

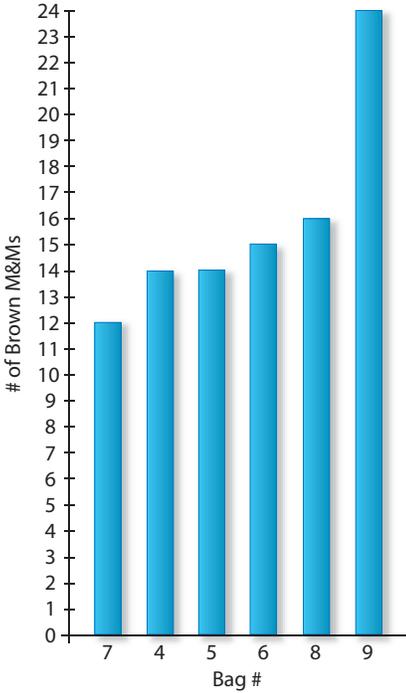
Selected ACE: *Data Distributions*

Investigation 1: #13, 17

Investigation 2: #3, 7

Investigation 3: #8

Investigation 4: #2

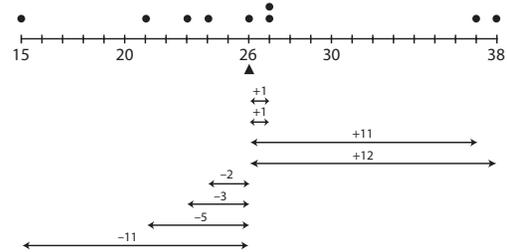
ACE Problem	Possible solution														
Investigation 1															
<p>13.</p> <p>a. The table below shows the data for the brown candies from Bags 4 – 9 of Exercise 1. Make an ordered value bar graph and a line plot for these data.</p> <p>Brown M&M's</p> <table border="1" data-bbox="240 783 745 909"> <thead> <tr> <th>Bag #</th> <th>4</th> <th>5</th> <th>6</th> <th>7</th> <th>8</th> <th>9</th> </tr> </thead> <tbody> <tr> <td>Number of Brown Candies</td> <td>14</td> <td>14</td> <td>15</td> <td>12</td> <td>16</td> <td>24</td> </tr> </tbody> </table> <p>b. What are the minimum and maximum values?</p> <p>c. What is the range?</p> <p>d. Are there gaps or clusters of data? Explain.</p> <p>e. Would an ordered value bar graph or a line plot better represent the data? Explain.</p>	Bag #	4	5	6	7	8	9	Number of Brown Candies	14	14	15	12	16	24	<p>13.</p> <p>a. Note: A value bar graph shows the number of M&M's in each bag. An <i>ordered</i> value bar graph shows the same thing, but the values are arranged in increasing order. These values are not the bag numbers, which are in essence just names. The bags could just as well have been named "A", "B", "C", etc. The values which have to be ordered are the numbers of brown candies. These values range from 12 to 24.</p>  <p>b. Not answered here.</p> <p>c. Not answered here.</p> <p>d. Not answered here.</p>
Bag #	4	5	6	7	8	9									
Number of Brown Candies	14	14	15	12	16	24									

	<p>e. A line plot shows how frequently each value (# M&M's) occurs. This is helpful in looking for clusters of data, or unusual values. Thus a line plot could be helpful here in locating which values occur most frequently, where most values are clustered, which values are typical/unusual, and where significant gaps occur. However, with only 6 pieces of data either graph would give the same information.</p>
<p>17.</p> <p>a. Describe any trends or patterns in immigration to the United states from Asia from 1820 to 2000 using the graph below. (See student text for graph.)</p> <p>b. Write two comparison statements about the trends from Mexico to the United states (Exercises 8 – 11) and from Asia to the United States from 1820 to 2000.</p> <p>c. Look back at Graph 2 in Problem 1.2. As the trend for immigration from Europe was decreasing from 1961 to 2000, what happened to the trends for immigration from Mexico and Asia?</p>	<p>17.</p> <p>a. As a percent of total U.S. immigration, immigration from Asia was too small to be noted from 1820 to 1860, then fairly constant from 1860 to 1960 (less than 5%), and then increased dramatically from 1960 to 2000.</p> <p>Note: We should be cautious about deducing that the raw numbers of Asian immigrants follow the same pattern; to make that deduction we would have to know what the total immigration was for each decade. For example we can not say for sure that 5% of $x < 35\%$ of y, without knowing the values of x and y.</p> <p>b. We can compare the graphs in Exercise 10 and 17 (this exercise) because they both record immigration from a region as a percent of the total immigration to U.S. The over all pattern of dramatic increase is very similar, but there are differences also. Mexican immigration began its increase slightly earlier than Asian</p>

	<p>immigration; Mexican immigration did not increase quite as much as Asian immigration in the decades 1970 to 1990; Mexican immigration is a smaller part of total immigration for the most recent figures (in 2000); Asian immigration may have peaked in 1981 - 1990.</p> <p>Note: We could deduce that there were more Asian immigrants than Mexican immigrants in 1991 to 2000, in terms of raw numbers as well as percents, because both are percents of the same total immigration for that decade.</p> <p>c. Not answered here.</p>
--	---

Investigation 2

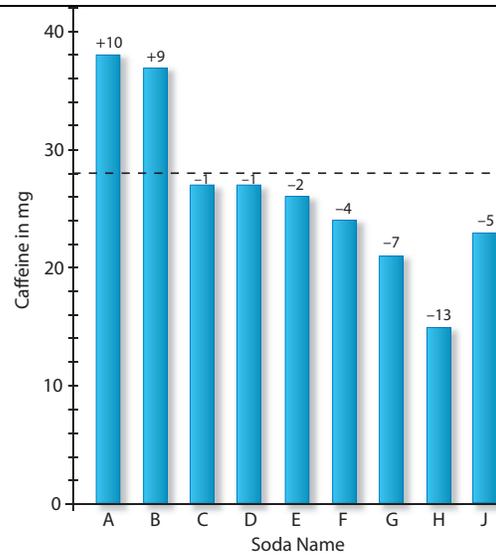
<p>3. a. What is the mean amount of caffeine in the soda drinks?</p> <table border="1" data-bbox="240 1104 889 1388"> <tr> <td>Name</td> <td>A</td> <td>B</td> <td>C</td> <td>D</td> <td>E</td> <td>F</td> <td>G</td> <td>H</td> <td>J</td> </tr> <tr> <td>Caffeine in 8 oz.</td> <td>38</td> <td>37</td> <td>27</td> <td>27</td> <td>26</td> <td>24</td> <td>21</td> <td>15</td> <td>23</td> </tr> </table>	Name	A	B	C	D	E	F	G	H	J	Caffeine in 8 oz.	38	37	27	27	26	24	21	15	23	<p>3. Note: there are several ways to think about finding the mean. Three of these are shown below. These are</p> <ul style="list-style-type: none"> <input type="checkbox"/> Balancing <input type="checkbox"/> Sharing <input type="checkbox"/> Using an Algorithm <p>a. We could think of <i>balancing</i> the distribution of caffeine values. A line plot is a convenient way to show the distribution. The idea of balancing is similar to thinking of a teeter-totter. Students can make a quick estimate of the balance point and then use the exact values shown in the line plot to check. (This is a useful method for making an estimate when the exact values are not all known. See <i>Samples and Populations</i>.)</p>
Name	A	B	C	D	E	F	G	H	J												
Caffeine in 8 oz.	38	37	27	27	26	24	21	15	23												
<p>b. Make a line plot for the soda drinks. c. What is the mean amount of caffeine in the other drinks?</p> <table border="1" data-bbox="240 1530 889 1738"> <tr> <td>Name</td> <td>A</td> <td>B</td> <td>C</td> <td>D</td> <td>Tea A</td> <td>Tea B</td> <td>Coffee</td> <td>Cocoa</td> <td>Juice</td> </tr> <tr> <td>Caffeine in 8 oz.</td> <td>77</td> <td>70</td> <td>25</td> <td>21</td> <td>19</td> <td>10</td> <td>83</td> <td>2</td> <td>33</td> </tr> </table>	Name	A	B	C	D	Tea A	Tea B	Coffee	Cocoa	Juice	Caffeine in 8 oz.	77	70	25	21	19	10	83	2	33	
Name	A	B	C	D	Tea A	Tea B	Coffee	Cocoa	Juice												
Caffeine in 8 oz.	77	70	25	21	19	10	83	2	33												
<p>d. Make a line plot for the other drinks. e. Write three statements comparing the amount of caffeine in soda and other drinks.</p>																					



From the above graph we can see that the estimated mean of 26 is too low, because we have a total difference of +25 above the estimated mean, and a total difference of only -21 below the estimated mean.

OR,

We could think of *sharing* the amounts of caffeine, taking from higher amounts to add to lesser amounts. A bar graph would make the idea of sharing clear. The bar graph below has a horizontal line drawn across at 28. The bars are marked to show how values exceed or fall short of 28 mg of caffeine. We can see that the horizontal line has been set too high because we only have +19 mg excess caffeine from the first two bars to "share" with other values below 28 mg of caffeine. By trial and error we can find a horizontal line that makes the "sharing" process come out evenly.



OR,

We could use the algorithm. The algorithm has the advantage of giving an exact answer for the mean. (Perhaps a disadvantage is that students don't have a picture of how this mean relates to the rest of the distribution of values.)

$$(38 + 37 + 2 \times 27 + 26 + 24 + 21 + 15 + 23) = 238.$$

$$238 / 9 = 26.4 \text{ (approx.)}$$

b. See above.

c. Not answered here.

d. Not answered here.

e.

Students could compare means, and use this measure of center to say that the typical "other" drink has a higher caffeine content.

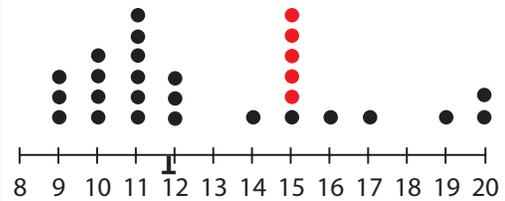
Or they might note that the mean for "other" drinks is affected by three very high values, so that the distribution for "other", shown as a line plot, has a very different shape from the distribution for soda drinks. They might choose

	<p>to use the median as a measure of center, instead of mean. Notice that mean and median are alike for the soda drinks, but quite different for the “other” drinks.</p> <ul style="list-style-type: none"> □ They might say that the “other” drinks show much more variability, and they might measure this variability by using the range, which is 81 for “other” and 23 for “soda.” This large variability for “other” drinks makes any attempt to say what is “typical” very unreliable. □ They might comment on significant gaps that appear in the “other” distribution.
<p>7.</p> <p>a. Compare the three sets of data. Which group of students has longer names? Explain your reasoning. (See student text for graphs.)</p> <p>b. Look at the distribution for 30 students in the U.S. Suppose the data for the six names with 13 letters were each changed to 16 letters.</p> <ul style="list-style-type: none"> i. Draw a plot showing this change. ii. Will this change affect the median name length? Explain. iii. Will this change affect the mean name length? Explain. 	<p>7.</p> <p>a. Since the question is about “longer” names, students might compare the means of medians, and use these measures of center to say which data set has a longer “typical” name. Clearly the center is higher for Russian names.</p> <p>OR,</p> <p>Students might focus on the longest names, the maximum values in the data sets. Again the Russian set of names has the highest maximum, so the absolute longest name in all three sets is a Russian name. Note: this is often an unreliable way to decide on “longest,” since the maximum for any particular set is only one value, and may be very unlike other values in the set, giving a false overall impression.</p> <p>OR,</p> <p>Students might choose a benchmark, such as 15 letters, and say that more than half the Russian names are greater than or equal to 15 letters long, while only about 20% of</p>

Japanese and U.S. names are as long as this. (More on this idea of benchmarks in the next Investigation.)

b.

- i. If we move the 6 pieces of data from 13 letters to 16 letters then the distribution changes its overall shape, from having a generally mound-shaped distribution to having a shape with two distinct mounds.



- ii. Notice that the 6 pieces of data that have been moved were already above the median. Moving them three units right does not change where the “middle” or median of the distribution is. What is important for calculating the median is the order of the data, and the position of the middle piece of data in this order, not how far above (or below) the middle any particular group of data values are.
- iii. Not answered here.

Investigation 3

8. Use the line plots and table below. How much slower are the Trial 1 reaction times for non-dominant hands than the Trial 1 reaction times for dominant hands? Explain. (See student text for graphs and table.)

8. Students have several ways to make a comparison. They might compare **measures of center**. From the graph we can see that the mean reaction time for the dominant hand is about 1.05 seconds, while the mean for the non-dominant hand is 1.3 seconds. Typically the dominant hand is 0.25

	<p>seconds faster. If we compare the medians the dominant hand is 0.2 seconds faster. (We can make more exact comparisons from the table.) The mean is higher than the median for the non-dominant hand because of the influence of 3 unusually slow times (slower than 2 seconds.)</p> <p>OR,</p> <p>we might compare the maximum (slowest) values for each distribution. This would be a poor way to compare. The maximum values are about the same for both distributions, but we can see that the non-dominant times distribution is clearly shifted right of the dominant hand times.</p> <p>OR,</p> <p>They might compare clusters, as a way of addressing typical times. The dominant times are clearly clustered around 1 second, while the non-dominant times seem to have two clusters, around 0.8 seconds and around 1.2 seconds. This is not a very conclusive comparison because the non-dominant times are more variable, and not so clearly clustered around a single value.</p> <p>Or,</p> <p>We might choose a benchmark such as 1.4 seconds. We can say that only 4 times (out of 40) are equal to or slower than 1.4 seconds for the dominant hand, while 15 times (out of 40) are equal to or slower than 1.4 seconds.</p>
<p>Investigation 4</p>	
<p>2. a. The three pairs of line plots below display data about 50 wood roller coasters. Means and medians are marked on each graph.</p>	<p>2. a. As in #8 investigation 3 we have several ways to make comparisons. Below are comparisons of Maximum</p>

(See student text for graphs.)

- a. Write three statements comparing wood roller coasters built before 1960 with wood roller coasters built in 1960 or later.
- b. Hector says that there are too few roller coasters to make comparisons. Do you agree with Hector? Explain.

Drop for the two time periods. The methods used to make the comparisons are

- Comparing centers
- Comparing variability
- Comparing to the same benchmark

(These same methods can be used to make comparisons of Maximum Heights and Top Speeds.)

Comparing Centers:

Both mean and median are greater for the later wood coasters. We can deduce that the typical wood coaster from the later era (1960-2004) has a greater maximum drop.

Comparing variability:

The range for the later coasters is $215 - 35 = 180$ feet. The range for the earlier coasters is $95 - 10 = 85$ feet. From this we could deduce that the later coasters are more variable, BUT this range value is very much influenced by the very unusual value of 215 feet for the later coasters. If we exclude that value the range for the later coasters would be $155 - 35$ or 120 feet, which is still a larger range value than for the earlier coasters.

OR, we could compare clusters. We can see that most of the later wood coasters cluster between 70 and 100 feet, while there is no evident cluster for the earlier coasters.

Both of these ways of thinking about how *spread out* the data are for the two eras would lead us to conclude that the later era shows more variability. BUT there is so little data in the 1902 - 1959 set that it would be impossible for clusters to form. This makes judging variability very

problematic.

Comparing to a benchmark:

We could say that half the later coasters had maximum drops greater than or equal to 88 feet, while only 1 (out of 10) of the earlier coasters had a drop as great as this.

b. Not answered here.