<u>Tuesday, September 1</u>: Chapter 2: Data

Def: _____ is systematically recorded information, whether numbers or labels, together with its context.

- Note that "data" is plural and "datum" is singular.

The context of the data can be described by answering the following questions (the W's):

- ______ was measured?
- ______ was measured?
- ______ were the data collected?
- ______ was the data collected?
- ______ was the study performed?
- ______ was the study performed?

Of these questions, the first two are the most important:

<u>Who?</u>
In most data tables, each row corresponds to an individual being measured.

Name	Height (in)	Hair Color	Brand of Car	# of Siblings
Joe	68	Brown	Ford	2
Sally	63	Red	Honda	1
etc.				

Def: Each individual being measured is called a ______.

Def: The data values recorded about each individual are called ___________________.

<u>What?</u>
The characteristics recorded about each individual are called _____________. These are recorded in the columns of a data table.

Def: A ______________ is any characteristic whose value may change from one individual to another.

Def: a variable is _________________ (or <u>qualitative</u>) if the possible responses fall into categories.

Def: a variable is _________________(or <u>numerical</u>) if the possible responses measure the quantity of that variable.

- Note: quantitative variables usually include <u>units</u>, which tell how the variable was measured. For example, if you are told the weight of an animal is 12, you wouldn't know very much until you were informed of the unit (e.g. tons or milligrams).
- Note: observations of categorical data are usually recorded with words (e.g. Honda, brown), but can also be recorded with numbers. Zip codes are an example. Living in the 85755 zip code isn't necessarily better than living in the 85704 zip code, even though it is higher numerically. In cases like these, the numbers are just labels for different categories.
- Note: Many variables can be used as a categorical variable or a quantitative variable. For example, scores on the AIMS test are recorded numerically, but also placed into categories such as "meets" and "exceeds".
- Note: some variables are neither, such as your school ID number. It doesn't measure anything, so it isn't quantitative. It can be considered categorical like zip codes, but there would only be one per category. These types of variables are called ______________ variables.

<u>How?</u>
How the data were collected is extremely important. The ____________ of data collection (e.g. sample surveys, observational studies, experiments) greatly influences the kinds of conclusions we can draw.

<u>Where? When?</u>
Knowing where and when the data was collected can also make a big difference. The results of a poll about the president's job performance will certainly differ depending on where and when the poll is performed.

<u>Why?</u>
Knowing why the study was performed is also important. If a study reveals that a Advil is the best pain reliever, but the study was performed by the makers of Advil, you should be suspicious. The purpose of the study was probably to help sell more medicine, not to be an objective study of pain relief.

Identify the W's and describe the variables.

Scientists at a major pharmaceutical firm conducted an experiment to study the effectiveness of an herbal compound to treat the common cold. They exposed each patient to a cold virus, then gave them either the herbal compound or a placebo. Several days later they assessed each patient's condition using a cold severity scale ranging 0-5. They found no evidence of the "benefits" of the compound (BVD 2.13).

<u>Chapter 3</u>: Displaying and Describing Categorical Data

The first rule of data analysis is: make a picture. The second rule of data analysis is: make a picture. The third rule of data analysis is: ______________________

Def: A ___________________ displays the possible values of a variable and how often it takes them.

Depending on the type of data and what you are looking for, there are a number of ways to organize and display the distribution.

When a variable is categorical, the first step is usually to construct a _______________________, which displays the possible categories and the number of observations in each category.

Make	**Frequency**
American	
Asian	
European	
Total	

Def: The _________________________ for a particular category is the fraction or proportion of the time that the category appears in the data set.

Make	**Relative Frequency**
American	
Asian	
European	
Total	

Note: the total relative frequency should be 1 (or 100%) except for rounding error.

There are several ways to display categorical data.

<u>Bar Charts and Relative Frequency Bar Charts</u>:

Def: The _____________________ says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents. For example, when making a ________________, the rectangles should all be the same width so only the height determines the area.

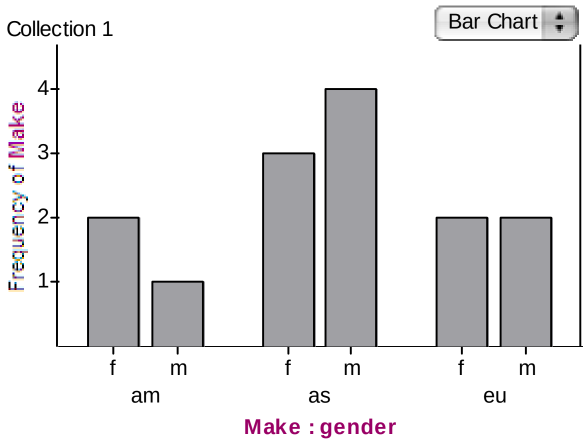
- The horizontal axis should include the variable name and the possible categories. The bars should have some space between them to indicate they are freestanding and can be arranged in any order.
- The vertical axis can be frequency or relative frequency and should include a numeric scale starting at 0
- Relative frequency bar charts make it easier to compare multiple distributions, especially when the sample sizes are different

What is wrong with the following graph?

PIE CHARTS: are useful when comparing categories that form "parts of the whole".

- Categories must not overlap (each subject must be in only one category)
  - The Direct TV ad shouldn't be converted into a pie chart since some channels are in all three categories!
- Label the variable and the categories

What if we wanted to compare the distribution of "Make" for males and females? We can use multiple Pie Charts or a Comparative Bar Chart (such as the one below)



HW #10: Suggested Reading (7-15, 20-24); Problems page 16 (3,5,7,11), page 36 (5,7,9,13,14)

<u>Thursday, September 3</u>: Chapter 3: Displaying and Describing Categorical Data

When we look at 2 categorical variables at the same time, we often arrange the counts into a ________ ________________, also called a ___________________________.

For example, a recent article in the Journal of the American Medical Association (April 10, 2002, vol 287, no 14) reports the results of a study designed to see if the herb, St. John's wort, is effective in treating moderately severe cases of depression. The study involved 338 subjects who were being treated for major depression. The subjects were randomly assigned to receive one of three treatments: St. John's wort (an herb), Zoloft (a prescription drug) or placebo (a fake pill) for an 8-week period.

	St. John's Wort	Zoloft	Placebo	Total
Full Response	27	27	37	91
Partial Response	16	26	13	55
No Response	70	56	66	192
Total	113	109	116	338

Describe the W's.

The margins of the table (bottom and right) give the totals for each variable. These totals are called the ______________________________________ of the 2 variables.

The _____________ in the table give the count (frequency) of each combination of the two variables.

- How many people took the Placebo and had no response?

- It is also possible to use relative frequencies, but there are several ways this can be done.
    - 66/338 is the proportion of people who took Placebo AND had no response (total percent)
    - 66/116 is the proportion of Placebo users that had no response (column percent)
    - 66/192 is the proportion of people with no response that took Placebo (row percent)

Did the responses of the subjects depend on which medication was given? To determine this, we can compare the _____________________________________ for each medication. A conditional distribution looks at the distribution of one variable when the value of the other variable is fixed.

For St. John's Wort,

- 27/113 = 24% had a full response
- 16/113 = 14% had a partial response
- 70/113 = 62% had no response.

Calculate the conditional distributions for Zoloft and Placebo.

Make segmented bar charts of all 3 conditional distributions so they can be easily compared.

Do the results look the same for all 3 medications?

If the distribution of one variable (the response) is the same for all categories of the other variable (the medication), we say the variables are _________________ or that they have <u>no association</u>. If two variables are independent, then knowing the value of one variable will not help you predict the value of the other variable. For example, if knowing what medication a person received doesn't help you predict their response, then the variables are independent.

If the distributions are not the same for each category, we say the variables are _________________ or that they have an <u>association</u>. For example, if knowing what medication a person received helps you predict their response, then the variables have an association.

Simpson's paradox: Sometimes it is inappropriate to average proportions.

- For example, suppose you are a baseball manager and need to choose a hitter for a certain situation. Player A has 33 hits in 103 at bats (.320) and Player B has 45 hits in 151 at bats (.298). It seems obvious that Player A would be the better choice.
- However, against right handed pitchers, Player A has 28 hits in 81 at bats (.346) and Player B has 12 hits in 32 at bats (.375). Player B seems better in this case.
- Also, against left handed pitchers, Player A has 5 hits in 22 at bats (.227) and Player B has 33 hits in 119 at bats (.277). Player B seems better in this case also.

Player	Overall	vs. Right	vs. Left
A	.320 (33/103)	.346 (28/81)	.227 (5/22)
B	.298 (45/151)	.375 (12/32)	.277 (33/119)

- How can this be? It seems like hitting against left handed pitchers is more difficult. And, even though player B is better in both cases, since player B has many more at-bats against left handed pitchers, his overall average is lower.
- Fortunately, Simpson's paradox is fairly rare, however.

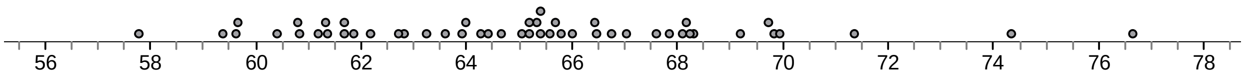
HW #11: Suggested Reading (24-34), Problems page 39 (19,21,23,27,31,39)

<u>Monday, September 7</u>: No School!

<u>Tuesday, September 8</u>: Chapter 4: Displaying Quantitative Data

There are several good ways to display the distribution of a quantitative variable.

The most basic is called a <u>DOTPLOT</u>. Here is an example using heights of 50 students.



Note: For all graphs, it is important to label the axis and scale clearly!

Dotplots are very good for displaying small data sets, however, when there are a large number of observations, <u>HISTOGRAMS</u> are a better choice. Histograms are similar to bar charts, but for quantitative data. As with bar charts, the possible values of the variable are plotted on the horizontal axis and the frequencies are expressed as the heights of the bars. However, the bars in a histogram should touch, indicating there we are not leaving out any possible values.

Since this data is quantitative, there are no "natural" categories to place the data. In this case, we will define our own categories, called <u>CLASSES or BINS</u>. There is no perfect way to create bins, but bins should always be the same length and never overlap or leave any gaps. Ideally, you should use between 4-10 bins that start and end at "nice" values.

What bins could we use with this data?

What if an observation falls exactly on a boundary?? As a convention, we will put boundary values into the upper bin. For example, suppose you decided on bin lengths of 2 inches starting at 56. Then, the boundaries would be 56", 58", 60", etc. The bin from 56-58 would be $56 \leq x < 58$ or 56-<58 and an observation of 58" would fall into the 58-<60 bin.

Graph the histogram:

- Make sure you label axes and scales. The vertical axis should always start at 0.
- The bars in a histogram should touch (unlike bar charts for categorical data)

Relative Frequency Histograms:
A relative frequency histogram looks the same as a regular histogram, except that it will have relative frequency (percent of the total) rather than frequency (number of observations) on the vertical axis. Relative frequency histograms are particularly useful for comparing distributions with different sample sizes, since the vertical axis will be on the same scale.

Note: a common mistake is to make a "histogram" by using one bar for each observation where the height of the bar represents the value of the variable for that case.

Note: Changing the bin widths can make a big difference. Check out the histogram applet at: http://www.rossmanchance.com/applets/

Note: There are methods for creating histograms with unequal classes, but we are skipping them. If a frequency distribution has non-bounded classes, such as "12 or more", a histogram cannot be made (think of the area principle).

A STEM-AND-LEAF plot is another way to display a relatively small numerical data set. In a stem-and-leaf plot, the STEM is the first part of the number and the LEAF is the last part of the number.

ex: freshman male weights {97,102,117,128,130,132,139,147,154,162,166,189,225}

Freshman Male Weights

9	7
10	2
11	7
12	8
13	029
14	7
15	4
16	26
17	
18	9
19	
20	
21	
22	5

key: 14 | 7 = 147 pounds

- The numbers to the left of the line are the stems (hundreds and tens digits) and the numbers to the right of the line are the leafs (units digits).
- You must include a key (with units) and a label/title
- Leaves should be single digits (no commas)

- It is best if the leafs are in numerical order, but it is not required.
- Stemplots will look very similar to a dotplot or histogram of the same data, but a stemplot preserves the individual data values

Back-to-back stemplots are useful for comparing distributions. For example, given the following female weights, make a back-to-back stemplot to compare female and male weights.

Freshman Female Weights = {93, 99, 100, 104, 109, 111, 113, 113, 121, 125, 126, 128, 142, 159, 185}

When a data set is very compact, it is often useful to ________________ to stretch the display to investigate the shape. This is sometimes called a ________________.

ex: body temperatures: {96.3, 97.6, 97.8, 97.9, 98.1, 98.1, 98.3, 98.5, 98.6, 98.6, 98.7, 98.8, 99.0, 99.5}

When a data set is very spread out, it is often useful to ______________ the data to shrink the display.

ex: grocery bills: {10.53, 13.67, 15.01, 18.30, 20.89, 27.07, 32.82, 37.57, 52.36}

The 4 Key Features of a Distribution are: ____

Words used to describe the __________ of a distribution:

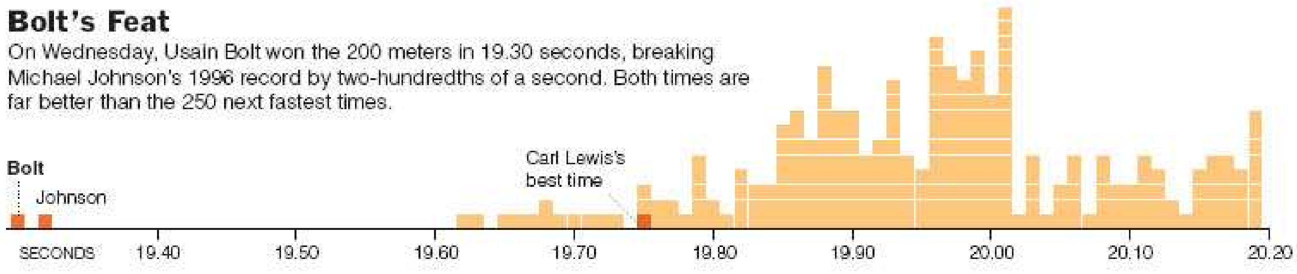
There is no specific definition of "unusual values", but here are some things to consider:

Def: ____________: data values that fall out of the pattern of the rest of the distribution

Def: ____________: isolated groups of values

Def: ____________: large spaces between values

Example:



http://www.nytimes.com/interactive/2008/08/20/sports/olympics/20080820-bolt-graphic.html?scp=1&sq=Bolts%20Feat&st=cse#

It is extremely important to investigate outliers!

“All outliers should be taken seriously and should be investigated thoroughly for explanations. Automatic outlier-rejection schemes (such as throw out all data beyond 4 sample standard deviations from the sample mean) are particularly dangerous.

“The classic case of automatic outlier rejection becoming automatic information rejection was the South Pole ozone depletion problem. Ozone depletion over the South Pole would have been detected years earlier except for the fact that the satellite data recording the low ozone readings had outlier-rejection code that automatically screened out the "outliers" (that is, the low ozone readings) before the analysis was conducted. Such inadvertent (and incorrect) purging went on for years. It was not until ground-based South Pole readings started detecting low ozone readings that someone decided to double-check as to why the satellite had not picked up this fact--it had, but it had gotten thrown out!

“The best attitude is that outliers are our "friends", outliers are trying to tell us something, and we should not stop until we are comfortable in the explanation for each outlier.”

(source: http://www.itl.nist.gov/div898/handbook/eda/section3/histogr8.htm)

HW #12: SR (45-53), problems page 65 (3, 7, 13, 15, 17, 21-histogram!, 25, 27) – ignore center and spread for this assignment.

Note: the graph in the answer to #15 is incorrect.

<u>Thursday, September 10</u>: Chapter 4/5 Displaying and Describing Numerical Data

Sometimes it is useful to plot data over time so we can notice trends.

For example, consider the following data for the death rates at 2 hospitals:

Month	Hospital A	Hospital B
1	7	4
2	6	3
3	5	4
4	5	5
5	6	5
6	5	6
7	4	5
8	5	5
9	4	6
10	3	7

Make a dotplot for each hospital. Which hospital seems better?

Make a timeplot for each hospital. Which hospital seems better?

What information does the timeplot tell you that the dotplots did not?

Note: Timeplots are not appropriate for some data, such as graphing the heights of this class. However, a timeplot would be nice for graphing the change in height for one student over time.

In the last lesson, we learned that to describe a distribution you should address Shape, Center, Spread, and Unusual Values. We also learned how to describe shape, but not much about center and spread. In this lesson, we will learn numerical ways to describe center and spread.

Def: The __________ of a data set is the middle value (a measure of center). That is, it divides the data set in half so that there are an equal number of observations above the median and below the median.

To find the median:

1. put the data in order
2. find the value of (n+1)/2 (n = the number of observations in the data set)
3. if n is odd, the median will be the (n+1)/2 value in the ordered list
4. if n is even, (n+1)/2 will be between two observations. The average of those two observations is the median.

- ex: 1, 13, 9, 5, 17, 23, 14

- ex: 12, 17, 5, 19, 23, 39

Def: a ______________________________ is a measure that is not affected by outliers or skewness.

- In the previous example, what would happen to the median if the 39 was changed to 399?

Def: the ____________ of a data set is the difference between the maximum and minimum values (a measure of spread).

- ex: 12, 18, 19, 23, 25

- Note: the range is 1 number, not 2!
- Is the range a resistant measure of spread?

A more resistant measure of spread is called the ______________________________ (IQR). The IQR measures the range of the middle half of the data, instead of the entire data set, which means that outliers should have no effect on the IQR.

To find the middle half of the data, we must identify the __________________.

- The first quartile, Q_1, is the value which separates the lower 25% of the data from the upper 75%.
- The third quartile, Q_3, is the value which separates the lower 75% of the data from the upper 25%.
- The median could be called Q_2, since it separates the lower 50% of the data from the upper 50%, but it usually is just called the median.

To find the quartiles, split the data set into two halves at the median (*if there is an odd number of observations, include the median in both halves*). Then, find the median of each half. The median of the lower half is Q_1 and the median of the upper half is Q_3.

Finally, to find the IQR, subtract $Q_3 - Q_1$.

Find the median, quartiles, range, and interquartile range:

1	2
2	
3	7
4	
5	17
6	2356
7	122456678
8	3367889
9	1168
10	0

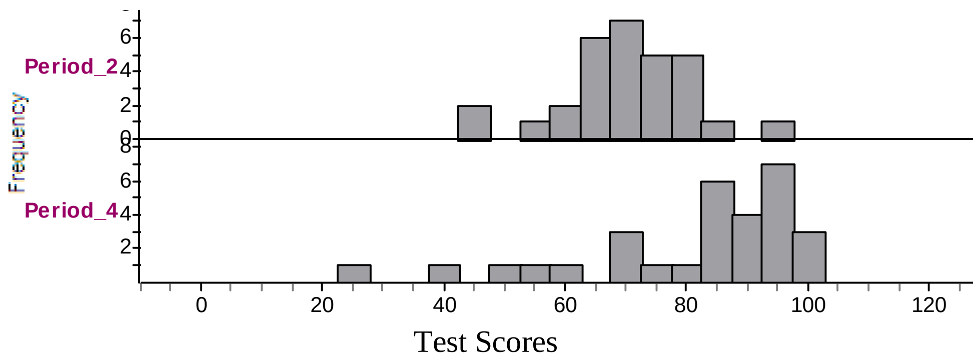
Test Scores
3 | 4 = 34%

Note: There are several different acceptable ways to calculate the quartiles. The TI-83 and some other books exclude the median from both halves when there is an odd number of observations instead of including it.

Comparing Distributions (again):

Whenever you are asked to compare distributions you must COMPARE shape, center, spread and unusual values and do it in the context of the problem. To compare you should use phrases like "the male distribution has a *higher* center and is *more* spread out than the female distribution"

Compare the distributions of test scores for the 2 classes below.



AP Question:

HW #13: SR (55-61, 73-76), Problems page 67 (18, 23, 26, 28, 29, 32, 34)

<u>Monday, September 14</u>: Chapter 5: Describing Distributions Numerically

Def: The ______________________________ of a distribution is a list of the minimum value, Q1, median value, Q3, and maximum value.

A _________________ is a graphical display that uses the 5# summary to picture a distribution.

- Note: the length of the box = _________________________________
- Note: the length of the entire plot = _________________

If a distribution has ______________, they are usually marked separately

To determine if there are outliers we place boundaries (fences) around the main part of the data.

- lower fence = Q1 – 1.5 IQR
- upper fence = Q3 + 1.5 IQR
- Any observations that are outside of these fences are considered outliers.
- If there are outliers, the whisker extends to the most extreme observation that is not an outlier.
- Are there outliers in the test score example from yesterday? Draw a boxplot of the data.

Boxplots are nice because we can quickly identify the center (median) and spread (range and IQR). We can also identify symmetry or skewness, but not how many modes the distribution has. Boxplots are especially useful for comparing distributions, but make sure they are on the same scale!

Cumulative Relative Frequency Distributions

Instead of wanting to know what percent of the data falls into a particular class, we often want to know what percent falls below a certain value. To make this possible, we will compute the ______________________________ for each class, which is the sum of the relative frequency of that class and all the classes below it.

For example, consider the following frequency distribution of test scores:

Score	Frequency	Relative Frequency	Cumulative Relative Frequency
0-<10	2	.01	
10-<20	7	.035	
20-<30	9	.045	
30-<40	12	.06	
40-<50	23	.115	
50-<60	34	.17	
60-<70	48	.24	
70-<80	39	.195	
80-<90	19	.095	
90-<100	7	.035	
Total	200	1	

- What proportion scored less than 30?
- Less than 90?
- What proportion scored at least 40?
- At least 70?
- What proportion scored 45?
- About what proportion scored 50-<70?

Def: The graph of a cumulative relative frequency distribution is called an ___________.

- About what proportion scored less that 45?

- What score separates the lower half of the scores from the upper half of scores?

Def: the P^{th} __________________ of a distribution is the value in the distribution such that P percent of the observations lie at that level or below.

- What is the 30^{th} percentile for this distribution?

- What is the 70^{th} percentile?

- What is the median?

- Make a boxplot of the test scores:

HW #14: SR (77-81), Problems page 92 (13, 15, 20, 24, 33, 35, 47)

<u>Tuesday, September 15</u>: Chapter 5: Describing Distributions Numerically

When a distribution is skewed, the median and IQR are the best way to describe center and spread since they are resistant measures. However, when a distribution is roughly symmetric, there are alternatives.

To describe the center, we can use the ___________________ (or average). To find it, simply add up all the observations and divide by the sample size: $\bar{x} = \frac{\sum x}{n}$

Notation:

- In our book they use the letter y to represent the observations in a data set.
- However, it is more common to use $\bar{x}$ for the sample mean, so that is what we will do.
- The sample mean is called "x-bar". Anytime there is a bar over a variable, it indicates a mean. Thus, $\bar{y}$ would indicate the mean of the values of y.
- It is called a "sample" mean since it is computed from the observations in a sample. If we had conducted a census and had all the members of the population, we would use μ to denote the population (true) mean. We use $\bar{x}$ to estimate μ since we almost always are using sample data.
- The sample size is always denoted n.

Graphically, the mean of a distribution is located at the balancing point of the distribution. The mean is the value that balances the deviations from the mean.

Is the mean a resistant measure of center? How is it affected by outliers and skewness?

Another way to measure the spread of distribution is to estimate the average deviation from the mean. This quantity is called the ____.

Find and interpret the standard deviation for the following distribution of Home Runs:

1, 13, 11, 1, 5, 9, 6, 5, 8, 11

Note: If we had the entire population, we would use σ to denote the population standard deviation (true SD). In the calculation we would use μ instead of $\overline{x}$ in the numerator and n in the denominator instead of n-1: $\sigma_x = \sqrt{\frac{\sum(x-\mu)^2}{n}}$

Note: The square of the sample standard deviation is called the ______________________________. It will come in handy later.

What if the last value in the data set above was 41 instead of 11? Calculate and interpret the new SD. Is the standard deviation a resistant measure of spread?

Using the TI-83

- Entering data in a list: *Stat: Edit*
  {4, 9, 11, 13, 13, 15, 15, 16, 16, 16, 17, 17, 18, 19, 20, 20}

- Making histograms: *Stat Plot: Plot 1* graph #3
  -*zoom: 9 zoomstat* makes a nice window
  -*trace* will show class boundaries and frequencies
  -*window: Xscl* changes bar width

- Making boxplots: *Stat Plot* graph #4, #5
  -#4 shows outliers and #5 does not
  -*zoom: 9 zoomstat* makes a nice window
  -*trace* will show 5 number summary and outliers
  -the TI-83 uses a different method to calculate quartiles when n is odd (excludes the median)

- Calculating summary statistics: after data has been entered into L1, choose *Stat: calc:1-var stats, L1*.

- Other operations with lists: Suppose you had data in L1 and you wanted to add 5 to each observation. In the header for L2, enter L1+5 and press *Enter*.

- To sort a data in a list: Stat: Edit: SortA or SortD

For the following sample of IQ scores, calculate the mean and standard deviation. Then, sketch a dotplot and then on the axis label the mean, mean +/- 1 SD, 2 SD, 3 SD. Finally, calculate the percentage of the data that is within each set of boundaries.
{73, 79, 85, 91, 98, 101, 102, 103, 105, 108, 110, 111, 112, 117, 118, 119, 121, 122, 131, 136}

Matching Distributions Activity

HW#15 SR (82-88), Problems page 91 (5, 6, 9 (check with calculator), 11, 25, 27, 28, 29)

<u>Thursday, September 17</u>: Chapter 6: The Standard Deviation as a Ruler

<u>Measures of Location (using the standard deviation as a ruler)</u>:

Suppose that a professional soccer team has the money to sign one additional player and they are considering adding either a goalie or a forward. The goalie has a 90% save percentage and the forward averages 1.2 goals a game.

In this league, the typical goalie saves 86% of shots on average with a standard deviation of 5% while the typical forward scores 0.9 goals per game on average with a standard deviation of 0.2. Who is the better player at his position?

The goalie is 4% higher than the average and the forward is 0.3 goals higher than average. But, since we are comparing different units, we cannot just say the goalie is better since 4 > 0.3. To make comparisons possible, we consider where each player falls in their respective distributions.

The goalie is $\frac{90-86}{5} = 0.8$ standard deviations above the goalie mean.
The forward is $\frac{1.2-0.9}{0.2} = 1.5$ standard deviations above the forward mean.

When we are using the standard deviation as a ruler to measure how far an observation is above or below the mean, we are using a ______________________________, or <u>z-score</u>.

$$z = \frac{x - \overline{x}}{s}$$

Using standardized scores has many advantages. Since z-scores have no units, we can compare values that are measured on different scales, with different units, or from different populations.

Suppose that the distribution of male heights has a mean of 69 inches with a standard deviation of 3 inches and the distribution of female heights has a mean of 64 inches with a standard deviation of 2.5 inches.

Who is taller, relatively speaking, a 65 inch male or a 60 inch female?

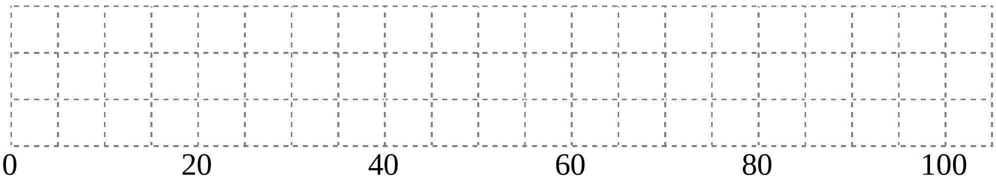
An exclusive club only allows members who are especially tall. The height requirement for women is 72 inches. What is the equivalent requirement be for men?

Shifting and Rescaling Data

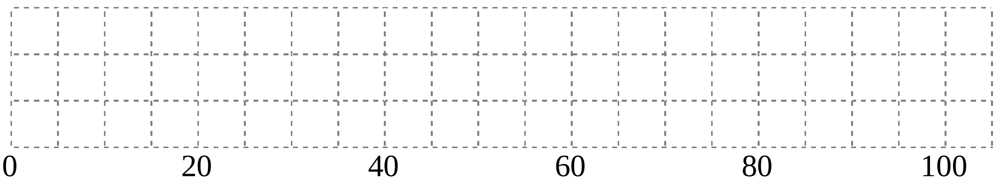
Suppose that I took a random sample of 9 students and recorded their score on a recent quiz. There scores are listed below:

30, 35, 36, 40, 40, 43, 45, 45, 50

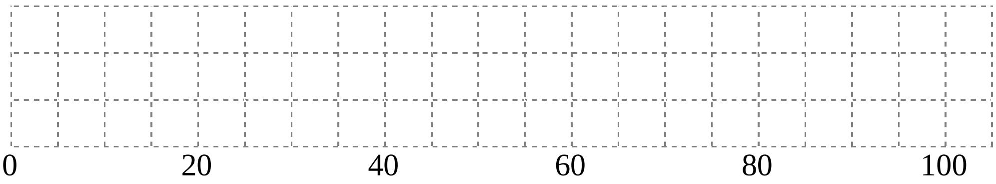
Sketch a dotplot for this distribution and list all the summary statistics for this data set (mean, standard deviation, median, quartiles, IQR, range).



Now, suppose that I was feeling especially generous and added 5 points to each score. Sketch the new distribution (same scale!) and recalculate the summary statistics. Which ones changed? Which did not? Did the shape change?



Now, go back to the original data and assume that the quiz was out of 50 points. To convert these scores to percents, we could multiply each of them by 2 (since 2 x 50 = 100). Sketch the new distribution (same scale!) and recalculate the summary statistics. Which ones changed? Which did not? Did the shape change?



A Manhattan Taxi cab company studied the lengths of taxi rides and computed the following statistics (in miles): mean = 4.8, standard deviation = 4.2, median = 3.6, Q1 = 1.8, Q3 = 5.9, IQR = 4.1

If they converted the mileage measurements to km, what will the summary statistics be in km? (1 mile = 1.62 km)

If the cost of a taxi ride is $2.50 plus $3.20 per mile, what are the summary statistics for the distribution of costs?

Note: When we convert measurements to z-scores, we are simply shifting and rescaling the data. Subtracting the mean shifts the center (mean) of the distribution to 0. Dividing by the standard deviation rescales the data by making the spread (standard deviation) equal to 1. Neither of these operations changes the shape of the distribution.

AP Question:
HW #16: SR (102-107), problems page 123 (2-5, 7, 10, 12, 14)

<u>Monday, September 21</u>: Review chapters 2-6
AP Question:
HW #17: Problems page 131 (1, 7, 10, 12, 17, 18, 19, 21, 33, 35, 38)

<u>Tuesday, September 22</u>: Test chapters 2-6