Unit 1: Exploring One-Variable Data

Introduction:

Statistics and statistical methods let us collect, describe, analyze, and draw conclusions about data. Understanding **variability** is a key to understanding data. Imagine if a population of students who all had exactly the same height and weight. This population would have *no variability.* Obviously, this is an unrealistic population, and it would be very easy for a statistician to study. Suppose that every member of another population of homeowners all paid identical property tax, identical mortgages, made identical salaries, and all had the same opinion on a property tax increase. If this were the case for the population, a statistician could easily collect data and draw conclusions about the population – it wouldn't matter how many people the statistician surveyed, because there is no variability in this population.

It is incredibly rare to encounter a population with no variability, and in order to collect, describe, and analyze data, as well as to draw conclusions about that data, we have to understand variability. Some variations in populations may be random, while others may not, so drawing conclusions based on data is a really important (and often difficult) process. Data that we collect, and numbers in general may convey meaningful information, and they may not – and statistics can help us understand where there may be useful information in data, and where we might encounter what looks like useful information which is actually junk (we will talk more about this when we are comparing data sets).

In addition, there are two main avenues for the use of statistics: **descriptive statistics** and **inferential statistics**. Hopefully, the names of these two makes the difference between them obvious. **Descriptive Statistics** uses and analyzes data to provide information about the group from which the data was collected (it simply summarizes information about a sample). **Inferential Statistics** uses statistical tools to infer (or make conclusions) about a population by analyzing the data in a sample.

Summary:

- **Variability**: How much the data in a sample or population varies.
- **Descriptive Statistics**: Using data to summarize or describe the sample that our data is taken directly from (we are describing only the individuals we have collected data from).
- **Inferential Statistics**: Using statistical tools to infer (or make conclusions) about a population by analyzing the data in a sample.

1.1 The Language of Variation

Objectives:

- Articulate the difference between categorical and numerical data.
- Articulate the difference between discrete and continuous random variables.
- Identify categorical data, numerical data, discrete variables, and continuous variables.
- Become fluent in the language of statistics.

Similar to algebra, we do use variables to represent quantities in statistics. We will be using two main types of variables:

- **Categorical Variables** (also referred to as *Qualitative Variables*) – variables that describe categories or distinct, fixed values based on qualities or properties of the data
- **Quantitative Variables** (also referred to as *Numeric Variables*) – variables that represent numeric quantities. There are two types of quantitative variables: ***discrete*** and ***continuous***
    - **Discrete Variables** represent numbers that <u>*cannot*</u> be infinitely subdivided – for example, the number of students in a class would be a discrete variable. You could have 27 or 28 students in a class, but not 27.375 students.
    - **Continuous Variables** represent numeric quantities that can be infinitely subdivided – distances, weights, and ages are all *continuous* quantities.

A data set consisting of observations on a single attribute is a **univariate data set**. This essentially means that there is one variable quantity. A univariate data set is **categorical** (or **qualitative**) if the individual observations are categorical responses. A univariate data set is **numerical** (or **quantitative**) if each observation is a numeric quantity.

Example 1: Classify each of the following attributes as either categorical or quantitative.

a) Make and model of a car purchased by a customer.

b) State of birth for someone born in the United States.

c) Price of a pair of shoes.

d) Height of the players on a basketball team.

Answers:

a) Categorical b) Categorical c) Quantitative d) Quantitative

Example 2: Classify each of the following as either categorical or quantitative variables. For quantitative variables, identify if the variable is discrete or continuous.

a) The month in which a person is born.

b) The mass of peas harvested from a pea plant in an experiment.

c) The amount of change in a student's pocket.

d) The number of apps on students' phones.

e) The brand of shoes college athletes wear.

f) The time it takes an athlete to run 1000 meters.

Answers:

a) Categorical b) Quantitative, Continuous c) Quantitative, Discrete
d) Quantitative, Discrete e) Categorical f) Quantitative, Continuous

The Basic Language of Statistics:

Like most disciplines, there is a lot of jargon* associated with statistics, and we have to become very well versed with it in order to work with and understand statistics. The introduction to categorical and numerical variables, and discrete or continuous data is an example of this vocabulary. It is essential that you memorize the terms on the following page and get used to using them every day in class – using the right term in the right context will go a long way to helping you understand the material in statistics better.

*While jargon has a negative connotation, colloquially, it simply means the technical language associated with a field. Statistics <u>*needs*</u> very specific jargon to succinctly describe very complex ideas – this is why fields that are highly technical (law, engineering, etc.) often have *jargon* associated with them.

Definitions:

- **Population:** The entire collection of individuals or objects that are being studied.
    - **Sample:** A subset of the population, selected for study/observation.
    - **Census:** A sample consisting of the *entire population*.
- **Variable:** A characteristic whose value can change.
- **Data:** A collection of observations on a variable (or variables). The values or measurements about the variables being observed.
- **Parameter:** A characteristic or fact of a *population*.
- **Statistic:** A characteristic or fact of a *sample*.

Example 3: The student senate at a university with 16,000 students is interested in the proportion of students who favor a change in the grading system to allow for plus and minus grades (e.g. C+, C, C-, rather than just C). Four hundred students are interviewed to determine their attitude toward this proposed change.

What is the population of interest?

What group of students constitutes the sample in this problem?

Example 4: Representatives of the insurance industry wished to investigate the monetary loss resulting from fire damage to structures in Greenville, California, in the Dixie Fire of 2021. From the set of all 250 structures that burned in Greenville, 80 structures were selected for inspection. Describe the population and sample for this problem.

Example 5: The supervisors of San Mateo county are interested in the proportion of property owners who support the construction of a sewer system. Because it is too costly to contact all 100,000 property owners, a survey of 500 owners selected at random (…this will be key later) is undertaken. Describe the population and sample for this problem.

Summary:

- **Univariate Data Sets** consists of observations of one **variable**. The data can be **categorical** or **numerical** (which can also be referred to as *qualitative* or *quantitative*, respectively).
- **Categorical Variables** (also referred to as *Qualitative Variables*) – variables that describe categories or distinct, fixed values based on qualities or properties of the data
- **Quantitative Variables** (also referred to as *Numeric Variables*) – variables that represent numeric quantities. There are two types of quantitative variables: ***discrete*** and ***continuous***
    - **Discrete Variables** represent numbers that <u>*cannot*</u> be infinitely subdivided – for example, the number of students in a class would be a discrete variable. You could have 27 or 28 students in a class, but not 27.375 students.
    - **Continuous Variables** represent numeric quantities that can be infinitely subdivided – distances, weights, and ages are all *continuous* quantities.
- **Vocabulary** of statistics is vitally important, and many terms seem to describe similar quantities (**population** and **census**) for example. It is essential that you make this jargon a part of your working vocabulary so that you get a better grasp on statistics in general.

<u>Definitions:</u>

- **Population:** The <u>entire</u> collection of individuals or objects that are being studied.
    - **Sample:** A subset of the population, selected for study/observation.
    - **Census:** A sample consisting of the <u>*entire population*</u>.
- **Variable:** A characteristic whose value can change.
- **Data:** A collection of observations on a variable (or variables). The values or measurements about the variables being observed.
- **Parameter:** A characteristic or fact of a *population*.

Checkpoint, Section 1.1:

True/False Questions

1. A statistic is a characteristic of a population.

2. A sample is the set of all possible data values for a given subject under consideration.

3. A population is part of a sample.

4. A population refers to the entire set of data values for a subject under consideration; a sample is a subset of the population.

Fill-in-the-Blank Questions

1. A (parameter, statistic) _____ is a characteristic of the sample.

2. A sample of an entire population is called a (sample, population, census) _____.

3. A subset of a population is called a (census, sample, small population) _____.

Short Answer Questions

Classify each of the following attributes as either qualitative or quantitative. For those that are quantitative, determine whether they are discrete or continuous.

(a) Where you go on vacation

(b) The distance from your home to the nearest grocery store

(c) The number of classes you take per school year

(d) The tuition for your classes

(e) The type of calculator you use

(f) Movie ratings (i.e., R, PG)

(g) Political party preferences

(h) Weight of sumo wrestlers

(i) Amount of money (in dollars) won playing poker

(j) Number of correct answers on a quiz

1.1 Homework

1. A consumer group decides to perform crash tests to determine the safety of a new model of car – the 2022 Tesla Model S. To determine the severity of the damage to this type of car resulting from a 15 mph crash into a concrete barrier, the research group tests five cars of this type and assesses the damage. Describe the population and the sample in the problem.

2. The student council at SI is interested in what proportion of SI students are interested in a set of themes for Spirit Week. If there are a total of 1584 students at SI and they poll one hundred fifty students to determine their attitudes about the themes.
    a. What is the population of interest?
    b. What group of students constitutes the sample in this problem?

3. The article **"Brain Shunt Tested to Treat Alzheimer's" (San Francisco Chronicle, October 23, 2002)** summarizes the results of a study in the journal *Neurology*. Doctors at Stanford Medical Center were interested in determining whether a new surgical approach would improve outcomes for Alzheimer's patients. The process involved surgically implanting a shunt to drain excess fluid from a fluid-filled cavity that cushions the brain. Eleven patients had shunts implanted and were followed for a year; they received quarterly tests of memory function. Another group of patients were treated with the standard care protocols for Alzheimer's, this was the comparison group. After collecting and analyzing the data from the study, the researchers concluded, "…results suggested the treated patients essentially held their own in the cognitive tests while the patients in the control group steadily declined. However, the study was too small to produce conclusive statistical evidence.
    a. What were the researchers trying to learn? What questions motivated their research?
    b. Do you think that the study was conducted in a reasonable way? What additional information would you want in order to evaluate this study?

4. In a study on whether taking a garlic supplement reduced the risk of getting a cold, participants were assigned to either a garlic supplement group or a group that did not receive a supplement **("Garlic for the Common Cold", *Cochrane Database of Systematic Reviews*, 2009)**. Based on the study, researchers concluded that the proportion of people taking a garlic supplement who get a cold is lower than the proportion of people not taking the supplement who get a cold.
    a. What were the researchers trying to learn? What questions motivated their research?
    b. Do you think that the study was conducted in a reasonable way? What additional information would you want in order to evaluate this study?

5. Classify each of the following variables as either categorical or numerical. For those that are numerical, determine if they are discrete or continuous.
    a. Number of students in a class of 28 who turn their AP Stats homework in on time.
    b. Weight of the next baby born at a particular hospital.
    c. Month of birth of the next baby born at a particular hospital.
    d. Amount of fluid in mL dispensed by a machine used to fill soda bottles in a factory.
    e. Birth order classification (only child, firstborn, middle child, lastborn) of students in an AP Stats class.

6. SI Administration is interested in knowing the opinions of all 150 faculty members on their opinion of the quantity of work they assign to students. They ask every faculty member to assess how many minutes of homework they believe they assign on average every week. All teachers at the school participate in the survey.
    a. What is the variable that is being tested? Is it qualitative or quantitative? If it is quantitative, is it discrete or continuous?
    b. What is the population of interest? What is the sample?
    c. Is this sample a census? Briefly explain why it is or is not a census.

1.2 Representing Categorical Variables with Tables and Charts

Objectives:

- Interpret categorical data on a frequency distribution and bar chart.
- Construct a bar chart.
- Identify and interpret limitations of different methods of displaying categorical data.

No doubt you have seen data displayed in a huge variety of ways, and we will be discussing some of those presentation methods in this section. You have likely seen **bar graphs** and **pie charts** as methods of displaying data, and there are several others. There are many others, but we will principally emphasize the **bar chart** because of the fact that it is one of the best ways to clearly and unambiguously display data. Of course, we will also look at **tables** as a way of displaying data. All of these methods are generally referred to as **frequency distributions for categorical data**. These are just various forms of tables that show both categories and frequencies (or relative frequencies).

- **Frequency:** How often a particular category appears in a data set.
- **Relative Frequency:** How often a particular category appears in a data set *out of the total number of observations.*

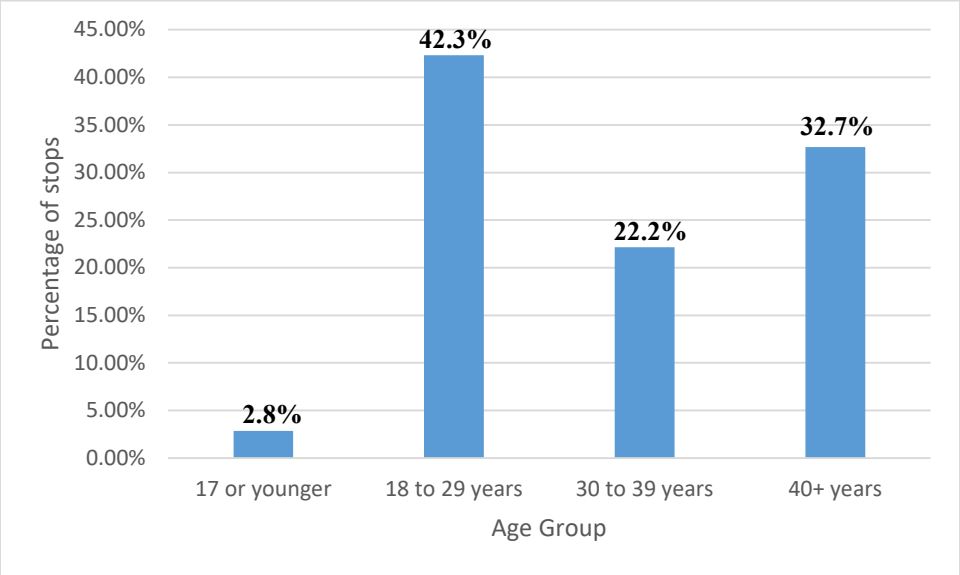
$$\text{Relative Frequency} = \frac{\text{frequency}}{\text{total number of observations}}$$

- **Frequency Distribution for Categorical Data:** A table that shows all categories for the variable along with the frequencies for each category.
- **Relative Frequency Distribution for Categorical Data:** A table that shows all categories for the variable along with the relative frequencies for each category.

Example 1: In 2014, Stanford researchers collected data on traffic stops over a 13-month time period by the Oakland Police Department. There were a total of 28,119 stops (vehicle, pedestrian, bicycle, other) that were recorded by 510 OPD officers. The following table shows the percentages of people stopped in various age groups. Below is a **frequency table** of the data.

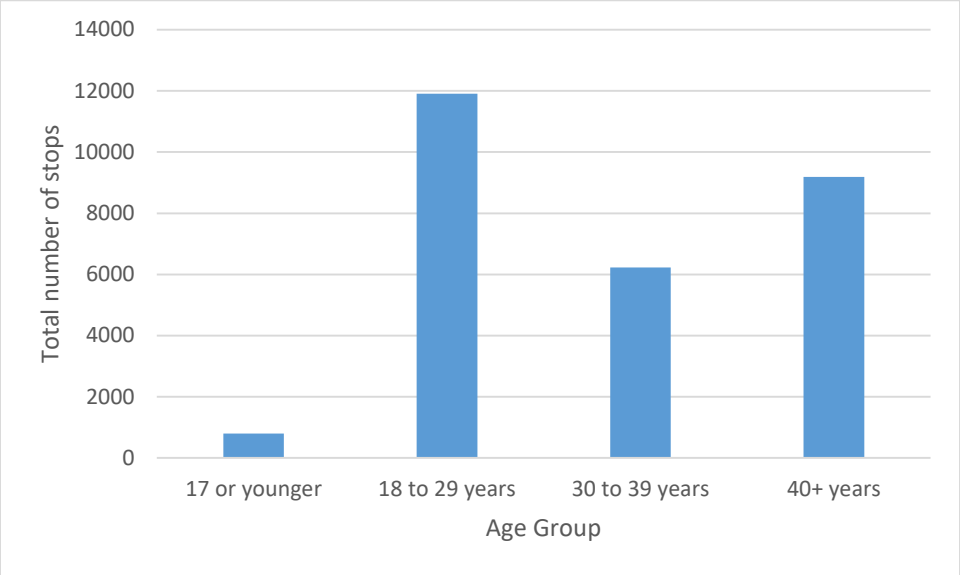
Age Group	Percentage of All Stops	Raw Number
17 years or younger	2.8%	801
18 to 29 years	42.3%	11,904
30 to 39 years	22.2%	6,229
40 years or older	32.7%	9,185

Below is a *relative frequency distribution* (because we are showing the percentages rather than the raw data.

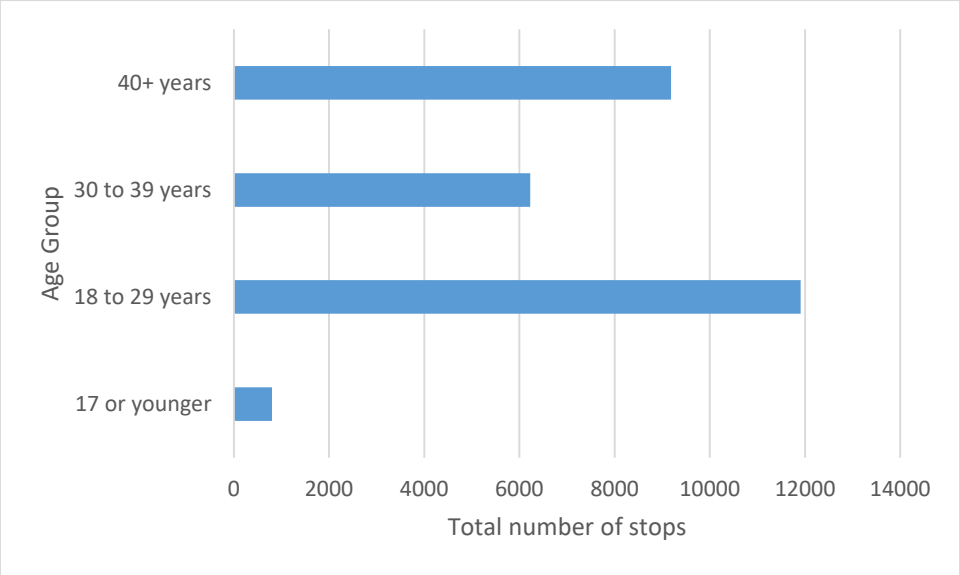


Note that we will not always have the exact percentages labeled (but you can get a good estimate by looking at the axis). In addition, a limitation of the relative frequency chart is that you cannot get an idea of the total number of observations – that is simply not displayed; you only know the amounts relative to one another.

Here is a *frequency distribution* for the same data. Note that the chart looks essentially the same, but the axes are different, because we are showing the actual numbers of stops, rather than the percentage of stops.



Bar charts are sometimes displayed horizontally as well:



Regardless of whether you draw the chart horizontally or vertically, there are several things you must do when creating a bar chart:

- Scale appropriately – if one column has four times the data of another one, the column should be four times as high (or long). The columns should retain the same width.
- Label your axes and your categories – this should be self-explanatory, make sure your data is well-labeled.
- The bars should be **separate from each other**. That is, each category should have a distinct and separate bar (this will change slightly when we start comparing multiple sets of data within a given category).

Example 2: Open your bag of candy and count the number and color of each piece of candy in the bag. Create a *frequency distribution* using the table below, and use that to draw two bar charts (one for the frequency, the second for the relative frequency).

Category							**Total**
Frequency							
Rel. Freq.							

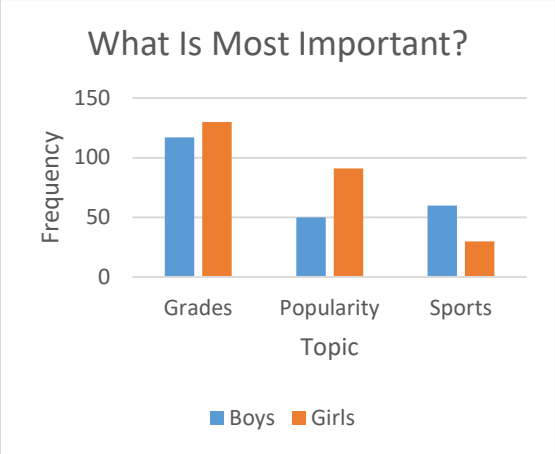
Comparative Bar Charts for Categorical Data:

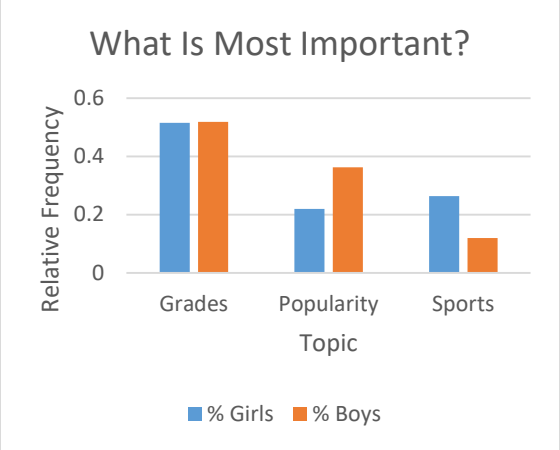
Example 2: Let's compare some groups of different sizes. Here is data from a survey of 227 boys and 251 girls in grades 4 through 6. Each student was asked what he or she thought was most important: getting good grades, being popular, or being good at sports. The resulting data, from the paper "The Role of Sport as a Social Determinant for Children" (Research Quarterly for Exercise and Sport [1992]:418 – 424), are summarized in the accompanying table:

	Boys		Girls	
Most Important	**Frequency**	**Relative Frequency**	**Frequency**	**Relative Frequency**
Grades	117	0.515	130	0.518
Popularity	50	0.22	91	0.363
Sports	60	0.264	30	0.12
Total	**227**	**0.999**	**251**	**1.001**

Let's look at the two following comparative bar charts…one constructed off the frequencies and one constructed off the relative frequencies.

Which one of these is better way of displaying the data? Explain your rationale.





Hopefully, you noticed that the overall number of boys and girls involved in the survey was **different** (227 boys and 251 girls). This would mean that the frequency chart could be deceptive – if you simply look at the bars, it looks like girls value grades much more than boys. But if you look at the relative frequency, you will notice that the percentages are much more similar. The difference you see in the frequency chart is due to the fact that more girls were surveyed.

Obviously, if you have variables you are comparing (in this case, boys and girls), you would want the two samples to be as close in size as possible to avoid this issue, but sometimes the reality of data collection constrains what we can do.

Also note that this is technically **bivariate data** (two variables are being studied). We will discuss this in much greater detail in the next unit.

There are other ways of displaying categorical data in charts. These are included for the sake of completeness – they do not show up on the AP Test (except for the occasional pie chart).

Similar to Bar Charts:

- **Stem Plot:** Essentially, just a bar chart, but it uses a line segment instead of a bar. This is hardly worth mentioning as a separate method, and it has all the same advantages as bar charts, and it is not used very frequently.

Area Based Charts (only for relative frequency):

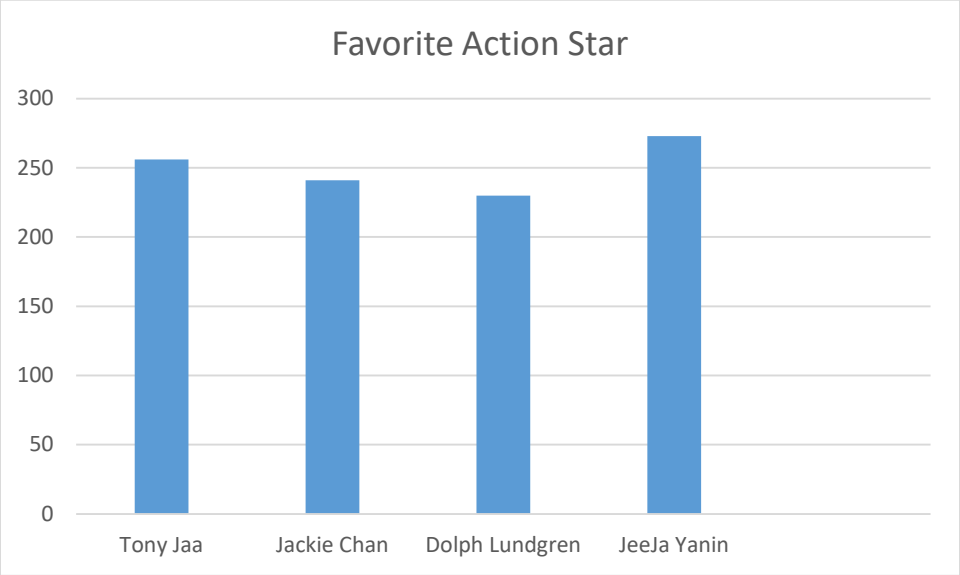
- **Pie Charts:** They can only be used to show relative frequency, and essentially they are simply circles with "slices" out of them representing the relative frequency of categories)
- There are other area-based charts, but they are never seen on the AP test, and rarely used in other contexts (if you want to look them up, look for **Treemaps** or **Waffle Charts**).

Trendy (but not terribly useful for statistical analysis):

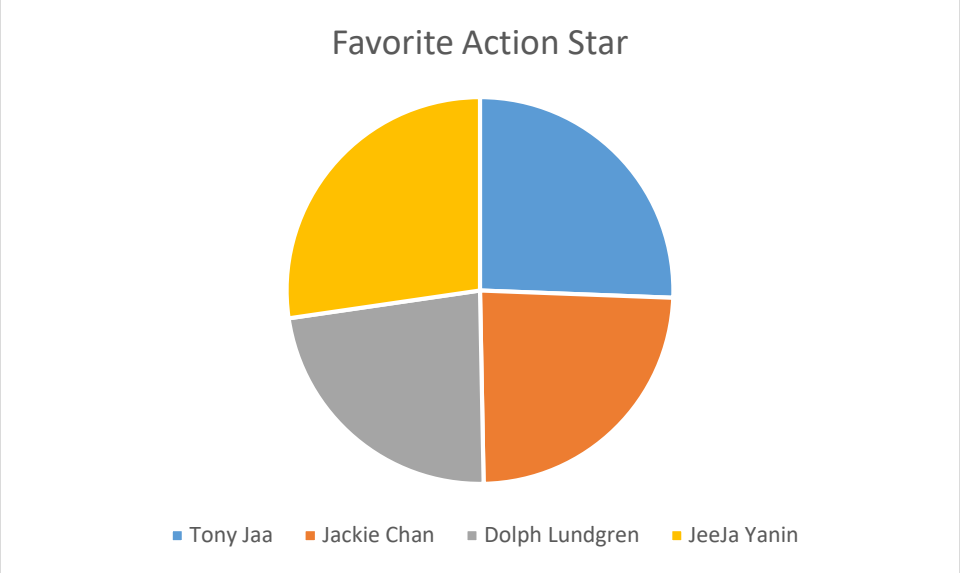
- **Word Clouds:** A currently popular method for analyzing text is the word cloud, where word frequency is counted, and then the most frequently used words are placed on a page with larger fonts.

The area-based and the word-cloud have similar limitations. If the categories have similar values, it becomes difficult to visually distinguish between the areas in question. Consider the following comparison of some fictional data displayed on both a bar chart and a pie graph.

1000 people are asked, “Who is your favorite action star?” and the results are displayed in the charts below:

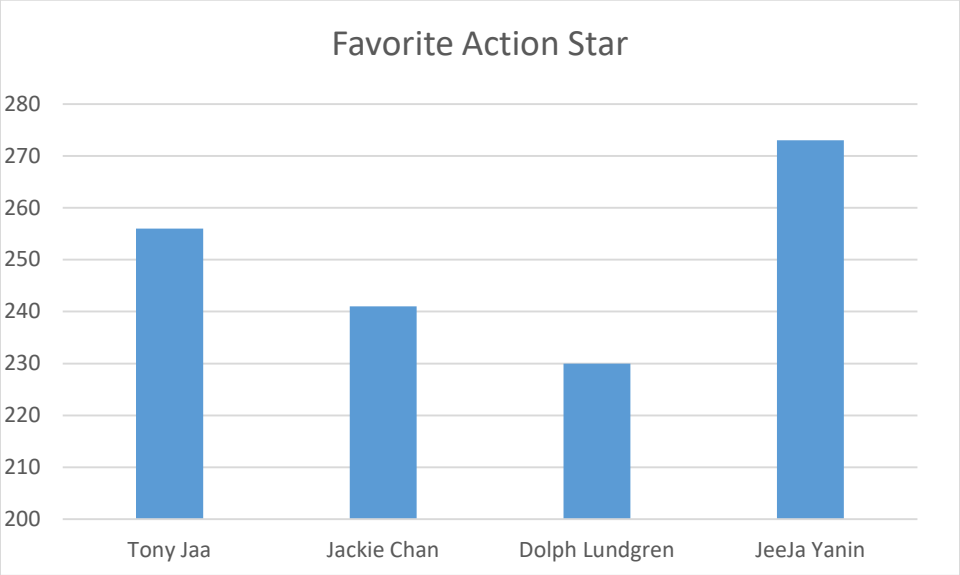


Note that it is relatively easy to assess the different heights, even though the categories are all fairly similar (the relative frequencies are 0.256, 0.241, 0.230, and 0.273, respectively). Even though all the categories are close to 25%, you can clearly see the difference, and you can tell with a fair degree of certainty how many data points are in each category.



With this pie chart, it is a bit more difficult to get an accurate assessment of comparative values, especially when the values are close. Pie charts are not good for displaying statistical data generally, and are often deceptive (whether they are intended to be or not).

This is not to say that bar graphs are perfect. Look at the same data as before, displayed on this bar graph. What do you notice in particular about this chart?



You should always be careful when both presenting and interpreting data. Presentation methods can be misleading, even if you do not intend them to be. Great care needs to be taken in how you present your data! (In fact, certain programs, like MS Excel, will sometimes automatically change the axes, assuming you want to highlight a difference that may not actually be that important. Be careful with the "automatic" features of data display software – they often make misleading graphs.)

Summary of Bar Charts and other Visual Displays:

- There are many different ways of displaying data, but many of the common ways can be misleading.
- **Frequency Distributions, Relative Frequency Distributions,** and **Bar Charts** are effective ways of displaying and reading categorical data.
- **Relative Frequency Distributions** and **Frequency Distributions** are simply the frequencies per category (or relative frequencies) displayed on a data table.

$$\text{Relative Frequency} = \frac{\text{frequency}}{\text{total number of observations}}$$

Frequency and Cumulative Frequency Distributions

Remember that *frequency* is simply the number of times a category appears in a data set. *Relative frequency* is the fraction of values from a category out of all the observed values.

Example 3: Here are some data on the average hours spent studying per day by a sample group of college students.

Average Hours of Study	Frequency
0-2	45
3-4	35
5-6	29
7-10	21

The frequency of students who study 5-6 hours a day is:

(a) 45 (b) 35 (c) 29
(d) 21 (e) 130

The relative frequency of students who study 5-6 hours a day is:

(a) 0.133 (b) 0.223 (c) 0.300
(d) 0.323 (e) 0.600

Statistics also has the terms **cumulative frequency** and **cumulative relative frequency**, where

- **Cumulative frequency** is the sum of frequencies
- **Cumulative relative frequency** is the sum of relative frequencies.

What this means is that you *add up* the values in a given range to generate the cumulative value. If you are graphing a complete cumulative frequency, then you have to add *all of the values* under a given category to generate the number for the next category.

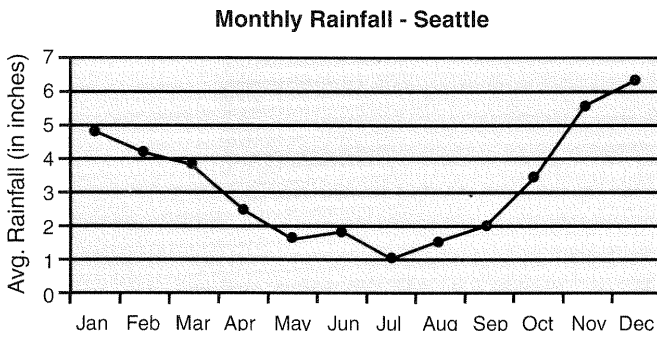
This means that cumulative frequency graphs **are always increasing!**

Example 4: Let's go back to our first example. What is the cumulative frequency of students who study 5-6 hours a day?

The cumulative relative frequency of students who study 4 hours a day **or fewer** is

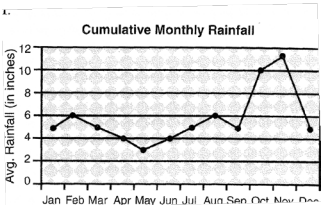
a) 0.385 b) 0.500 c) 0.615 d) 0.690 e) 0.810

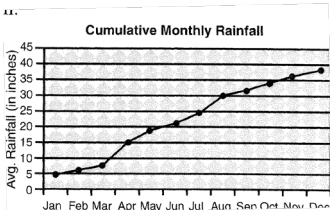
Example 5: Below is a line graph of monthly rainfall (in inches) in Seattle, WA. Note that the line goes *up and down*, so it's **not** a cumulative graph. Each heavy dot on the line shows the rainfall for the month written below that dot. The lines that connect the dots help you see the general month-to-month trends.

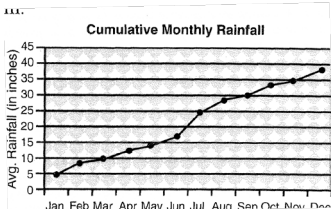


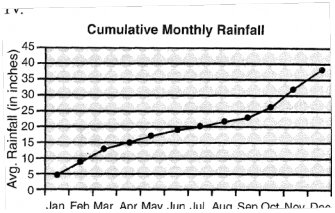
Which of the following cumulative graphs is consistent with the line graph of monthly rainfall display above?

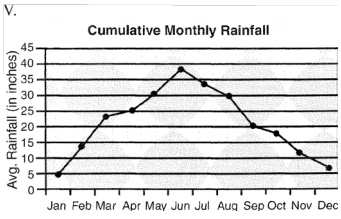
(a) Graph I (b) Graph II
(c) Graph III (d) Graph IV
(e) Graph V











Cumulative Frequency Summary:

- **Cumulative Frequency Graphs:** Add up the values for the previous categories to generate the value for the next category.
- **Cumulative Relative Frequency Graph:** This is the exact same as above, but you are adding up proportions rather than values, so the final value should always be 1 (because you have 100% of the data represented).
- **These graphs are always increasing!**
- **Interpreting intervals:** Just add up the interval in question. Remember, "at least" and "at most" include the value that was stated.
    - For example, "at least 4 hours studying" would be 4 hours or more studying.

Checkpoint 1.2

True/False Questions

1. The sum of the relative frequencies in a relative frequency distribution should always equal 1.
2. Numerical data can be placed into categories.
3. Categorical data can be classified into discrete and continuous data.

Fill-in-the-Blank Questions

1. The (frequency, relative frequency) _____ is the number of occurrences of a measurement or a data value.

2. Data such as gender, eye color, ethnicity, etc., are classified as (continuous, discrete, categorical) _____ data.

Multiple Choice

1. If the first five classes of a frequency distribution have a cumulative frequency of 50 from a sample of 58, the sixth and last class of the frequency distribution must have a frequency count of

   (a) 58 (b) 50 (c) 7 (d) 8

1.2 Homework

1. The Gallup report **"More Americans Say Real Estate Is Best Long-Term Investment" (gallup.com, April 2016)** included data from a poll of 1015 adults. Their responses to the question, "What do you think is the best long-term investment?" are summarized in the relative frequency distribution given below.

Response	Relative Frequency
Real Estate	0.35
Stocks and Mutual Funds	0.22
Gold	0.17
Savings	0.15
Bonds	0.07
Other	0.04

   a. Use this information to create a bar chart for the response data.
   b. Write a couple of sentences commenting on the distribution of responses to the question posed.

2. The National Confectioners Association asked 1006 adults the following question: "Do you set aside a personal stash of Halloween candy?" 55% of those surveyed responded no, 41% responded yes, and 4% either did not answer or said they did not know **(*USA Today*, October 22, 2009)**. Use this information to construct a pie chart to display the results of this survey.

3. A student survey on e-books was conducted in 2012 **(*The Chronicle of Higher Education*, August 23, 2013)**. 1588 students participated in the survey, and they were asked to indicate their level of agreement with the following statement:

   "I would like to be able to get all my textbooks in digital form."

   the results are summarized in the table below:

Response	Percent in Category
Strongly Disagree	19.1%
Disagree	27.5%
Agree	26.3%
Strongly Agree	16.0%
Don't Know	11.1%

   a. Create a bar chart to display this data.
   b. Write a headline that would be appropriate for a newspaper article that summarized these results.

4. During 2017, Gallup conducted a survey of adult Americans and asked the following question: "What was the main reason you decided to enroll in the school or college where you completed your highest level of education?" **("Why Higher Ed?"g Gallup, Inc., January 2018)**. The responses are summarized in the table below:

Response	Percent in Category
Location	28%
Access/Affordability	22%
School Reputation and Fit	20%
Good Job or Career	19%
Learning and Knowledge	5%
Family or Social Expectations	4%
Other/No Response	2%

   a. Create a pie chart to summarize these data.
   b. Create a bar chart to summarize these data.
   c. Which of these charts, pie chart or bar chart, best summarizes the important information? Explain your reasons.

5. Box Office Mojo (boxofficemojo.com) tracks movie ticket sales. Ticket sales for the top 5 movies of 2021 are shown in the table below:

Movie	2021 Ticket Sales (millions of dollars)
Spider-Man: No Way Home	$573.0
Shang-Chi and the Legend of the Ten Rings	$224.5
Venom: Let There Be Carnage	$212.6
Black Widow	$183.7
F9: The Fast Saga	$173.0

   a. Create a bar chart with this data.
   b. Comment on anything you note in the displayed data.

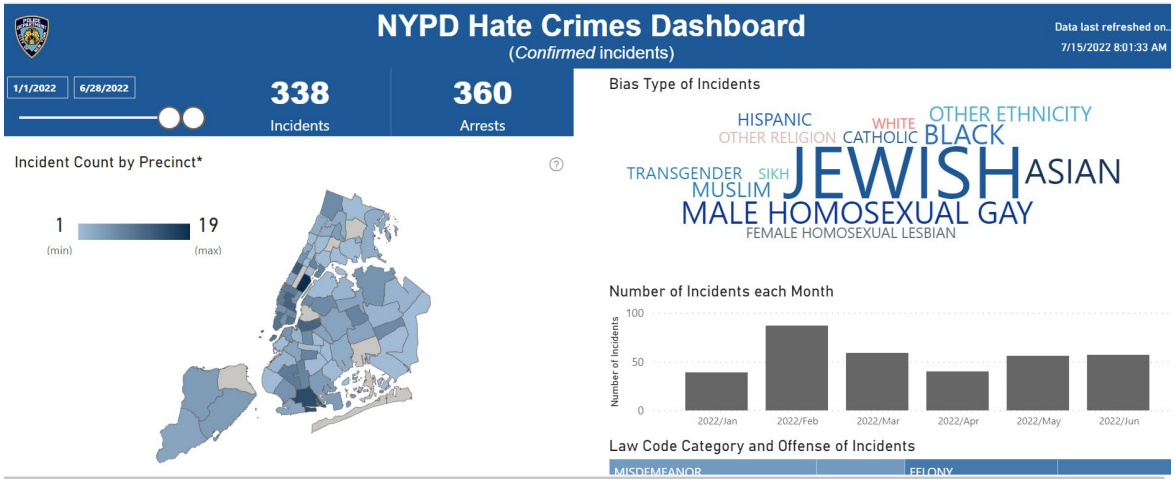
6. The following table shows the distribution of scores on a final exam for and elementary statistics course with a large section of students.

Classes for Exam Scores	Number of Students
Under 40	6
40 - 49	6
50 - 59	13
60 - 69	18
70 - 79	40
80 - 89	12
90 and above	5

a. How many total students took this final exam?
b. How many students were in the 60-69 class?
c. What is the cumulative frequency for the class 60-69?
d. What is the cumulative relative frequency for the class 60-69?
e. Create a line graph for these data.
f. Create a cumulative relative frequency line graph for these data.

A quick note on "Wordclouds" – these are becoming increasingly popular methods of displaying data, but they are profoundly limited because of a number of factors. Essentially, for Wordclouds, the font size changes based on the frequency of occurrence of keywords, but they can also be used to track other data. In fact, the NYPD Hate Crimes Dashboard has a Wordcloud for the number of incidents of Bias Type of Incidents on their home page (https://app.powerbigov.us/view?)

The Wordcloud is interactive, so hovering over the individual word will get you the actual number of incidences of crimes as well.



Let's look at a close-up of this data (current for 7/15/22):

If you look at the font, Black and Asian appear to be roughly the same size. But the actual data indicates that there were 35 incidents with Blacks as the victims of bias crimes and 51 incidents with Asians as the victims of bias crimes (the number of Jewish victims of bias crimes is 149, thus the prominent font and central location). Those two numbers seem relatively similar, but one is actually almost 1.5 times the other, and it is difficult to notice that from the font size. Further conflating the issue is the relative size of the population – for example, if the Asian population was 1.5 times the Black population, then this may not be much of a difference at all, but we do not have that information in the chart. Given that current demographics of New York City (in 2022) have roughly 20.2% of the population are black while 11.8% are Asian, we can see that this Wordcloud does not allow us a good analysis of the data (the Jewish demographic is roughly 13% of the population by way of comparison). I am sure this is not an intentional distortion of data, but it does show the limitation of using something "popular" to display important data.

1.3 Representing Quantitative Variables with Graphs

Objectives:

- Identify continuous and discrete variables.
- Identify the different ways of displaying quantitative data.
- Construct a stem-and-leaf plot.
- Construct a dotplot.
- Construct a histogram.

Recall that for quantitative (numerical) data can be categorized as *discrete* or *continuous.*

- A **continuous variable** can encompass all values in a given range (it is infinitely sub-divisible). Things like height, weight, length, and time are continuous variables. Often, if a quantity is *measureable*, it is continuous.
- A **discrete variable** can only have fixed (often, but not always, integer) values. Things like the number of students in a room, number of mistakes on a test, or the amount of change in a pocket would be discrete variables. One easy way to recognize this is if a quantity is *countable*, it is discrete.

Review Example 1: Identify the following as discrete or continuous variables.

a) The number of kittens in a litter of newborn kittens.

b) The volume of water in milliliters in a test tube.

c) The height a roller coaster car has ascended on its tracks.

d) The mass of each kitten in a litter of kittens.

e) The number of crows in a murder of crows.

f) The cost of a new pair of shoes.

Answers:

a) Discrete b) Continuous c) Continuous d) Continuous
e) Discrete f) Discrete

We are going to look at several ways of displaying these numerical variables. The key with all of these ways is that the give us a look at the general "shape" of the data – where there are gaps, peaks, large groups or clusters, etc. We will talk more about this analysis process in the next section. The three main ways we will look at data are as follows:

- **Dotplots:** Basically, just a series of dots *above* a number line, with each dot representing a single data point.
- **Stem-and-Leaf Display:** Arranging the data in a list where the "stem" (first portion of the number) and "leaves" (second portion of the number) are arranged in a vertical list.
- **Histogram:** Similar to a bar chart for categorical data, the numerical data is arranged in to groups and plotted as bars on an *x*-*y* axis system.

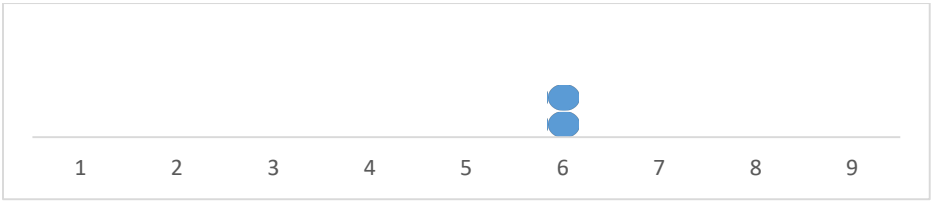
Suppose we have the set of data listed below, and we want to make a dot plot from it. It is the data from 15 randomly-selected SI seniors concerning how many honors and AP classes they took in their junior and senior years (in total):

Number of Honors/APs	0	1	2	3	4	5	6	7	8	9
Students	0	1	0	0	4	6	2	0	1	1

First we draw a number line, and label the axis:

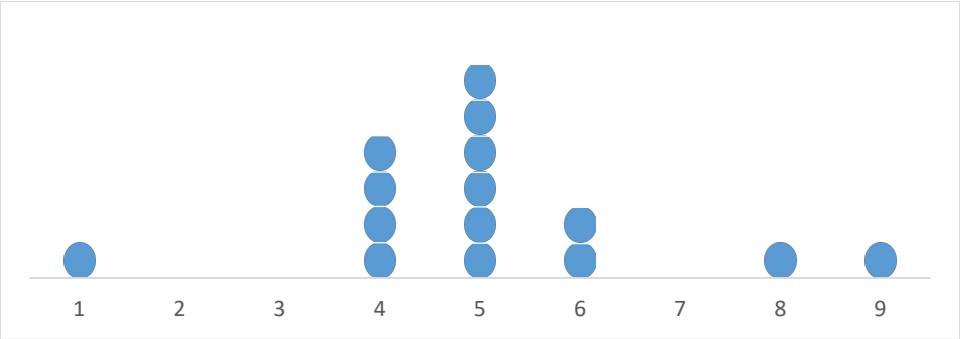
Number of AP/Honors Classes for SI Students

Then place a number of dots above each number corresponding to the number of students. For example, we have 2 students who have taken 6 classes, so we put two dots above the 6.



Number of AP/Honors Classes for SI Students

Just fill out a number of dots for each category. These dots are fairly large, but they do not have to be. The program I am using also puts them fairly close together, but a bit of space between dots is good. Try to be consistent with the space between however. For example, every column with four dots should have the same height as every other group of four dots.



Number of AP/Honors Classes for SI Students

Example 2: Now we will collect the same data anonymously from the class, and you will construct your own dotplot.

Number of Honors/APs	1	2	3	4	5	6	7	8	9	10	11	12
Students												

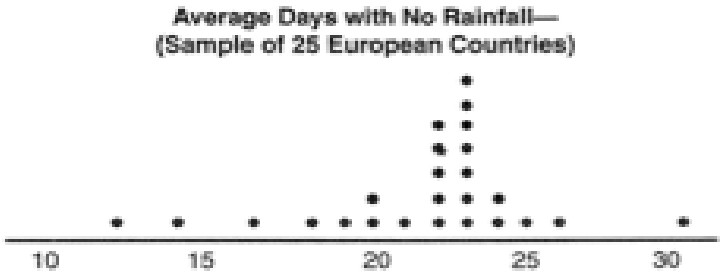
Construct the dotplot in the space provided below.

Dotplots are not incredibly common, and they are mostly used for relatively small data sets (otherwise, it gets very tedious drawing that many dots).

Dotplot Summary:

- **When to use:** Use these when we have small quantitative (numerical) data sets.
- **How to Construct One:**
  - Draw a horizontal line and mark it with an appropriate measurement scale.
  - Locate each value in the data set along the measurement scale, and represent it by a dot. If there are two or more observations with the same value, stack the dots vertically.

Example 3: Use the dotplot below to answer the following questions.



a) How many countries had 22 days with no rainfall?

b) What was the highest number of days without rainfall for any country? The lowest?

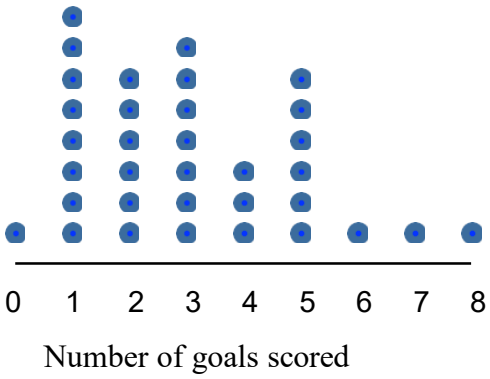
c) What was the number of days without rainfall for the largest number of countries?

d) How many countries had at least 23 days without rainfall? What percentage of the countries is this?

Example 4: How good was the 2008 US women's soccer team? Here are data on the number of goals scored by the team in each of the 34 games played during the 2008 season.

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

Here is a dotplot of the data:



What proportion of games had 4 or more goals scored?

(a) 3 (b) 9% (c) 26% (d) 35% (e) 12

Stem-and-Leaf Display:

When we have moderately-sized data sets, stem-and-leaf displays are a common way of showing the data. Essentially, you pick the highest place value digit (or digits) as the "stem" and order those vertically. Then arrange the data "leaves" (the remaining portion of the number) as lists next to the stem. It is not essential that all the "leaves" are in numeric order, but it is a nicer way to present data if the leaves are in order as well. Include a "key" that shows the units for stem and leaves somewhere in the display.

This is more easily understood by example.

Example 5: The accompanying data on daily protein intake (in grams of protein per kilogram of body weight) for 20 competitive athletes was obtained from a plot in the article "A Comparison of Plasma Glutamine Concentration in Athletes from Different Sports" (Medicine and Science in Sports and Exercise [1998]:1693 – 1697):

1.4	2.2	2.7	1.4	2.3
1.7	2.3	1.5	1.8	2.8
1.8	1.9	2.0	2.3	1.5
1.9	1.7	1.8	1.6	3.0

I tend to order the values in a list. I would choose the "stem" to be the ones digit, and the "leaves" as the tenths digit.
1.4, 1.4, 1.5, 1.5, 1.6, 1.7, 1.7, 1.8, 1.8, 1.8, 1.9, 1.9, 2.0, 2.2, 2.3, 2.3, 2.3, 2.7, 2.8, 3.0

Stem	Leaves
1	445567788899
2	0233378
3	0

Key: 1|4 =1.4 grams of protein/kg body mass

Or

Stem: Ones
Leaves: Tenths

Some people will put commas in between the individual data points, but generally if you have single digits, you simply write a list.

This particular display doesn't really differentiate enough, because there is a lot of data for each stem. So we can break up the stem into "Low" (leaves = 0,1,2,3,4) and "High" (leaves = 5,6,7,8,9).

Stem	Leaves
1L	44
1H	5567788899
2L	02333
2H	78
3L	0

Key: 1|4 =1.4 grams of protein/kg body mass

Or

Stem: Ones
Leaves: Tenths

Example 6: Below is a list of in-state tuition average for 12 states. We will show a couple of different ways to make a stem and leaf plot with this data:

9,179	6,880	11,106	7,011	9,263	10,701
6,262	6,944	7,577	11,669	11,504	11,321

If we use the Stem = Thousands, we would need 2 digit stems, and we never skip spots in our stems, even if there is no data in that part of the display (it is important to see these gaps).

Stem	Leaves
06	880, 262, 944
07	011, 577
08	
09	179, 263
10	701
11	106, 669, 504, 321

Stem: Thousands

Leaves: Ones

Notice that we did not put the leaves in order, and that the "ones" for the leaves consist of three digits. We separated the leaves with commas because of that.

We could also "truncate" the data – that is, drop off the end numbers and reduce the leaves to single digits. Very little is lost in treating the data this way, though we would need to acknowledge that we truncated the data. We could also round normally, but this is less commonly done in statistics.

Stem	Leaves
06	829
07	05
08	
09	12
10	7
11	1653

Stem: Thousands

Leaves: Hundreds

Notice that now, the 09 stem (for example) would refer to the quantities \$9,100 and \$9,200 – we do not have the same actual numbers that are in our original data, but it is a quick and easy way to display large numbers efficiently.

Oftentimes, however, we have two data groups that we want to compare, so it is helpful to be able to make one stem-and-leaf display with two data sets. One set will be on the left of the stem, the other will be on the right of the stem.

Back-to-Back Stem-and-Leaf Displays for Numerical Data

Example 7: The data below show the standardized (converted to a common scale of births for every 1000 females ages 15-19) birth rates for teenagers in 2008 by state. The states have been divided into east and west. The following back-to-back stem plot is shown on the right.

Western States	Birth Rate	Eastern States	Birth Rate
Alaska	55	Alabama	64
Arizona	71	Connecticut	45
Arkansas	78	Delaware	63
California	57	Florida	59
Colorado	57	Georgia	69
Hawaii	45	Illinois	59
Idaho	55	Indiana	63
Iowa	57	Kentucky	68
Kansas	64	Maine	42
Louisiana	59	Maryland	43
Minnesota	54	Massachusetts	48
Missouri	62	Michigan	55
Montana	44	Mississippi	62
Nebraska	60	New Hampshire	36
Nevada	75	New Jersey	37
New Mexico	59	New York	41
North Dakota	42	North Carolina	72
Oklahoma	70	Ohio	60
Oregon	65	Pennsylvania	56
South Dakota	46	Rhode Island	64
Texas	64	South Carolina	58
Utah	52	Tennessee	70
Washington	58	Vermont	32
Wyoming	49	Virginia	50
		West Virginia	52
		Wisconsin	55

Western States	Eastern States
	|3|2
	|3|67
42	|4|123
965	|4|58
42	|5|02
99877755	|5|556899
4420	|6|023344
5	|6|89
10	|7|02
85	|7|

Note that when we observe the back-to-back stem-and-leaf display, we can more easily compare the two datasets. We will talk more about how to compare data sets and analyzing their shapes in the next section.

To create the back-to-back display, just add a "leaf" column on the left side of the "stem" as well. Make sure to label both leaves, and provide a key as well.

Stem-and-Leaf Display Summary:

- **When to use:** Numerical data sets with a small to moderate number of observations (does not work well with very large data sets).
- **How to Construct:**
  - Select one or more leading digits for the stem values. The trailing digits (or sometimes just the first one of the trailing digits) become the leaves.
  - List the possible stem values in a vertical column.
  - Record the leaf for every observation beside the corresponding stem value.
  - Indicate the units for stems and leaves somewhere in the display (a Key).

Histograms:

Histograms are a very common way of displaying quantitative (numerical) data. They are incredibly useful for displaying large data sets. We can use them to display both discrete and continuous data.

These look very similar to bar graphs, but there are some significant and important differences. Given that the data is numerical and not categorical, we are using an *x*-*y* axis system, so the bars should be placed right next to one another, because we are looking at numbers rather than distinct categories.

There are slight variations for the process for discrete vs continuous data:

Discrete

1. Draw a horizontal axis and mark the possible values for the variable.
2. Draw a vertical axis and mark frequency/relative frequency.
3. For each possible value draw a rectangle **centered above** the axis value.

Continuous

1. Draw a horizontal axis and mark the possible values for the variable.
2. Draw a vertical axis and mark frequency/relative frequency.
4. Determine a region width and draw rectangles of the appropriate width marked on the axis.

These are most easily seen by creating some histograms.

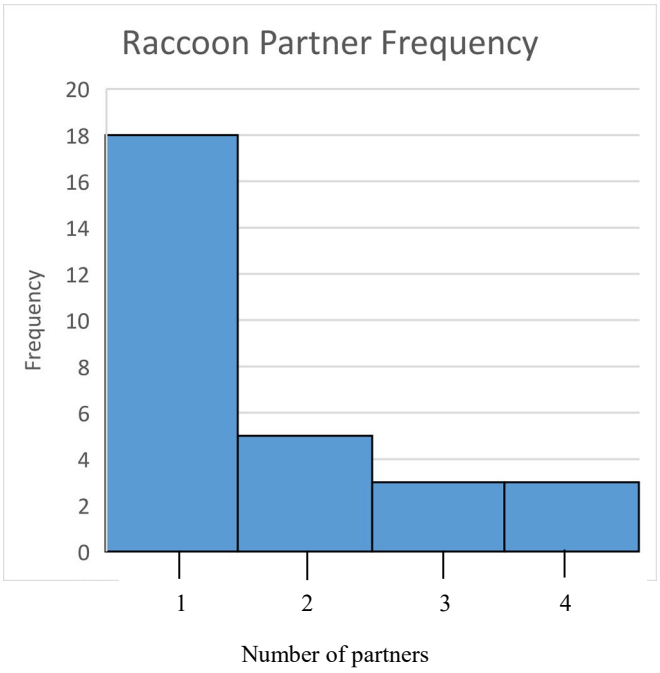
Example 8: The authors of the article "Behavioral Aspects of the Raccoon Mating System: Determinants of Consortship Success" (Animal Behaviour [1999]: 593 – 601) monitored raccoons in southern Texas during three mating seasons in an effort to describe mating behavior. Twenty-nine female raccoons were observed, and the number of male partners during the time the female was accepting partners (generally 1 to 4 days each year) was recorded for each female. The resulting data were as follows:

1	3	2	1	1	4	2	4	1	1	1	3
1	1	1	1	2	2	1	1	4	1	1	2
1	1	1	1	3							

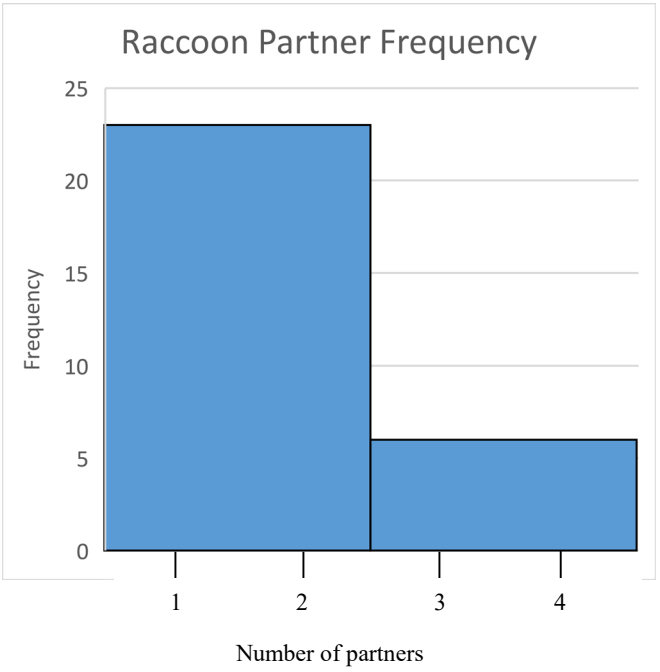
You should count up each of the discrete values:

1 partner	18 occurrences
2 partners	5 occurrences
3 partners	3 occurrences
4 partners	3 occurrences

Below is a histogram for the data:



If we wanted our bars wider, (for example, grouping 1 and 2 together, and 3 and 4 together), we would do the same process, but we would have 23 occurrences in the 1 to 2 range, and 6 occurrences in the 3 to 4 range:



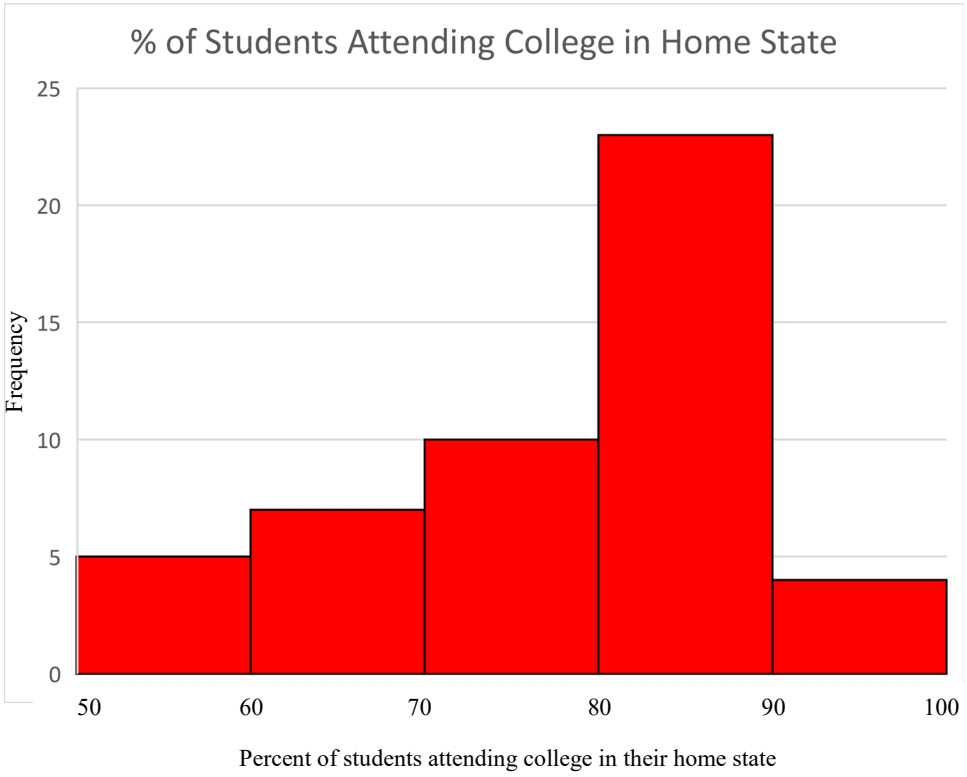
When we use continuous data to construct a histogram, the decisions for the cutoffs for where to put the data is important. For example, if we are shipping packages, and the package weights vary from 0 to 25 pounds, we might consider intervals of 5 pounds. We would generally have categories of 0 to 5 pounds, 5 to 10 pounds, etc.

But where should we put the package that is exactly 5 pounds? To clear this up, we typically have intervals that are from 0 to <5 (just less than 5), 5 to <10, etc. We use the less than sign just as a quick substitute for the phrase "less than".

So if you see data regions on a histogram, assume you go "low value" to "less than High value".

Note that these would correspond to "half-open intervals" from Precalculus/Calculus. For example, the above regions would be $x \in [0,5)$, and so on.

Example 9: Below is a histogram of the data of the percent of students (by state) who attend college in their home state.



Notice the first column – there are 5 values in it and the range is from 50% to just under 60%. That means there are 5 states that have somewhere from 50 to 60% of students from that state attending college in that state.

We can also use our calculator to generate histograms. Below is a list of maximum wind speeds in meters per second in Hong Kong for each year in a 45-year period.

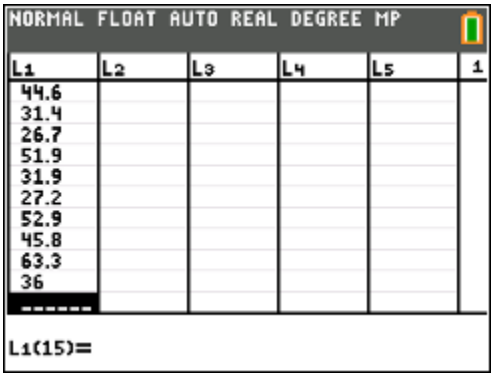
30.3	39.0	33.6	38.6	44.6	31.4	26.7	51.9	31.9	27.2	52.9	45.8	63.3
36.0	64.0	31.4	42.2	41.1	37.0	34.4	35.5	62.2	30.3	40.0	36.0	39.4
34.4	28.3	39.1	55.0	35.0	28.8	25.7	62.7	32.4	31.9	37.5	31.5	32.0
35.5	37.5	41.0	37.5	48.6	28.1							

Enter these values into a list on your calculator:

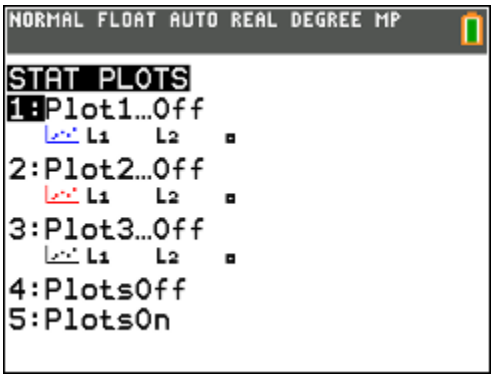
- Press the STAT key
- Select 1: Edit…
- Start typing in values in the list, hitting enter after each value.
- Once you have entered them, hit 2nd, then quit (the mode key)

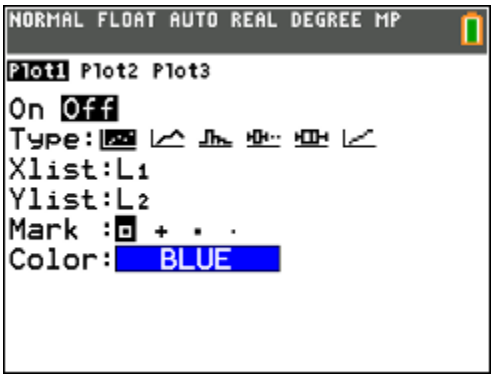
Be very careful typing in your data – it is very hard to find a mistake in a list after you have entered it. Trying to check back and forth between a list on the calculator and a table of data is very annoying!

Entering your list should look like this:

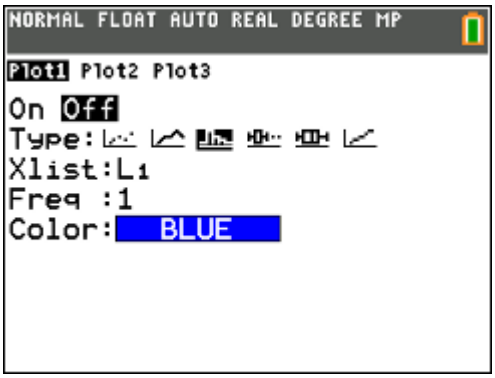


Right above the "y=" button on the calculator (on the upper left corner), is the stat plot option – hit 2nd then y= to open this menu: Then hit enter on 1:





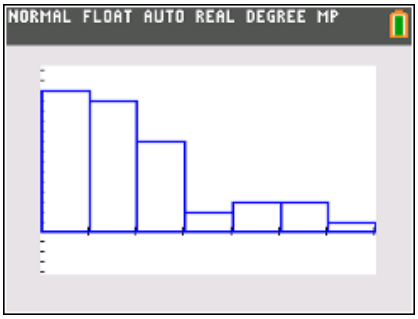
Switch the plot to "On" by moving the cursor over and hitting enter, then move the cursor to the 3rd option in the "Type" row to turn on a histogram:



Notice that it defaulted to L_1, which is where we had entered our data. In order to graph this, we would just hit the graph button, like any other function, but with Statistics, the window is very important, so we have to go to the "zoom" menu and use option 9: Zoom Stat

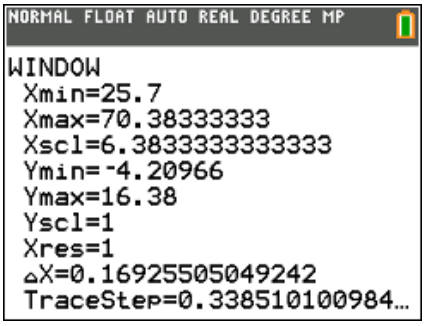
Important: Always use 9:Zoom Stat first!

You can adjust the window afterwards if necessary, but getting the right initial window saves a lot of time and trouble.



This gave us a relatively nice looking histogram, but we should check out what the calculator chose as column widths. It automatically starts at the minimum value and makes intervals based on the range (from min to max), so our intervals will sometimes be a bit weird.

If we hit the window key, we see the following:



Seeing this, I might decide to have my histograms go from 20 to 70 (since none of my values exceed 70) with an interval width of 10. Here's what I'd enter to do that:

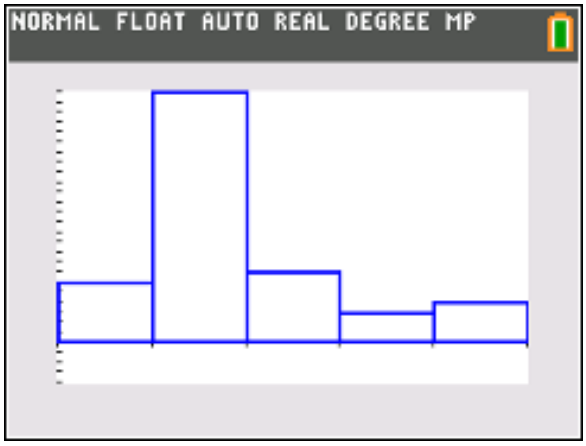
Xmin=20

Xmax=70

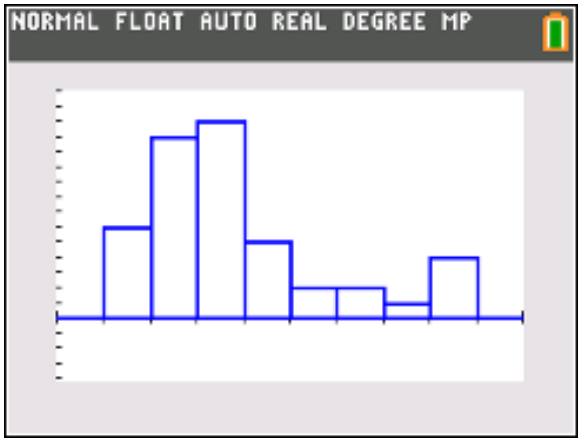
Xscl=10

As this could change the heights of some of my bars, I may need to adjust the Ymax as well. I set mine to 25.

Here is the histogram we get as a result:



But suppose I want my column widths to be 5 rather than 10? I will simply go into my "window" menu and change the Xscl (*x*-scale) to 5 instead. This gives me the histogram below (I also lowered my Ymax to 15, because having narrower columns mean that fewer values will be in each column):



Histogram Summary:

- **When to use:** Numerical data sets with a large number of observations.
- **Calculators:** Get familiar with the LIST and StatPlot functions, and ZoomStat is really important.
- **Remember:** There are subtle differences between the process for discrete and continuous data.
    - **Discrete:** The column is centered over its value on the *x*-axis.
    - **Continuous:** The column widths are from the low value to just less than the high value (from 5 to <10, for example)

Checkpoint 1.3

Multiple Choice

1. Which of the following statements are true?

   I. Stem-and-leaf displays are useful both quantitative and categorical data sets.
   II. Stem-and-leaf displays are equally useful for small and very large data sets.
   III. Stem-and-leaf displays can show symmetry, gaps, clusters, and outliers.

(a) I only (b) II only (c) III only (d) I and II (e) I and III

Questions 2 and 3 refer to the plot below:

.40	33
.41	0
.42	6778
.43	278
.44	667
.45	99
.46	8

This stem-and-leaf plot is for the slugging percentage of all National League teams as of midseason in 2007, shown here as proportions.

2. What values for slugging average have a frequency of at least 2? Note that the definition of slugging percentage is given below:

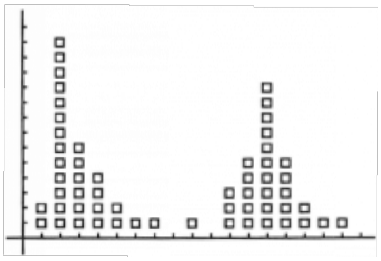
$$\text{Slugging percentage} = \frac{\#\text{ of bases reached}}{\text{Total \# of times at bat}}$$

Where the "# of times at bat" has a specific definition that excludes walks and certain other situations.

(a) .432 only
(b) .403 and .427
(c) .403, .427 and .459
(d) .403, .427, .459 and .466
(e) .403, .427, .459 and .446

3. How many baseball teams are there in the National League?

(a) 7
(b) 14
(c) 16
(d) 23
(e) You cannot tell from this display.



4. The dot plot shows a U-shaped distribution with which of the following characteristics?

   I. Bimodal
   II. Symmetric
   III. Gaps

(a) I only
(b) II only
(c) I and III
(d) I and II
(e) I, II, and III

1.3 Homework

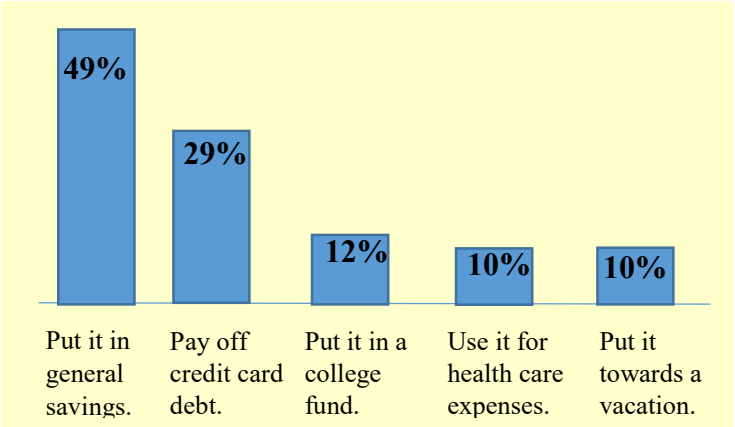
1. For the following numerical variables, determine if they are discrete or continuous.
   a. The number of thefts from a grocery store in a given week.
   b. The amount of weight that a 2-pound package of chicken decreases due to moisture loss before it is sold.
   c. The number of unpopped kernels of popcorn in a bag of microwave popcorn.
   d. The number of students in an AP Statistics class who are playing video games on an iPad rather than paying attention.
   e. The amount of time in a given 60-minute period that a randomly selected AP Statistics student is playing an iPad game.

2. In a survey of 150 people who recently purchased motorcycles, the following information was collected:

   Brand of motorcycle purchased
   Price of motorcycle with tax included
   Number of motorcycles previously purchased
   Color of motorcycle purchased
   Weight of motorcycle with all additional upgrades at purchase

   a. Which of these variables are categorical?
   b. Which of these variables are numerical and continuous?
   c. Which of these variables are numerical and discrete?
   d. Which type of graphical display would be more appropriate for the brand of motorcycle, a bar chart or a dotplot? Explain why.
   e. Which type of graphical display would be more appropriate for the number of motorcycles previously purchased, a bar chart or a histogram? Explain why.

3. Suppose you collected data from a group of 750 adults answering the question, "If you were given $1000, what would you do with it? You took the data that you collected and presented it in the bar graph below:



   a. Is the response to the question a qualitative or quantitative variable?
   b. Explain why a bar chart rather than a dotplot or histogram was used for this data.
   c. You must have made a mistake displaying your data with this bar chart. Explain how you know there is a mistake.

4. Below is a table of gas taxes by state rounded to the nearest tenth of a cent.

State	Gas Tax	State	Gas Tax	State	Gas Tax
Alaska	9.0	Massachusetts	24.0	Montana	32.8
Hawaii	16.0	Wyoming	24.0	Wisconsin	32.9
Virginia	16.2	Kansas	24.0	Idaho	33.0
Missouri	17.4	Arkansas	24.8	Florida	34.4
Mississippi	18.4	Connecticut	25.0	Rhode Island	35.1
New Mexico	18.9	Kentucky	26.0	West Virginia	35.7
Arizona	19.0	Michigan	26.3	Oregon	36.0
Oklahoma	20.0	Tennessee	27.4	North Carolina	36.4
Texas	20.0	Georgia	27.9	Maryland	36.9
Louisiana	20.1	Minnesota	28.6	Ohio	38.5
South Carolina	22.8	Indiana	30.0	Illinois	39.1
Delaware	23.0	South Dakota	30.0	New York	40.5
North Dakota	23.0	Iowa	30.5	New Jersey	41.4
Colorado	23.3	Nebraska	30.6	Washington	52.0
Nevada	23.8	Vermont	30.7	California	53.3
New Hampshire	23.8	Maine	31.4	Pennsylvania	58.6
Alabama	24	Utah	31.8		

a. Create a stem-and-leaf plot of the gas taxes using stem of tens.
b. Create a truncated stem-and-leaf plot using the stem of tens.
c. Use your calculator to make a histogram of the gas taxes using 10 cent intervals. Sketch the histogram.
d. Use your calculator to make a histogram of the gas taxes using 5 cent intervals. Sketch the histogram.

5. Below is a table of scores (out of 20 points) for an AP Statistics Test.

15	14	9
16	8	3
12	20	10
4	8	18
9	14	11
18	12	9
12	7	10
18	11	4
17	9	19
8	5	20
		15

a. Create a dotplot of this data.
b. Create a stem-and-leaf display of this data using stems of 0L, 0H, 1L, 1H, and 2L.
c. Use your calculator to create a histogram of the data using interval widths of 4.99, and starting at 0. Sketch the histogram.
d. How does the histogram compare to the stem-and-leaf display?

6. The data on annual maximum wind speed in meters per second in Hong Kong for each year for the past forty-five years are given below. These data appeared in the journal ***Renewable Energy*** **(March, 2007).**

30.3	39.0	33.6	38.6	44.6	31.4	26.7	51.9	31.9	27.2	52.9	45.8	63.3
36.0	64.0	31.4	42.2	41.1	37.0	34.4	35.5	62.2	30.3	40.0	36.0	39.4
34.4	28.3	39.1	55.0	35.0	28.8	25.7	62.7	32.4	31.9	37.5	31.5	32.0
35.5	37.5	41.0	37.5	48.6	28.1							

   a. Create a histogram for this data.
   b. Is there a potential problem with this dataset if it is being used to advocate for power generation using wind turbines in Hong Kong? What data would you have collected?

7. The accompanying relative frequency table is based on data from the **2016 College Bound Seniors Report (collegeboard.org).**

Critical Reading SAT Exam Scores	Relative Frequency for Male Students	Relative Frequency for Female Students
200 to <300	0.049	0.038
300 to <400	0.151	0.156
400 to <500	0.308	0.332
500 to <600	0.285	0.286
600 to <700	0.155	0.144
700 to <800	0.052	0.044

   a. Create a relative frequency histogram for SAT critical reading scores for male students.
   b. Create a relative frequency histogram for SAT critical reading scores for female students.
   c. Based on the histograms you created from a. and b., comment on similarities and differences in the distributions of the scores for females and males.

1.4 Describing the Distribution of a Quantitative Variable

Objectives:

- Identify features of datasets
- Describe distributions for qualitative variables using statistical language.
- Describe histograms, dotplots, and stem-and-leaf displays using Shape, Outliers, Center, and Spread ("SOCS").

When we started graphing in the previous section, we mentioned several times that we would be looking at these graphs to try and interpret the datasets that we used to construct them. The AP test (and statisticians in general) has a standard set of vocabulary that they use when analyzing and describing the distributions of a quantitative (numerical) variable.

Official AP Stats vocabulary for properly describing statistical plots - ALWAYS mention these characteristics when *describing a plot*!

The four things you should always mention when describing a graph, plot of data, or quantitative variable are as follows:

- **S**hape – the general shape or "look" of the graph including skew, clusters, and gaps
- **O**utliers – points that are very far from the rest of the dataset
- **C**enter – where the graph is centered (often mean and/or median)
- **S**pread – how spread out the data is, including range, interquartile range, and standard deviation.

Always remember your "S.O.C.S."!

Shape

Here is some vocabulary we use to describe the **shape** of statistical plots.

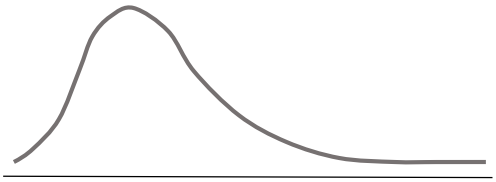
- One characterization of the general shape is related to the number of peaks, or **modes**.
  - **unimodal** – single peak
  - **bimodal** – two peaks
  - **multimodal** – many peaks
- A plot is **symmetric** if there is a vertical line of symmetry such that the part of the plot to the left of the line is a mirror image of the part to the right.

Proceeding to the right from the peak of a unimodal plot, we move into what is called the **upper tail** of the plot. Going in the opposite direction moves us into the **lower tail**.

If a unimodal plot is not symmetric, then we call that **skewed**.

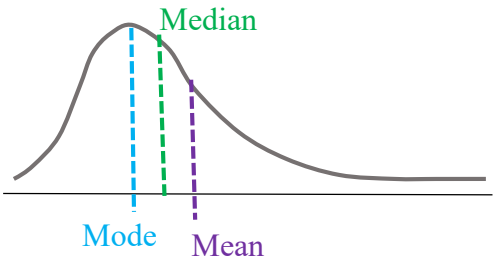
Approximately (or roughly) **Normal:** certain symmetric graphs that match up with a normal distribution are called "approximately normal" (much more on this later).

Skewed Right (or positively skewed)

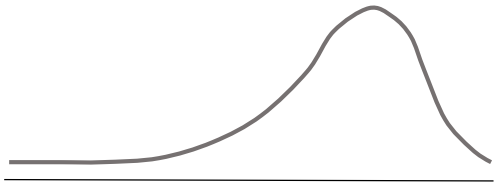


If the upper tail of the plot (the right tail) stretches much farther than the lower tail (the left tail), then the distribution is **Skewed Right** (or positively skewed).

- Mean is right of the Median
- Mode (peak) is on the left
- Median is between mean and mode

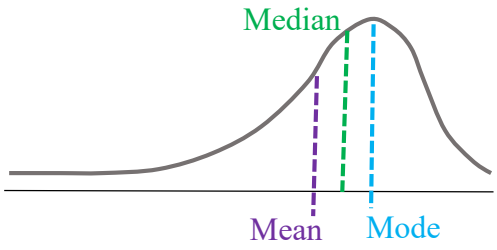


Skewed Left (or negatively skewed)



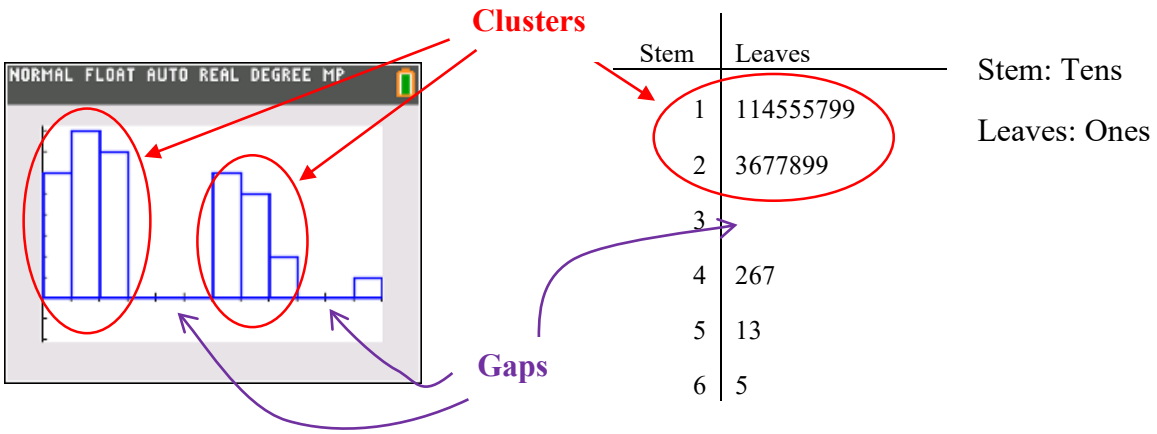
If the lower tail of the plot (the left tail) stretches much farther than the upper tail (the right tail), then the distribution is **Skewed Left** (or negatively skewed).

- Mean is left of the Median
- Mode (peak) is on the right
- Median is between mean and mode



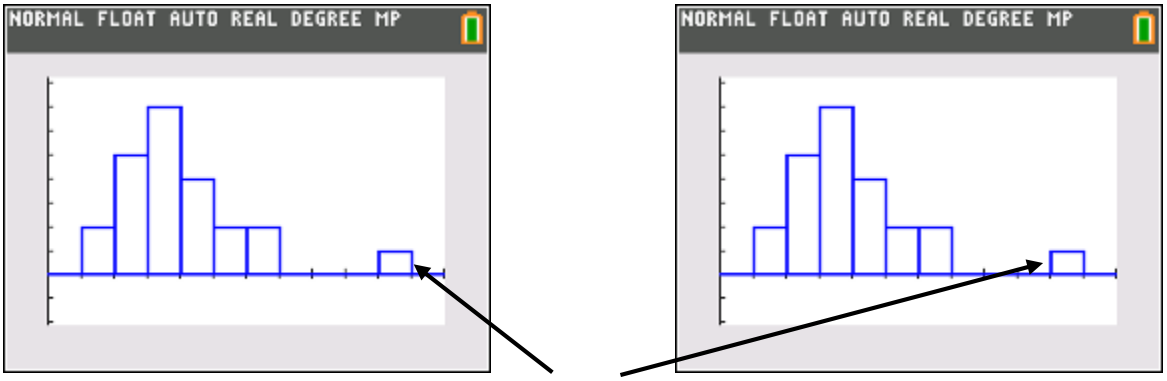
Other things that we look for in graphs are "clusters" and "gaps":

- **Gap:** a region on a distribution where there is no data.
- **Clusters:** a lot of data concentrated in a small region, often separated by gaps.



Outliers

Outliers are points that lie very far out from a data set. For example, if most of the students in a class are between 5' 4" and 6' 0", but there is one student who is 7' 2", that student's height is an **outlier**. We will address an official definition of this in a subsequent chapter. For now, if a point looks "far away" from the rest of the data, it is likely an outlier.



Likely Outliers

Center

This is simply a measure of where the data is centered – visually, this is where the majority of the data is clustered – the peak and/or near the peak of the graph. There are three main terms dealing with this:

- **Mean:** the average value from the dataset – add up all of the data-points, and divide by the number of observations. Mean is often denoted by $\bar{x}$
  - $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Median:** the middle number in a data set, even numbered set the mean is calculated by averaging the 2 middle numbers (though it could be any value in between the two numbers)
- **Mode:** the most frequently occurring value in a data set. In a graph, it is the peak (or peaks).

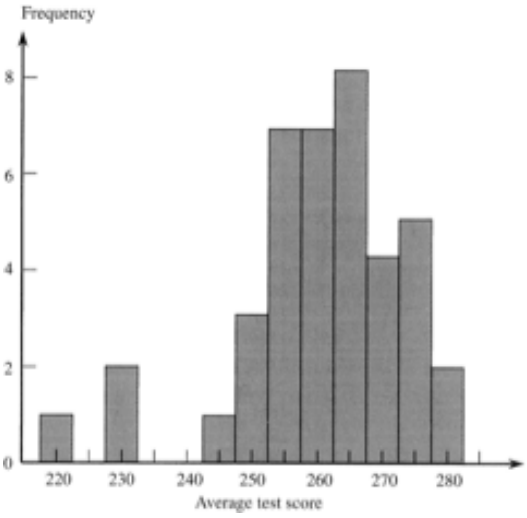
Spread

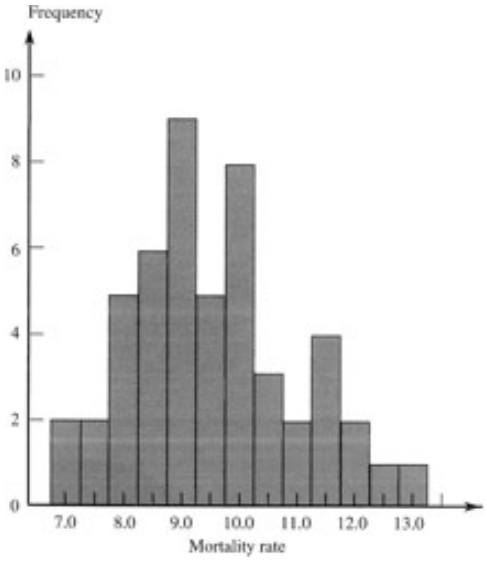
This is just a measure of how spread out the data is; that is the distance from the lowest to the highest data points.

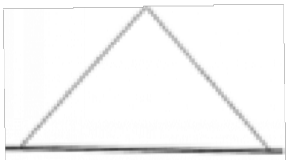
- **Spread:** the scope of values from smallest to largest; for example, "the spread of the data is 8 – 27".
- **Range:** the difference between the largest and smallest value; for example, "the range of the data is 19".
- Spread is often referred to as **variability**.

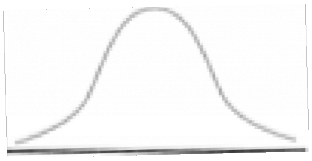
Example 1: How would you describe the **shape** of these plots:

Western States	Eastern States
	3 | 2
	3 | 67
42	4 | 123
965	4 | 58
42	5 | 02
99877755	5 | 556899
4420	6 | 023344
5	6 | 89
10	7 | 02
85	7 |

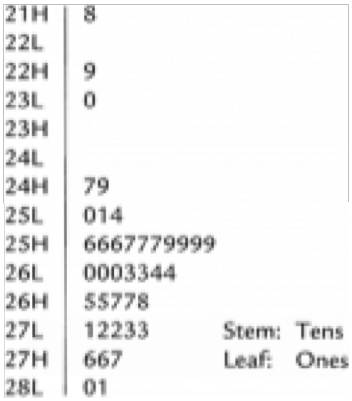






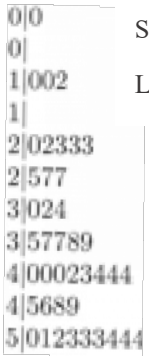


Example 2: Use SOCS to describe the shape of this plot:



This appears to be roughly symmetric (possible slight left skew), unimodal, has gaps, no outliers (218 is relatively close to the rest of the data), spread from 218 to 281, range 63

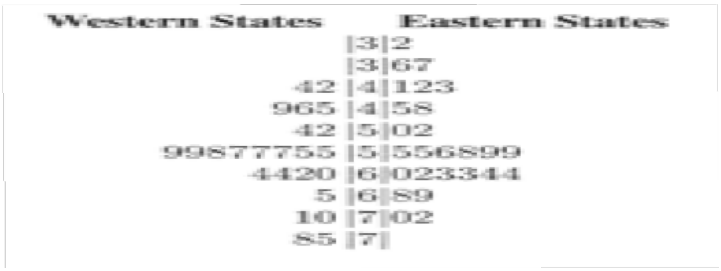
Example 3: Use SOCS to describe the shape of this plot:



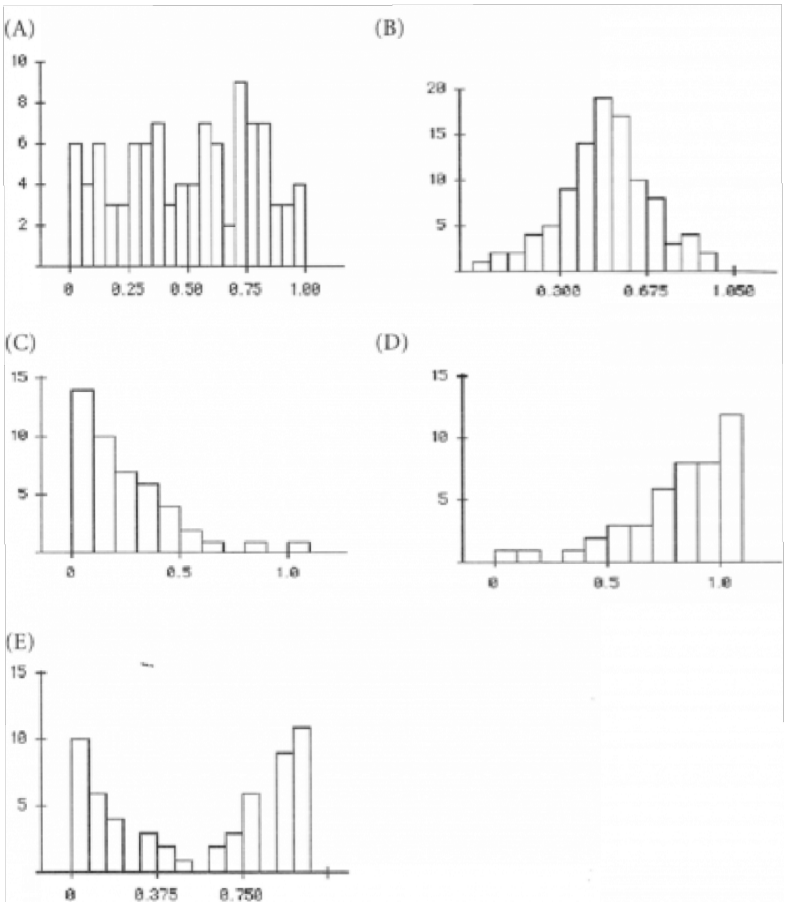
Stem: Tens

Leaves: Ones

Example 4: Use SOCS to describe the shape of this dotplot:



Example 5: For which of the following is the mean greater than the median?



Answer: **C** because this graph is positively (right) skewed, so the mean is greater than the median.

Summary:

- **Remember your "SOCS":** Use "Shape, Outliers, Center, and Spread" to describe distributions.
- **Shape:** Symmetric, Left-skewed, Right-skewed, Roughly (or approximately) Normal, Clusters and Gaps.
    - **Left-skewed**: The tail of the data is on the left.
    - **Right-skewed**: The tail of the data is on the right.
- **Outliers:** Data points that are very far from the bulk of the data.
- **Center:** Measures of center are mean, median, and mode.
- **Spread:** Measures of *variability* are center and spread.
    - **Spread:** the span of values from lowest to highest in a dataset.
    - **Range:** the difference between the highest and lowest values in a dataset.

Checkpoint 1.4

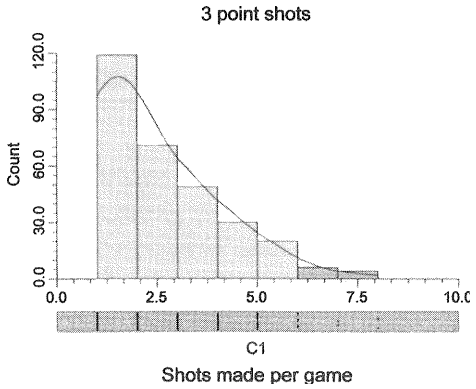
1. Which of the following are more likely to be skewed to the right than skewed to the left?
   I. Household incomes
   II. Home prices
   III. Age of teenage drivers

(a) II only (b) I and II (c) I and III (d) II and III (e) I, II, and III

2. The graphical display with the relative frequencies along the vertical axis for quantitative data is

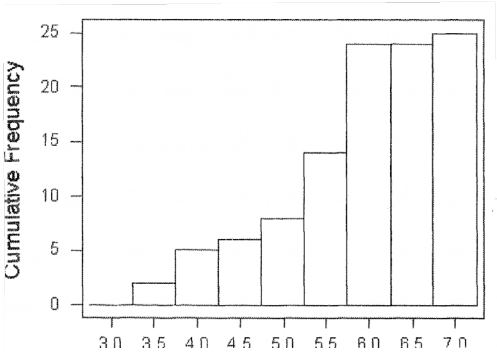
   (a) the pie chart
   (b) the bar chart
   (c) the histogram
   (d) all of the above

3. The histogram below was obtained from data from 300 basketball games in a junior high school basketball league. It represents the number of three-point baskets made in each game. A researcher takes an SRS of sample size n = 30 and computes the mean of each sample. Which of the following best describes the shape of the sampling distribution?



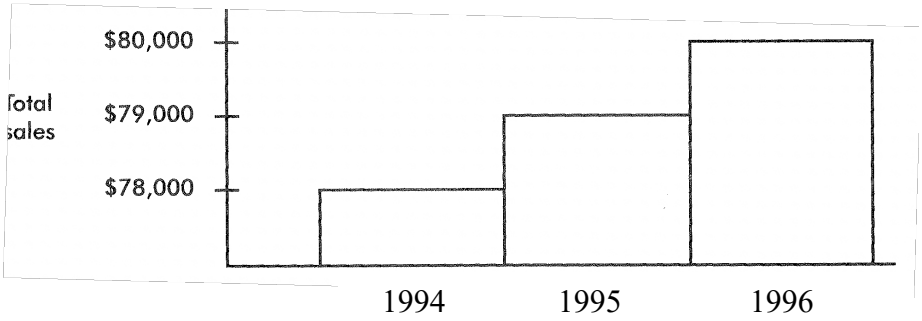
   (a) Skewed to the left
   (b) Skewed to the right
   (c) Uniformly distributed
   (d) Normally distributed
   (e) Bimodal and awesome

4. The lengths (in innings) of 25 randomly selected Little League baseball games were recorded, and a cumulative frequency histogram was created from the results. What is the best conclusion that can be made from the graph?



(a) The median game length is 5 innings.
(b) Fourteen games lasted 5.5 innings.
(c) A majority of the games lasted 6 or more innings.
(d) The distribution of game lengths is severely skewed left.
(e) Games lasting more than 6 innings occurred least frequently.

5. Consider the following histogram:



Which of the following statements are true?

I. Total sales in 1995 were two times the total sales in 1994, while total sales in 1996 were three times the 1994 total.
II. The choice of labeling for the vertical axis results in a misleading sales picture.
III. A histogram showing the same information, but this time with a vertical axis starting at $78,000, would be less misleading.

(a) I only (b) II only (c) III only (d) II and III (e) None of the above

6. A statement is made that more students are purchasing graphing calculators than any other type of calculator. Which measure is being used here?

   (a) Mean
   (b) Median
   (c) Mode
   (d) None of the above

7. What type of distribution is described by the following information?

Mean = 5.5 Median = 5.3 Mode = 5.4

   (a) Negatively skewed
   (b) Roughly Symmetric
   (c) Bimodal
   (d) Positively skewed
   (e) Answer depends on the spread of the data

8. The blood pressure reading for any age group is normally distributed. The normal distribution is symmetric and mound-shaped. Therefore, the correct measure of central tendency to use for blood pressure is:

   (a) the mean
   (b) the median
   (c) the mode
   (d) the skew
   (e) (a) - (c)

1.4 Homework

1. The data below is a frequency distribution for the number of times that a person who had successfully quit smoking had attempted to quit before.

Number of Attempts	Frequency
0	778
1	306
2	274
3	156
4	65
5	58
6	50
7	45
8	36
9	34
10	11

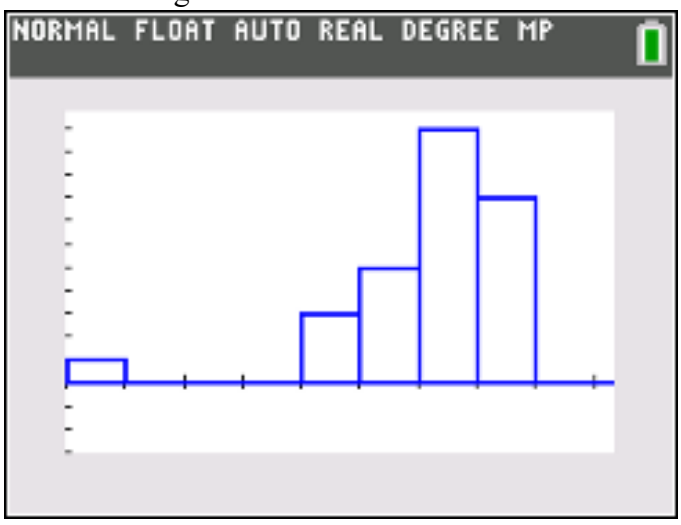
   a. Construct a histogram with this data.
   b. Describe the data using "Shape, Outliers, Center, and Spread".

2. Below is a table of scores (out of 20 points) for an AP Statistics Test.

15	14	9
16	8	3
12	20	10
4	8	18
9	14	11
18	12	9
12	7	10
18	11	4
17	9	19
8	5	20
		15

   a. Create a histogram of this data using intervals of 0 to <3, 3 to <6, etc.
   b. Describe the data using "S.O.C.S."

3. A histogram of a distribution is given below:



Note, the *x*- and *y*-axis scales are 1 unit per mark.

   a. Would you describe the graph as left-skewed, right-skewed, or roughly symmetrical?
   b. Do there appear to be any outliers? Gaps? If there are, describe them.
   c. The actual data used to construct the histogram is included below:

      1 5 5 5 6 6 6 6 6 7 7
      7 7 7 7 7 7 7 7 7 8 8
      8 8 8 8 8 8

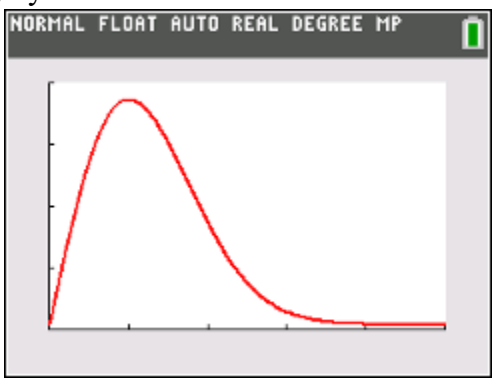
      What is the range? The spread?
   d. What is the median? The mean? The mode?

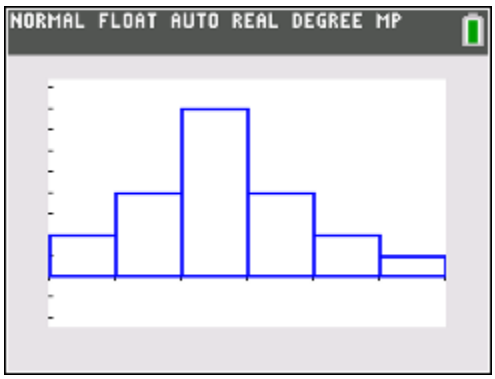
4. The stem-and-leaf plot below is the data measuring time on task for a particular AP Statistics class in minutes out of an eighty-minute period for a class of 27 students:

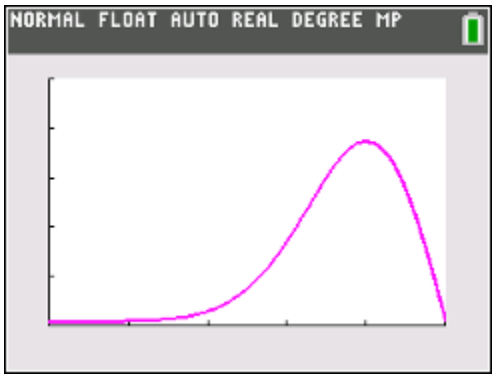
Stem	Leaves
0	4 8
1	2 3 3 4 5 6 7 7 7 8 8 9
2	1 1 1 2 2 5 5 6 6
3	1 1 3 7 8
4	6 7 9
5	
6	
7	9

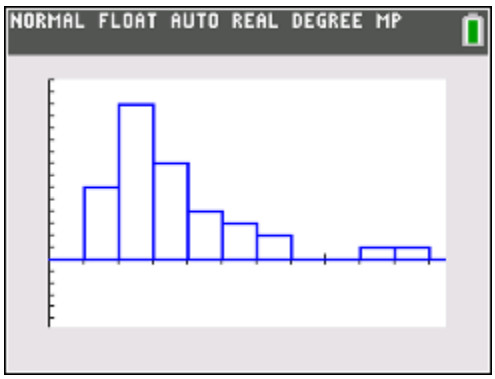
Describe the data in terms of "Shape, Outliers, Center, and Spread".

5. Given the distributions below identify each as positively skewed, negatively skewed, or roughly symmetric.

a. 

b. 

c. 

d. 

6. **Consumer Reports Health (consumerreports.org/health)** reported the sodium content in milligrams per serving of 11 different brands of peanut butter. The data are listed below:

   120 50 140 120 150 150 150 65 170 250 110

   a. Create a dotplot for the data above.
   b. Find the mean, median, and mode for the data.

7. The article **"The Wedding Industry's Pricey Little Secret" (June 12, 2013, slate.com)** stated that the widely reported average wedding cost is grossly misleading. The article reports that in 2012, the average wedding cost was $27,427, while the median wedding cost was $18,086.
   a. What does the large difference between the mean and the median tell you about the distribution in wedding costs for 2012?
   b. Do you agree with the author of the article that reporting the average wedding cost is grossly misleading? Explain why or why not.
   c. Why might the "wedding industry" report the average rather than the median?
   d. The article also states that "the proportion of couples who spent the 'average' or more was actually a minority." Do you agree with this statement? Explain why or why not using the mean and the median value.

1.5 Summary Statistics for a Quantitative Variable and Graphical Representations

Objectives:

- Represent a dataset with a "box-and-whisker" plot (also known as a "boxplot")
- Calculate mean, median, Q1, Q3, interquartile range, and outliers.
- Make a 5 number summary for a boxplot

Recall that the *median* is the middle number in a dataset, and the *mean* is the average of the dataset.

The *median* is an important statistic because it is resistant to change even if the data is skewed or there are outliers. This is referred to as a **resistant statistic**.

If you have ever heard about home values in the Bay Area (or other locations), you may have noticed they always refer to the *median home price*. This is because if we add a few ridiculously expensive houses to the dataset of the rest of the houses, it can skew the mean quite a lot.

You may have experienced this with a grade; if you have good grades on several assignments, and then you get a 0 on one because you failed to turn in an assignment, your overall average grade can be affected quite dramatically.

Example 1: Suppose you have 5 grades in a class on 5 tests. They are 84, 89, 90, 91, and 93.

a) Determine the mean and median of this data set.

b) Suppose you get caught cheating on the 6^{th} test and you get a 0 on that test. Determine the mean and median of this new data set: 0, 84, 89, 90, 91, 93.

c) What do you notice about the change from part a) to part b)?

Answers:

a) The mean is $\left(\frac{84+89+90+91+93}{5}\right) = 89.4$, the median is 90 (the middle number in the list).

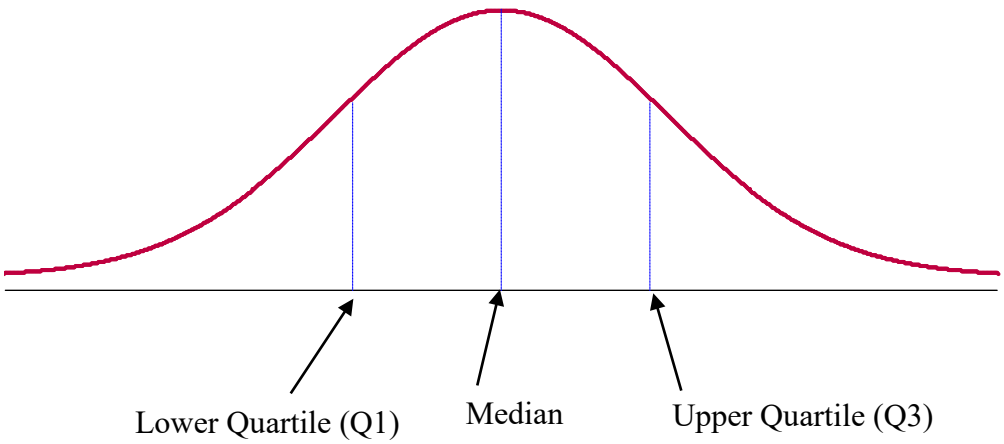
b) The new mean is $\left(\frac{0+84+89+90+91+93}{5}\right) = 74.5$, the median is 89.5 (halfway between the middle two numbers: 89 and 90).

c) The outlier value (0) skewed the data set to the left, and as a result, the mean fell dramatically, with little effect on the median.

As you can see, the median is resistant to change, even if we add an extreme outlier. For larger datasets, the median can be even more resistant. But we still need additional ways of describing the shape and spread of the data.

Therefore, we will be looking at a number of additional ways of breaking up the data to describe the shape, center, and spread. In addition to the idea of the median and mean, we will be looking at **quartiles**. While the *median* breaks the data into a higher half and lower half, the **quartiles** break the data into quarters. This can be very useful in looking at how clumped up or spread out the data is (the *variability*).

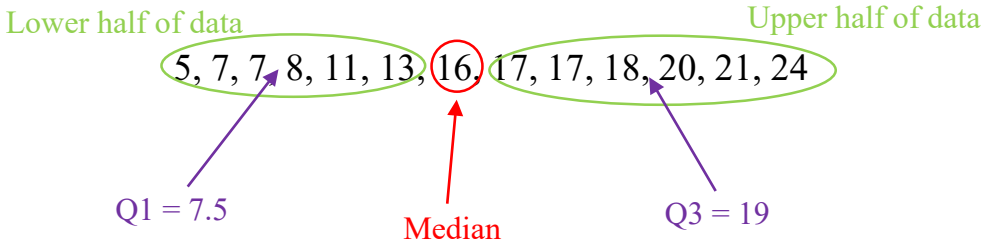
- **Lower quartile (Q1):** The median of the lower half of the sample (cutoff for bottom 25% of data), sometimes called the **first quartile**.
- **Upper quartile (Q3):** The median of the upper half of the sample (cutoff for top 25%...or bottom 75% of data), sometimes called the **third quartile**.
- The **interquartile range** (**IQR**), a resistant measure of variability, is given by **IQR** = upper quartile – lower quartile
- The **IQR** is the *middle* 50% of the data



Note: If the number on observations, *n*, in your data set is odd, then the median will be a number in your list. This ***is excluded*** when calculating Q1 and Q3 and is not considered to be in the either half of the data for this calculation – it is still a part of the dataset, though.

Example 2: Let's take a look at the dataset below and find the median, Q1, Q3, and the IQR

Data: 5, 7, 7, 8, 11, 13, 16, 17, 17, 18, 20, 21, 24



- Median = 16; since there are 13 data points, the 7th number in the list is the median
- Q1 = 7.5; the median of the lower half of the dataset.
  - Since there were only 6 values, Q1 is halfway between the 3rd and 4th value.
- Q3 = 19; the median of the upper half of the dataset.
  - Since there were only 6 values, Q3 is halfway between the 3rd and 4th value of the upper half of the data (10th and 11th overall).
- IQR = Q3 – Q1 = 11.5. This means the middle 50% of the data is spread out over a range of 11.5.

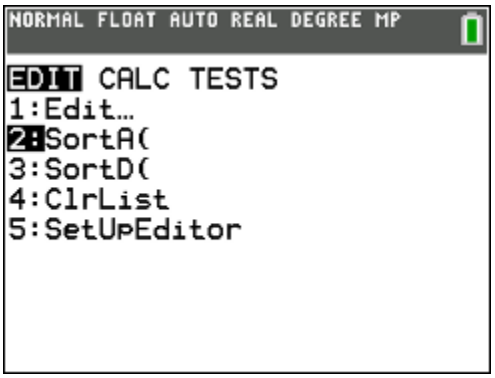
Example 3: The Oregon Department of Health services publishes cost-to-charge ratios for hospitals in Oregon on its web site. The cost-to-charge ratio is computed as the ratio of the actual cost of care to what the hospital actually bills for care, and the ratio is usually expressed as a percentage. A cost-to-charge ratio of 60% means that the actual cost is 60% of what was billed. The ratios for 31 hospitals in Oregon for inpatient services in 2002 were:

68	76	60	88	69	80	75	67	71	100	63
71	74	64	48	100	72	65	50	72	100	63
54	60	75	57	74	84	83	62	45		

Find the median, lower quartile, upper quartile, and IQR.

In order to address this, we should put this list in order. We can either write the list down, or we can put the list on our calculator and have it order the data for us. Note that there are 31 values in the list.

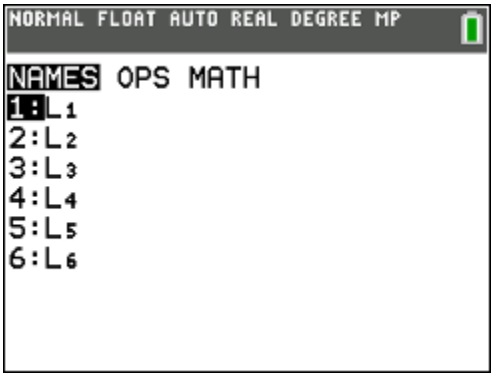
Once you put all the values in your list, go back to the main screen of your calculator and hit the **stat** button. We will want to choose option 2:SortA(. This stands for "sort ascending"; that is, sort the list from low numbers to high numbers.

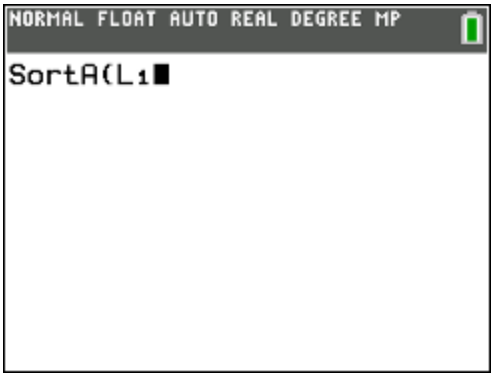


After you have selected this, you need to tell the calculator what list to use. To do this, note that above the **stat** button is the word list.

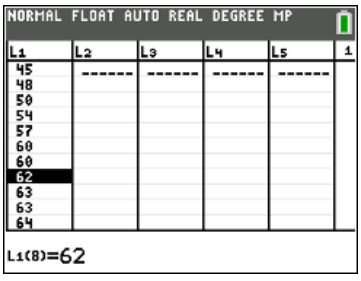
Now hit **2nd** **stat**, and the list menu will open. Select the list into which you have entered your data, and hit enter.

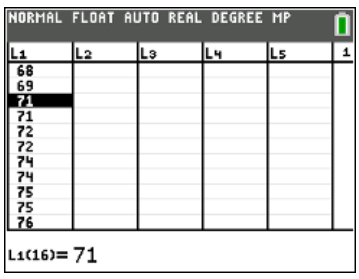
Now hit **2nd** **stat**, and the list menu will open. Select the list into which you have entered your data, and hit enter. Just hit enter on the main screen after this and your list is sorted.

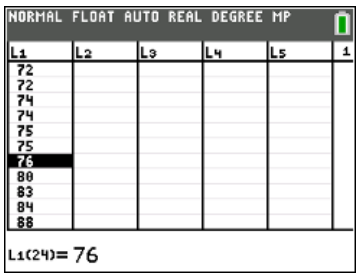




We know that there are 31 numbers in the list, so the 16th number will be the median. This means that there are 15 numbers below and above the median, so Q1 is the 8th number in the list, and Q3 is the 24th number. Since the lists are labeled, simply scroll through your list and find them.



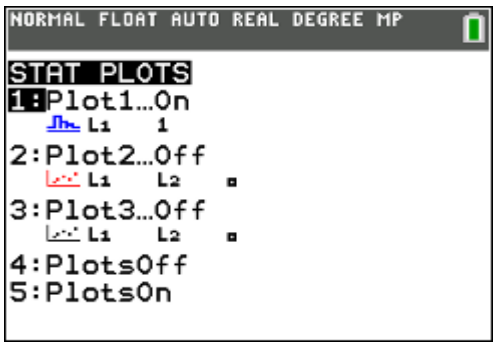


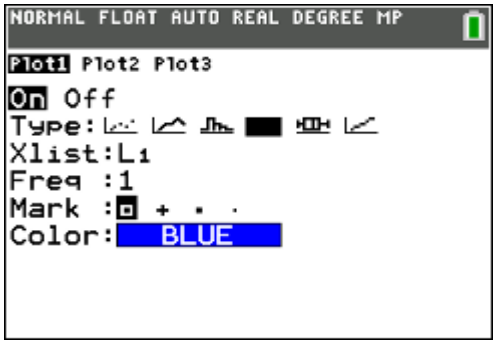


Q1 = 62 median = 71 Q3 = 76

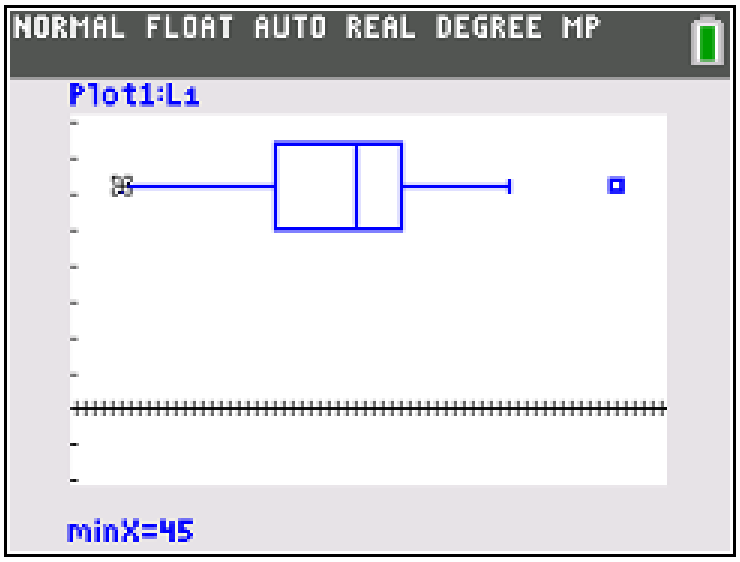
IQR = 76 – 62 = 14

But this seems like a lot of work, and it turns out that there is a really good way to display this data. We can make what is called a "box-plot" (these are sometimes called "box-and-whisker plots"). To do this, we will use the stat plot function on the calculator. Highlight the 4th option in the "Type:" list, and this will create a box plot. You could also use the 5th option, but the 4th one will show any outliers for us.





Don't forget that ZoomStat is important!



Each region of the graph represents 25% of the data.

- The left "whisker" contains the bottom 25% of the data.
- The left "box" contains the lower middle 25% (26-50%)
- The right box contains the upper middle 25% (51-75%)
- The right "whisker" contains the top 25% (76-100%)
- The entire box is the middle 50% of the data, which encompasses the IQR.

If you use the trace button on your calculator, you can scroll through all of these values by moving the cursor left and right!

But how could we identify that the data point 100 is actually an outlier? The calculator identified it for us, but we can calculate outliers quite easily.

- **Outlier:** Any point that is more than 1.5 times the IQR away from the nearest quartile.
  - **Low Outliers:** Any data point < Q1 – 1.5(IQR)
  - **High Outliers:** Any data point > Q3 + 1.5(IQR)

Example 4: Using the same data from Example 3, calculate the thresholds for your outliers, and determine if any of the values in the list are outliers.

68	76	60	88	69	80	75	67	71	100	63
71	74	64	48	100	72	65	50	72	100	63
54	60	75	57	74	84	83	62	45		

We had previously identified the values for Q1, Q3, and the IQR:

Q1 = 62

Q3 = 76

IQR = 14

Q1 – 1.5(IQR) = 62 – 1.5(14) = 41

Q3 + 1.5(IQR) = 76 + 1.5(14) = 97

Therefore, anything below 41 or above 97 is considered an outlier. So there are only two outliers, both values of 100 in the list.

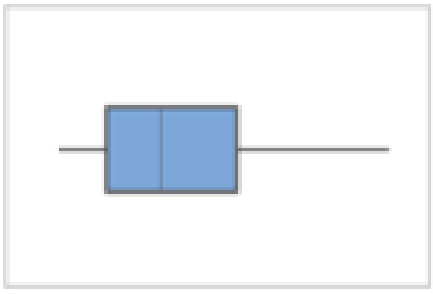
Note: If an outlier is more than 3 x IQR from the nearest quartile, it is sometimes called an *extreme outlier*. Otherwise, it is referred to as a *mild outlier*. These are not used on the AP test, but they are sometimes used by statistics professors and in day-to-day use of statistics.

The 5-Number Summary

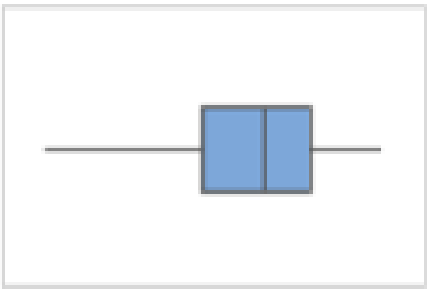
A very quick way of summarizing data is called the **5-Number Summary**. This is simply listing the *minimum value*, Q1, the median, Q3, and the *maximum value*. These are also the 5 numbers you need to make a boxplot.

- **Minimum Value:** The lowest value in a dataset.
- **Maximum Value:** The highest value in a dataset.

We can also tell if our data is skewed from looking at a boxplot. If the right whisker is very long, the distribution is right-skewed (positively skewed). If the left whisker is very long, then the distribution is left-skewed (negatively skewed).



Right-skewed



Left-skewed

Sometimes, however, we will need to draw a boxplot on our own. Just follow these steps, and you can create your own boxplot or modified boxplot. A modified boxplot is just one that shows outliers.

Drawing a Boxplot (or Modified Boxplot)

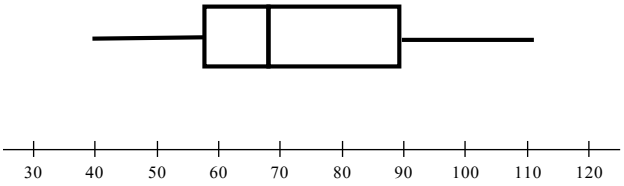
- Draw a horizontal (or vertical) axis with a scale.
- Make a rectangular box with the left and right edge matching Q1 and Q3, and a segment representing the median. Pay attention to your scale!
- Draw horizontal (or vertical) line segments out to your minimum and maximum values.
    - If you are making a modified box plot, stop the whiskers at the lowest and highest values **excluding the outliers** and then plot the outliers as dots.

Example 5: 20 students take a 500 question test online, and then are given practice to help them improve. After the practice, they take the 500 question test again, and their score increase is recorded to generate the dataset below. Use this to draw a boxplot and create a 5-number summary. Identify outliers, if there are any.

40 43 48 55 55 60 60 60 64 66 68 70 75

85 85 93 95 96 100 110

Since there are 20 numbers, the median is halfway between the 10^{th} and 11^{th} value, Q1 is halfway between the 5^{th} and 6^{th} value, and Q3 is halfway between the 15^{th} and 16^{th} values.



Minimum = 40

Q1 = 57.5

Median = 67

Q3 = 89

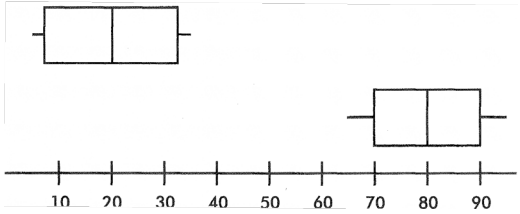
Maximum = 110

IQR = 31.5 Q1 – 1.5(IQR) = 10.25 Q3 + 1.5(IQR) = 136.25

Since all values are between 10.25 and 136.25, there are no outliers.

If you have two (or more) boxplots on the same axis, then you can compare the boxplots to each other. Just make sure you know what the plot represents!

Example 6: Consider the following parallel boxplots illustrating the daily temperatures (in degrees Fahrenheit) of an upstate New York city during January and July.



Which of the following are true statements?

I. The ranges are the same.
II. The interquartile ranges are the same.
III. Because of symmetry, the medians are the same.

(a) I only
(b) II only
(c) I and II
(d) I and III
(e) II and III

Summary:

- **Resistant Statistic:** values that do not change much even if there are outliers (the median and the IQR are both resistant statistics).
    - Mean, Standard Deviation, and Range are not resistant statistics.
- **5-Number Summary:** this consists of the minimum, Q1, the median, Q3, and the maximum.
    - We use these values to construct boxplots and modified boxplots.
- **Outliers:** these occur at values that are more than 1.5 times the IQR from the nearest quartile.
- **Calculator Work:** You can use your calculator (via stat plot) to graph boxplots and modified boxplots.
    - Option 9 from the Zoom menu is very helpful (9:ZoomStat).

Checkpoint 1.5

1. When a set of data has suspect outliers, which of the following are preferred measures of central tendency and of variability?

   (a) mean and standard deviation
   (b) mean and variance
   (c) mean and range
   (d) median and range
   (e) median and interquartile range

2. You have a distribution summarizing the number of days in the past two months (60 days) that an individual watched TV. The median number of days = 25, the lower quartile = 21, and upper quartile = 42. Given this information, which of the following statements is true?

   (a) The distribution is roughly normally distributed.
   (b) The threshold value for the upper outlier is 31.5.
   (c) This distribution is skewed to the right.
   (d) There are no outliers in this distribution.
   (e) The lower outlier threshold is 10.

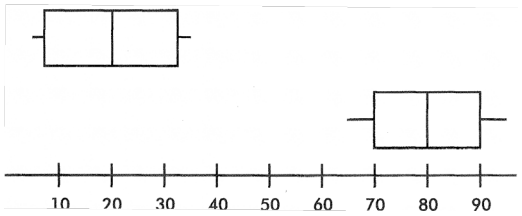
3. The vertical sides on the box of a horizontal box plot are located at

   (a) the minimum value and the first quartile of the data set
   (b) the minimum value and the maximum value of the data set
   (c) the third quartile and the maximum value of the data set
   (d) the first quartile and the third quartile of the data set
   (e) the median value and mean value of the data set

4. The first 115 Kentucky Derby winners by color of horse were as follows: roan, 1; gray, 4; chestnut, 36; bay, 53; dark bay, 17; and black, 4. Which of the following visual displays is most appropriate?

   (a) Bar chart
   (b) Histogram
   (c) Stem-and-Leaf Display
   (d) Boxplot
   (e) Time plot

5. Consider the following parallel boxplots illustrating the daily temperatures (in degrees Fahrenheit) of an upstate New York city during January and July.



Which of the following are true statements?

I. The ranges are the same.
II. The interquartile ranges are the same.
III. Because of symmetry, the medians are the same.

(a) I only
(b) II only
(c) I and II
(d) I and III
(e) II and III

1.5 Homework

1. Based on a large national sample of working adults, the **U.S. Census Bureau** reports the following information on commute times (1-way) to work for those who do not work at home. The 5-number summary for the data collected is listed below:

   Minimum: 2 minutes

   Lower Quartile: 7 minutes

   Median: 18 minutes

   Upper Quartile: 31 minutes

   Maximum: 205 minutes

   It was also reported that the mean for these data was 22.4 minutes.

   a. Is the commute time likely to be left-skewed, right-skewed, or roughly symmetric based on the statistics provided?
   b. Construct a boxplot for the data.
   c. Demonstrate whether there must be any outliers in the data.

2. Fiber content in grams per serving for 18 high fiber cereals **(consumerreports.com)** were collected and are provided below:

   7 10 10 7 8 7 12 12 8 13 10 8

   12 7 14 7 8 8

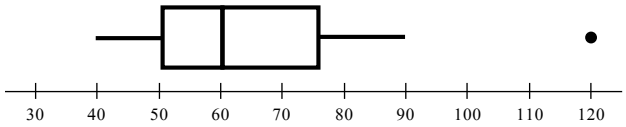
   a. Find the 5-number summary for the fiber content data set. Use that to construct a boxplot for this data.
   b. Explain why the minimum value for this data set and the value for the lower quartile are equal.

3. ***Consumer Reports*** also recorded the sugar content in grams per serving for the same cereals as in problem 2 **(consumerreports.com)**. These data are listed below:

   11 6 14 13 0 18 9 10 19 6 10 17
   10 10 0 9 5 11

   a. Calculate the median, Q1, Q3 and the interquartile range for this data.
   b. Construct a boxplot for the data.
   c. Are there any outliers in this data? How do you know?

4. Below is a boxplot for a certain data set.



   a. Identify the minimum, Q1, the median, Q3, the maximum, and the range.
   b. Identify any outliers.
   c. Demonstrate why the point you identified is an outlier.

5. Below is a table of scores (out of 20 points) for an AP Statistics Test.

15	14	9
16	8	3
12	20	10
4	8	18
9	14	11
18	12	9
12	7	10
18	11	4
17	9	19
8	5	20
		15

   a. Find the minimum, lower quartile, median, upper quartile, and maximum.
   b. Find the interquartile range and identify any outliers.
   c. Sketch a boxplot representing this data.

1.6 The Normal Distribution and the Empirical Rule

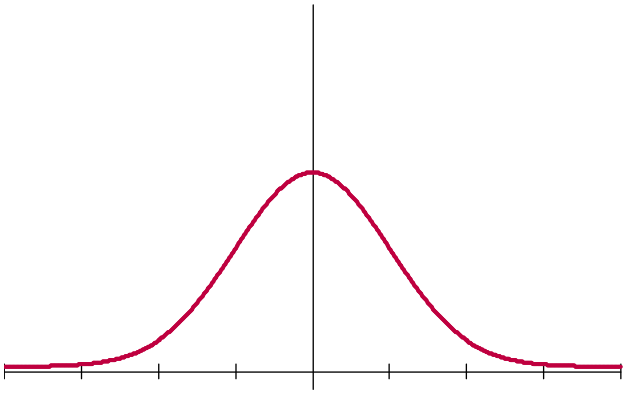
Objectives:

- Calculate deviations and standard deviations.
- Sketch a normal distribution curve.
- Apply the empirical rule to normally distributed datasets.
- Calculate *z*-scores and percentiles.

In section 1.4, we mentioned that curves could be described as "approximately normal". It is time to discuss what we mean by "normal". **Normal Distributions** form a huge part of statistics, and we will discuss it at length over several chapters. For now, we need to know 4 main things:

- It is *bell-shaped*: the curve has what is near-universally referred to as a "bell shape".
- It is *symmetrical*: it has a vertical line of symmetry at the center.
- It has an *infinite base*: that is, the curve stretches from negative infinity to positive infinity.
- The *mean* is at the peak at the center. This is because the graph is symmetrical.

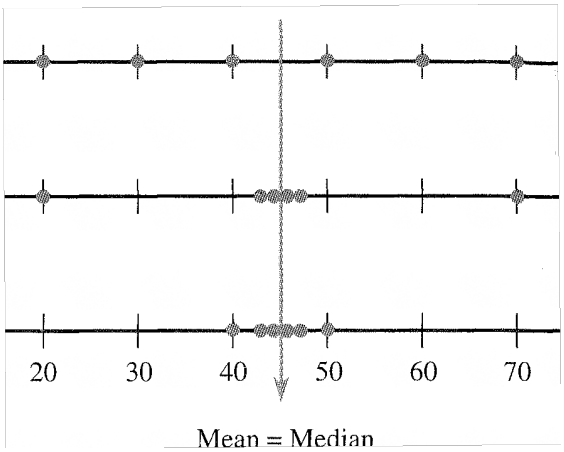
Here is an example of the normal distribution curve on an *x*-*y* axis system:



Many different parameters for populations follow a normal or approximately normal distribution, which makes this a very powerful statistical tool. To understand its utility, however, we need to understand a different measure of variability known as the *standard deviation*.

But to get an understanding of this, we need to look at the idea of *deviation* in general.

If we take a look at the three sets of data below (represented by points on a number line), we could calculate the mean and median quite easily. It turns out that all three of these have identical means and medians (45).



What do you notice that is different about the three data sets?

Hopefully, it is pretty obvious to you that they have different *spread* and *range*. But another way to talk about the variability of the data is to talk about **deviation**.

- **Deviation:** the difference between a data point and the mean for the dataset.
  - All this means is "how far" a data point is from the mean.
  - The *n* **deviations from the sample mean** are the differences $(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_n - \bar{x})$
  - Remember, $\bar{x}$ is the mean for a dataset.

Example 1: Research by the Food and Drug Administration (FDA) shows that acrylamide (a possible cancer-causing substance) forms in high-carbohydrate foods cooked at high temperatures and that acrylamide level can vary widely even within the same brand of food (Associated Press, December 6, 2002). FDA scientists analyzed McDonald's French fries purchased at seven different locations and found the following acrylamide levels:

497 193 328 155 326 245 270

Calculate the mean, median, and *deviations* for this data set. What to the deviations sum up to?

Answers:

Mean = 287.7 Median = 270

Value:	497	193	328	155	326	245	270
Deviation:	209.3	–94.7	40.3	–132.7	38.3	–42.7	–17.7

Sum of Deviations: 0 (if you add up the rounded values, you get 0.1, but the actual sum is 0)

You might have noticed that the deviations will always add up to 0. This is because we generate them from finding how far each is from the average. And since the average was generated by adding up all the values and dividing by the number of values, you will always get a sum of 0.

(If you want a more detailed look at the algebra behind this, just ask your teacher, but it essentially hinges on the fact that you are subtracting out the average the same number of times as the data set, which essentially gives you each data point minus itself when you add them all up).

Of course, seeing a list of deviations can get you a sense of the variability of the data, but this quickly becomes unmanageable with larger data sets. We need a kind of "average deviation" to get a sense of the variability. But as we saw above, if we try to average the deviation, we would always get zero. To handle this, we need something much more akin to the distance formula from algebra.

Recall that with the distance formula, the negative "distances" were squared using the Pythagorean theorem, and then we could get the distance between two points. Applying an analogous (but not identical) process gets us two quantities: **Sample Variance** and **Sample Standard Deviation.**

- **Sample Variance:** denoted by s^2 is the sum of the squared deviations divided by $(n-1)$

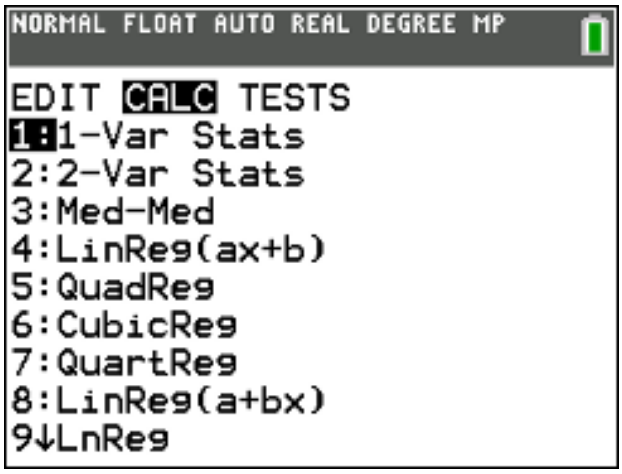
$$s^2 = \frac{\sum\left(x-\bar{x}\right)^2}{n-1}$$

- **Sample Standard Deviation:** denoted by *s* is simply the positive square root of the sample variance.

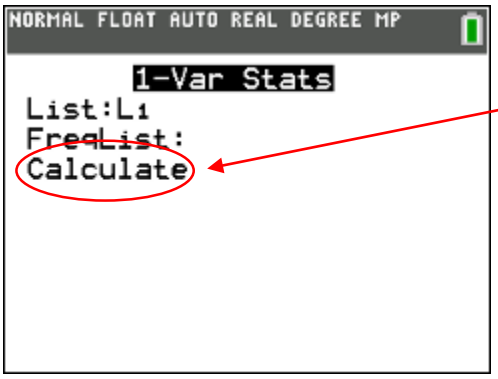
There is a lot of debate over why the denominator is $n - 1$ and not n (everyone agrees that it should be $n - 1$, there are just competing reasons why it is that), so we won't go into any of the reasoning here, but suffice it to say that it is important to use $n - 1$ for *sample variance* and *sample standard deviation*.

Obviously, calculating these values by hand is tedious and time-consuming, and it is relatively easy to make mistakes. So we let our calculators do a lot of this work for us. We still should be able to do it by ourselves if necessary, but once you go past 5 or 6 data points, always reach for your calculator.

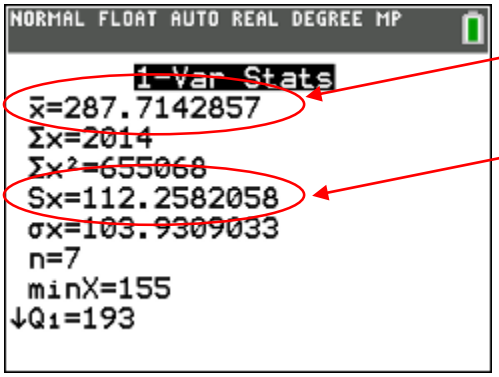
If we go back to the **stat** button on our calculator, and move over to the Calc Menu, and then we will select 1: 1–Var Stats:



- **Make sure you are using the list where your data is (in this case, L_1)**
- Always leave **"FreqList"** blank.
- Then simply highlight Calculate and click **enter**



Highlight this and click enter.



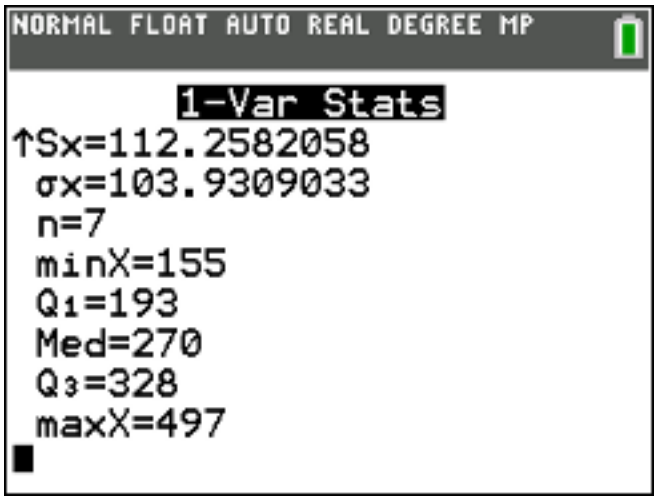
This is the *mean*.

This is the *sample standard deviation*.

If you need the *sample variance*, simply square this number.

The value underneath your *sample standard deviation*, is the *population standard deviation* – more on this later.

You might also notice "minX" and "Q1" at the bottom. Keep scrolling down, and you get your 5-number summary.



- The standard deviation can informally be interpreted as the size of a "typical" or "representative" deviation from the mean.
- We can't say whether s is large or small until we compare it to another data set.
- There are measures of variability for the entire population that are analogous to s^2 and s for a sample. These measures are called the **population variance** and the **population standard deviation** and are σ^2 and σ, respectively.

Example 2: Consider the following back-to-back stem-and-leaf display:

73	2	
642	3	37
7	4	246
9300	5	7
9920	6	0039
943	7	0299
8	8	349
	9	8

Key: 8|3 means 8.3 kg

Which of the following are true statements?

I. The distributions have the same mean.
II. The distributions have the same range.
III. The distributions have the same variance.

(a) II only (b) I and II (c) I and III (d) II and III (e) I, II, and III

You definitely want to enter these two lists in your calculator, and make sure that you translate the data correctly. It is easy to make a small mistake and transpose digits or transpose numbers from one list to another. When you run the 1-Variable Stats on your calculator, you find that the range and variance are the same, but the means are different, so the answer is (d).

Example 3: A sample was taken of the salaries of 20 employees of a large company. The following are the salaries (in thousands of dollars) for this year. For convenience, the data are ordered.

28 31 34 35 37 41 42 42 42 47
49 51 52 52 60 61 67 72 75 77

Suppose each employee in the company receives a $3,000 raise for next year (each employee's salary is increased by $3,000). The standard deviation of the salaries for the employees will

a) be unchanged.
b) increase by $3,000.
c) be multiplied by $3,000.
d) increase by $3,000.
e) None of the above.

How would the standard deviation change if each employee were given a 25% raise?

Answers:
If each value increases by $3 thousand, the average for the data set would go up by 3 as well. Since, in calculating the deviation, there would be no change in the deviations, so the standard deviation will **a) be unchanged**.

If each value was increased by 25%, each value would change by a different amount, so the deviations will change. Calculating standard deviations for both samples and comparing them, we can see that the standard deviation also increased by 25% (much more on this later).

A very important way of analyzing individual data points is by understanding just how far from the mean a data point is. But just knowing how far something is from the mean doesn't necessarily tell you that much. Knowing how many *standard deviations* from the mean a data point is can be much more useful (especially if the data is distributed normally). A ***z*–score** is a measure of how many standard deviations from the mean a data point is.

z–scores

- It is positive or negative depending on whether the observed value is abo The ***z*–score** associated with a particular value is given by the following:

$$z-\text{score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \qquad \text{or} \qquad z-\text{score} = \frac{x-\bar{x}}{s}$$

- The ***z*–score** tells us how many standard deviations *an observed value* is from the mean.
- It is positive or negative according to whether the value lies above or below the mean.
- A more positive or more negative $z-\text{score}$ is further from the mean.
    - Anything more than a $z-\text{score}$ of 2 (or less than –2) is actually quite rare in a population (only 2.5% of a normal population falls above 2, and the same amount falls below –2)
    - Much more on this with *the Empirical Rule.*

Example 4: The average cost per ounce for glass cleaner is 7.7 cents with a standard deviation of 2.5 cents. What is the *z*–score of Windex with a cost of 10.1 cents per ounce?

a) 0.96
b) 1.31
c) 1.94
d) 4.04
e) None of these

$z-\text{score} = \frac{10.1-7.7}{2.5} = 0.96$, so the answer is a).

Note that this would mean this particular cost for Windex is 0.96 standard deviations above the mean price.

Example 5: A student took two national aptitude tests in the course of applying for admission to colleges. The national average and standard deviation were 475 and 100 respectively, for the first test and 30 and 8, respectively, for the second test. The student scored 625 on the first test and 45 on the second test. Use $z-\text{scores}$ to determine on which exam the student performed better.

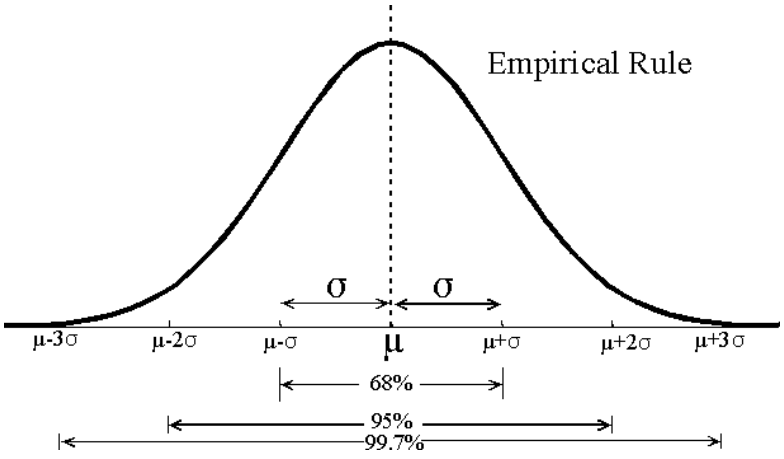
Answer: Calculating the $z-\text{score}$ for each test score, we get the following:

Test 1: $z-\text{score} = 1.5$ Test 2: $z-\text{score} = 1.875$

Since the $z-\text{score}$ was higher on the second test, the student performed better on the second exam.

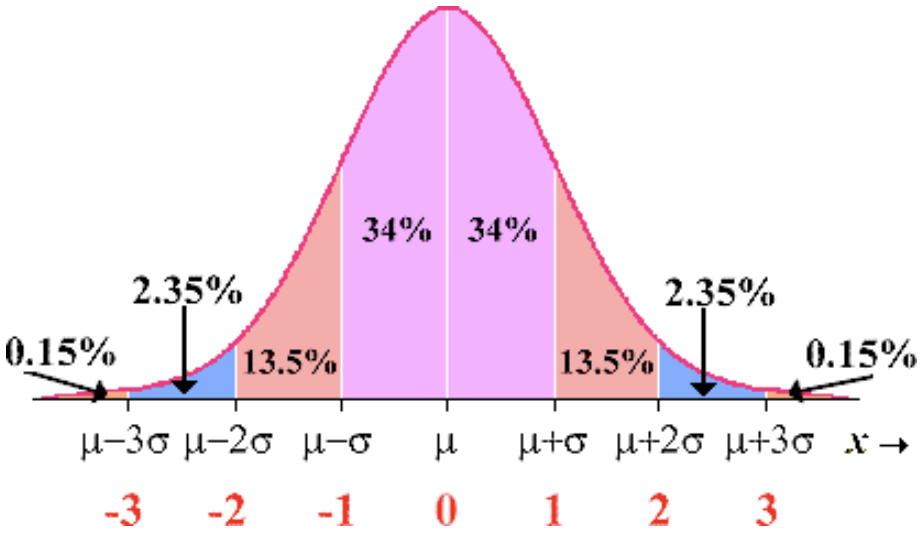
The Empirical Rule (the 68 – 95 – 99.7 Rule)

If the histogram of values in a data set can be reasonably well approximated by a normal curve, then...



- Approximately 68% of the observations are within 1 standard deviation of the mean.
- Approximately 95% of the observations are within 2 standard deviation of the mean.
- Approximately 99.7% of the observations are within 3 standard deviation of the mean.

Another graphical view of this rule looks at the percentages on each side of the mean ...



You can see the 34% on each side one standard deviation from the mean totals to 68%, two standard deviations from the mean total to 95%, etc.

You can also see the z – scores across the bottom of the curve.

Example 6: In a study investigating the effect of car speed on accident severity, 5000 reports of fatal automobile accidents were examined, and the vehicle speed at impact was recorded for each one. It was determined that the average speed was 42 mph and that the standard deviation was 15 mph. In addition, a histogram revealed that vehicle speed at impact could be described by a normal curve.

a) Roughly what proportion of vehicle speeds were between 27 and 57mph?

b) Roughly what proportion of vehicle speeds exceed 57mph?

Example 7: In a certain southwestern city, the air pollution index averages 62.5 during the year with a standard deviation of 18.0. Assuming the empirical rule is appropriate, the index falls within what interval 95% of the time?

a) (8.5, 116.5)
b) (26.5, 98.5)
c) (44.5, 80.5)
d) (45.4, 79.6)
e) There is insufficient information to answer this question.

We need to find the values for $z-$scores of ± 2. This is two standard deviations up and down from the mean, so ± 18.0. Adding that to the mean gets us c) (44.5, 80.5).

Example 8: Suppose a population of individuals has a mean weight of 160 lbs, with a population standard deviation of 30 lbs. According to the empirical rule, what percent of the population would be between 130 and 220 lbs.?

a) 10% b) 68% c) 81.5% d) 95% e) 99.7%

Percentiles:

Another way of talking about data is using **percentiles**. A percentile simply refers to what percentage of the data lies below a given data point. For example, if you had a test score in the 95th %-ile (a common abbreviation for percentile), your score was higher than 95% of the scores in the sample.

- **Percentile (%-ile):** For any particular number r between 0 and 100, the r th percentile is a value such that r % of the observations in the data set fall below that value (i.e. to the left of that value, or $z-$score).

Percentiles are a common way of measuring where an individual data point lies in reference to the other points in a sample. If you have ever taken standardized tests, you may have noticed a percentile score associated with the score you got on the test. Now, hopefully, you know what that score actually means.

Common percentiles are associated with $z-$scores because of the empirical rule and symmetry of the normal curve. Ones that you should quickly recognize are listed here:

- $z-$score $= -2$ 2^{nd} percentile
- $z-$score $= -1$ 16^{th} percentile
- $z-$score $= 0$ 50^{th} percentile
- $z-$score $= 1$ 84^{th} percentile
- $z-$score $= 2$ 98^{th} percentile

Ex7 Given that a sample is approximately normal with a mean of 50 and a standard deviation of 5, the value for the 84th percentile for this distribution is

(a) 45 (b) 55 (c) 60 (d) 65

Ex8 Given that a sample is approximately bell-shaped with a mean of 60 and a standard deviation of 3, the approximate value for the 16th percentile for this distribution is

(a) 54 (b) 66 (c) 63 (d) 57

Ex9 Given that a sample is approximately bell-shaped with a mean of 60 and a standard deviation of 3, the approximate value for the 98th percentile for this distribution is

(a) 63 (b) 66 (c) 69 (d) 57

Summary:

- **Deviation:** the difference between a data point and the mean for the dataset.
- **Sample Variance:** denoted by s^2 is the sum of the squared deviations divided by $(n-1)$

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

- **Sample Standard Deviation:** denoted by s is simply the positive square root of the sample variance.
- The ***z*-score** associated with a particular value is given by the following:

$$z-\text{score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \qquad \text{or} \qquad z-\text{score} = \frac{x-\bar{x}}{s}$$

- The ***z*-score** tells us how many standard deviations *an observed value* is from the mean.
- **Percentile (%-ile):** refers to what percentage of the data lies below a given data point.
- **The Empirical Rule:** The 68 – 95 – 99.7 Rule
  - 68% of normally distributed data is within ± 1 standard deviations of the mean.
  - 95% of normally distributed data is within ± 2 standard deviations of the mean.
  - 99.7% of normally distributed data is within ± 3 standard deviations of the mean.

Checkpoint 1.6

1. Each of the following data sets is a population and has a mean of 40.

   I. {38, 43, 47, 27, and 45}
   II. {41, 40, 39, 42, and 38}
   III. {59, 41, 53, 17, and 30}

   Estimate their population standard deviations and list them from smallest to largest according to standard deviation size.

   (a) I, II, III (b) III, II, I (c) I, III, II (d) II, I, III (e) III, I, II

2. If the standard deviation of a set of observations is 0, you can conclude

   a. that there is no relationship between the observations.
   b. that the average value is 0.
   c. the all the observations are the same value.
   d. that a mistake in arithmetic has been made.
   e. none of the above

3. A population of bolts has a normal distribution with a mean thickness of 20 millimeters, with a population standard deviation of .01 millimeters. Give, in millimeters, a minimum and maximum thickness that will include 95% of the population of bolts.

   a. 19.98 to 20.02 millimeters
   b. 19.99 to 20.01 millimeters
   c. 19.97 to 20.03 millimeters
   d. 19.8 to 20.2 millimeters
   e. These can't be accurately computed.

4. If a sample has a mean of 100 and a standard deviation of a 6, what is the value in the data set that corresponds to a *z*–score of 2?

   (a) 88 (b) 94 (c) 92 (d) 112

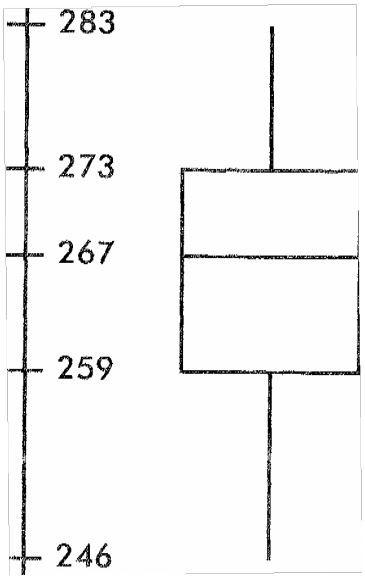
5. A final statistics exam had a mean of 70 and a variance of 25. If Bruce earned an 80 on his exam, what is his *z*–score?

   (a) -2 (b) 10 (c) 0.4 (d) 2

6. A *z*–score is called a standardized score because you can:

    a. translate any *x* value into a *z*–score.
    b. translate any *x* value from a normal distribution into a *z*–score.
    c. translate *z*–scores into a proportion, a percentile, or a probability of the normal curve.
    d. use *z*–scores to find the area between a *z*–score and the mean, or the area below a *z*–score.
    e. use them to compare *x* values to a universal standard, in this case, the standard normal distribution.

7. The 2005 NAEP Trial State Assessment calculated a state-by-state average mathematics proficiency score for students in the eighth grade. The resulting boxplot is



Which of the following are true statements?

I. The state score 273 has a percentile ranking of 75.
II. The mean state score is 267.
III. The lowest state score is 246.

(a) III only
(b) I and II
(c) I and III
(d) II and III
(e) I, II, and III

1.6 Homework

1. The average playing time of CDs in Mr. Maychrowitz's collection is 48 minutes, and he calculated the standard deviation to be 6 minutes. He also finds that the playing time is an approximately normal distribution.
    a. What are the values 1 standard deviations above and below the mean? What are 2 standard deviations above and below the mean.
    b. What proportion of this collection is between 42 and 54 minutes? Between 36 and 60 minutes?
    c. Mr. Maychrowitz has a Rush album that is approximately 66 minutes long. How likely is it that he has many other CDs of this length in his collection?

2. For a set of data on commute times in the Bay Area the mean commute time is found to be 31.5 minutes with a standard deviation of 26 minutes. Given that the commute time cannot be negative, with this mean and standard deviation could you assume that the commute times in the Bay Area are normally distributed? Explain why or why not.

3. Mr. Maychrowitz likes to correct tests before school. He finds that on average it takes him 58 minutes to correct a set of AP Statistics tests. He also knows that his correcting times are normally distributed and have a standard deviation of 8 minutes. School starts at 9 am.
    a. If he has a set of AP Statistics test to correct, how early should he get to school to have a roughly 98 percent chance of finishing before school starts?
    b. Suppose he has another set of tests on another day, and he starts grading at 8:10 am. In what percentile rank would the time to correct have to fall for him to finish before class starts?

4. Suppose that the distribution of scores on a national exam can be described by a normal curve with a mean of 509 and the 16^{th} percentile is 415.
    a. What is the 84^{th} percentile?
    b. What is the approximate value of the standard deviation for the exam scores?
    c. What *z*-score would be associated with an exam score of 600? Of 480?
    d. Do you think that there are many scores below 227? Explain.

5. A biologist tests the response times of a population of rats to various stimuli. Part of the data she collected is represented by the statistics below:

    Stimulus 1: Mean of 4.0 seconds, standard deviation of 1.2 seconds
    Stimulus 2: Mean of 1.8 seconds, standard deviation of 0.6 seconds.

    If a rat has a response time of 3.0 seconds to stimulus 1 and 1.2 seconds to stimulus 2, on which test did the rat respond more rapidly compared to other rats in the population?

1.7 Comparing Distributions of Quantitative Variables

Objectives:

- Use histograms, boxplots, and stem-and-leaf displays to compare two sets of independent data.
- Compare independent datasets using mean, standard deviation, IQR, or median.

Using all of the tools that we have acquired in the previous sections, we can start to compare independent datasets using what we know about the different statistics.

It is important that our datasets are independent of one another for this type of analysis. In addition, just as we can use the individual statistics to find information about a dataset, we can then compare that information about two (or more) datasets and draw conclusions.

When comparing graphs, we also need to use our "SOCS" analysis. Just as we can find Shape, Outliers, Center, and Spread for a graph, we can use that to compare two graphs.

For example, if I have two classes taking a statistics exam, I could calculate the mean for both classes, and I could use that to determine which class, on average, did better. If I looked at the spread instead (or in addition to), I could tell which class had grades that were more widely dispersed or tightly clustered.

Example 1: The average apple has a diameter of 3.25 inches with a standard deviation of 0.5 inch. The average orange has a diameter of 4.5 inches and has a standard deviation of 1 inch. If I have an apple with a diameter of 4 inches and an orange with a diameter of 5.5 inches, which fruit is largest compared to others of its kind?

Apple:

$\overline{x_a} = 3.25$ inches

$s_a = 0.5$ inches

Our apple has $x_a = 4$ inches

This gives a $z-$score of $z = 1.5$

Orange:

$\overline{x_o} = 4.5$ inches

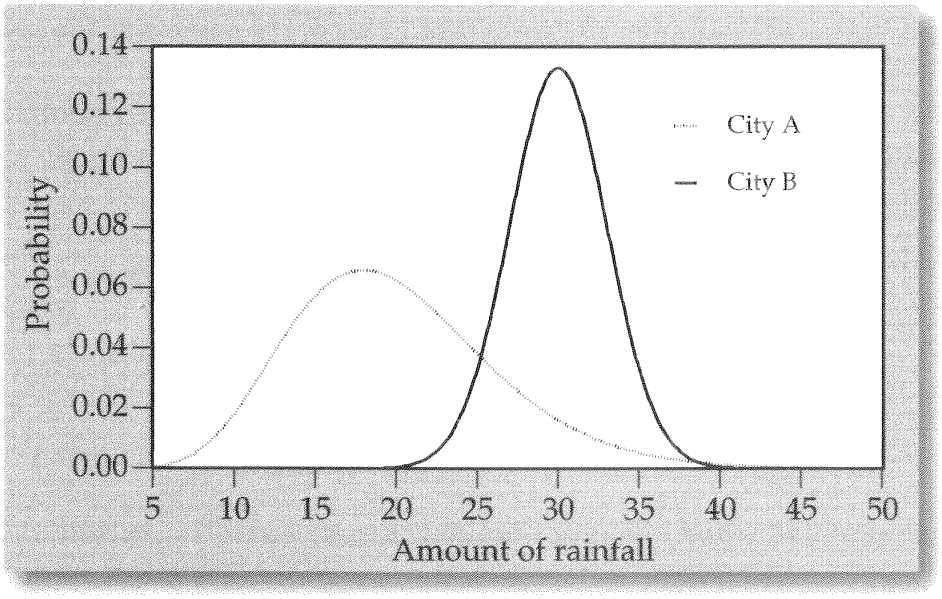
$s_o = 1.0$ inches

Our orange has $x_o = 5.5$ inches

This gives a $z-$score of $z = 1.0$

Since the apple has a higher $z-$score, it is more *standard deviations* away from its mean. This would mean the apple is larger relative to other apples.

Example 2: The following graph summarizes the data collected on annual rainfall in two cities for the past 150 years.



Which of the following conclusions can be made from this graph?

(a) The cities have different mean annual rainfalls, but the range of their annual rainfalls is approximately the same.
(b) On average, City B gets more rain than city A, but has a smaller range of annual rainfall.
(c) On average, City B gets more rain than city A, but has a larger range of annual rainfall.
(d) On average, City A gets more rain than city B, but has a smaller range of annual rainfall.
(e) On average, City A gets more rain than city B, but has a larger range of annual rainfall.

Example 3: Batteries of 2 brands are compared. Brand A has a mean life of 48 months and a standard deviation of 2 months. Brand B has a mean of 48 months and a standard deviation of 6 months. Which brand would you say is the better choice? Why?

Answer:

They both have the same mean value, but Brand A has a smaller standard deviation (2 months). That means that 95% of the batteries produced by Brand A should last between 44 and 52 months.

Brand B with its standard deviation of 6 months means that 95% of its batteries will last between 36 and 60 months.

I would choose brand A, because you are much more likely to get a lifespan close to the stated lifespan of the battery. With Brand B, it would not be unlikely to get a battery life as low as 36 months (though you could get one that is as high as 60 months).

Important: as far as the AP test is concerned, the analysis is more important than the conclusion.

In the example above, if you had made the same argument, but concluded that you would choose brand B because of the chance of getting the battery with the longer lifespan. They will not judge that you essentially are more of a "gambler" with your conclusion, because your analysis and comparison are correct.

Example 4: Test 1 has a mean of 128 and $s = 34$. Test 2 has a mean of 86 and $s = 18$. Test 3 has a mean of 15 and $s = 5$. Which of these scores is the highest relative score? Test 1 score of 144 or Test 2 score of 90 or Test 3 score of 18.

Example 5: A biologist recorded the ages in months of 55 black bears, and also recorded whether each bear was male or female. The data are shown in the back-to-back stemplot below.

Female		Male
4	15	
	14	
2 0	13	
	12	
	11	
	10	
4	9	
2	8	
1	7	0
1 0	6	6
8 3 2	5	3 5
6 5 4 2 1	4	6
8 8 5 4 3 2	3	2 4 7 8
9 7 6 2 1 0 0	2	0 0 1 2 2 3 4 4 4 5 6 7 7 9
	1	5 7 8

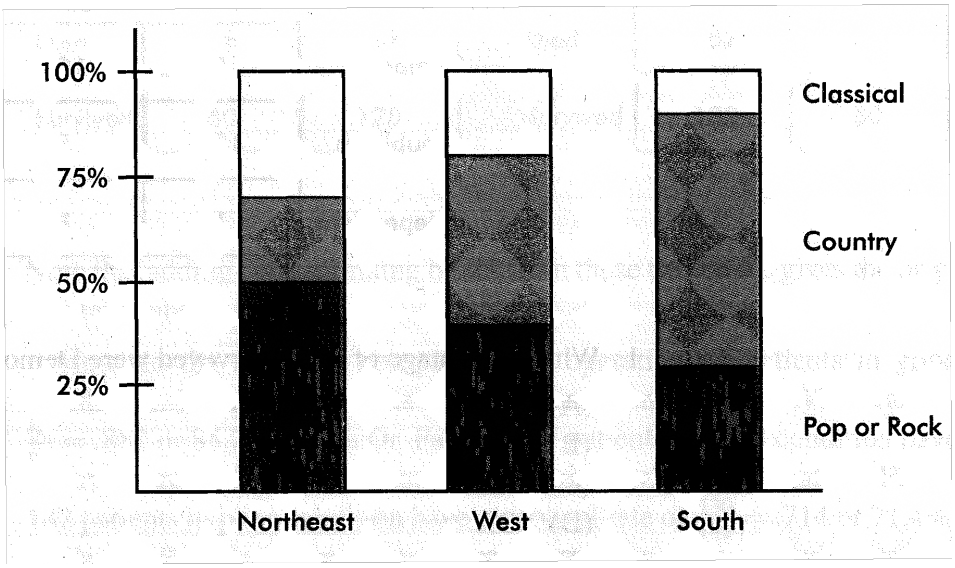
7|0 represents 70 months

Which of the following statements is true?

a) The mean age and the range of ages are greater for male bears than they are for female bears.
b) The maximum age and the median age are both greater for female bears than they are for male bears.
c) The median age is the same for both male and female bears.
d) The median age is greater for female bears than for male bears, and the range of ages is greater for male bears than for male bears.
e) Both the male and female bears have outliers in their respective datasets.

Checkpoint 1.7

A study of music preferences in 3 geographic regions resulted in the following segmented bar chart:



Use the information in the chart to determine the answers to questions 1 – 4.

1. Which of the following is the greatest?

   (a) The number of people in the Northeast who prefer pop or rock
   (b) The number of people in the West who prefer classical
   (c) The number of people in the South who prefer country
   (d) The above are all equal
   (e) It is impossible to determine the answer without knowing the actual numbers of people involved.

2. All three bars have a height of 100%.

   (a) This is a coincidence.
   (b) This happened because each bar shows a complete distribution.
   (c) This happened because there are three bars each divided into three segments.
   (d) This happened because of the nature of musical patterns.
   (e) None of the above is true.

3. What percentage of those surveyed from the Northeast prefer country music?

   (a) 20% (b) 30% (c) 40% (d) 50% (e) 70%

4. Which of the following is the greatest?

   (a) Percentage of those from the Northeast who prefer classical
   (b) Percentage of those from the West who prefer country
   (c) Percentage of those from the South who prefer pop or rock
   (d) The above are all equal
   (e) It is impossible to determine the answer without knowing the actual numbers of people involved.

1.7 Homework

1. You are applying for college and you want the better of your ACT or SAT score to go to the colleges. You know the following information about the standardized tests.

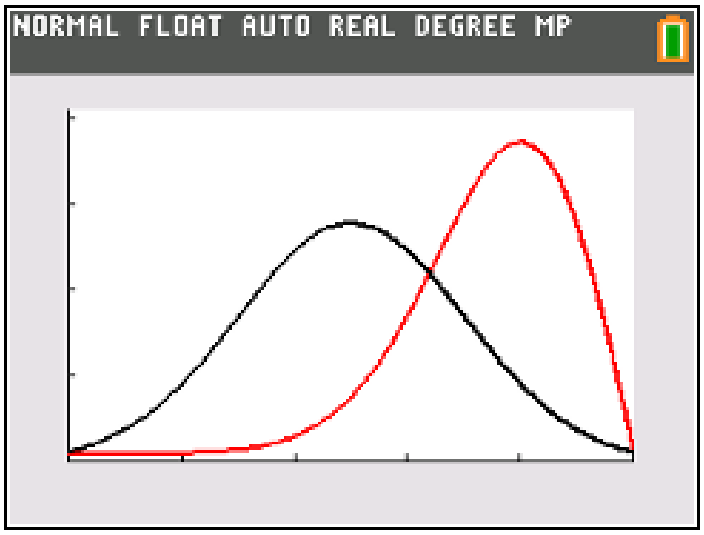
   Both tests are normally distributed

   SAT Mean: 1060 SAT Standard Deviation: 217

   ACT Mean: 21 ACT Standard Deviation: 5

   a. If you score a 30 on the ACT and a 1400 on the SAT, on which test did you do better compared to the rest of the population taking the test.
   b. If your friend took both tests and scored a 1200 on the SAT and a 25 on the ACT which test would you advise him to send to colleges?

2. Below are two distributions of data:



Note: The scales on the *x*- and *y*-axes are both 1

   a. Approximate the value for the mean and median for both graphs.
   b. Compare the skew or lack of skew for both graphs.
   c. Suppose the black graph represents test scores of a population without additional preparation, and the red graph represents test scores of a population with tutoring before the test. Based on the mean, median, and skew difference do you believe that the tutoring made a difference? Explain.

3. Consider the following back-to-back stem-and-leaf displays:

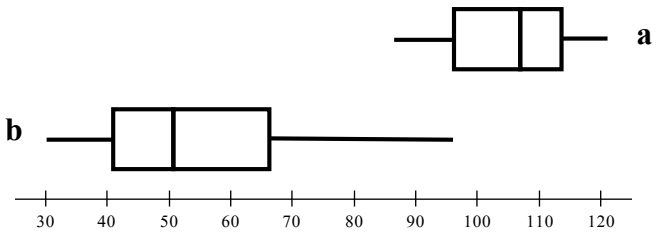
73	2	
642	3	37
7	4	246
9300	5	7
9920	6	0039
943	7	0299
8	8	349
	9	8

Stem: Tens

Leaves: Ones

a. Find the range and spread for both data sets.
b. Find the median of both datasets.
c. Do there appear to be any obvious outliers in either data set?
d. Suppose the two separate sets of data represent wait times in minutes in different urgent care clinics. Based on the data, does it seem like one clinic would be a better choice for urgent care? Explain.

4. Given the two boxplots below, labeled a and b:



a. Find the 5-number summaries for both data sets, and compare them using "S.O.C.S."
b. Determine whether each of the following statements are true or false:
   i. Data set **a** has a smaller range and interquartile range than data set **b**.
   ii. Data set **b** seems to be skewed left, while data set **a** is roughly symmetric.
   iii. The median for data set **b** is less than half the value of the median for data set **a.**
   iv. The maximum for data set **b** is below the value of the minimum for data set **a.**
   v. The maximum for data set **b** is equal to the value of the lower quartile for data set **a.**

Unit 1 Practice Test: