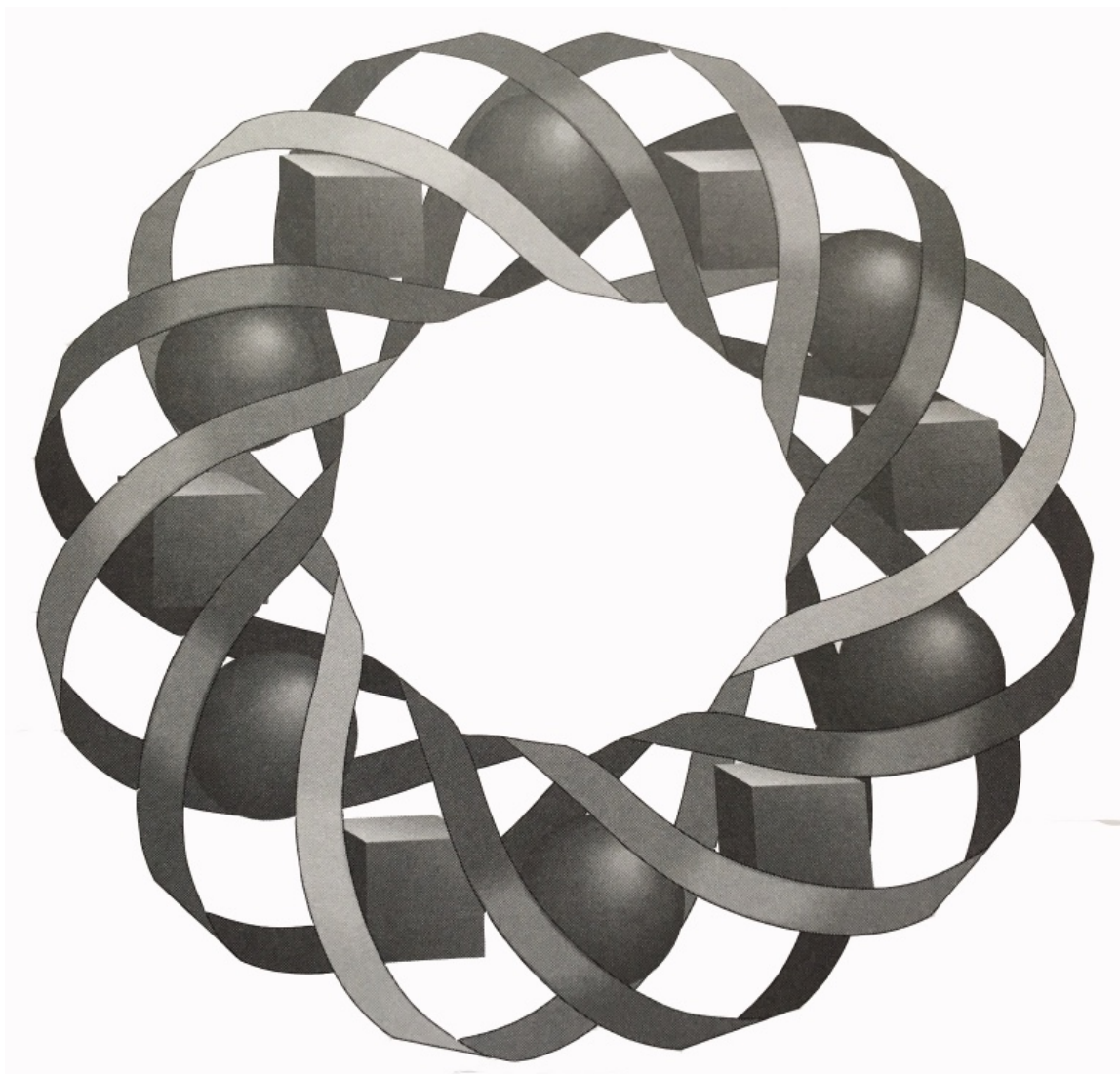# When to Deviate from A Mean Task

You can't argue with statistics. Or can you? In this module, you look at several different strategies for presenting—and interpreting—statistical information.

Randy Carspecken • *Pete Stabio*

When to Deviate from a Mean Task

Introduction

At Two Eagle High School, the senior class is responsible for writing and editing the annual yearbook. This year's seniors want to include stories about events such as the recycling drive and the Academic Bowl. Using statistics reported for each story, the seniors claim their class is the best in the school. The other classes, however, are unhappy with this assessment. They want to do their own analysis of the data. In this module, you develop some statistical methods for responding to the seniors' claim.

Activity 1

Statistics are used by all sorts of people and in all forms of communication. They can be used to explain points of view, promote special interests, and compare specific quantities. A statistical summary of data can be represented in many different ways, often in graphs or charts. Two of the most commonly used methods for displaying statistics are **histograms** (a type of bar graph) and **pie charts**.

Mathematics Note

Histograms display information using rectangles or bars. The bars must be of uniform width and the scales must have uniform intervals.

For example, Figure **1** shows a histogram of the number of movies attended by Two Eagle students in September. The horizontal axis shows the number of movies attended, while the vertical axis shows the number of students in each category.
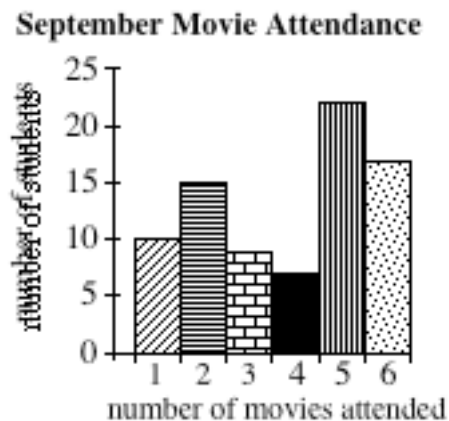

Figure 1: Histogram of September movie attendance

Exploration

The seniors on the Two Eagle yearbook staff want to include the first semester honor roll in their annual. They chose the graph shown in Figure **2** to display their statistics.

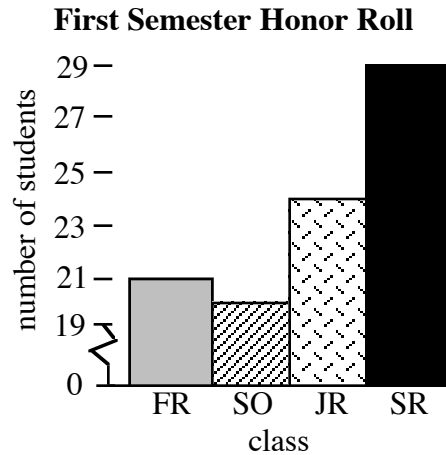**First Semester Honor Roll**

Figure 2: A histogram of the first semester honor roll

Notice the break in the vertical axis. In this case, the break indicates that the numbers from 0 to 19 are not shown on the graph.

Mathematics Note

A **frequency table** consists of two columns. One column contains a data item; the other displays the number of observed occurrences of that item. The number of occurrences of any particular data item is its **frequency**.

For example, the histogram in Figure **1** was generated from the frequency table shown in Figure **3**. This table summarizes the results of a survey of 80 students.

Number of Movies	Frequency
1	10
2	15
3	9
4	7
5	22
6	17
Total	80

Figure 3: Frequency table for September movie attendance

---

a. Use the data presented in Figure **2** to create the corresponding frequency table.

b. Use your frequency table from Part **a** to create a histogram with a vertical axis that begins at 0 and ends at 30, without a break. Compare this new graph with the original in Figure **2**. **Note:** Save your frequency table and histogram for use later in this module.

c. Two Eagle High has a total of 104 freshmen, 81 sophomores, 110 juniors, and 128 seniors. Create another histogram to show the percentage of honor-roll students in each class. Compare this histogram with the one shown in Figure **2**.

Discussion

a. How accurately does the graph in Figure **2** portray the number of students from each class on the honor roll?

b. Compare your graphs from Parts **b** and **c** of the exploration. Which graph do you prefer? Explain your response.

c. How could the statistics encountered in the exploration be used to favor a particular class?

Mathematics Note

Pie charts (or circle graphs) consist of circular regions divided into sectors, each representing a percentage of the whole. For example, the pie chart in Figure **4** illustrates the same data as the histogram in Figure **1**.
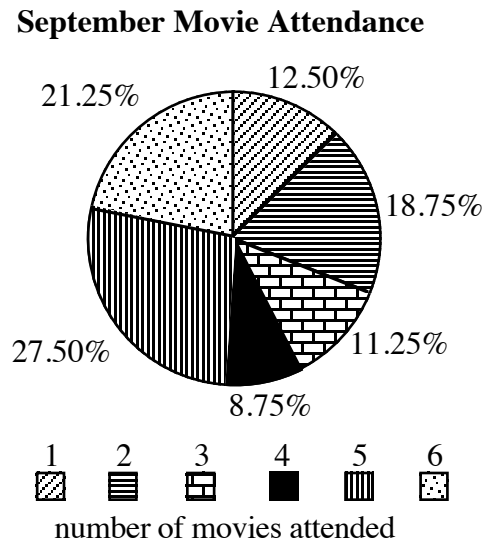
September Movie Attendance

Figure 4: Pie chart of September movie attendance

Like the corresponding histogram, the pie chart in Figure **4** can be generated using a frequency table. The percentages are determined by dividing each frequency by the total. Then, to determine the measure of the central angle that should represent each frequency, these percentages are multiplied by 360°.

For example, 10 students attended 1 movie in September. This represents 10/80 or 12.5% of the total. The measure of the central angle for the corresponding sector of the circle is $0.125 \cdot 360° = 45°$.

d. The sum of the percentages in the pie chart in Figure **4** is 100%. How could you round each percentage to the nearest whole number, while maintaining a sum of 100%?

Assignment

1.1 **a.** Use the frequency table from Part **a** of the exploration to construct a pie chart.

b. Compare the pie chart to the histogram in Figure **2**. What advantages or disadvantages do you observe with the pie chart?

c. If the number of students in the sophomore class doubled, while the number of sophomores on the honor roll remained the same, would you need to adjust your pie chart? Explain your response.

d. Upon seeing your pie chart, one student commented, "The seniors had the largest percentage of their class on the honor roll." Explain why this is a misinterpretation of the pie chart.

1.2 In order to display additional information on the same graph, a histogram can be "stacked." Both histograms below were created using information from the same survey.
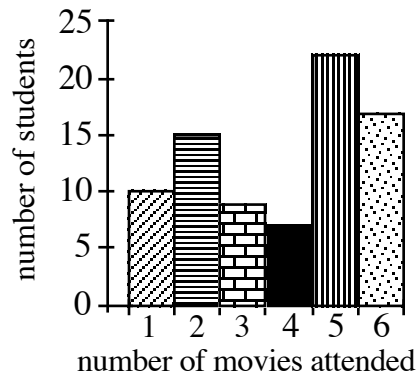

September Movie Attendance
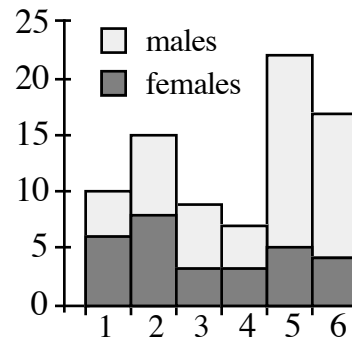
a. Use the stacked histogram on the right to make an interpretation of the data that was not possible with the histogram on the left.

b. Explain why it might be difficult to use the stacked histogram to compare the numbers of male students in each attendance category.

70

1.3 Your frequency table from Part **a** of the exploration describes the number of students from each class on the first semester honor roll. Two Eagle High has a total of 104 freshmen, 81 sophomores, 110 juniors, and 128 seniors.

a. Create a stacked histogram that displays both these sets of information.

b. Using your stacked histogram, can you identify the class with the largest number of students on the first semester honor roll? Explain your response.

c. Using your stacked histogram, can you identify the class with the greatest percentage of students on the first semester honor roll? Explain your response.

1.4 The following graph summarizes the results of Two Eagle High's recycling drive. The yearbook staff plans to use this graph to show that the seniors collected more aluminum than any other class.

Aluminum Collected
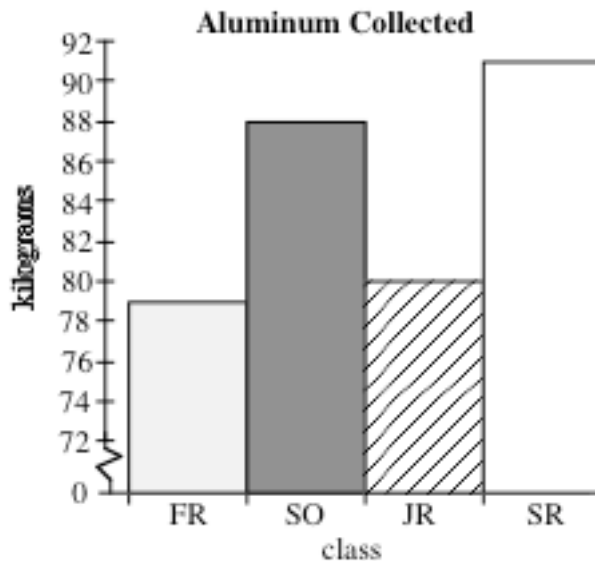
a. How could you change the histogram to represent the freshman class more positively?

b. Two Eagle High has a total of 104 freshmen, 81 sophomores, 110 juniors, and 128 seniors. Using this information, create a graph which shows that a class other than the seniors was most effective in collecting aluminum for the recycling drive.

1.5   The numbers of males and females inducted in the Two Eagle High School Honor Society are displayed in the table below.

Class	Gender	No. of Applicants	No. Inducted
Junior	male	20	12
	female	10	5
Senior	male	20	6
	female	40	16

a. Using the data in this table, create a graph to support each of the following statements.

1. The females at Two Eagle High are better students than the males.

2. The males at Two Eagle High are better students than the females.

3. The honor society's selection process favored seniors.

4. The honor society's selection process favored juniors.

b. Do you think it is possible to use accurate statistics to support an inaccurate or deceptive position? Explain your response.

* * * * *

1.6   The following table shows the land area and 1990 population for Alaska, Texas, California, and Montana.

State	Land Area (km^2)	1990 Population
Alaska	1,477,267	550,043
California	403,970	29,760,021
Montana	376,991	799,065
Texas	678,358	16,986,510

a. 1. Create a pie chart of the information on land area.

2. Use your chart to compare the areas of each of the following pairs of states: Alaska and Texas; Alaska and California; Alaska and Montana

b. 1. Create a pie chart of the information on population.

2. Use this chart to compare the populations of each of the following pairs of states: California and Texas; California and Alaska; California and Montana.

1.7   The following table shows the land area and 1990 population of the five most densely populated states.

State	Land Area (km^2)	1990 Population
Connecticut	12,550	3,287,116
Maryland	25,316	4,781,468
Massachusetts	20,300	6,016,425
New Jersey	19,215	7,730,188
Rhode Island	2,706	1,003,464

a.   Jody claims that Rhode Island has the most people per square kilometer because it has the smallest area. Is her claim correct? Explain your response.

b.   Use a histogram to compare the population densities of the five states.

1.8   Use each of the following types of graphs to emphasize a characteristic of your state.

a.   a pie chart

b.   a histogram

* * * * * * * * * *

Research Project

Design and write a survey on a topic of interest to your school. The survey should consist of no more than 10 carefully worded questions.

Give the survey to 30 people in your school.

After completing Activities **1–4** of this module, analyze the collected data using appropriate graphs and statistical measurements.

Prepare an oral presentation that describes the conclusions you drew from the results of the survey.

Activity 2

When reporting statistical information, the elements of an entire data set are often represented by a single number. For example, instead of listing the size of each class in your school, you might give the school's mean class size. In this case, the mean (or average) class size is a single statistic "representing" all class sizes. Statistics such as the mean, median, and mode are **measures of central tendency**.

Exploration

The Two Eagle High Physical Education Department administers a fitness test before allowing students to participate in sports. This test includes a check of the student's resting and active pulse rates. In this exploration, you collect and analyze data on pulse rates for three levels of activity.

a. After sitting quietly for a period of five minutes, measure and record your pulse rate (beats per minute).

b. Exercise for one minute. Immediately after exercising, measure and record your pulse rate again.

c. Rest from your exercise for a period of two minutes. Measure and record your pulse rate one more time.

d. Organize the class data in a table. **Note:** Save this table for use in the assignment.

e. For each level of activity, determine a representative pulse rate for the class.

Discussion

a. What were the representative pulse rates for your class? How did you choose these numbers?

b. How did changing the level of activity affect the representative pulse rate?

c. In discussing an individual's pulse rate, what additional information might be important?

d. Do you think that your class statistics are representative of the entire school?

Assignment

2.1 The following table shows the number of siblings reported by nine students at Two Eagle High School.

Student	Number of Siblings
A	1
B	0
C	2
D	2
E	0
F	3
G	2
H	6
I	1

    a. Which measure of central tendency best represents the number of siblings per student? Justify your choice.

    b. Create a frequency table of this data. For example, because students B and E each have 0 siblings, the frequency for 0 siblings is 2.

    c. Use your frequency table to create the corresponding histogram. Represent the number of siblings on the horizontal axis and the frequency on the vertical axis.

2.2 Measures of central tendency are often used to describe an entire set of data. In situations involving small data sets, measures of central tendency can be misleading.

    a. Identify four people in your community with different occupations. Make a table that displays each person's occupation and estimated annual income. Determine the mean, median, and mode of the income data.

    b. Add the occupation and estimated annual income of a professional athlete to your table. Determine the mean, median, and mode of the income data for this new group.

    c. Does the mean income you calculated in Part **a** accurately represent the incomes of the original group? Explain your response.

    d. Does the mean income you calculated in Part **b** accurately represent the incomes of the new group? Explain your response.

    e. Which of the three measures of central tendency—mean, median, or mode—is most affected by data values very different from the majority of values? Which of the three measures of central tendency is affected the least?

2.3   Describe a set of data in which the mode does not exist.

2.4   Describe some possible pitfalls associated with using a measure of central tendency to describe a set of data.

Mathematics Note

A **stem-and-leaf plot** displays the values in a data set. To simplify interpretation, data is usually ordered and a legend is included.

For example, the stem-and-leaf plot in Figure **5** displays the number of hours worked by the Two Eagle yearbook staff during the month of September.

```
0 | 1257
1 | 4688
2 | 03357
3 | 788999
4 | 1145
5 |
6 | 7          1 | 4   represents 14 hr
```
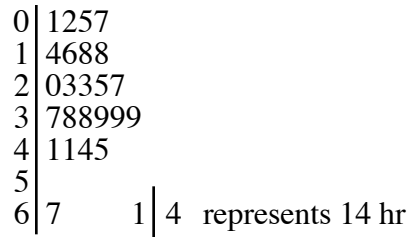
Figure 5: Hours worked by Two Eagle yearbook staff

In the row that reads 2Ω03357, the stem is 2. The numbers 0, 3, 3, 5, and 7 are leaves. This row represents the data points 20, 23, 23, 25, and 27 hr. The stem 5 has no leaves. This indicates that none of the staff worked from 50 hr to 59 hr.

A **back-to-back stem-and-leaf plot** can be used to compare two sets of data. The stems are placed in the center column of the plot. One set of leaves is placed to the left of the stems, the other set to the right.

For example, the back-to-back stem-and-leaf plot in Figure **6** displays the same information on work hours as Figure **5**. It also shows how those hours were distributed between the boys and the girls.

```
Girls       |   | Boys
      52 | 0 | 17
      84 | 1 | 68
      53 | 2 | 037
     997 | 3 | 889
       5 | 4 | 114
         | 5 |          4 | 1 | 6   represents 16 hr for boys
       7 | 6 |                      and 14 hr for girls
```
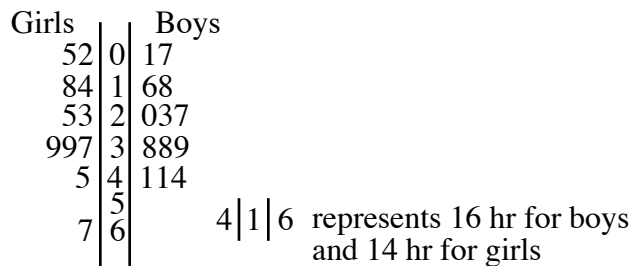
Figure 6: Hours worked by girls and boys on the yearbook staff

2.5   The **range** of a data set is the difference between the greatest data value and the least data value. The range provides one indication of how widely the data varies.

   a.   Determine the range of the data in Figure **5**.

   b.   Can the range of a set of data be negative? Explain your response.

2.6   The heights in centimeters of 10 students at Two Eagle High are listed below. The heights of females are underlined.

   <u>153</u>, 157, 163, <u>165</u>, 166, <u>169</u>, 170, <u>173</u>, <u>176</u>, 185

   a.   Determine the mean, median, and mode of the data.

   b.   Construct a stem-and-leaf plot of the data, using the first two digits of the heights as the stems and the last digits as the leaves.

   c.   Find the range of heights.

   d.   Create a histogram of the data using the following intervals for heights in centimeters: [150, 159], [160, 169], [170, 179], and [180, 189].

   e.   Separate the data into two categories: heights of female students and heights of male students. Use the data to make a back-to-back stem-and-leaf plot.

   f.   Use the back-to-back stem-and-leaf plot to create a stacked histogram.

2.7   Advertisers often omit important information in their sales pitches. One advertisement in the Two Eagle yearbook claims that "after more than a million miles of testing, our automobiles had a repair-free record."

   a.   Does the "repair-free" statistic alone convince you that the cars are reliable? Describe one way in which a "million miles of testing" could fail to prove reliability.

   b.   To convince you that the cars are reliable, which measure of central tendency should the advertisement mention?

2.8   **a.**   Based on the pulse-rate data you collected in the exploration, create an advertisement about the physical fitness of America's youth. Make any claim you can support.

   b.   What are some of the weaknesses of your claim?

<p style="text-align:center">* * * * *</p>

2.9   **a.** Use measures of central tendency to compare the New England states with the Mountain states.

	State	Land Area (km^2)	1990 Population
New England States	Connecticut	12,550	3,287,116
	Maine	79,939	1,227,928
	Massachusetts	20,300	6,016,425
	New Hampshire	23,231	1,109,252
	Rhode Island	2,706	1,003,464
	Vermont	23,956	562,758
Mountain States	Arizona	296,400	3,665,228
	Colorado	268,660	3,294,394
	Idaho	214,235	1,006,749
	Montana	376,991	799,065
	Nevada	284,397	1,201,833
	New Mexico	314,334	1,515,069
	Utah	212,816	1,722,850
	Wyoming	251,501	453,588

b. Would you rather live in the New England states or the Mountain states? Use the statistics from Part **a** support your choice.

2.10   **a.** Create a back-to-back stem-and-leaf plot to represent the land areas of the states in the following two regions. (You may want to round these values before creating your stem-and-leaf plot.)

	State	Land Area (km^2)
Region 1	Alabama	131,443
	Florida	139,852
	Georgia	150,010
	Mississippi	121,506
	North Carolina	126,180
	South Carolina	77,988
	Tennessee	106,759
Region 2	Illinois	143,987
	Indiana	92,904
	Iowa	144,716
	Michigan	147,136
	Ohio	106,067
	Wisconsin	140,673

b. Compare the land areas for the two regions.

＊ ＊ ＊ ＊ ＊ ＊ ＊ ＊ ＊ ＊

Activity 3

To help identify potential athletes, the Athletic Department would like a profile of all Two Eagle High students. Since larger hands give basketball or volleyball players an advantage, the coaches of those two teams are especially interested in student hand spans. A hand span is the distance from the end of the thumb to the end of the little finger when the two are stretched as far apart as possible.

Exploration

The hand spans, in centimeters, of 22 sophomores at Two Eagle High School are listed below (hand spans of females are underlined): 15.4, 23.4, 19.9, <u>14.8</u>, 17.6, <u>19.3</u>, 20.7, 22.3, <u>20.0</u>, <u>16.4</u>, 18.1, <u>19.6</u>, <u>21.9</u>, 24.1, <u>17.5</u>, 23.9, 25.2, <u>18.5</u>, 24.1, <u>17.6</u>, <u>15.8</u>, 23.4.

a. **1.** Order the hand-span data from least to greatest.

  2. Determine the median of the entire set of data.

  3. Determine the median of the "upper half" of the data. This number is the **upper quartile**.

  4. Determine the median of the "lower half" of the data. This number is the **lower quartile**.

  5. Calculate the difference between the upper and lower quartiles. This difference is the **interquartile range**.

  6. Data values that are more than 1.5 times the interquartile range above the upper quartile or 1.5 times the interquartile range below the lower quartile are **outliers**. Identify the outliers, if any, in the hand-span data.

Mathematics Note

A **box-and-whisker plot** (or box plot) is a way to represent a set of data graphically. A box plot displays the median, lower quartile, and upper quartile of a data set, as well as any outliers. It may be constructed either horizontally or vertically.

A box-and-whisker plot can be created using the following steps.

- Order the data from least to greatest.
- Draw a scale that covers the range of data.
- Draw a rectangular box that stretches from the lower quartile to the upper quartile. (The width of the box is not significant.)
- Place a bar inside the box to mark the median.
- Find the interquartile range.
- Determine if there are any outliers.
- Mark any outliers with asterisks.
- Draw the **whiskers** (segments that connect each end of the box to the farthest data values which are *not* outliers).

For example, the following data set describes the number of minutes spent doing mathematics homework by 30 Two Eagle sophomores: $0, 4, 5, 9, 9, 9, 10, 16, 16, 18, 18, 18, 18, 19, 21, 21, 21, 21, 22, 22, 22, 23, 24, 26, 26, 27, 29, 29, 39, 45$.

For this data set, the lower quartile (LQ) is 16, the median (M) is 21, the upper quartile (UQ) is 24, and the interquartile range is $24 - 16 = 8$. The outliers are values less than $16 - (1.5)8 = 4$, or greater than $24 + (1.5)8 = 36$. A box plot for this data is shown in Figure **7**.
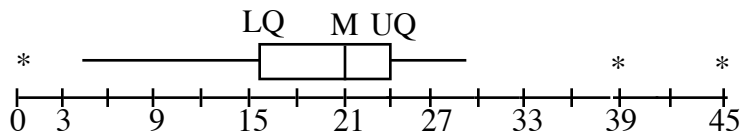

Figure 7: Box-and-whisker plot

b. Create three different box-and-whisker plots of the hand-span data. One should display the data for all 22 students, one should display just the boys' hand spans, and one should display just the girls' hand spans. Compare the three plots. **Note:** Save these plots for use in the assignment.

c. Measure and record the span of your dominant hand in centimeters.

d. Add your hand-span measurement to the hand spans of all 22 students. Create a box-and-whisker plot for this new data set. Compare this plot to the plot for all 22 students from Part **b**.

e.   Add your hand-span measurement to either the boys' or the girls' data. Create a box-and-whisker plot for this new set of data. Compare this plot to the corresponding plot from Part **b**.

Discussion

a.   What would have more impact on a box-and-whisker plot—adding a data value within the interquartile range, or adding an outlier? Explain your response.

b.   What advantages are there to using the median instead of the mean when describing a data set?

c.   Is the mean of a data set readily identifiable in a box-and-whisker plot? Explain your response.

d.   What are some of the advantages and disadvantages of using a box-and-whisker plot to illustrate data?

e.   A box-and-whisker plot naturally leads to a so-called "five-point analysis" of a data set. What do you think these five points are?

Assignment

3.1   Use the box-and-whisker plots from Part **b** of the exploration to help answer the following questions about box-and-whisker plots in general. Explain your response for each one.

   a.   Are the lengths of the whiskers related to the interquartile range?

   b.   How would adding an outlier to the data affect the plot?

   c.   How would adding a data point within the interquartile range affect the plot?

3.2   Using as many statistical measures as you can, describe how your hand span compares to each of the following:

   a.   the hand spans of all 22 Two Eagle sophomores

   b.   the hand spans of the sophomore girls

   c.   the hand spans of the sophomore boys.

3.3   The world's tallest man, Robert Wadlow, had a hand span of about 37 cm. If this value were added to the data set in the exploration, would it be considered an outlier?

3.4   The following graph shows box plots for the number of sit-ups performed by three different classes at Two Eagle High.
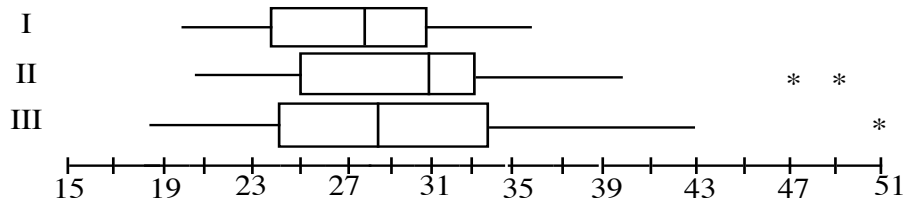
a.   Which class had the person who did the most sit-ups?

b.   For which class did the number of sit-ups vary the least?

c.   Which class had the greatest upper quartile?

d.   In which class did the bottom 25% perform the most sit-ups?

e.   Which class has the greatest number of students?

3.5   Create a set of data with at least 16 items in which the mean and the median have the same value and there is at least one outlier. Explain the process you used to create the data set. Make a box plot of your data.

* * * * *

3.6   A survey conducted at two different universities, A and B, produced the following box plots.

Age of Students
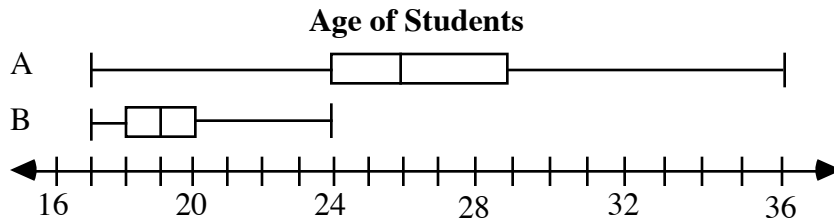
a.   Use the information in the plots to compare the students at university A to the students at university B.

b.   Describe some information which may have been determined from the data but is not shown in the box plots.

3.7   a.   Create a box plot for each of the following sets of data.

1.   {5,5,5,5,5,30,30,30,30,30,70,70,70,70,70,95,95,95,95,95}

2.   {5,6,7,28,37,41,50,61,68,76,89,92,95}

b.   Compare the two plots.

c.   Explain why a five-point analysis using box plots sometimes can be misleading.

d.   Identify a type of graph that would better show the differences in the two data sets. Create the corresponding graphs.

* * * * * * * * * *

Activity 4

Measures of central tendency are not always useful when comparing data sets. If the mean, median, and a mode of a data set are the same, or relatively close, other statistical measures are needed to provide a better picture of the data.

Exploration 1

Every year, Two Eagle High sponsors a competition between classes called the Academic Bowl. The rules for the contest are listed below.

- Each class may enter only one team.

- Each team must consist of four students from one class.

- Each match is scored on a 100-point scale.

- Team scores are determined by averaging the scores of the four team members.

- Team members may work independently or cooperatively.

- Team members may change after any match.

The teams meet five times during the year, with the overall winner advancing to a state competition. Based on the information in Table **1**, the seniors believe that they should be declared the winners.

Table 1: Results of the Academic Bowl

	Freshmen	Sophomores	Juniors	Seniors
Match 1	100	74	82	99
Match 2	60	90	72	66
Match 3	77	85	88	85
Match 4	94	82	89	100
Match 5	85	85	85	66

You have been selected as a judge for this year's contest. How will you choose a winner?

- **a.** Enter the data from Table **1** in a spreadsheet. Record the scores for each team in its own column. **Note:** Save this spreadsheet for use in Exploration **2**.

- **b.** Create a histogram for each team's scores. To simplify comparisons, all four histograms should be the same size and have the same scales.

- **c.** Use your histograms along with the measures of central tendency (mean, median, and mode) to determine the winner of the Academic Bowl.

- **d.** Describe the process you used to choose a winner.

Discussion 1

a. Which class did you select as the winner? What criteria did you use?

b. Using the histograms you created in Part **b** of Exploration **1**, compare the range in scores for each class.

c. Would the difference in the range of scores for each class provide a reasonable criterion for selecting a winner? Explain your response.

d. Which of the rules for the competition might affect the range of scores? Explain your response.

Exploration 2

In Exploration **1**, you found that measures of central tendency alone may not provide enough information to select a winner for the Academic Bowl. Along with the mean, median, or mode, the variation among scores might be used to help determine a winner.

Mathematics Note

There are several methods for measuring variation in a set of data. One of these measures, **mean absolute deviation**, describes the average distance from the mean for the numbers in a data set. For a set of data with n members represented by x_1, x_2, …, x_n, the mean absolute deviation is given by:

$$\frac{|x_1 - \mu| + |x_2 - \mu| + \cdots + |x_n - \mu|}{n}$$

where the notation $|\;|$ represents absolute value and μ, the lowercase Greek letter *mu*, is the mean of the data.

For example, consider the five match scores for the seniors in the Academic Bowl. The mean (μ) of these scores is 83.2. The distance from the mean for each score—in other words, the absolute value of the difference between the score and the mean—is shown in Figure **8**. The mean absolute deviation of the scores is the mean of these distances: $68.8/5 \approx 13.8$.

Score	Distance from Mean
99	15.8
66	17.2
85	1.8
100	16.8
66	17.2
Total	68.8

Figure 8: Distance from mean of the seniors' scores

---

a. Use the data in Table **1** to determine the mean absolute deviation for each team's scores.

b. Create a set of scores using any five values from 50 to 100. Experiment with these values to obtain a set of scores with a mean absolute deviation of 0. Record your data set.

c. Create another set of scores with any five values from 50 to 100. Experiment with these to obtain a set of scores with a mean absolute deviation as great as possible. Record your data set.

Discussion 2

a. Why is absolute value used in the calculation of mean absolute deviation?

b. Mean absolute deviation is defined as the average distance from the mean for the numbers in a data set. How does the formula given in the mathematics note illustrate this definition?

c. Figure **9** below shows a histogram of the number of parking tickets issued by a traffic officer in five days. The horizontal line indicates the mean number of tickets per day. How is the distance from the top of each bar to the line related to mean absolute deviation?
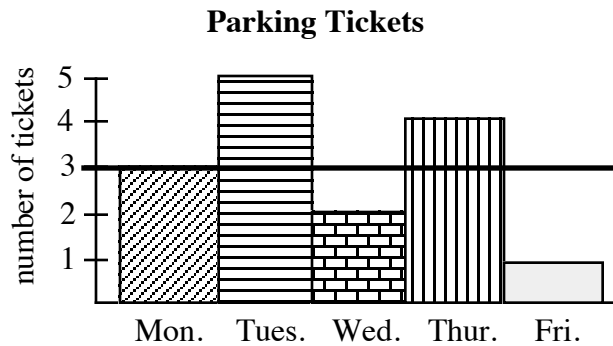
Parking Tickets

Figure 9: Histogram of parking tickets

d. How do the histograms from Exploration **1** reflect the fact that the mean absolute deviations for the teams are different?

e. How would considering the mean absolute deviation of each team's scores affect your choice for the winner of the Academic Bowl?

f. How might mean absolute deviation and measures of central tendency be used to predict a team's score in a sixth match?

g. Which team is least likely to score 95 in a sixth match? Explain your response.

Assignment

4.1 The following graphs display the scores for each member of the junior and sophomore teams in the fifth match of the Academic Bowl.

The members of the junior team scored 85, 85, 75, and 95, while all the members of the sophomore team scored 85.
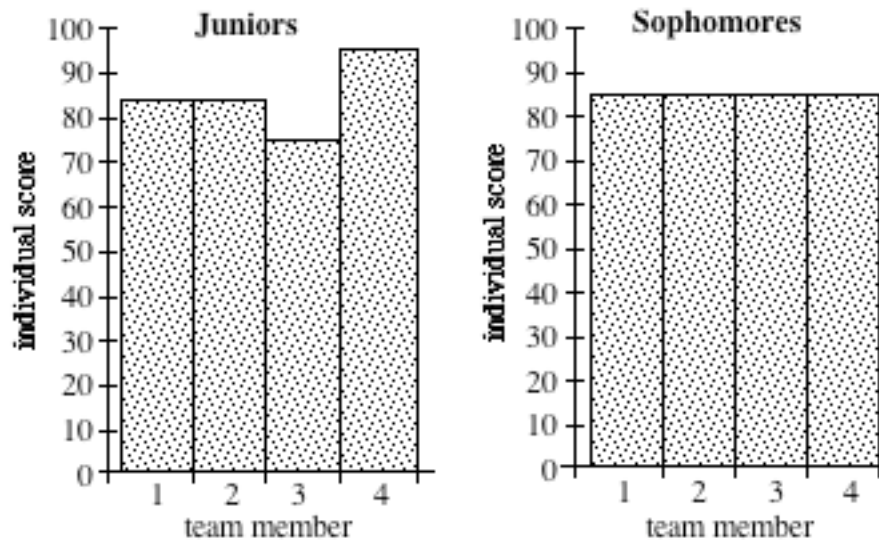
a. Can you differentiate between these two data sets using measures of central tendency alone? Explain your response.

b. Use the juniors' scores to describe how two unequal scores can have equivalent measures of deviation from the mean.

c. If team member 4 on the junior team scored 90 instead of 95, and team member 3 scored 80 instead of 75, how would this affect the mean absolute deviation? Explain why this occurs.

d. How would the mean absolute deviation be affected if a fifth member of the junior team scored 85? Explain your response.

4.2 The individual scores for the seniors in the fifth match of the Academic Bowl were 65, 99, 45, and 55. For the freshmen, the individual scores were 77, 93, 77, and 93.

a. Use this information, along with that given in Problem **4.1**, to determine the mean absolute deviation for the scores of each team in the fifth match.

b. Why can the mean absolute deviation for the sophomores' scores be determined without any calculations?

c. According to the rules for the Academic Bowl, team members may work independently or cooperatively. Describe the method which you think each team may have used. Explain your reasoning.

4.3   The two histograms below display the results of an academic bowl at a neighboring school. Each graph shows both the team scores and their distances from the mean for five matches.
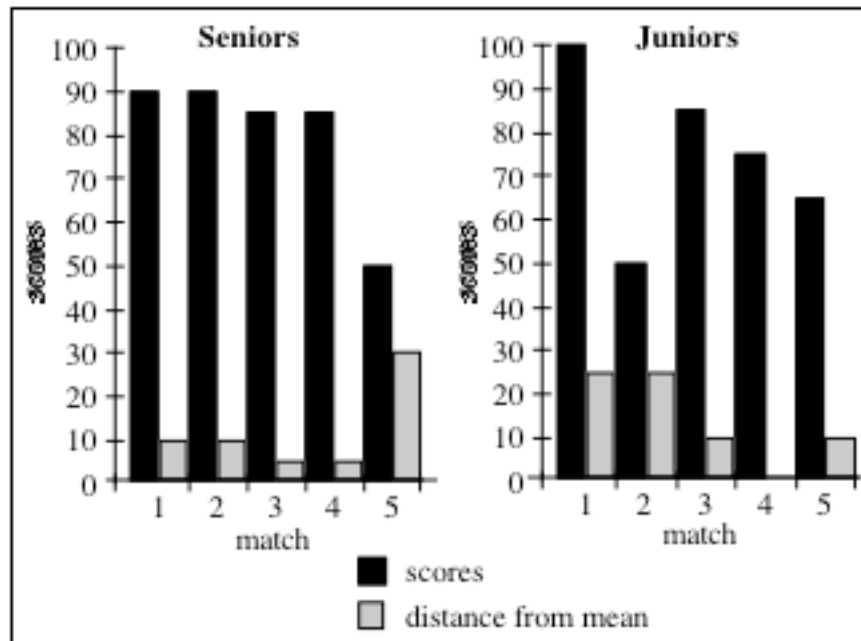
a. Identify the scores that deviate more than "an average amount" from the mean.

b. What statistical measure did you use to determine "an average amount" in Part **a**?

c. What is the significance of the missing bar in the fourth match for the juniors?

4.4   As described in the mathematics note, the mean of the Two Eagle seniors' five match scores is 83.2, with a mean absolute deviation of 13.8. The scores within 1 mean absolute deviation of the mean are contained in the interval [69.4, 97].

a. Describe how the interval [69.4, 97] was determined.

b. When is a data point 0 mean absolute deviations from the mean?

c. Determine the interval that contains scores within 2 mean absolute deviations of the mean. (Hint: Make sure that the extreme values in this interval make sense.)

4.5   The scores for the senior team in Problem **4.3** were 90, 89, 87, 86, and 51. The scores for the junior team were 100, 50, 85, 75, and 65.

    a. Calculate the mean absolute deviation for each team's scores.

    b. How does the mean absolute deviation describe the variation in scores for each team?

    c. For each team, identify all scores more than 1 mean absolute deviation from the mean.

4.6   The winning team in the Academic Bowl advances to the state competition. In the past, teams that score at least 80 points typically place among the top three at state. Select a winning team from Two Eagle High and justify your choice.

<p align="center">* * * * *</p>

4.7   The table below shows some statistics for two basketball players at Two Eagle High. Rita is the leading scorer for the girls' team and Bernard is the leading scorer for the boys' team.

Player	Rita	Bernard
Average Points per Game	25	25
Mean Absolute Deviation (in Points per Game)	5	15
Points Scored in Last Game	35	35

    a. Which player is the more consistent scorer? Explain your response.

    b. For each player, identify the range of values that are within 1 mean absolute deviation of the mean.

    c. Which player's performance in the last game is more impressive? Explain your response.

    d. Both the boys' and girls' teams score an average of 60 points per game, with the same mean absolute deviation. Compare the range of the average scoring of Rita's teammates to that of Bernard's teammates.

4.8   At the end of the cross country season, the coach presents an award to the most improved runner. This award goes to the athlete whose time in the sixth race of the season is most extraordinary when compared with the mean time of the previous five races. The table below shows the times in each race for the top three candidates. Which runner should receive the award? Justify your choice.

	Runner 1	Runner 2	Runner 3
Race 1	25 min, 30 sec	20 min, 10 sec	22 min, 15 sec
Race 2	23 min, 45 sec	19 min, 50 sec	25 min, 20 sec
Race 3	24 min, 15 sec	21 min, 30 sec	20 min, 15 sec
Race 4	22 min, 5 sec	19 min, 45 sec	24 min, 25 sec
Race 5	21 min, 10 sec	18 min, 40 sec	20 min, 10 sec
Race 6	21 min, 5 sec	19 min, 20 sec	21 min, 45 sec

* * * * * * * * * *

Activity 5

When considered along with measures of central tendency, mean absolute deviation can help you interpret a set of data. Another measure of the spread in a data set is **standard deviation**. Although mean absolute deviation and standard deviation are usually close in value, standard deviation is preferred by statisticians because of its mathematical properties.

Mathematics Note

Standard deviation is a common measure of the spread in a data set. It is calculated by first squaring the individual distance from the mean (the deviation) for each piece of data, then finding the mean of these squared deviations. The standard deviation is equal to the non-negative square root of the mean of the squared deviations.

Standard deviation is often represented by the symbol σ (the lowercase Greek letter *sigma*). For a population with n members represented by x_1, x_2, ..., x_n, where μ is the mean, the standard deviation is given by the formula below:

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}}$$

For example, consider the scores for the freshmen team in the fifth match of the Academic Bowl. The mean score (μ) is 85.25. The sum of the squared deviations may be calculated by using a spreadsheet, as shown in Figure **10**.

Score (x_n)	$(x_n - \mu)$	$(x_n - \mu)^2$
85	–0.25	0.06
74	–11.25	126.56
92	6.75	44.56
90	4.75	22.56
	Total	194.74

Figure 10: Determining standard deviation

The mean of the sum of the squared deviations is $194.74/4 \approx 48.69$. The standard deviation is $\sqrt{48.69} \approx 6.98$.

Discussion

a. How are the formulas for standard deviation and mean absolute deviation similar? How are they different?

b. How is the difference between the mean and each data point kept positive in the formula for standard deviation?

c. What is the purpose of taking the square root of the mean squared deviation?

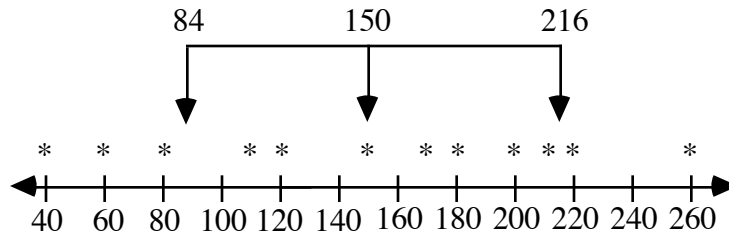d. Why is it important to consider only the non-negative square root?

Assignment

5.1 In the five matches of Two Eagle's Academic Bowl, the senior team scored 99, 66, 85, 100, and 66. Calculate the mean and the standard deviation for these scores.

5.2 Many calculators have built-in functions to compute the mean and standard deviation for a set of data.

Enter the Academic Bowl scores for the juniors (82, 72, 88, 89, and 85), the sophomores (74, 90, 85, 82, and 85), and the freshmen (100, 60, 77, 94, and 85) into your calculator.

a. Use a calculator function to find the mean for each team's scores. One symbol commonly used to represent the mean is $\bar{x}$. What notation does your calculator use?

b. Use a calculator function to find the standard deviation for each team's scores. One common notation for standard deviation is σ. What notation does your calculator use?

5.3   The past 12 drama performances at Two Eagle High attracted the following audiences: 220, 210, 180, 170, 120, 60, 40, 80, 110, 150, 200, and 260. This data is displayed in the line plot below. The mean for the 12 performances is identified by the middle arrow. The other two arrows indicate values 1 standard deviation from the mean.

```
        84            150            216
         |             |              |
         v             v              v

    *   *   *    * *    *   * *   * **       *
    +--+---+---+---+---+---+---+---+---+---+---+--->
    40  60  80  100 120 140 160 180 200 220 240 260
```

a. How many performances were attended by an audience more than 1 standard deviation from the mean?

b. In this data set, are there any audiences more than 2 standard deviations from the mean? Explain your response.

* * * * *

5.4   The following table shows the mass in kilograms for the players on a high school football team.

54	85	78	67	75
75	76	75	80	85
62	90	88	79	68
55	95	65	85	92
62	83	72	90	88

a. Find the mean mass for the players on the team.

b. Calculate the standard deviation for the mass of the players.

c. What percentage of the players have a mass more than 1 standard deviation from the mean?

d. Are there any players whose masses are more than 2 standard deviations from the mean? If so, identify their masses.

5.5   The interval [230, 450] represents values within 1 standard deviation of the mean for a data set.

a. Determine the mean of the data set.

b. Determine the interval that represents values within 2 standard deviations of the mean.

c. Determine the interval that represents values within 3 standard deviations of the mean.

* * * * * * * * * *

91

Summary Assessment

At Two Eagle High, the biology teacher has two classes: third period and fifth period. Both classes take the same tests. On previous tests, the average score in each class has been very close.

On the most recent test, the third-period class received the following scores: 52, 60, 66, 71, 72, 73, 75, 77, 78, 80, 82, 84, 85, 88, 89, and 90.

On the same test, the fifth-period class received these scores: 76, 77, 78, 79, 80, 80, 82, 83, 84, 85, 85, 88, 90, 90, 90, 92, 92, 96, 96, and 99.

1. Create a back-to-back stem-and-leaf plot of this data.

2. Determine the mean test score for each class.

3. Create a box plot of the data for each class and write a summary of your observations.

4.   **a.** Determine the standard deviation for the test scores from each class.

       b. What percentage of the scores in each class are within 1 standard deviation of the mean?

- **Histograms** display information using rectangles or bars. The bars must be of uniform width and the scales must have uniform intervals.
- **Pie charts** (or circle graphs) consist of circular regions divided into sectors that each represent a percentage of the whole.
- A **frequency table** consists of two columns. One column contains a data item; the other displays the number of observed occurrences of that item. The number of occurrences of any particular data item is its **frequency**.
- The **mean**, **median**, and **mode** are **measures of central tendency**. They are often described as "representative" of the numbers in a data set.
- A **stem-and-leaf plot** displays the values in a data set. To simplify interpretation, data is usually ordered and a legend is included. A **back-to-back stem-and-leaf plot** may be used to compare two sets of data.
- The **range** is a measure of spread found by subtracting the least data value from the greatest data value.
- The **lower quartile** is the median of the lower half of the data points. The **upper quartile** is the median of the upper half of the data points.
- The **interquartile range** is the difference between the upper quartile and the lower quartile.
- **Outliers** are extreme data values more than 1.5 times the interquartile range above the upper quartile or below the lower quartile.
- A **box-and-whisker plot** (or box plot) displays the median, lower quartile, upper quartile, and outliers (if any) of a data set. The **whiskers** are segments that connect each end of the box to the farthest values which are not outliers.
- **Mean absolute deviation** is a measure of spread that describes the average distance from the mean for the numbers in a data set. For a set of data with n items represented by x_1, x_2, ..., x_n, the mean absolute deviation is given by:

$$\frac{|x_1 - \mu| + |x_2 - \mu| + \cdots + |x_n - \mu|}{n}$$

where the notation $||$ represents absolute value and μ is the mean of the data.

- **Standard deviation** is another measure of spread. For a population with n members represented by x_1, x_2, ..., x_n, where μ represents the mean, the standard deviation is given by:

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}}$$

93