Unit 3: Statistics

Mathematics of Data Management

Lesson Outline

Big Picture

Students will:
- explore, analyse, interpret, and draw conclusions from one-variable data;.
- explore, analyse, interpret, and draw conclusions from two-variable data;
- investigate and evaluate validity of statistical summaries;
- culminating Investigation:
  - analyse, interpret, draw conclusions, and write a report of their research;
  - present summary of finding;
  - critique presentations of their peers.

Day	Lesson Title	Math Learning Goals	Expectations
1	Numerical Summaries – Measuring Centre *(Lesson Included)*	• Apply existing knowledge of measures of central tendency to solve a contextual problem involving discrete data • Demonstrate an understanding of the difference between “grouped” versus “ungrouped” (i.e., “raw”) data and how to apply measures of central tendency to each	D1.1,
2	Graphical Summaries – Exploring Shape and centre *(Lesson Included)*	• Recognise the importance of observing the frequency distribution of a variable as an initial step in one-variable analysis • Identify common shapes of distributions and to use the shape of a distribution as an indicator of the ‘nature’ of the data set (centre in this case) and the population that it represents	D1.1, D1.2, D1.3
3	Numerical Summaries – Measuring Variation *(Lesson Included)*	• Recognize the need to measure the level of variation that exists in a data set as a part of performing a detailed one-variable analysis • Interpret standard deviation as a measure of variation that shows how closely the data clusters to the middle of the data set	D1.1, D1.2, D1.5
4	Graphical Summaries – Exploring Shape and Variation *(Lesson Included)*	• Explore how graphical summaries reveal information about the variation that exists in the data • Use box plots to display data, and to describe the variation in the data set as revealed by this display	D1.1, D1.2, D1.3, D1.5
5	Introduction of the Culminating Investigation	• Interpret, analyse, and summarize data related to the study of the problem. • Draw conclusions from the analysis of the data, evaluate the strengths of the evidence, specify limitations, suggest follow-up problems or investigations. • Focus on one-variable analysis.	E1.4, E1.5
6	Sampling and Repeated Sampling	• Make inferences about a population from sample data. • Explore repeated sampling by taking samples of a given size from the population and calculating the sample mean • Understand that different samples will lead to different sample means and interpret the distribution of these means	D1.5

Day	Lesson Title	Math Learning Goals	Expectations
7	Understanding Confidence Intervals *(Lesson Included)*	• Understand the difference between point estimates and interval estimates of population parameters • Investigate and interpret confidence intervals using an iterative process that further extends their understanding of repeated sampling and its connection to the interpretation of confidence intervals	D1.4, D1.5
8–9	Analysing Two Variable Data	• Graph two numerical variables on a scatter plot. • Determine the appropriateness of a linear model to describe the relationship between two numerical attributes. • Recognize the meaning of the correlation coefficient, using a prepared investigation. • Compare a quantitative and a categorical variable, e.g., gender vs. Income, using appropriate displays, e.g., stacked box plots. • Compare two categorical variables, e.g., gender vs. colour-blindness, using a contingency or summary table and computing proportions.	D2.1, D2.3
10	Understanding Correlation	• Explore different types of relationships between two variables, e.g., the cause-and-effect relationship between the age of a tree and its diameter; the common-cause relationship between ice cream sales and forest fires over the course of a year; the accidental relationship between your age and the number of known planets in the universe. • Interpret statistical summaries to describe and compare the characteristics of two variable statistics.	D2.2, D2.5, E1.4, E1.5
11	Two Variable Data Exploration – Diabetes Exemplar	• Explore different type of relationships between two variables, e.g., the cause-and-effect relationship between the age of a tree and its diameter; the common-cause relationship between ice cream sales and forest fires over the course of a year; the accidental relationship between your age and the number of known planets in the universe. • Interpret statistical summaries to describe and compare the characteristics of two variable statistics.	D2.2, D2.5, E1.4, E1.5
12–13	Interpreting and Making Inferences	• Perform linear regression using technology to determine information about the correlation between variables. • Determine the effectiveness of a linear model on two variable statistics. • Investigate how statistical summaries can be used to misrepresent data. • Make inferences and justify conclusions from statistical summaries or case studies. • Communicate orally and in writing, using convincing arguments.	D2.2, D2.4, D2.5, E1.4, E1.5
14	Culminating Investigation	• Interpret, analyse, and summarize data related to the study of the problem. • Draw conclusions from the analysis of the data, evaluate the strengths of the evidence, specify limitations, suggest follow-up problems or investigations. • Focus on two-variable analysis.	E1.4, E1.5

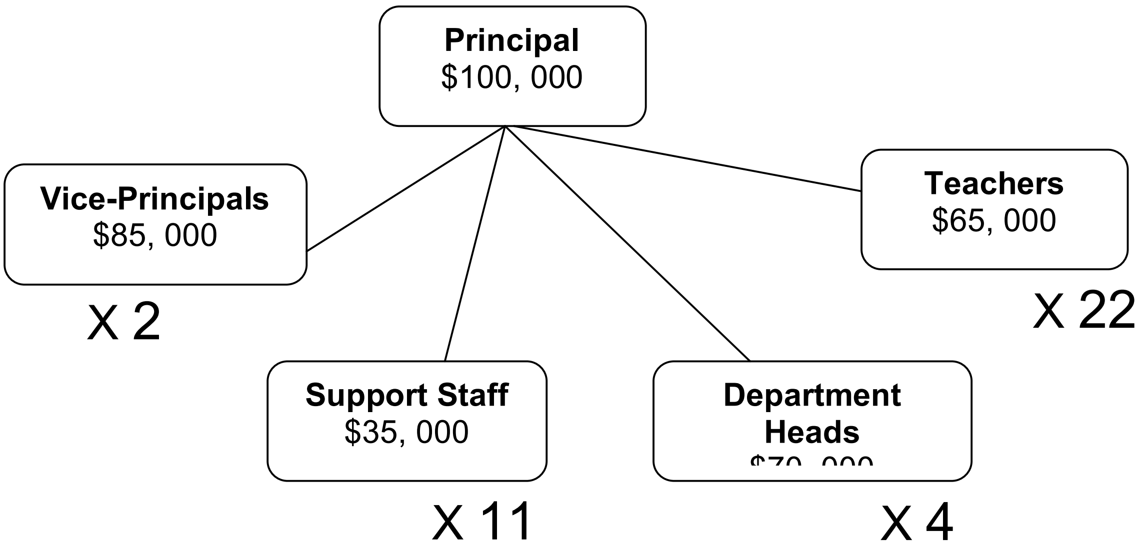
Day	Lesson Title	Math Learning Goals	Expectations
15	Assess Validity	• Interpret and assess statistics presented in the media (e.g., promote a certain point of view, advertising), including how they are used or misused to present a certain point of view. • Investigate interpretation by the media based on lack of knowledge of statistics, e.g., drug testing, false positives. • Examine data collection techniques and analysis in the media, e.g., sample size, bias, law of large numbers. • Scrapbook of statistical observations from the media.	D3.1, D3.2, E1.5
16–17	Culminating Investigation Related to Occupations	• Use journalism as an example to demonstrate applications of data management in an occupation. • Gather, interpret, and describe how the information collected in their project relates to an occupation, e.g., insurance, sports statistician, business analyst, medical researcher. • From their projects identify university programs that explore the applications.	D3.3, E1.3
18	Culminating Investigation	• Edit and compile a report that interpret, analyses, and summarizes data related to the study of the problem. • Draw conclusions from the analysis of the data, evaluate the strengths of the evidence, specify limitations, suggest follow-up problems or investigations.	E1.4, E1.5, E2.1
19–20	Jazz/Summative		
Reserve time 10 days	Culminating Investigation	• Present a summary of the culminating investigation to an audience of their peers. • Answer questions about the culminating investigation and respond to critiques. • Critique the mathematical work of others in a constructive manner.	E2.2, E2.3, E2.4

Unit 3: Day 1: Numerical Summaries – Measuring Centre		**MDM4U**
Minds On: 15 Action: 25 Consolidate:35 Total=75 min	**Math Learning Goals:** • Apply existing knowledge of measures of central tendency to solve a contextual problem involving discrete data • Demonstrate an understanding of the difference between "grouped" versus "ungrouped" (i.e., "raw") data and how to apply measures of central tendency to each	**Materials** • BLM 3.1.1 • BLM 3.1.2 • Chart paper • markers

		Assessment Opportunities
Minds On…	**Think/Pair/Share → Brainstorm** Pose the following question to encourage students to reflect on prior learning: "*When presented with a data set (e.g., the list of student heights in this classroom, class test results), what is the purpose of calculating measures of central tendency (i.e., mean, median, mode)?*" Ask what other information can we obtain from this data set? (Student responses may include: *maximum or minimum values, results are grouped around a particular value).* **Whole Class → Discussion** Hand out BLM 3.1.1. Set the context for the problem: Many workplaces, much like a high school, are made up of various employees who earn different salaries. Read the problem aloud while modeling the strategy of identifying and highlighting important information in text. Give students an opportunity to develop several examples where each measure is most appropriate.	Students may need some direction around the meaning of 'purpose' in this question. Prompting Questions: Why might we perform these calculations? What additional information is gained? Which measure is most appropriate for this problem? Assess prior knowledge of measures of central tendency – address misconceptions with students as they arise Reading strategy: Use think-aloud strategy to model connecting to personal experiences, identifying important information, and summarizing to check for understanding.
Action!	**Pairs → Investigation** Students apply their understanding of measures of central tendency to solve the problem with a partner, and present their findings on a chart paper. **Learning Skills/Teamwork/Checkbric: Watch for pairs that are not demonstrating effective teamwork skills. Encourage partners to share and compare their thinking until both are equally capable of presenting their solution.**	
Consolidate Debrief	**Pairs → Presentation** Select pairs of students to present their findings. Choose at one pair that has determined the difference between the median and mean calculations, one pair that have accurately determined the differences between the grouped vs. ungrouped calculation, and one pair that have identified the need for more information to determine who is correct. **Whole Class → Discussion** Lead the discussion as it arises from the student presentations. Opportunities should arise during the discussion to identify several important mathematical concepts which can be summarized by the teacher: (Refer to BLM 3.1.2) Also, the opportunity may arise to hint at the need for more than just measures of central tendency as a way to 'summarize' a data set.	
Exploration *Application* *Reflection*	**Home Activity or Further Classroom Consolidation** Students organize data provided into a frequency table (grouping) and then calculate the mean and median. For reflection: *If the data is continuous and must be organized into intervals, what value should you select to use in the calculation of the grouped mean and median? What are some pros and cons to your choice?*	Refer to BLM 3.1.2

3.1.1 A *Mean*ingful Money Problem

Imagine a small school that uses the following breakdown of employees. Each amount listed is the annual average salary made by a person in each role.



When at a meeting to discuss increases to the salaries, three numbers are used to describe the *average* salary at the school. Each employee claims to have a mathematical calculation to support their number.

Employee #1 claims that that average salary is $71, 000. Employee #2 claims that the average is $59, 125. Employee #3 states the value they believe the average is $65, 000. The discussion among the staff breaks down into an argument over who has the correct calculation.

What's going on here? How can all of these answers be accounted for? What errors have been made? Explain your thinking.

3.1.2 Teacher Supplement

Action:

Use probing questions to help students: (e.g., What calculations were performed by each person in the problem? What is different about the methods used? What special challenges are created when we use a measure of central tendency as the solitary representation of a data set?)

Note: As the chapter progresses and students develop new measures, they learn to use more than a single value and instead rely on a set of measures to effectively to describe a data set.

Pairs of students are expected to produce a summary on chart paper that details their solution and any strategies used. Assist pairs who have not identified the differences in the calculation methods used by the characters in the problem.

Consolidate Debrief:

The purpose of calculating measures of central tendency is to be able to describe a data set using only a single value. Draw out these ideas:

1. The difference between mean and median as measures of central tendency.
2. The difference between "grouped" and "ungrouped" data and how the calculations for mean and median are performed in each case.
3. The "grouped' data shows the potential values of the variable and the frequencies of those values in the data set. (This is the foundation for all one-variable analysis: that we need to consider the frequencies of the values that occur for a single variable.)

The "grouping" of raw data (sometimes called **microdata**) is a necessary procedure for students to learn and understand since it is the means by which we see frequencies appear in statistics. The analysis of a variable and the frequencies of the values that appear is the foundation of all one-variable analysis.

The typical calculation of mean that students already know ($\bar{x} = \frac{\sum x}{n}$) requires the data to be in its raw or ungrouped form.

The calculation of **mean for discrete grouped data** is similar to that of weighted mean: $\bar{x} = \frac{\sum xf}{\sum f}$ where students must find the product of each x value and its corresponding frequency, take the sum, and then divide by the sum of the frequencies.

3.1.2 Teacher Supplement (Continued)

Should the data be continuous, and therefore grouped into intervals, it is common practice to use the interval midpoint as the value of x.

E.g., The calculation of **mean for continuous data** grouped in the table below:

x	***Interval Midpoint***	***f***
$0 \le x < 5$	2.5	2
$5 \le x < 10$	7.5	11
$10 \le x < 15$	12.5	7

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{(2.5 \times 2) + (7.5 \times 11) + (12.5 \times 7)}{2 + 11 + 7} = \frac{175}{20} = 8.75$$

It is important to note that a mean calculated this way is only an approximation of the true mean since not every individual data value is known.

Depending on whether the data we work with is from a sample or is the population, we use different symbols to designate common measures. This is necessary since the calculation of a measure based on a sample (called a **statistic**) is a point estimate of the same measure of the population (called a **parameter**).

Home Activity or Further Classroom Consolidation:

Provide a data set for organization and a data set that would require students to calculate the mean and median when the data set is organized by intervals.

Unit 3: Day 2: Graphical Summaries – Exploring Shape and Centre

MDM4U

Time	Math Learning Goals	Materials
Minds On: 15 Action: 30 Consolidate:30 Total=75 min	**Math Learning Goals:** • Recognise the importance of observing the frequency distribution of a variable as an initial step in one-variable analysis • Identify common shapes of distributions and to use the shape of a distribution as an indicator of the 'nature' of the data set (centre in this case) and the population that it represents	**Materials** • BLM 3.2.1 • BLM 3.2.2 • Fathom™ Dynamic Data Software • Graphing calculators • Teacher-selected data sets

Assessment Opportunities

Minds On…	**Whole Group → Demonstration** Revisit one of the grouped data sets from yesterday. (This could be one used for yesterday's home activity.) Model how to **sketch** the distribution (i.e., frequency histogram) of the data from the frequency chart. Make special note of the horizontal axis as a continuous number line (even for discrete data) and that the height of the bars indicates the frequencies. Also, model the use of a vertical line through the centre of the distribution as a marking of the mean (calculated previously).	This Minds On… provides an opportunity to re-introduce the unit's focus on the frequencies of the values of a single variable and to reinforce that *frequency* is not a second variable being considered.
Action!	**Pairs → Exploration** Prepare a file that contains three data sets showing three different distributions: left-skewed, symmetric, and right-skewed. Distribute BLM 3.2.1. Students use Fathom™ to compare the three distributions as both a dot plot and a histogram. (Note: this activity can be adapted for use with graphing calculators) **Learning Skills/Teamwork/Checkbric: Check for students that may struggle with the technology – provide assistance and support as needed. A checklist could be used to record some observations of students working independently.** **Mathematical Process/Connecting/Observation/Mental Note: Circulate to assist students not making the connection between the shape and centre of the sample distribution and population.**	Some initial instruction around Fathom™ may be needed if students have not used previously ***Types of Distribution.ppt*** Advise students not to try to drag a variable onto the vertical axis of the graph since this is a picture of only one variable.
Consolidate Debrief	**Whole Group → Note Making** Teacher provides brief direct instruction explaining these key points: 1. Grouping and displaying a single variable as a distribution is an important aspect of analysis because we are provided with a rapid, general description of the data. 2. This technique is really only useful for quantitative data/variables. 3. Defining the common shapes of distributions, (e.g., mound-shaped, skewed, bi-modal) and important properties of these distributions. 4. If the data we have comes from a sample, then we might assume that the population has a similar shape.	Use probing questions to check for student understanding: How is this dot plot different than the scatter plots you've worked with in previous classes? Can you give me some reasons why you've drawn the mean line at this point on the graph?
Concept Practice *Skill Practice*	**Home Activity or Further Classroom Consolidation** Practice drawing and identifying the shapes of various distributions.	Encourage students to use the shape of the distribution to predict and mark the mean and to then test their prediction by calculating.

3.2.1 Comparing Distributions Using Technology

Use the data file provided and Fathom™ to complete the following activity. Be sure to record your sketches and comments in your notebook as you work.

1. Open the file provided.

2. C**reate a dot plot and histogram** for each variable. (To do this, drag an empty graph from the toolbar onto the workspace and then drag one of the variables to the horizontal axis of the graph. You can choose dot plot or histogram using the menu that appears in the top, right corner of your graph.)

****NOTE**: Fathom™ is very 'smart' and skips the grouping phase of one-variable analysis completely. Other programs, such as MS Excel cannot do this, and must be given grouped data in order to construct the histogram.

3. Sketch the two graphs for each variable – six in total – in your notebook.

4. Make some observations about each data set based on these graphs. What information can you obtain by comparing the two different plots for the same variable? What inferences can you make by using the same graph to compare all three variables? Record these observations in you notebook along with your sketches.

5. In your notebook, use a vertical line to estimate the value of the mean for each of the graphs. For which type of graph – dot plot or histogram – is this easier? Explain your thinking.

6. Use Fathom™ to calculate and draw the mean to check your estimates. (To do this, right-click on each graph and select **Plot Value**. A formula window will appear. Type *mean( )* into the window, insert the attribute name inside the brackets, and click the **OK** button.)

7. Make some inferences about the shape of a distribution and how it may be related to its centre. If the data provided came from a sample, how might you use these results to describe the overall population?

3.2.2 Teacher Supplement

One-variable analysis involves more than calculating the mean and median. Students are required to view a variable/data set through three *lenses*:

1. **Centre**: Calculations such as mean and median are the common measures of centre, however, students should also be encouraged to use reasoning when observing a distribution to estimate the centre based upon the shape.

2. **Variation**: Students consider not only the range of the data, but also the diversity or variability in the values that the variable assumes.

3. **Shape**: A graph provides a rapid general impression of the nature of the data set and the population that it represents. By considering shape, students are also able to make estimates about the measures of centre and variation, and to see how those measures are reflected in the distribution.

There are several common shapes of distributions. Due to variability data sets will rarely appear exactly as described here. Instead, the labels can be applied as generalizations of the multitude of distributions that have similar visual characteristics.

Bimodal Distribution: This distribution has high frequencies at the extreme values of the variable (x) and lower frequencies in the centre. The measures of centre are often inaccurate when calculated on a distribution of this shape. (Imagine two hills with a valley in between.)

Uniform Distribution: This distribution has similar (sometimes identical) frequencies for most or all of the possible values of the variable. (Imagine an invisible horizontal line stopping all of the bars of the histogram at the same height.)

Mound-shaped: This class of distributions has several sub-types. This distribution has frequencies that 'peak' centred at a small range of values, and the frequencies decrease as you move towards the extreme values of the variable.

Symmetric: This distribution is often called a 'bell-shape' since it looks like a bell. The frequencies peak in the centre of the distribution and decrease evenly on either side of the centre.

Right-Skew (a.k.a. Positive-Skew): This distribution peaks at lower values of x and trails off to the right.

Left-Skew (a.k.a. Negative-Skew): This distribution peaks at higher values of x and trails off to the left.

***Note**: There are many rules published about the relationship between the mean and median for the skewed distributions, (e.g., for right-skewed distributions the mean is always larger than the median). These rules are not always true!! Students should be encouraged to make these kinds of observations and generalizations, but cautioned that they are not true in every case.

Unit 3: Day 3: Numerical Summaries – Measuring Variation		**MDM4U**
Minds On: 15 Action: 45 Consolidate:20 Total=75 min	**Math Learning Goals:** • Recognize the need to measure the level of variation that exists in a data set as a part of performing a detailed one-variable analysis • Interpret standard deviation as a measure of variation that shows how closely the data clusters to the middle of the data set	**Materials** • BLM 3.3.1 • BLM 3.3.2 • Fathom™ Dynamic data Software • Graphing calculators • Chart paper, markers

		Assessment Opportunities
Minds On…	**Pairs → Brainstorming** Pose the question: *"When describing a collection or set, what does 'variation' mean? How would you show low variation vs. high variation numerically? Graphically?"* Students brainstorm and record ideas and graphs with their partner. **Whole Class → Discussion** Pairs share ideas during class discussion. Encourage students to come up and draw their 'graphs' of low/high variation. A reminder of **range** as a basic measure of variation can occur here.	Promote thinking of variation as a measure that involves the centre. Check understanding around distributions as displaying a single variable and its frequencies and correct as necessary.
Action!	**Small Groups → Investigation** Provides BLM 3.3.1. Students investigate and discuss the data sets provided, articulating which company is deserving of the best reputation. Students develop their own methods for answering the question. Encourage students to provide both graphical and numeric solutions to the problem. **Small Groups → Presentation** Select a few groups to present their findings. Focus their talk and the discussion that ensues around the method(s) that they developed to answer the problem. If no group provides the concept of deviation from the mean, the teacher may have to provide this method of thinking as their way. **Mathematical Process/Connecting/Observation/Mental Note: Problem solving is a major focus of today's lesson. Student should be encouraged to work with their partner to reason, communicate ideas, make connections, and apply knowledge and skills. The presentation of solutions to the class also promotes the sharing of ideas and strategies, as well as students talking about mathematics.**	Note that the problem has the mean as a 'reference point' to focus the students' discussions of variation Probing question: I see you have calculated the range, are there any other calculations that you could make to further determine who deserves the best reputation?
Consolidate Debrief	**Whole Group → Demonstration** Demonstrate calculation of the standard deviation for the two data sets **using technology** (e.g., Fathom™, graphing calculator) and explain its meaning as the **average amount of the deviation of each individual result from the mean**. Some explanation of standard deviation as one of several common measures of variation in a data set (e.g., range) is also required.	Note: Students do not manually calculate standard deviation but should focus on how the value produced by the technology can be interpreted/used in analysing variation. We are working towards creating a complete one-variable summary of the data/variable.
Differentiated *Application* *Concept Practice* *Reflection*	**Home Activity or Further Classroom Consolidation** Revisit the data sets provided to, or collected by students on Day 1. Use technology to calculate the range and standard deviation. For reflection: Make a journal entry to answer the following question: *What do you think are some of the benefits and challenges of using graphical and numerical summaries to describe and analyse data sets? Give examples to support your comments.*	The reflection journal could be collected in order to provide anecdotal assessment feedback.

3.3.1 Who has the better reputation?

Two companies that manufacture precision fuel nozzles are competing for a contract to work with NASA. Both companies pride themselves on producing the finest nozzles with an advertised diameter of 6 mm.

In order to compare the two companies, NASA collects a sample of 30 nozzles from each company. The company who has the **most reliable product** will be awarded the contract.

Company A						**Company B**					
6.28	6.42	5.52	6.09	5.71	6.18	5.87	6.07	6.18	5.76	6.13	5.65
5.80	6.10	6.09	6.06	6.11	5.95	6.03	6.01	6.14	6.03	6.52	6.84
6.25	6.10	6.02	6.16	5.61	5.97	5.76	6.88	5.77	5.26	6.01	5.96
5.92	5.89	6.11	5.56	5.70	5.63	5.28	5.82	6.13	6.37	5.64	5.85
6.13	5.94	6.17	6.14	5.80	5.97	5.88	6.46	6.35	6.52	6.37	5.70

Which company deserves the contract? Develop a mathematical method to justify your choice.

3.3.2 Teacher Supplement

Range and **standard deviation** are two measures applied in one-variable analysis to measure the level of variation or dispersion in the data set. Variation is one of the *lenses* that students are required to view a variable through during analysis.

In particular, standard deviation measures variation by considering how closely the data clusters around its centre. This measure relies on students having a strong understanding of the concept of a deviation.

A **deviation** is the difference between an individual value (x) in a distribution and the mean of the distribution. It is important to note that there will be a deviation calculated for each data point in the distribution.

The process for calculating standard deviation is as follows:

1. Deviations (one for each data) are calculated.
2. Deviations are squared.
3. The average of these squared deviations is calculated.
4. The result is square-rooted.

Standard deviation is therefore a measure that indicates the **average amount (measured in the units of x) that the data deviate from the mean**. The formula for standard deviation differs depending on if students are working with sample data or population data. This is due to the fact that less variation will be captured in a sample (since it is smaller than the population) and therefore the formula must compensate for the systematic underestimation of variation that would result.

__Note__: Understanding the formula for calculating standard deviation is important, but should not be the focus of student activities. Instead, emphasis should be placed on using technology to quickly calculate the value and then interpreting the meaning of the value in the context of the data set.

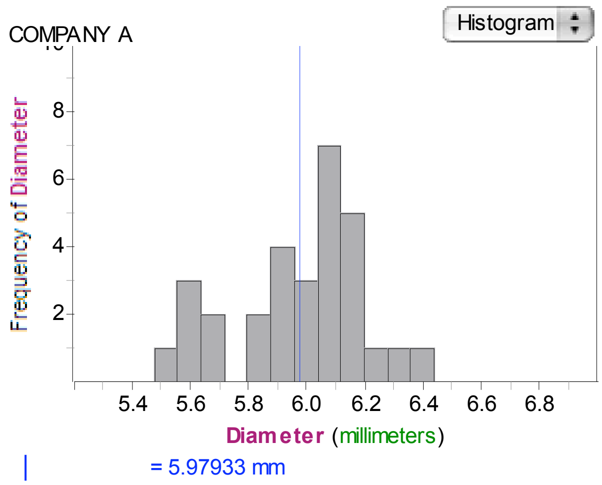
Unit 3: Day 4: Graphical Summaries – Exploring Shape and Variation

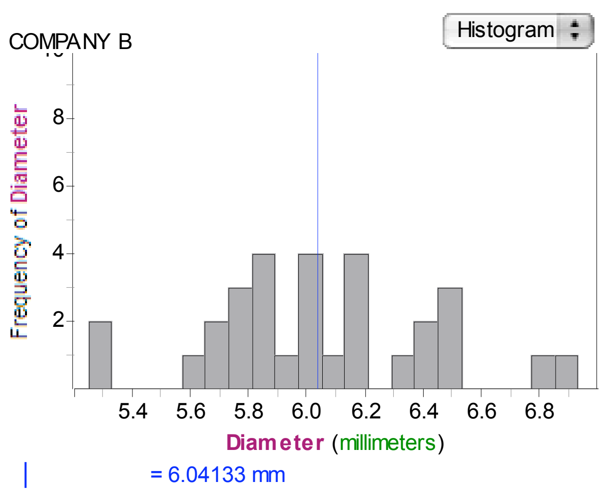
MDM4U

Timing	Math Learning Goals	Materials
Minds On: 15 Action: 40 Consolidate:20 Total=75 min	**Math Learning Goals:** • Explore how graphical summaries reveal information about the variation that exists in the data • Use box plots to display data, and to describe the variation in the data set as revealed by this display	**Materials** • BLM 3.4.1 • BLMs 3.4.2A, 3.4.2B • BLM 3.4.3 • BLM 3.4.4 • Fathom™ Dynamic Data Software • Graphing calculators

		Assessment Opportunities
Minds On…	**Whole Group → Discussion** Show the two histograms from previous day (BLM 3.4.1) and pose questions: "*We've previously calculated the range and standard deviation of these data sets. How are these measures of spread reflected in the distributions? Where can you 'see' the variation in the histograms?*" Encourage students to make connections between the value of the standard deviation and the shape of the graph. Indicate that there may be other ways to visualize the variation using an alternative graphical summary.	Advise students that varying heights of the bars of the histogram should not be interpreted as variation (a common misconception). Phrases such as 'tall and skinny graph' or 'short and fat graph' describe what is formally referred to as the *kurtosis* of the graph.
Action!	**Pairs → Exploration** Distribute BLM 3.4.2A and BLM 3.4.2B so that both partners in a pair get the same example. Students follow the instructions to construct a box-and-whisker plot by making the relevant calculations. Circulate to provide assistance as needed. **Small Group → Conference** Each pair joins with another pair that has different examples. Students are asked to compare their work and to discuss the similarities and differences revealed in the box plots. **Learning Skills/Teamwork/Checkbric: Circulate to encourage students to share their thinking with their partner and in their groups. Taking turns, active listening, and other teamwork skills could be emphasized as student work together.** **Mathematical Process/Selecting Tools and Computation Strategies/Observation/Mental Note: The variety of tools (e.g., Fathom™, graphing calculators) available to students is quite large; Circulate to encourage students to become proficient with these tools.**	A common error is To include the median value in either the upper or lower half of the data. Probing questions: How can the two box plots be compared to reveal differences in the two data sets? Does this graphical summary support or refute our conclusions from previous lessons?
Consolidate Debrief	**Whole group → Demonstration** Demonstrate the construction of a box plot for the two data sets **using technology** (e.g., Fathom™, graphing calculator) and explain the use of the box plot as an effective tool for summarizing a data set. Display both data sets side-by-side on the same plot. Place emphasis on switching comfortably between the histogram and the box plot for the same data set. Reiterate that the shape, centre and variation being described in a sample can be inferred to the population that the sample was taken from.	The IQR (Inter-Quartile Range) should now be added to the list of tools for one-variable analysis; review the complete set of tools/requirements (both graphical and numerical) for completing a one-variable analysis.
Differentiated Application	**Home Activity or Further Classroom Consolidation** Apply the skills learned to perform a complete one-variable analysis on one of the data sets collected for your culminating project. Use BLM 3.4.3 as a guide for the process.	Students' analyses could be collected to provide formative assessment based on the given criteria

3.4.1 Company A vs. Company B





3.4.2A Creating a Box-and-Whisker Plot

Use the following data set to complete the following investigation:

2006-2007 - Regular Season – Toronto Maple Leafs – All Skaters – Total goals
0, 0, 0, 0, 0, 0, 1, 1, 2, 3, 4, 6, 7, 8, 8, 10, 11, 12, 12, 13, 14, 15,15, 18, 20, 21, 21, 24, 27

1. What is the minimum value? Maximum?

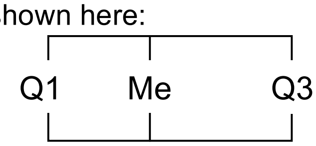
2. Find the median.

3. Ignoring the median value, look only at the upper half of the data. Calculate the median of the upper half. Name this value Q3.

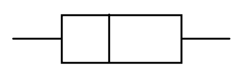
4. Once again ignoring the median value, look only at the lower half of the data. Calculate the median of the lower half. Name this value Q1.

5. On the number line at the bottom of this page, mark out an evenly-spaced scale and mark dots at each of the above values (i.e., min, Q1, median, Q3, and max).

6. Use vertical lines at the Q1, median, and Q3 values and two horizontal lines to make a box connecting these values as shown here:



7. Extend horizontal lines out from the edges of the box to the minimum and maximum points. (Like whiskers from a cat's face.)

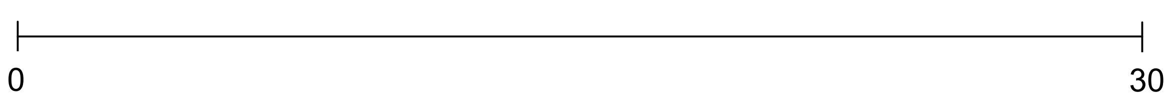


8. Answer the following based on this graphical summary:
    a. Approximately what % of the data are above the median?

    b. Approximately what % of the data lie within the box?

    c. Approximately what % of the data are in each whisker?

    d. The difference between Q3 and Q1 is called the IQR (Inter-Quartile Range). What does this calculate?



3.4.2B Creating a Box-and-Whisker Plot

Use the following data set to complete the following investigation:

2006-2007 - Regular Season – Montreal Canadiens – All Skaters – Total goals
0, 0, 0, 1, 1, 1, 2, 2, 3, 4, 5, 6, 6, 6, 9, 10, 11, 13, 16, 18, 20, 22, 22, 26, 30

1. What is the minimum value? Maximum?

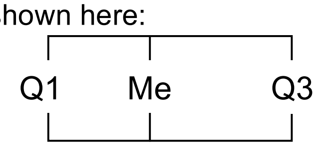
2. Find the median.

3. Ignoring the median value, look at only at the upper half of the data. Calculate the median of the upper half. Name this value Q3.

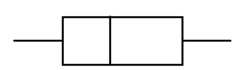
4. Once again ignoring the median value, look only at the lower half of the data. Calculate the median of the lower half. Name this value Q1.

5. On the number line below, mark out an evenly-spaced scale and mark dots at each of the above values (i.e., min, Q1, median, Q3, and max).

6. Use vertical lines at the Q1, median, and Q3 values and two horizontal lines to make a box connecting these values as shown here:



7. Extend horizontal lines out from the edges of the box to the minimum and maximum points. (Like whiskers from a cat's face.)

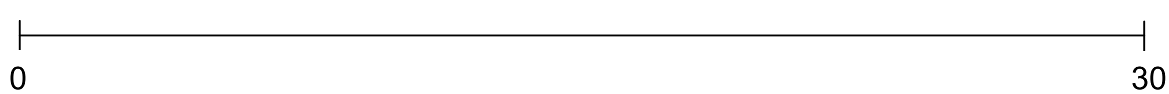


8. Answer the following based on this graphical summary:
   e. Approximately what % of the data are above the median?

   f. Approximately what % of the data lie within the box?

   g. Approximately what % of the data are in each whisker?

   h. The difference between Q3 and Q1 is called the IQR (Inter-Quartile Range). What does this calculate?



3.4.3 Culminating Project – One-Variable Analysis

Remember this quote from Patton, "[data analysis is] to make sense of massive amounts of data, reduce the volume of information, identify significant patterns, and construct a framework for communicating the essence of what the data reveal."

Activity:

1. Select one of the variables from your project. This could be one of the main ones and **must be quantitative**. At the start, the data should be "raw" or "ungrouped" – i.e., it should just be a list of values that have not yet been organized.

2. Start by organizing the data into a frequency table or stem-and-leaf-plot. (*These are the most basic of organizers.*)

3. Create a dot plot, histogram, and box-and-whisker plot of the data. (*These are possibilities for graphically summarizing the data from single variable.*) Which one do you think best represents the data? Explain.

4. Calculate the mean, median, variance, standard deviation, range, and inter-quartile range for the data. (*These represent the most basic calculations used to analyse a single variable. Together, they form a numerical summary.*) Organize these values in a table. Make some inferences about the variable based on the calculations.

5. Arrange all of your work onto two-page layout and print it out.

Assessment:

This activity is in preparation for your Culminating Investigation Presentation. Submit your work, to receive written comments (i.e., **formative** assessment) based on the criteria outlined below.

Consider the following criteria for assessing your summaries:
1. Organization of data and initial analysis of the variable are performed correctly – using correct terminology and notation.
2. Several meaningful (and related to your thesis/study) inferences are made based on the numerical and graphical summaries

Unit 3: Day 7: Understanding Confidence Intervals

MDM4U

Minds On: 15 Action: 35 Consolidate:25 Total=75 min	**Math Learning Goals:** • Understand the difference between point estimates and interval estimates of population parameters • Investigate and interpret confidence intervals using an iterative process that further extends their understanding of repeated sampling and its connection to the interpretation of confidence intervals	**Materials** • BLM 3.7.1 • BLM 3.7.2 • BLM 3.7.3 • BLM 3.7.4 • Fathom™ Dynamic Data Software

Assessment Opportunities

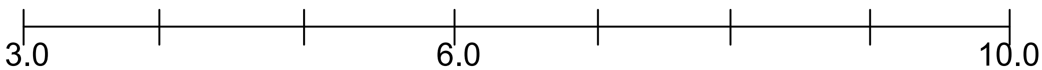
Minds On…	**Individual → Brainstorm** Provide BLM 3.7.1. Students individually record their thoughts prepare to share with the class. Take up answers while focusing on two concepts: 1. An estimation of the mean as an interval is as acceptable as a point estimate of the mean. 2. The interval needs to be wider in order to increase our confidence in the estimation, but we want to keep the interval as small as possible.	The sample mean is often confused with the population mean. Emphasize the relationship between these values: the sample mean acts as a single-value, point estimate of the population mean.
Action!	**Whole Class → Exploration** Provide each student (or pair of students) with BLM 3.7.3 and **one** of the 20 samples provided on BLM 3.7.2. Each sample from BLM 3.7.2 can only appear once in the room and all 20 samples must be used. Pose question: "*What is the smallest interval that can be applied to each of your sample means so that 19 out of the 20 intervals (i.e., 95%) contain the population mean?*" Lead students through the following iterative process: (Refer to BLM 3.7.4) 1. Provide an interval width (e.g. ± 0.9) 2. Students calculate and record their estimation interval 3. complete a whole-class check , record results using BLM 3.7.3 4. provide a smaller interval width (e.g. ± 0.8), and repeat the process. When the appropriate interval width is found (the last iteration to contain 19 out of the 20 samples), have the students record their intervals in a central location for comparison. **Mathematical Process/Problem Solving/Observation/Mental Note: Circulate to observe students as they use an iterative process as a mathematical strategy for solving a problem. Use probing questions to check for understanding.**	Iteration is not the most efficient of strategies, however the logical, experimental nature of this strategy is appropriate for mathematical inquiry. Probing Questions: Why do statisticians prefer interval estimates to point estimates? Why does the interval width have to increase to increase the confidence level? Key summary points can be demonstrated and/or verified using the estimation tool in Fathom™
Consolidate Debrief	**Whole Class → Summarizing** Provide brief direct instruction to summarize the key concepts learned for the day. (Refer to BLM 3.7.4) There are several possible 95% confidence intervals – this is because a confidence interval is centred around the sample mean, and the class is using 20 samples. (The class has created 20 different 95% confidence intervals. **Whole Group → Demonstration** Demonstrate the calculation of the estimates for one of the samples **using** Fathom™ which allows students to enter a mean, standard deviation and sample size, and then calculates the appropriate interval for any given confidence level.	If the population parameter is unknown statisticians determine the proper interval size using only the mean and the standard deviation of a single sample. This is not expected in this course.
Concept Practice	**Home Activity or Further Classroom Consolidation** Practice interpreting confidence intervals on assigned questions.	

3.7.1 A Sleepy Sample

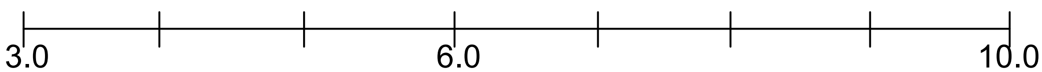
Imagine that, two minutes from now, every student in this class is going to be asked to tell us the number of hours of sleep he/she had last night; and that, three minutes from now, we will calculate the average (mean) number of hours of sleep.

1. Ask the person to your left and the person to your right how many hours of sleep they had last night. Based on this small sample, estimate the average for the whole class.

2. Use the scale below to mark off an interval that you believe with 80% confidence will include the mean number of hours of sleep for the whole class.



3. Use the scale below to mark off an interval that you believe with 99% confidence will include the mean number of hours of sleep for the whole class.



4. Which of the two intervals is wider? Explain why this happens.

5. Is there any benefit to using an interval to make an estimate of the mean as opposed to a single value? Explain your thinking.

3.7.2 Samples for Exploration

a sample

	value	<new>
1	4.813	
2	4.676	
3	4.480	
4	5.105	
5	4.038	
6	5.958	
7	5.577	
8	6.338	
9	3.810	
10	5.750	
11	4.260	
12	4.682	
13	4.817	
14	4.137	
15	4.635	

a sample

	value	<new>
1	4.767	
2	3.733	
3	4.635	
4	4.654	
5	5.153	
6	1.996	
7	6.322	
8	5.448	
9	4.349	
10	4.573	
11	4.687	
12	6.500	
13	5.781	
14	5.957	
15	4.136	

a sample

	value	<new>
1	3.357	
2	3.248	
3	6.440	
4	4.173	
5	4.858	
6	5.763	
7	5.802	
8	4.747	
9	2.990	
10	4.491	
11	4.631	
12	5.779	
13	5.880	
14	6.257	
15	5.334	

a sample

	value	<new>
1	6.474	
2	5.780	
3	4.463	
4	5.095	
5	5.235	
6	3.940	
7	5.336	
8	4.762	
9	6.040	
10	7.612	
11	5.332	
12	4.416	
13	4.671	
14	2.860	
15	6.679	

a sample

	value	<new>
1	4.214	
2	4.864	
3	6.217	
4	5.458	
5	5.475	
6	4.953	
7	6.082	
8	6.050	
9	3.962	
10	5.845	
11	3.414	
12	5.167	
13	4.687	
14	5.787	
15	4.087	

a sample

	value	<new>
1	3.665	
2	3.069	
3	4.292	
4	4.755	
5	4.889	
6	5.482	
7	6.361	
8	4.072	
9	5.170	
10	4.835	
11	5.159	
12	5.330	
13	6.495	
14	6.126	
15	6.021	

3.7.2 Samples for Exploration (Continued)

a sample

	value	<new>
1	3.205	
2	6.213	
3	4.384	
4	4.609	
5	6.006	
6	5.757	
7	4.759	
8	4.704	
9	4.814	
10	5.669	
11	5.180	
12	3.638	
13	3.586	
14	6.124	
15	3.514	

a sample

	value	<new>
1	3.874	
2	6.199	
3	6.961	
4	4.588	
5	7.216	
6	4.182	
7	4.129	
8	5.628	
9	5.932	
10	8.076	
11	4.216	
12	4.679	
13	4.107	
14	6.391	
15	3.412	

a sample

	value	<new>
1	3.121	
2	4.981	
3	5.686	
4	6.202	
5	4.370	
6	4.996	
7	5.871	
8	5.046	
9	5.889	
10	2.723	
11	4.368	
12	3.321	
13	5.701	
14	5.626	
15	5.476	

a sample

	value	<new>
1	4.320	
2	6.122	
3	3.968	
4	6.195	
5	4.167	
6	5.090	
7	5.150	
8	6.134	
9	3.824	
10	6.618	
11	6.519	
12	3.999	
13	6.317	
14	5.015	
15	4.685	

a sample

	value	<new>
1	4.363	
2	5.009	
3	2.656	
4	5.125	
5	3.912	
6	4.199	
7	4.143	
8	6.350	
9	5.631	
10	6.158	
11	5.916	
12	5.427	
13	5.284	
14	3.673	
15	5.809	

a sample

	value	<new>
1	5.519	
2	6.403	
3	3.418	
4	4.111	
5	4.282	
6	5.032	
7	6.285	
8	3.516	
9	4.507	
10	4.250	
11	4.732	
12	5.708	
13	5.086	
14	3.893	
15	4.775	

3.7.2 Samples for Exploration (Continued)

a sample

	value	<new>
1	6.172	
2	4.703	
3	6.603	
4	7.112	
5	4.509	
6	3.867	
7	3.789	
8	5.518	
9	4.623	
10	5.057	
11	4.609	
12	5.658	
13	5.967	
14	5.154	
15	5.890	

a sample

	value	<new>
1	6.813	
2	4.903	
3	7.566	
4	6.599	
5	4.620	
6	3.843	
7	5.914	
8	3.248	
9	6.806	
10	5.198	
11	6.376	
12	3.822	
13	3.890	
14	5.072	
15	5.272	

a sample

	value	<new>
1	6.508	
2	5.397	
3	5.223	
4	5.504	
5	4.266	
6	3.891	
7	4.316	
8	4.673	
9	5.324	
10	3.180	
11	5.777	
12	5.774	
13	6.250	
14	4.707	
15	4.216	

a sample

	value	<new>
1	6.314	
2	6.676	
3	4.999	
4	3.162	
5	5.991	
6	3.754	
7	4.731	
8	4.320	
9	4.942	
10	5.180	
11	5.085	
12	2.646	
13	5.254	
14	5.734	
15	4.088	

a sample

	value	<new>
1	5.196	
2	4.967	
3	5.961	
4	4.439	
5	5.272	
6	3.339	
7	6.271	
8	3.932	
9	4.026	
10	3.424	
11	4.211	
12	6.194	
13	6.339	
14	4.225	
15	4.500	

a sample

	value	<new>
1	5.227	
2	4.607	
3	6.368	
4	3.071	
5	6.143	
6	5.565	
7	6.319	
8	3.734	
9	5.088	
10	4.303	
11	6.175	
12	4.773	
13	7.037	
14	3.987	
15	4.619	

3.7.2 Samples for Exploration (Continued)

a sample

	value	<new>
1	6.249	
2	5.184	
3	5.032	
4	4.726	
5	6.219	
6	4.997	
7	5.938	
8	5.255	
9	3.439	
10	3.352	
11	4.532	
12	5.081	
13	4.923	
14	4.166	
15	5.661	

a sample

	value	<new>
1	6.079	
2	4.077	
3	5.218	
4	4.755	
5	3.459	
6	4.503	
7	5.396	
8	6.507	
9	6.443	
10	5.677	
11	6.017	
12	5.672	
13	5.310	
14	3.558	
15	6.238	

3.7.3 Student Summary Chart

My sample mean = __________. (This value is your point estimate of the population mean.)

We will begin with a wide interval (which will 'capture' the true population mean 100% of the time) and will slowly shrink it. Record your work below:

Trial #	Margin of error (provided by your teacher)	Interval estimate (with your sample mean in the middle)	Number of intervals in the class that 'capture' the population mean	Percent of intervals in the class that 'capture' the population mean
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				

When the number of intervals (in the room) that contain the true population mean changes from 19 to 18, then the confidence you have in your interval estimate has changed from 95% to 90%.

What is the smallest margin of error provided by your teacher where 95% of the intervals in the class 'captured' the population mean?

3.7.4 Teacher Supplement

Theoretical Background
The process (or set of methods) for using the characteristics of a random sample to describe the characteristics of the population is known as **statistical inference**. Values calculated using the sample data are known as **statistics**. The same values in the population are known as **parameters**. Since the population parameters are often unknown, then statistical inference is really about using statistics to **estimate** the true values of parameters.

One way to estimate a population parameter, such as the population mean, is to collect a random sample and calculate the sample mean. This sample mean is known as a **point estimate** (because it's a single number) of the population mean. We also know that larger samples will help make our point estimate more accurate.

The downside to point estimates is that we have no way of knowing if they are actually close to the true population parameter! It could be that, because of variability, the sample mean is 'way off' the true population value. So, an alternative solution is to use an interval estimation of the population parameter. This interval has the **single sample mean as its middle** and attempts to capture the true population parameter within it.

An **interval estimation** is helpful because it can be accompanied by a statement of confidence. Interval estimate + Confidence statement = Confidence Interval.
The confidence statement (often given as a percent) indicates the percent of confidence in the given size interval that will 'in the long run' capture the population parameter.
[Note: The confidence % ***does not*** *indicate a percent chance that the given interval captures the population parameter. (i.e., it is incorrect to say, "the population parameter has a 95% chance of falling within this interval.")*

For example, let's assume that we have a population that has a mean (of 5), which we don't know but trying to estimate. When a sample is collected, the sample mean is 4.7. We might choose a very large interval with the sample mean at the middle, for example $-9995.3 < \mu < 10004.7$, and be 100% confident that the interval would contain the true population mean. However, this is not an effective estimate. Alternatively, we could choose a very small interval, for example $4.69999 < \mu < 4.70001$, and would be nearly 0% confident that the interval would contain the true population mean.

The key is to find an appropriate balance in the relationship between the interval size and the confidence level.

Action: **An Iterative Process**
Student work is recorded on BLM 3.7.3.

1. Students begin by calculating the sample means. These sample means could act as point estimates of the population mean, however they will use the sample means as the middle value in each of their intervals during this activity.
2. The teacher acts as the facilitator for a whole-class iterative exploration.
3. In this scenario, the **population mean is known to be 5**. During this activity, students will be 'working backwards' with confidence intervals; in practice,

3.7.4 Teacher Supplement (Continued)

confidence intervals are used to estimate *unknown* population parameters. It is important to determine the smallest possible interval while still having confidence that it will contain the population mean.

4. Use the following interval widths (which can also be called margins of error) for the iterative process: ±0.9, ±0.1, ±0.7, ±0.65, ±0.6, ±0.55, ±0.5, and others if you wish to refine the process. Elicit the help of the class in determining the next, smaller, margin of error to try.
5. At each iterative step, students record (BLM 3.7.3) both the number and percent of intervals in the classroom that contain the population mean (5) (e.g., "The population mean lies within the interval that is ±? of the sample mean 15 times out of 20. Therefore, we are 75% confidant that this interval contains the true population mean.")
6. When the **appropriate interval width** is found (i.e., 19 out of 20 samples) students record their intervals in a central location for comparison. Note: when the number of samples in the class that 'captures the population mean' is 18 this means that the interval now captures the mean only 90% of the time -- 18 out of 20 in the class.)

Consolidate Debrief:

Summary of Key Concepts:

- A point estimate is good for estimating a population parameter (such as mean) but it may not be close to the true value due to random variation in the sample
- An interval estimate is a better estimate even though **we can never be sure that the interval contains the true value of the population parameter, we can be 95% (or 90%, or whatever%) sure that it does**.
- The middle of the confidence interval is always the value of the point estimate of the parameter (the sample mean), and the values of the endpoints of the intervals vary from sample to sample.
- The confidence level does not mean that the population parameter has a 95% chance of falling in the given interval but means that **95% of the time, this size interval will contain the true population mean**.
- The confidence interval can be reported using either an interval or as a margin of error (the half-width of the interval).
  (e.g., $46 < \mu <$ can also be written $\mu = 51 \pm$)
- **As the width of the interval decreases, the confidence level also decreases, because larger intervals will contain the unknown population parameter more often.** During today's activity, we attempted to find the smallest interval we could that still 'worked' 95% of the time!
- This concept can be extended to estimates of population proportions (such as in opinion or election polls); (e.g., the statement, "45% of consumers prefer Brand X, ±3%, 19 times out of 20," is interpreted as "95% of the time, the proportion of the population that prefer Brand X is between 42% and 48%")