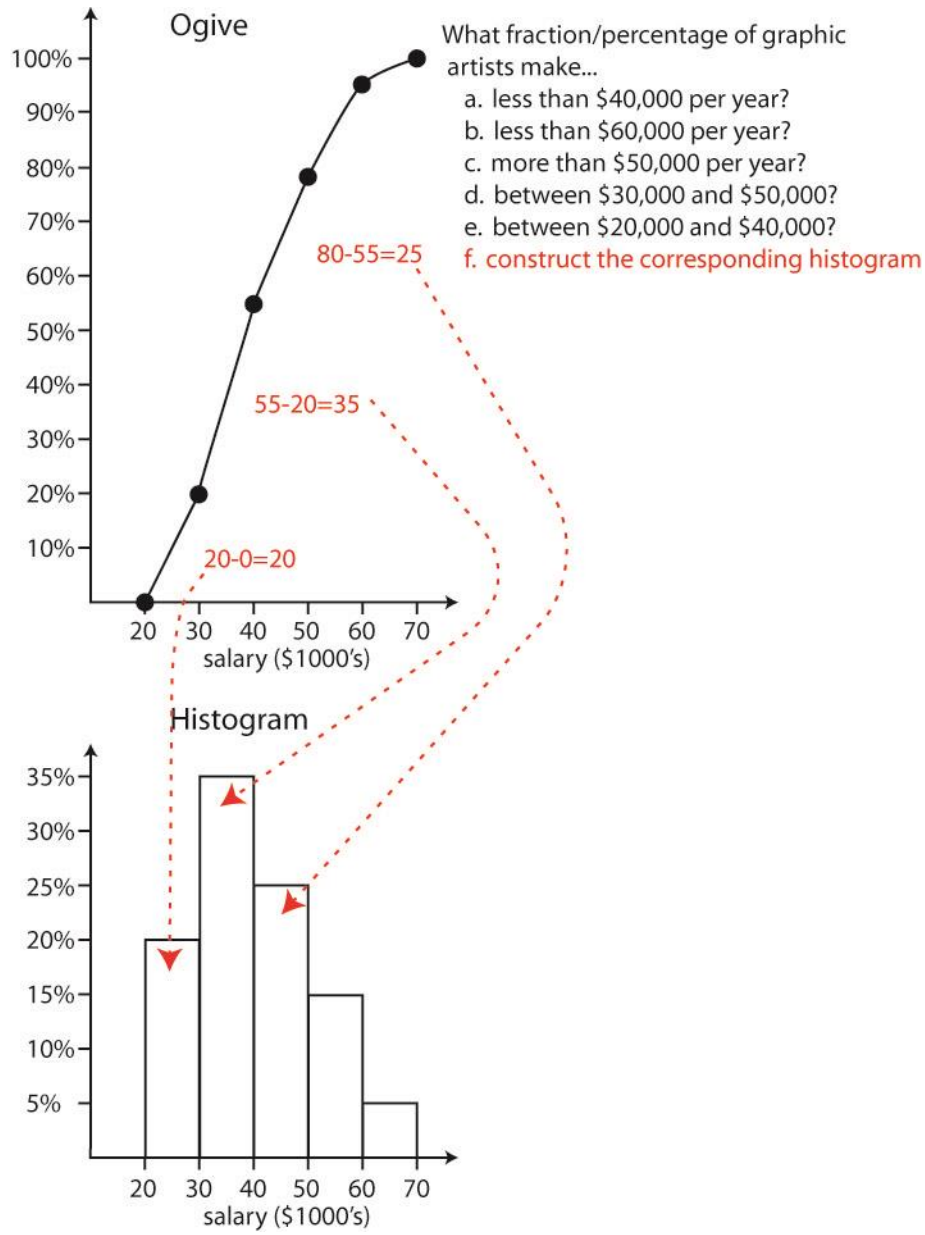


- What fraction/percentage of gazelles run ...
- a. between 28 and 30 mph?
 - b. between 22 and 26 mph?
 - c. less than 24 mph?
 - d. more than 26 mph?
 - e. graph the corresponding ogive

3. **Example** - Consider the histogram shown above on the right which describes the speed of a sample of gazelles. The corresponding ogive is shown on the right.

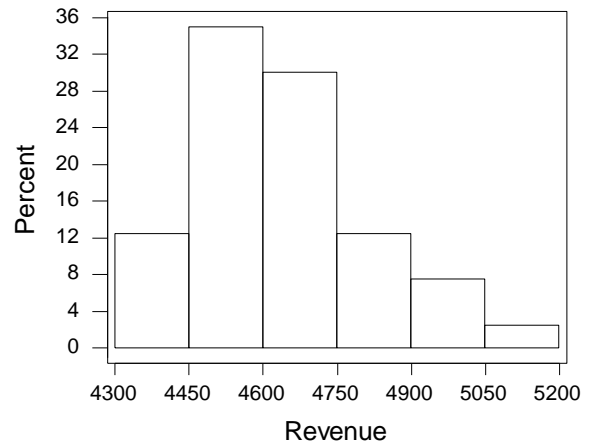


4. **Example** - Consider the *ogive* shown below which describes the salary of mid-career graphic artists.

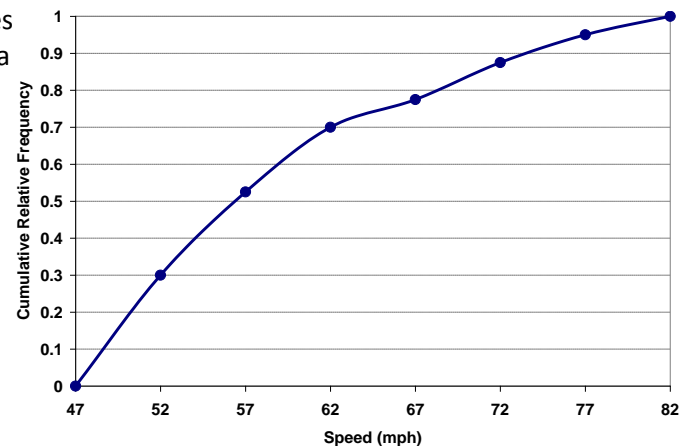


Homework 2.9

1. Consider the histogram shown below which summarizes the total amount of money that was received by a restaurant (revenue) on 40 Friday nights. What percentage of Friday nights have revenue of (a) \$4900 to \$5050, (b) less than \$4600, (c) more than \$4750? (d) Graph the corresponding ogive.

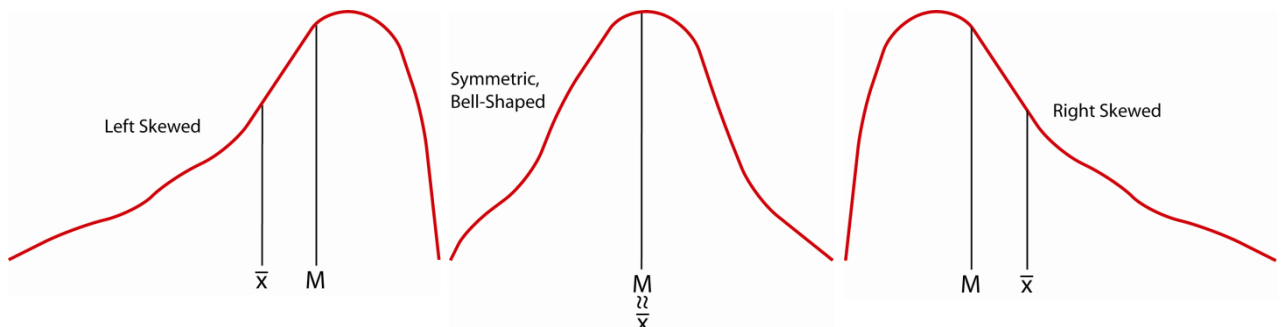


2. Consider the ogive shown below that summarizes the speeds of a random selection cars along a stretch of highway with a posted speed limit of 55 mph. Guidelines specify that cars traveling 57 mph or less are not speeding, cars traveling more than 57 mph and up to 67 mph are classified as Class 1 Speeding, while those traveling more than 67 mph are classified as Class 2 Speeding. What percentage of cars are (a) not speeding, (b) Class 1 Speeding, (c) Class 2 speeding. (d) Draw the corresponding Histogram



2.13 – Detecting Skew Numerically

1. **Detecting Skew Numerically** – The difference between the *mean* and the *median* tells us about the *skew* of the data as shown in the figures below.



2. Thus, table one the right summarizes this.

Notes:

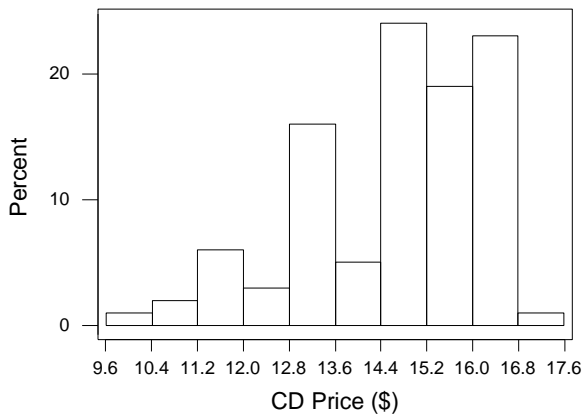
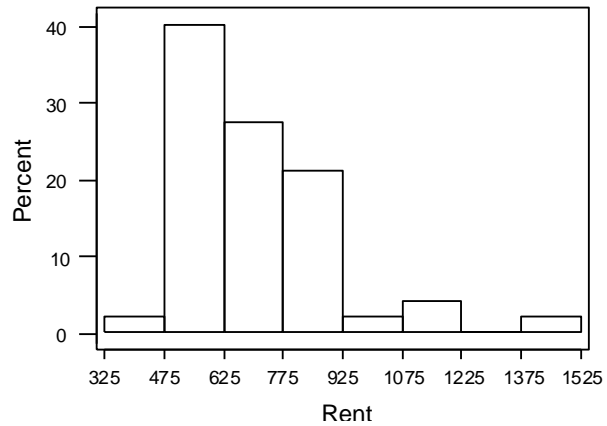
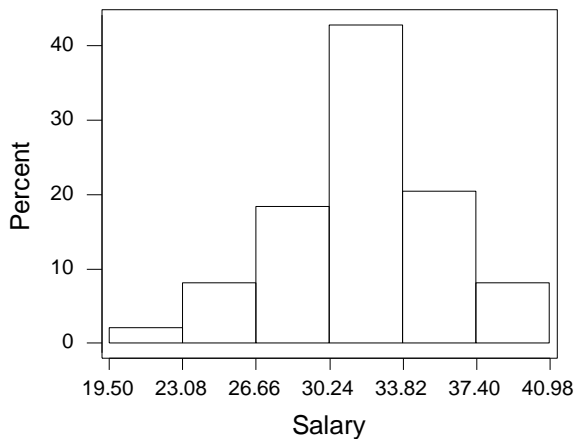
- The larger the difference, the more skew in the data (right or left skewed).
- If there is little difference, the data has a *symmetric* distribution.
- It does take a bit of judgment to determine if the difference between the Mean and Median is close to zero. Detailed knowledge of the problem at hand is useful in detecting skew numerically.
- We remember that outliers influence the mean more than the median. So, if the mean is *larger* than the median then the sample probably has some extra large data values and the histogram will appear *skewed right*. If the mean is *smaller* than the median then the sample probably has extra small data values and the histogram will appear *skewed left*.

Mean – Median		Interpretation
negative	$\bar{x} < M$	Left Skew
about zero	$\bar{x} \approx M$	Symmetric
positive	$\bar{x} > M$	Right Skew

3. **Example** – Estimation of skew using the mean and median from three examples considered earlier. The values of the percentage differences cannot be used to make judgments about the skew, but are used here for comparison purposes.

Data	Mean	Median	Difference	% Difference	Skew
Salaries	\$32,048.00	\$32,150.00	-\$102.00	-0.3%	Symmetric
Rent	\$706.70	\$675.00	\$31.70	+4.7%	Right
CD Price	\$14.65	\$15.01	-\$0.36	-2.4%	Left

For comparison, the histograms are shown below:

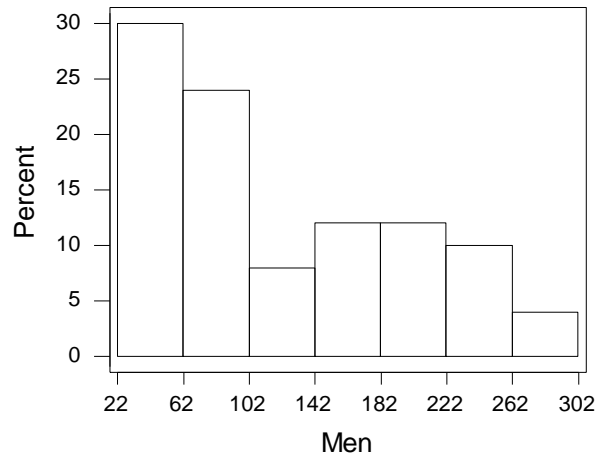
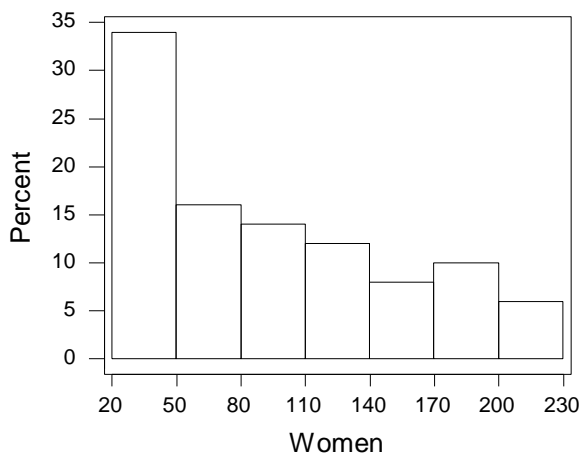


4. **Example** – Estimation of skew using the mean and median from 5-Word Word Search Problem. Men and Women VSU students were timed as described in Chapter 1. The statistics shown below were generated by Minitab.

As the table below shows, the women’s data shows that the mean is about 13.9% larger than the median while the men’s data shows that the mean is 34.2% larger than the median. Thus, we would expect both data sets to be skewed to the right with a much heavier skew in the men’s data.

Data	Mean	Median	Difference	% Difference	Skew
Women	93.49	82.09	11.4	13.9%	Right
Men	121.60	90.60	31.0	34.2%	Right

For comparison, the histograms are shown below. Both histograms show strong right skew.



Homework 2.10

- Compare the mean and median from the results of Homework 2.3 and compare to the corresponding histograms from Homework 2.7.
- The median weight of an apple is 5.6 oz. and the average is 6.5 oz. What type of skew is present?
- The average gas mileage for a Nissan Pathfinder is 15.6 mpg and the median is 17.2 mpg. What type of skew is present?
- Which sample has the (a) least skew, (b) most skew, (c) most left skew, (d) most right skew, (e) least left skew, (f) least right skew, (g) most symmetry, (h) least symmetry?

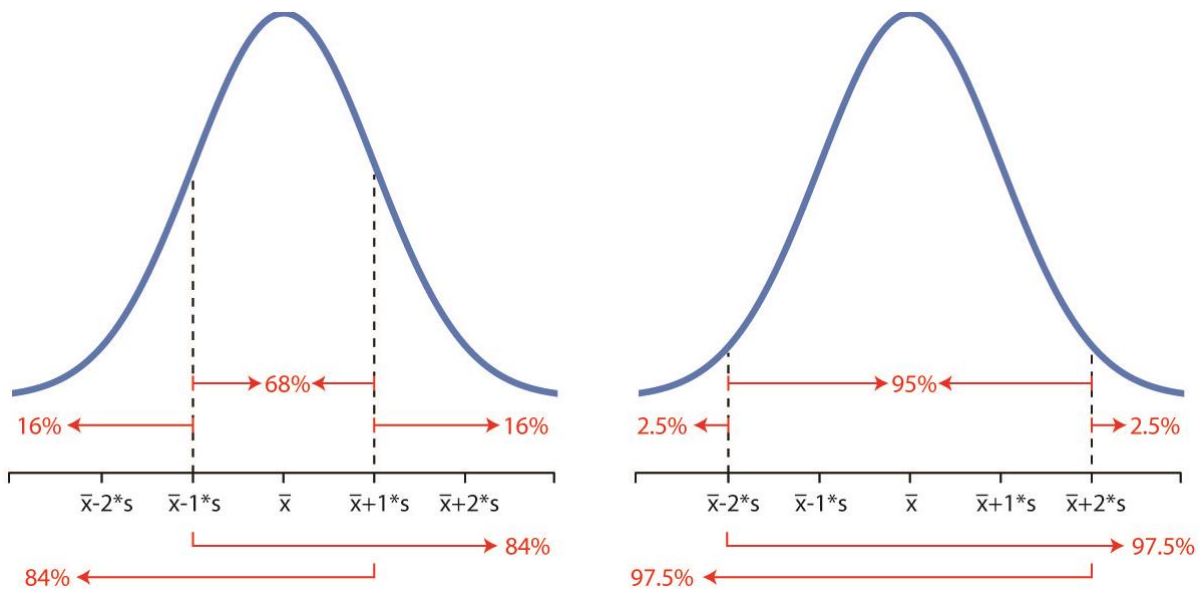
Sample	Average	Median
1	27.4	25.3
2	128.6	130.2
3	67.3	63.1
4	20.2	28.6

2.14 – Empirical Rule

1. **Empirical Rule** – If data is *bell-shaped (mound-shaped)*, then Empirical Rule states that:

- Approximately 68% of data should fall within ± 1 standard deviation of the average, $\bar{x} \pm 1s$
- Approximately 95% of data should fall within ± 2 standard deviations of the average, $\bar{x} \pm 2s$
- Approximately 99.7% (*i.e.* almost all) of the data should fall within ± 3 standard deviations of the average, $\bar{x} \pm 3s$

2. **Diagrams of 68% rule and 95% rule:**



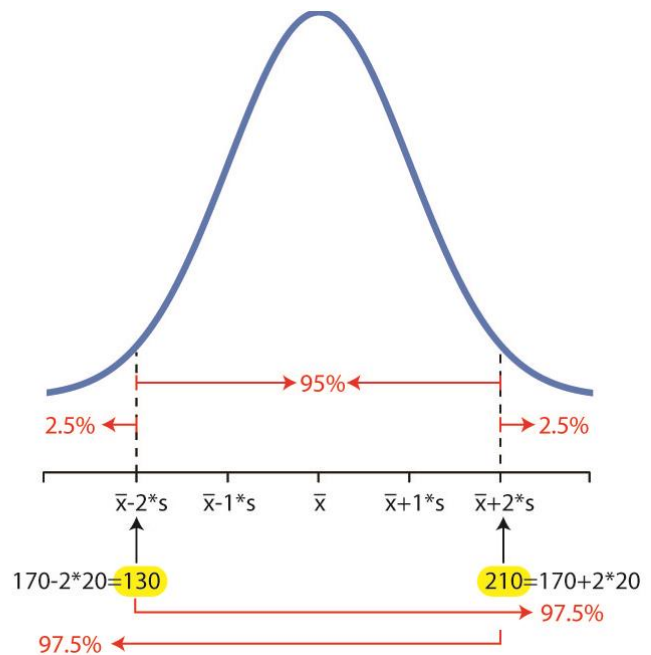
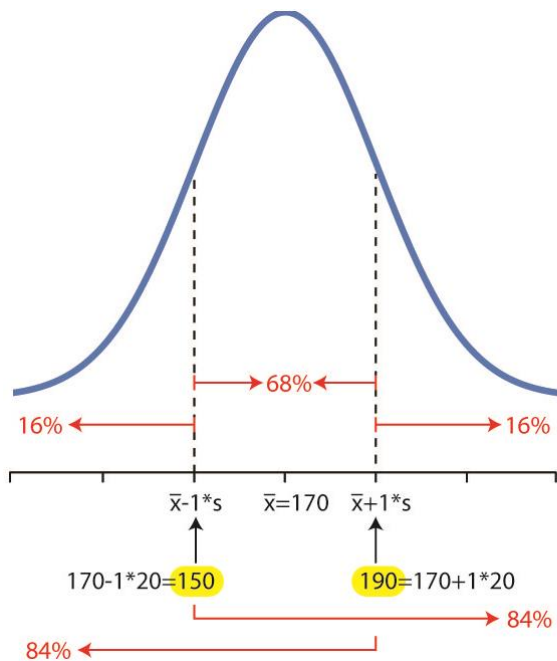
- In the “68%” graph on the left, we see where the *center* 68% of the data should fall. Remembering that the graph must show 100% of the data, we see that $100\% - 68\% = 32\%$. Now, this *left over* 32% should be split equally between the two tails, 16% each. Thus, 68% rule allows us to break the data up into three groups: 16%, 68%, 16%. We also note that $16\% + 68\% = 84\%$, so that the 68% rule also breaks the data into 84%, 16% and 16%, 84%.
- The “95%” graph on the right shows that the data can be broken up into these sets: 2.5%, 95%, 2.5%; and 97.5%, 2.5%; and 2.5%, 97.5%.

3. **Example** – The distribution of the weight of men is bell-shaped with a mean (average) of 170 lbs. and standard deviation of 20 lbs.

- a. 68% of men have weight between what two values?
- b. 16% of men have weight below what value?
- c. 84% of men have weight above what value?
- d. 16% of men have weight above what value?
- e. 84% of men have weight below what value?
- f. 50% of men have weight below what value?
- g. 95% of men have weight between what two values?
- h. 2.5% of men have weight below what value?
- i. 97.5% of men have weight above what value?
- j. 2.5% of men have weight above what value?
- k. 97.5% of men have weight below what value?
- l. 99.7% of men have weight between what two values?

What percentage of men have weight...

- m. between 150 lbs. and 190 lbs.?
- n. more than 150 lbs.?
- o. less than 150 lbs.?
- p. more than 190 lbs.?
- q. less than 190 lbs.?
- r. more than 170 lbs.?
- s. between 130 lbs. and 210 lbs.?
- t. less than 210 lbs.?
- u. less than 130 lbs.?
- v. more than 130 lbs.?
- w. more than 210 lbs.?
- x. between 110 lbs. and 230 lbs.?



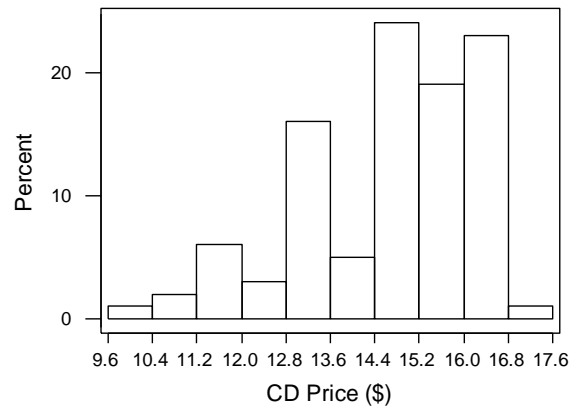
Homework 2.11

1. Consider the data from Homework 2.3. Does the Empirical Rule apply for either data set? If so, calculate the actual percentage of data that falls with one standard deviation of the average. Also, calculate for two and three standard deviations. Make a chart of these values. In other words, calculate the boundaries for the 68% empirical rule. Next, pull the data up and then count the number of data values that fall in this range, then make this a percentage by dividing by the sample size. Now, repeat for the other two empirical rules. In a sense you are “testing” the empirical rule against one of your data sets.
2. The average price of a used Cadillac manufactured in 2003 is \$22,000 with a standard deviation of \$3000. The distribution of used Cadillac prices is bell-shaped. What percentage of Cadillac’s cost...
 - a. less than \$19,000?
 - b. less than \$25,000?
 - c. less than \$22,000?
 - d. more than \$19,000?
 - e. more than \$25,000?
 - f. more than \$28,000?
 - g. more than \$16,000?
 - h. less than \$28,000?
 - i. less than \$16,000?
 - j. between \$16,000 and \$28,000?
 - k. between \$19,000 and \$25,000?
 - l. between \$13,000 and \$31,000?
3. The average time it takes a worker to perform a task on an assembly line is 38 seconds with a standard deviation of 4 seconds. The distribution of times is bell-shaped.
 - a. 2.5% of the time, a worker takes more than ____ seconds.
 - b. 2.5% of the time, a worker takes less than ____ seconds.
 - c. 68% of the time, a worker takes a time between ____ seconds and ____ seconds.
 - d. 84% of the time, a worker takes less than ____ seconds.
 - e. 97.5% of the time, a worker takes more than ____ seconds.
 - f. 50% of the time, a worker takes more than ____ seconds.
 - g. 84% of the time, a worker takes more than ____ seconds.
 - h. 95% of the time, a worker takes a time between ____ second and ____ seconds.
 - i. Almost all workers take between ____ seconds and ____ seconds.

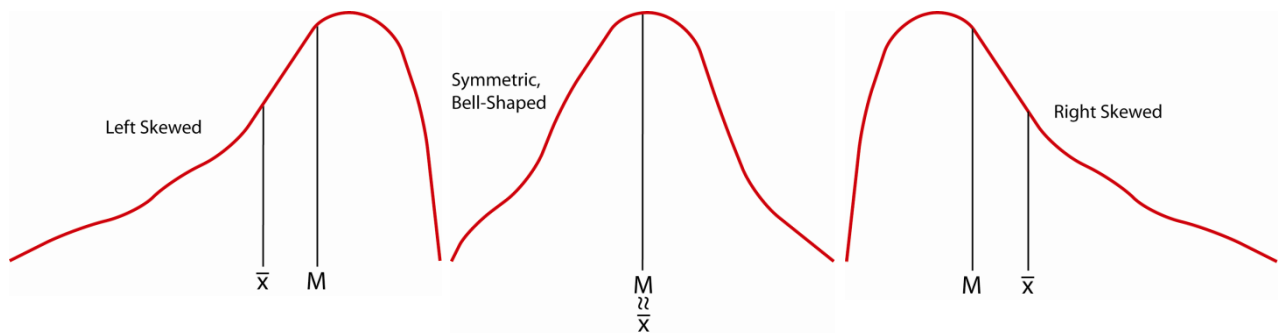
2.15 – Estimating Statistics from a Histogram

1. **Estimating the Median from a Histogram** – To estimate the *median* from a histogram showing percents, simply add up the boxes from left to right. The first time your addition goes above 50% stop adding. The median is *in* the box that made the addition go above 50%. From there, make a good guess (perhaps the middle of the box). Note that you could estimate the quartiles and other percentiles in a similar manner.

Example – In the histogram shown to the right, we might estimate that the first box contains 2% of the data, the second box contains 3%, the third contains 6%, the fourth contains 4%, the fifth contains 15%, the sixth contains 5%. Now, add these boxes (2+3+6+4+15+5) to get 35%. Since we have not reached 50%, we need to add the next box, the seventh, which contains about 23%. Now the total is 58% (35+23). This tells us that the median CD price is between \$14.40 and \$15.20. We might estimate that the median then is the midpoint between these two numbers, \$14.80. Note that the actual median is \$15.01. What we have done is *estimate* the median.



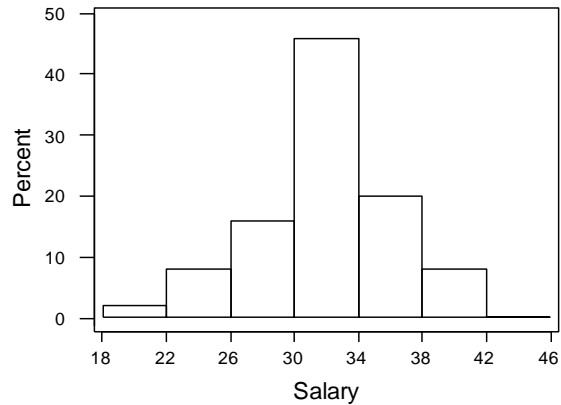
2. **Estimating the Location of the Mean from a Histogram** – For symmetric data, the mean and median are approximately the same (center figure below). For skewed data, the mean gets pulled in the direction of the skew. The figure on the left shows left skewed data, it has a few *very small values* (relatively) which *pull* the sensitive mean to the left of the median. Similarly, the figure on the right shows right skewed data and how the mean is pulled towards the larger values.



Example – Consider the CD Price histogram shown above. It is clearly left skewed so that we could say that the mean is less than the median. Thus, the mean is less than \$14.80

3. **Estimating the Range from a Histogram** – Simply subtract the lower class limit for the first class from the upper class limit from the last class. Note that this will usually be an overestimate of the range.

Example – Consider the Salary histogram shown at right. We would estimate the Range to be $42 - 18 = 24$. Note that the actual range is 21.5.



4. **Estimating the Standard Deviation from a Histogram** – If the data is bell-shaped, then we can use the empirical rule.

The Empirical Rule says that almost all (99.7%) of data should fall, $\bar{x} \pm 3s$. Thus, the distance between $\bar{x} - 3s$ and $\bar{x} + 3s$ is about the *estimated range*. Thus *half the estimated range* should approximately be equal to $3s$. Note that in practice, you might use a value between 2 and 3 depending on your sample size.

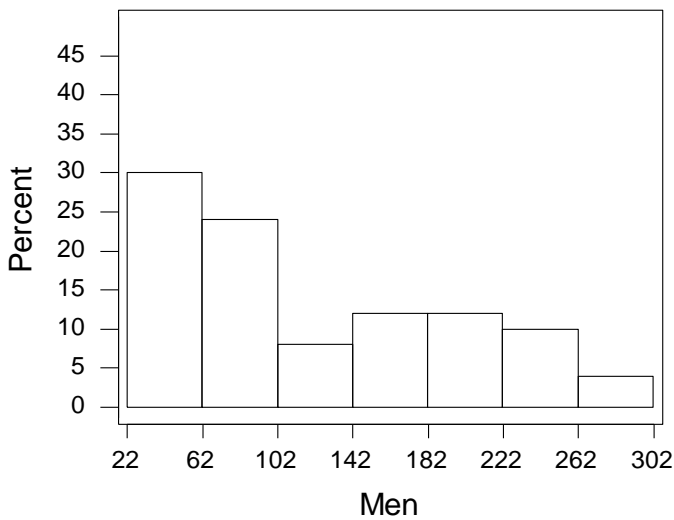
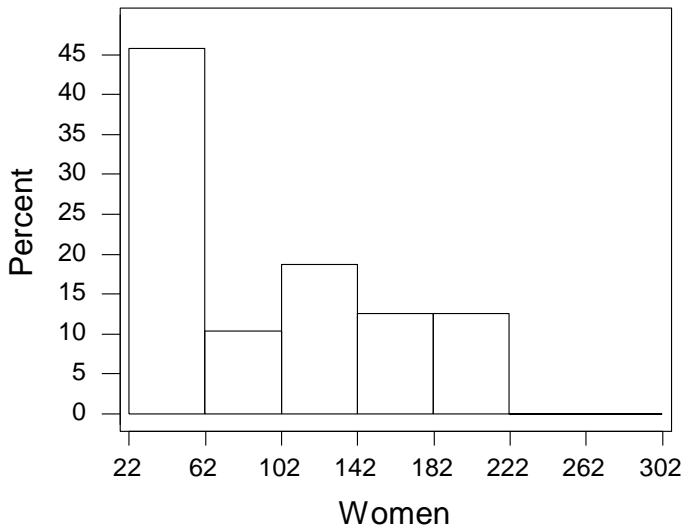
Example – Consider the Salary histogram shown above, at right. The histogram appears to have a bell-shaped distribution. Thus, we will assume the Empirical Rule applies. The estimated range is 24, and half of that is 12. Thus, $3s = 12$ and $s = 4$. Note that the actual value is $s = 4.311$.

Example - You ask the manager of a call center what the standard deviation of the length of calls is. Her response is, "I don't know, but most of the time it is between 5 and 13 minutes." You ask the manager what she means by "most." She responds, "about 95%." Assume the length of calls is bell-shaped. Estimate the standard deviation.

Since the manager said, "about 95%", then 95% of calls are between 5 and 13 minutes, $\bar{x} \pm 2s$. So, half the distance between these two values should be $2s$. Thus, $2s = \frac{13-5}{2} = 4$ and $s=2$.

Another way to think of this: \bar{x} should be half way between 5 and 13. So, $\bar{x} = 9$. Now, using the 95% empirical rule: $9 - 2s = 5$ and $9 + 2s = 13$. What value of s makes both these statements true? Answer: $s=2$.

5. **Example** – Consider the word search problem consider earlier. When we are comparing two data sets, it can be useful to graph both histograms on the same scale as shown below.



As we can see from the histograms, about 45% of women took less than a minute (62 seconds), while only 30% of men could complete the puzzle this fast. Also, all women finished the puzzle in less than 222 seconds (3.7 minutes) while about 14% of men took longer than this.

Homework 2.12

- Pick a set of histograms from Homework 2.6. Estimate the median and range. If either data set is symmetric, then estimate the standard deviation. Show these values in a table along with the actual values. Do the estimates seem reasonable?

2.16 – Making a Box Plot

1. **Box Plot** - A Box plot is a graphical summary of quantitative data. Similar to the histogram, it shows the center, spread, and skew of the data. It is a graph of the first and third quartiles, the median, *fences*, *whiskers*, and *outliers*.

2. Steps to Make a Box Plot

a. Compute: Q_1 , M , Q_3 , IQR

b. Compute *fence* locations:

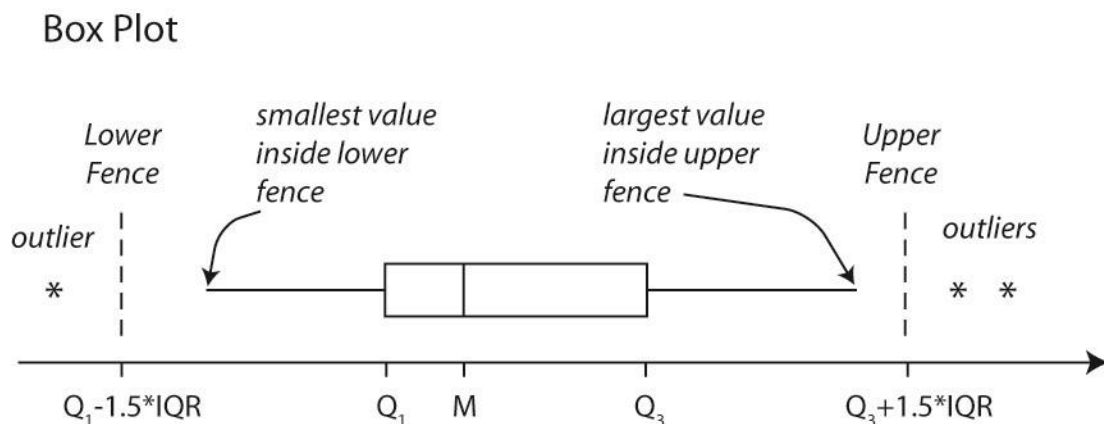
$$\text{Lower Fence} = Q_1 - 1.5 * IQR$$

$$\text{Upper Fence} = Q_3 + 1.5 * IQR$$

c. Determine *whisker* locations. As shown in the figure below, the whiskers are the lines coming out of either end of the box. The left whisker extends to the smallest value inside the lower fence and the right whisker extends to the largest value inside the upper fence.

d. Determine whether there are any outliers. Any values *outside* the fences are outliers and are marked with a symbol, for instance, “*”.

e. Draw the box plot as shown in the figure below. Note that the fences are usually not actually drawn, they are just understood to be there. Below, we have drawn them as dashed lines simply for emphasis.



Note that an alternate form of the boxplot draws the left whisker all the way to the minimum and the right whisker to the maximum value. We will not use this approach.

3. **Example** – This CS starting salary data was considered previously.

19.5	26.8	29.1	30.5	31.5	32.2	33.2	33.9	35.4	37.4
24.2	27.4	29.1	30.9	31.9	32.4	33.3	34.8	36.1	39.2
24.8	27.8	29.6	30.9	32.1	32.5	33.3	34.9	36.1	39.7
25.3	27.8	30.1	31.1	32.1	32.8	33.4	35.0	36.3	40.2
25.7	28.6	30.3	31.4	32.1	33.1	33.4	35.3	36.9	41.0

To make a box plot:

- a. Calculate: Q_1, M, Q_3, IQR (see Section 2.6, items 3 and 4 if necessary)

$$Q_1 = 29.6, M = 32.15, Q_3 = 34.9, IQR = 34.9 - 29.6 = 5.3$$

- b. Compute fence locations:

$$\text{Lower Fence} = Q_1 - 1.5 * IQR = 29.6 - 1.5 * 5.3 = 21.65$$

$$\text{Upper Fence} = Q_3 + 1.5 * IQR = 34.9 + 1.5 * 5.3 = 42.85$$

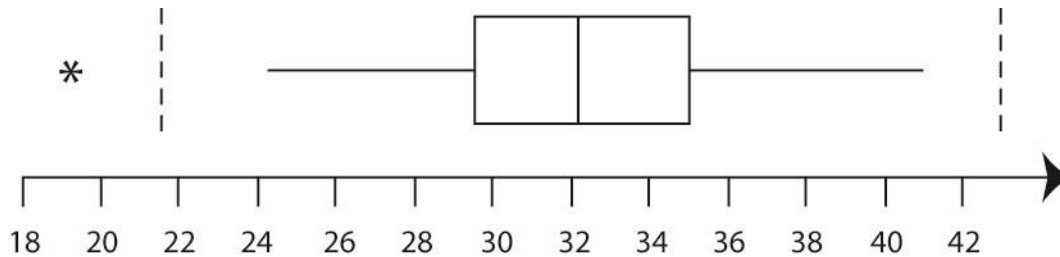
- c. Determine Whisker Locations:

Lower = 24.2 (smallest value *more than* 21.65)

Upper = 41.0 (largest value *less than* 42.85)

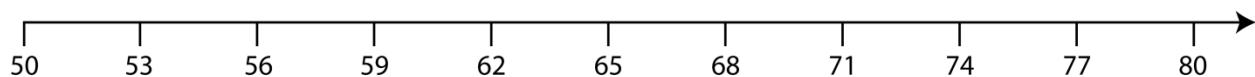
- d. Determine whether there are any outliers: There is one outlier, 19.5 which is below the Lower Fence, 21.65

- e. Draw Box Plot as shown below (fences shown as dashed lines, for emphasis):



4. **Box Plot Example** - A sample of the speed of cars on an interstate reveals that: $Q_1 = 65, M = 67, Q_3 = 71$.

- (a) Construct a Box Plot (we will not be able to put the whiskers on the Box Plot because we do not have the actual data). (b) Draw the fences. (c) Would these speeds be considered outliers: 75, 70, 53, 82, 77? Classify these data values as *Possible Outliers (O)* or *Not an Outlier (N)* using the Box Plot.



Homework 2.13

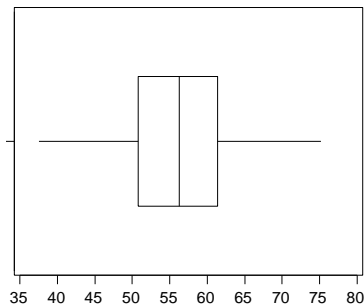
1. Consider the data in Homework 2.2. Build a Box plot by hand for each dataset. Hint: first compute the numerical descriptive statistics on your calculator.

2.17 – Box Plot Shapes and Skew

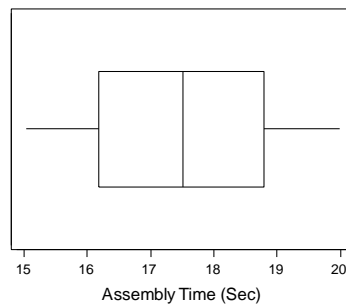
This section provides guidelines for determining symmetry or skewness in a Box Plot:

1. If whiskers are approximately the same length and median is approximately in the middle of the box then the data is *symmetric*. If the box is significantly *thinner* than the whiskers, then this is an indication that the data may be bell-shaped. If the whiskers and each half of the box are similar in size, then the data may be more uniform in distribution.

Symmetric, bell-shaped, no outliers

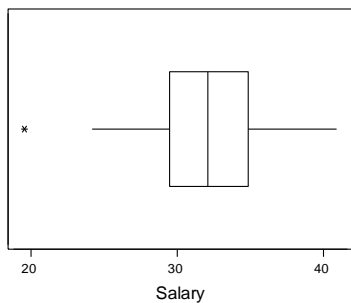


Symmetric, uniform, no outliers

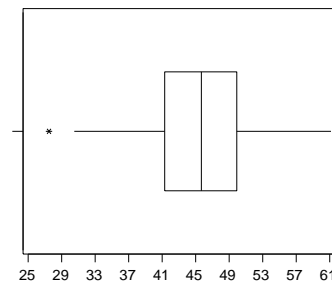


What if there is exactly one outlier? Sometimes, we ignore it, unless it is very large (or small). So, in the examples below, we would say that this data is symmetric.

Symmetric, probably bell shaped, 1 outlier on the low side

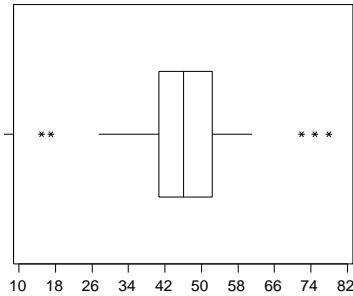


Symmetric, bell shaped, 1 outlier on the low side

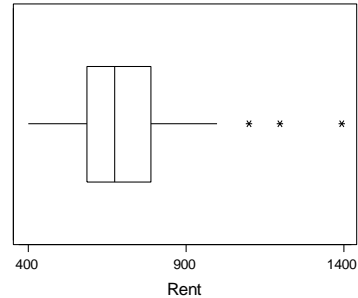


What if there are several outliers? If there are a roughly equal number of small and large outliers, at similar distances from the whiskers, I usually would say that this data is symmetric as in the example on the left. However, on the right is a case where we would probably say that the data is right skewed, thus we are influenced by the three outliers on the high side.

Symmetric, bell shaped, 2 outliers on the low side and three on the high side

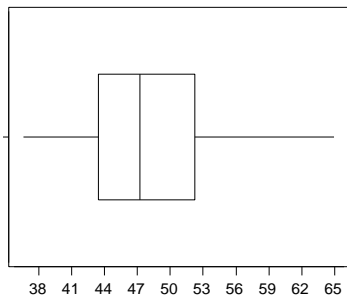


Right Skewed, 3 outliers on the high side.

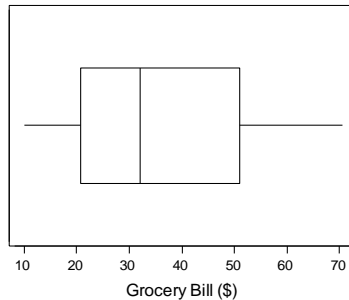


- If the whiskers are distinctly different lengths *or* the median is not in the middle of the box, then the data is *skewed*. If the length to the right of the median is longer than the length to the left of the median, then the data is *right skewed*, otherwise it is *left skewed*.

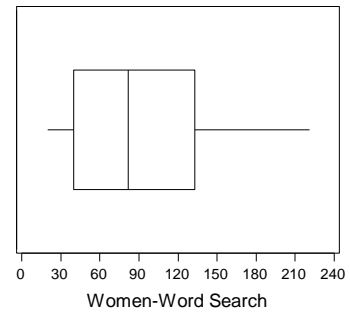
Right Skew, no outliers



Right Skewed, no outliers

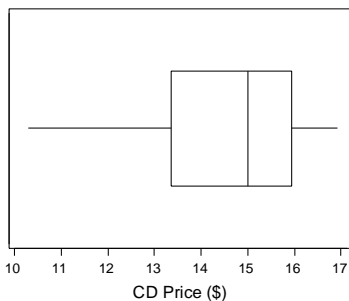


Right Skewed, no outliers

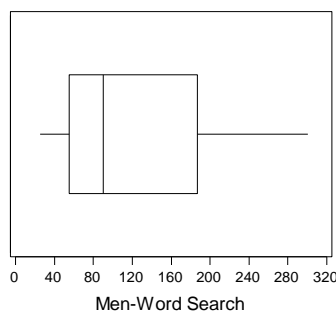


- If the whiskers are different lengths *and* the median is not in the middle of the box, then the data is usually *highly skewed*.

Strong Left Skew, no outliers

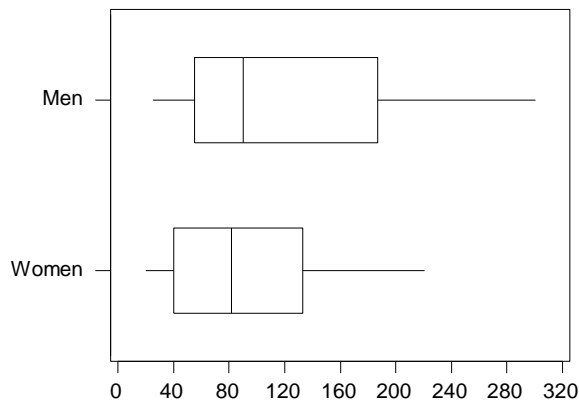


Strong Right Skew, no outliers



4. **Example** – When we are comparing two samples, we usually make the two box plots on the same graph so that we can compare them. The two box plots below are for the word search problem discussed earlier. We could summarize this data as follows:

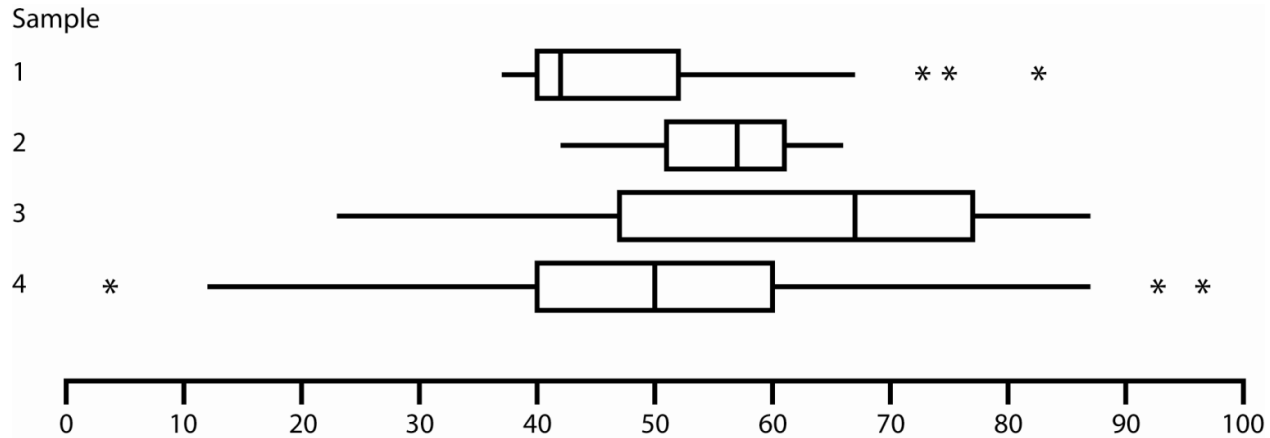
The median for the men’s data is larger than the women’s data, though it is hard to see from the graph, unless closely inspected. Thus, there is some indication that men take longer than women. From the graph, these medians appear to be about 80 and 90 seconds, respectively. The men’s data has more spread than the women’s data as measured by the range (estimated from graph). We also see that the men’s data is more skewed than the women’s data. The men’s box plot shows that about 25% of men took longer than 3 minutes (180 seconds), while a smaller fraction of the women took this long (less than 25%). There were no outliers indicated by the box plots for either the men’s or women’s data.



Homework 2.14

1. Compare (center, spread, skew, outliers) the box plots from Homework 2.14.

Consider the box plots shown below to answer next three questions.



2. Which sample has the
 - a. most variability?
 - b. least variability?
 - c. most skew?
 - d. least skew?
 - e. largest percentage of data greater than 50?
 - f. largest percentage of data greater than 70?
 - g. largest median?
 - h. smallest median?
 - i. smallest first quartile?
 - j. largest first quartile?
 - k. Classify the skew of each sample
 - l. most right skew?
 - m. most left skew?
3. Which statement is correct?
 - a. The median of Sample 1 is about the same as the first quartile for Sample 2
 - b. The IQR for Sample 4 is larger than the IQR for Sample 3
 - c. Half of the data in Sample 3 is larger than all the data in Sample 2
 - d. At least 25% of the data in Sample 4 is larger than all the data in Sample 2
4. Consider Sample 2. The fence on the high side is located at what value (approximately)?
5. Which sample has the most variability in the center-most 50% of the data?
6. Which sample has a mean that is significantly larger than its median?

2.18 – Making a Box Plot on TI-83 Calculator

The steps below show how to use the TI 83/84 calculator to construct a box plot that shows outliers. Assume that you have a dataset in L1. Next,

1. Enter this data in L1 (or some other list):

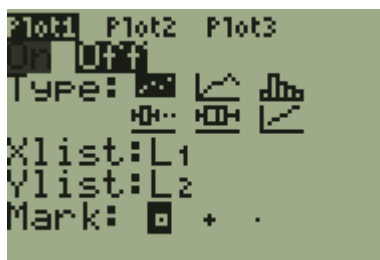
7, 39, 13, 9, 25, 8, 22, 0, 2, 18, 2, 30, 7, 35, 12, 15, 8, 6, 5, 29, 0, 11, 39, 16, 15

2. Enter this data in L1 (or some other list):

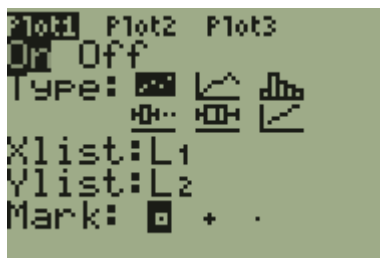
- a. Choose: *Stat Plot* by pressing the 2nd key and then they Y= (upper left, top row) key.



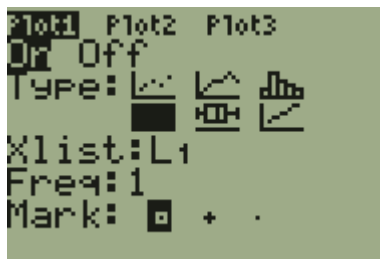
- b. Press the 1 key and then *Enter*.



- c. Use the left-arrow key to highlight *On*, if necessary, and press *Enter*.



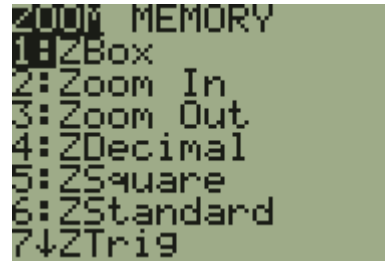
- d. Use the down-arrow key to move to *Type*. Then use the right-arrow key to choose the first box plot icon (1st item on second row). Press *Enter*.



- e. Use the down-arrow key to move to *Xlist*. Enter the correct list if it is not already shown.
- f. Ignore *Mark*, one should be selected. This is the symbol that will be used to display any outliers.

3. Display the boxplot

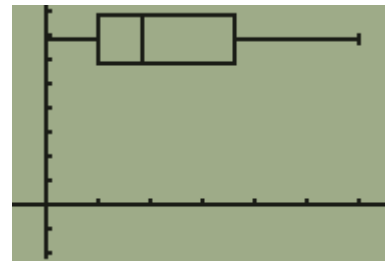
a. Press the *Zoom* key (top row, middle).



b. Use the down-arrow key to scroll down to 9 (ZoomStat).



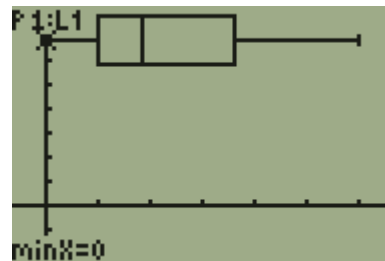
c. Press *Enter*. The boxplot is displayed.



4. Investigate the boxplot

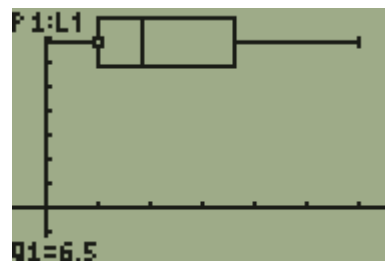
a. Press the *Trace* key (top row). This will place an icon on the extreme left of the boxplot, either an outlier or the end point of the left whisker.

In this case, the text below tells us that endpoint of the left whisker is the minimum data value and has the value 0.



b. Use the right-arrow key to scroll to the right. Information on the next item is displayed.

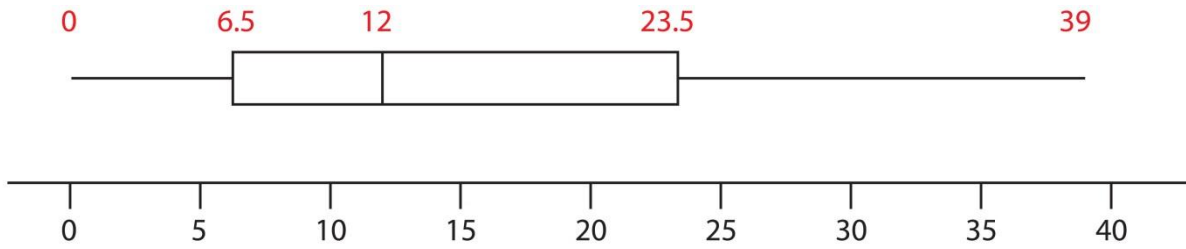
In this case, the text below tells us that the first quartile is 6.5.



c. Continue scrolling to the right to reveal the value of the median, third quartile, endpoint of the right whisker, and outliers if there are any.

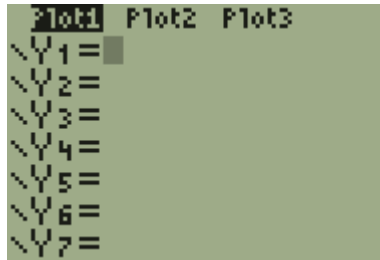
5. Suppose I ask you to draw a *boxplot* with the axes properly labeled from this graph.

Tracing the boxplot, we arrive at:



6. If your graph does not display correctly, try these things to fix it:

- a. Press the $Y=$ key (top row). There should be nothing to the right of any of the equals signs. If there is, delete it.



- b. Choose *Stat Plot* by pressing the 2^{nd} key and then the $Y=$ key. Make sure Plot 1 is *On* and all other Plots are *Off*.



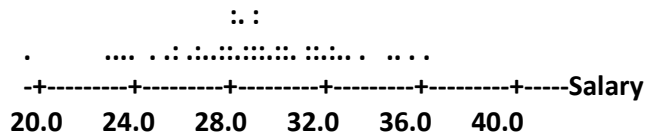
Homework 2.15

1. Consider the data from Homework 2.3. Use your calculator to make a box plot of each data set.

2.19 – Other Graphical Techniques

1. **Dotplot** - Plot the data on the real number line using a circle or an 'X'. Stack any repeat data values.

Example: Salary data previously considered.



2. **Stem-and-Leaf Plot** - A graphical representation of the data that preserves the actual data values.

Steps:

1. Sort the data.
2. Choose stems, arrange vertically in smallest to largest.
3. Attach leaves to stems smallest to largest.

Example - Salary data previously considered.

Stem	Leaves
1	9
2	445567778999
3	00000111122222223333333445556666799
4	01

2.20 – Z-scores

1. **Z-score** – A z-score can be calculated for each data value. It is a way of *standardizing* a data value so that we can get a relative view of how large or small it is. If you calculate the z-score for each item in a sample, then the resulting z-scores will have mean 0 and standard deviation 1. In general, we can calculate a z-score for each data value:

$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, 2, \dots, n$$

Note that we take each data value, subtract the *center* (\bar{x}) and divide by the *spread* (s).

2. **Interpretation:** a *z-score* is the *number* of standard deviations that a data value is away from the mean. If the z-score is *negative*, then it tells the number of standard deviations that the data value is *below* the mean. If the z-score is *positive*, then it tells the number of standard deviations that the data value is *above* the mean.

Example – Suppose a data set has an average of 20 and standard deviation of 5.

- The data value 30 has z-score of 2: $z = \frac{x - \bar{x}}{s} = \frac{30 - 20}{5} = 2$

Thus, we see that the value 30 is 2 standard deviations above the mean:

$$30 = \bar{x} + 2s = 20 + (2 * 5).$$

- The data value 15 has z-score -1: $z = \frac{x - \bar{x}}{s} = \frac{15 - 20}{5} = -1.$

Thus, the data value 15 is 1 standard deviation below the mean:

$$15 = \bar{x} + 1s = 20 - (1 * 5).$$

3. **Z-scores and Outlier Detection** – If a z-score is greater than three in absolute value, then it is a possible outlier.

If $|z| > 3$ then x is a *Possible Outlier*.

Where does this rule come from? If the data is bell-shaped, then the Empirical Rule applies. In this case, a z-score greater than two in absolute value may be an outlier as we remember that the Empirical Rule states that only 5% of the time will a value fall outside $\bar{x} \pm 2s$. Also, remember that only 0.3% of the time will a value fall outside $\bar{x} \pm 3s$. Thus, a z-score greater than 3 in absolute value is an even stronger indicator of an outlier, in the presence of symmetric data. For this course, we will use the value 3.

4. Z-score Examples

- a. **Example** – Consider the word search problem considered previously which compares men and women to see how long they take to solve the problem. The (partial) descriptive statistics are repeated here:

Variable	N	Mean	StDev
Women	50	93.49	60.35
Men	50	121.6	76.70

The data and z-scores are shown below:

Women	Men	z-Women	z-Men	Women	Men	z-Women	z-Men
20.2	25.6	-1.2	-1.3	85.2	89.9	-0.1	-0.4
20.8	30.6	-1.2	-1.2	88.5	99.1	-0.1	-0.3
24.5	34.8	-1.1	-1.1	96.7	105.9	0.1	-0.2
24.8	35.6	-1.1	-1.1	100.0	107.8	0.1	-0.2
26.2	35.9	-1.1	-1.1	105.5	120.6	0.2	0.0
28.9	41.5	-1.1	-1.0	104.2	129.2	0.2	0.1
33.3	40.8	-1.0	-1.1	103.7	151.7	0.2	0.4
36.2	44.8	-0.9	-1.0	113.5	160.4	0.3	0.5
34.0	45.1	-1.0	-1.0	122.1	166.0	0.5	0.6
40.3	45.9	-0.9	-1.0	119.6	169.2	0.4	0.6
39.1	55.6	-0.9	-0.9	120.9	180.3	0.5	0.8
39.7	53.3	-0.9	-0.9	124.6	181.0	0.5	0.8
38.8	54.6	-0.9	-0.9	128.0	185.7	0.6	0.8
41.6	57.2	-0.9	-0.8	149.0	191.2	0.9	0.9
45.5	58.9	-0.8	-0.8	154.0	199.5	1.0	1.0
46.0	64.9	-0.8	-0.7	164.9	199.9	1.2	1.0
46.6	63.2	-0.8	-0.8	169.5	208.6	1.3	1.1
50.5	75.2	-0.7	-0.6	180.9	208.1	1.4	1.1
50.2	74.7	-0.7	-0.6	179.1	226.7	1.4	1.4
53.6	76.2	-0.7	-0.6	184.7	227.9	1.5	1.4
60.3	80.0	-0.6	-0.5	189.9	240.7	1.6	1.6
59.6	85.8	-0.6	-0.5	200.6	251.2	1.8	1.7
59.9	88.9	-0.6	-0.4	198.7	255.5	1.7	1.7
59.4	91.3	-0.6	-0.4	210.3	271.1	1.9	1.9
79.0	89.5	-0.2	-0.4	221.4	300.9	2.1	2.3

When the data is sorted (as above), it is easy to look for outliers using z-scores. We simply scan the beginning and end of the list looking for z-scores of more than 3 in absolute value. In this case, we see that there are no outliers indicated by the z-scores. We see that the smallest men’s and women’s data values have z-scores of -1.2 and -1.3, respectively. The largest data values have z-scores of 2.1 and 2.3, respectively. In this case, the results are the same as when we looked at the boxplots (Section 2.17): no outliers. In general, this is not the case. Box plots will show outliers more frequently than z-scores. For a data value to be an outlier with respect to z-scores, it must really be large (or small).

- b. **Example** - In a class of 32 students the average test score was 70 with a standard deviation of 10. Would a score of 45 be considered an outlier?
- c. **Example** – A machine produces bolts with an average diameter of 10 mm and standard deviation of 1/4 mm. (a) Would a bolt with diameter 10.5 mm be considered an outlier? (b) Would a bolt with diameter

10.8 mm be considered an outlier?

5. **Comparing Things Measured on Different Scales** - Z-scores are useful to compare things measured on different scales. For instance, suppose that one person scored a 550 on the quantitative portion of the SAT and another scored 14 on the mathematics portion of the ACT test. How would we determine who got the higher score? The solution is to use z-scores.

Similarly, how would we determine when it was much hotter it is in Fairbanks, Alaska than it is in Valdosta, Georgia, relatively. The two cities have different average temperatures (and standard deviations) so how would we compare a temperature of 75 in Fairbanks and 90 in Valdosta in May? The solution, again, is to use z-scores.

- a. **Example** - The mean score on the SAT-Quantitative is 500 with a standard deviation of 50. The mean score on the ACT-Quantitative is 11 with a standard deviation of 2. Jane scored 550 on the SAT and Mary scored a 14 on the ACT. Which person got the higher score?

$$z_{\text{Jane}} = \frac{550 - 500}{50} = 1, \quad z_{\text{Mary}} = \frac{14 - 11}{2} = 1.5$$

So we see that Jane scored 1 standard deviation above the average and Mary scored 1.5 standard deviations above the average. Thus, Mary had the higher score.

- b. **Example** - The average high temperature in Fairbanks, AK in January is -1 with a standard deviation of 10. The average high temperature in Valdosta, GA is 62 degrees with a standard deviation of 5 degrees. In relative terms, which would city would be colder: Fairbanks at -12 degrees or Valdosta at 51 degrees?

$$z_{\text{Fairbanks}} = \frac{-12 - (-1)}{10} = -1.1, \quad z_{\text{Valdosta}} = \frac{51 - 62}{5} = -2.2$$

Thus, Fairbanks is 1.1 standard deviations below the average and Valdosta is 2.2 standard deviations below the average. Thus, Valdosta is colder in relative terms.

Homework 2.16

1. Use the data from Homework 2.1 to find the z-scores of the smallest and largest data values. Are there outliers? Does this correspond with what the box plots revealed?
2. Consider this data: 4.5, 6.2, 8.2, 9.1, 10.2, 10.5, 11.7, 11.9, 13.4. The data value 8.2 is how many standard deviations away from the mean? It is above or below the mean?
3. The average on a Biology exam is 63 with a standard deviation of 10. The average on a Math exam is 148 with a standard deviation of 30. Bob scores a 73 on the Biology exam and 168 on the Math exam. Which exam did he score higher on, relative to the rest of the class?

Ch. 2 Appendix – Percentage Differences

- Introduction** – An important part of statistics is being able to communicate them in writing or verbally. An important aspect of that is the ability to compare two numbers, relatively. One way to do this is to use percentage differences.
 - For instance, we might like to make statements such as this: “The Honda Accord gets 13% *less* gas mileage than the Honda Prelude.”
- Comparing** – When we want to compare two values, say X and Y with percentage differences, we can do this in two ways. We can compare X to Y or we can compare Y to X.
 - In the example above, we are comparing the Accord *to the Prelude*. We can also compare the Prelude *to the Accord*, for instance, “The Honda Prelude gets 15% higher gas mileage than the Honda Accord.”
- Formula** – When comparing X to Y, the percentage difference is $\frac{X-Y}{Y} * 100$. Notice in the formula that you compute the *difference* between the two values and then divide by the value of the item that you are *comparing to*.
- Example** – Suppose that we have the following data on two Honda models:

Model	mpg
Prelude	32
Accord	26

Compare the gas mileage using percent differences

Prelude compared to Accord:

$$\frac{32 - 26}{26} * 100 = 23.1\%$$

The Prelude gets about 23% better gas mileage than the Accord

Accord compared to Prelude:

$$\frac{26 - 32}{32} * 100 = -18.8\%$$

The Accord gets about 19% worse gas mileage compared to the Prelude

Note that it is your choice as the researcher as to how you compare things. There will always be two ways: you can compare X to Y or you can compare Y to X. Your words must be clear about the comparison, though.

5. **Example** - Suppose a new copy of a text book costs \$200 and a used copy costs \$100. These two statements are equivalent:

- The new copy of the book costs 100% more than the used copy. $\frac{200-100}{100} * 100 = 100\%$
- The used copy of the text costs 50% less than new copy. $\frac{100-200}{200} * 100 = -50\%$

Note that sometimes, when there is a large difference between the two values we are comparing, we make slightly different, but equivalent statements:

- The new copy of the book is twice the cost of the used copy. $\frac{200}{100} = 2$
or The new copy costs two times more the used copy.
- The used copy is half the cost of the new copy. $\frac{100}{200} = 0.5$

Homework 2.17

1. Consider the data shown below. In later sections, we will define the terms *Average* and *Median*, but for now, just think of them as two descriptive statistics. (a) Compare the Average to the Median for each dataset. In other words, how much larger or smaller is the Average as compared to the Median? (b) Write a sentence for each of the four results from part a. (c) How many times larger is the Median than the Average for Sample 4? Write a sentence which describes your result.

Sample	Average	Median
1	27.4 gal	25.3 gal
2	128.6 mph	130.2 mph
3	67.33 m	63.1 m
4	\$20.2	\$28.6

2. The average time for men to complete a puzzle is 1.8 minutes and the average time for women is 1.2 minutes. Fill in the blanks to these sentences:
- Men take _____ times longer to complete the puzzle than women.
 - Men take _____ longer to complete the puzzle as compared to women.
 - Women take about _____ the time men take to complete the puzzle.
 - Women take _____ less time to complete the puzzle as compared to men.