Georgia
Standards of Excellence
Frameworks

Mathematics

GSE Algebra II/ Advanced Algebra

Unit 7: Inferences & Conclusions from Data

Georgia Department of Education

Richard Woods, Georgia's School Superintendent
"Educating Georgia's Future"

Unit 7
Inferences and Conclusions from Data

Table of Contents

***Revised standards indicated in bold red font.**

OVERVIEW

In this unit students will:

- Describe and compare distributions by using the correct measure of center and spread, and identifying outliers (extreme data points) and their effect on the data set
- Use the mean and standard deviation of the data set to fit it to a normal distribution where appropriate
- Estimate and interpret areas under a normal curve using calculators, spreadsheets or tables
- Design simulations of random sampling: assign digits in appropriate proportions for events, carry out the simulation using random number generators and random number tables and explain the outcomes in context of the population and the known proportions
- Design and evaluate sample surveys, experiments and observational studies with randomization and discuss the importance of randomization in these processes
- Conduct simulations of random sampling to gather sample means and proportions. Explain what the results mean about variability in a population and use results to calculate margins of error
- Generate data simulating application of two treatments and use the results to evaluate significance of differences
- Read and explain in context data from outside reports

Although the units in this instructional framework emphasize key standards and big ideas at specific times of the year, routine topics such as estimation, mental computation, and basic computation facts should be addressed on an ongoing basis. Ideas related to the eight process standards should be addressed constantly as well. To assure that this unit is taught with the appropriate emphasis, depth, and rigor, it is important that the tasks listed under “Evidence of Learning” be reviewed early in the planning process. A variety of resources should be utilized to supplement this unit. This unit provides much needed content information, but excellent learning activities as well. The tasks in this unit illustrate the types of learning activities that should be utilized from a variety of sources.

STANDARDS ADDRESSED IN THIS UNIT

Mathematical standards are interwoven and should be addressed throughout the year in as many different units and activities as possible in order to emphasize the natural connections that exist among mathematical topics.

KEY & RELATED STANDARDS

Interpreting Categorical and Quantitative Data
Summarize, represent, and interpret data on a single count or measurement variable

MGSE9-12.S.ID.2 **Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, ~~mean absolute deviation~~, standard deviation) of two or more different data sets.**

MGSE9-12.S.ID.4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

Making Inferences and Justifying Conclusions
Understand and evaluate random processes underlying statistical experiments

MGSE9-12.S.IC.1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population**.**

MGSE9-12.S.IC.2 Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation. *For example, a model says a spinning coin falls heads up with probability 0.5. Would a result of 5 tails in a row cause you to question the model?*

Make inferences and justify conclusions from sample surveys, experiments, and observational studies

MGSE9-12.S.IC.3 Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

MGSE9-12.S.IC.4 Use data from a sample survey to estimate a population mean or proportion develop a margin of error through the use of simulation models for random sampling.

MGSE9-12.S.IC.5 Use data from a randomized experiment to compare two treatments; use simulations to decide if differences between parameters are significant.

MGSE9-12.S.IC.6 **Evaluate reports based on data.** ***For example, determining quantitative or categorical data; collection methods; biases or flaws in data.***

RELATED STANDARDS

MGSE7.SP.1. Understand that statistics can be used to gain information about a population by examining a sample of the population; generalizations about a population from a sample are valid only if the sample is representative of that population. Understand that random sampling tends to produce representative samples and support valid inferences.

MGSE7.SP.2.Use data from a random sample to draw inferences about a population with an unknown characteristic of interest. Generate multiple samples (or simulated samples) of the same size to gauge the variation in estimates or predictions.

MGSE7.SP.3. Informally assess the degree of visual overlap of two numerical data distributions with similar variability, measuring the difference between the centers by expressing it as a multiple of a measure of variability.

MGSE7.SP.4. Use measures of center and measures of variability for numerical data from random samples to draw informal comparative inferences about two populations

MGSE9-12.S.ID.1 Represent data with plots on the real number line (dot plots, histograms, and boxplots).

MGSE9-12.S.ID.3 Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

STANDARDS FOR MATHEMATICAL PRACTICE

Refer to the Comprehensive Course Overview for more detailed information about the Standards for Mathematical Practice.

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.
8. Look for and express regularity in repeated reasoning.

ENDURING UNDERSTANDINGS

- Understand how to choose summary statistics that are appropriate to the characteristics of the data distribution, such as the shape of the distribution or the existence of outliers.

- Recognize that only some data are described well by a normal distribution.

- Understand how the normal distribution uses area to make estimates of probabilities

- Compare theoretical and empirical results to evaluate the effectiveness of a treatment.

- Understand the way in which data is collected determines the scope and nature of the conclusions that can be drawn from the data.

- Understand how to use statistics as a way of dealing with, but not eliminating, variability of results from experiments and inherent randomness.

- Understand how to use the margin of error to find a confidence interval.

ESSENTIAL QUESTIONS

- How do I choose summary statistics that are appropriate to the data distribution?
- How can I find a standard deviation?
- How do I decide if the normal distribution describes a set of data?
- When do I use the normal distribution to estimate probabilities?
- How can I find the sampling distribution of a sample proportion?
- How can I find the sampling distribution of a sample mean?
- How do I use theoretical and empirical results to determine if a treatment was effective?
- How does the way I collected data effect the conclusions that can be drawn?
- How do I use statistics to explain the variability and randomness in a set of data?
- How do I interpret the margin of error of a confidence interval?
- How do I use a margin of error to find a confidence interval?

CONCEPTS AND SKILLS TO MAINTAIN

In order for students to be successful, the following skills and concepts need to be maintained:
- Determine whether data is categorical or quantitative (univariate or bivariate)
- Know how to compute the mean, median, interquartile range, and mean absolute deviation by hand in simple cases and using technology with larger data sets
- Determine whether a set of data contains outliers.
- Create a graphical representation of a data set
- Be able to use graphing technology
- Describe center and spread of a data set

- Describe various ways of collecting data

SELECT TERMS AND SYMBOLS

The following terms and symbols are often misunderstood. These concepts are not an inclusive list and should not be taught in isolation. However, due to evidence of frequent difficulty and misunderstanding associated with these concepts, instructors should pay particular attention to them and how their students are able to explain and apply them.

The definitions below are for teacher reference only and are not to be memorized by the students. Students should explore these concepts using models and real life examples. Students should understand the concepts involved and be able to recognize and/or demonstrate them with words, models, pictures, or numbers.

The websites below are interactive and include a math glossary suitable for high school children.
Note – At the high school level, different sources use different definitions. Please preview any website for alignment to the definitions given in the frameworks
http://www.amathsdictionaryforkids.com/
This web site has activities to help students more fully understand and retain new vocabulary.
http://intermath.coe.uga.edu/dictnary/homepg.asp

Definitions and activities for these and other terms can be found on the Intermath website. Intermath is geared towards middle and high school students.

- **Center.** Measures of center refer to the summary measures used to describe the most "typical" value in a set of data. The two most common measures of center are median and the mean.

- **Central Limit Theorem.** Choose a simple random sample of size n from any population with mean μ and standard deviation σ. When n is large (at least 30), the sampling distribution of the sample mean $\overline{x}$ is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Choose a simple random sample of size n from a large population with population parameter p having some characteristic of interest. Then the sampling distribution of the sample proportion $\hat{p}$ is approximately normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$. This approximation becomes more and more accurate as the sample size n increases, and it is generally considered valid if the population is much larger than the sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. The CLT allows us to use normal calculations to determine probabilities about sample proportions and sample means obtained from populations that are not normally distributed.

- **Confidence Interval** is an interval for a parameter, calculated from the data, usually in the form *estimate* $\pm$ *margin of error.* The confidence level gives the probability that the interval will capture the true parameter value in repeated samples.

- **Empirical Rule.** If a distribution is normal, then approximately

  68% of the data will be located within one standard deviation symmetric to the mean

  95% of the data will be located within two standard deviations symmetric to the mean

  99.7% of the data will be located within three standard deviations symmetric to the mean

- **Margin of Error.** The value in the confidence interval that says how accurate we believe our estimate of the parameter to be. The margin of error is comprised of the product of the *z-score* and the standard deviation (or standard error of the estimate). The margin of error can be decreased by increasing the sample size or decreasing the confidence level.

- **Mean absolute deviation.** A measure of variation in a set of numerical data, computed by adding the distances between each data value and the mean, then dividing by the number of data values. Example: For the data set {2, 3, 6, 7, 10, 12, 14, 15, 22, 120}, the mean absolute deviation is 20.

- **Parameters.** These are numerical values that describe the population. The population mean is symbolically represented by the parameter μ_X. The population standard deviation is symbolically represented by the parameter σ_X.

- **Population.** The entire set of items from which data can be selected.

- **Random.** Events are random when individual outcomes are uncertain. However, there is a regular distribution of outcomes in a large number of repetitions.

- **Sample.** A subset, or portion, of the population.

- **Sample Mean**. A statistic measuring the average of the observations in the sample. It is written as $\bar{x}$. The mean of the population, a parameter, is written as μ.

- **Sample Proportion.** A statistic indicating the proportion of successes in a particular sample. It is written as $\hat{p}$. The population proportion, a parameter, is written as *p*.

- **Sampling Distribution.** A statistics is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

- **Sampling Variability.** The fact that the value of a statistic varies in repeated random sampling.

- **Shape**. The shape of a distribution is described by symmetry, number of peaks, direction of skew, or uniformity.
  - **Symmetry**- A symmetric distribution can be divided at the center so that each half is a mirror image of the other.
  - **Number of Peaks**- Distributions can have few or many peaks. Distributions with one clear peak are called unimodal and distributions with two clear peaks are called bimodal. Unimodal distributions are sometimes called bell-shaped.
  - **Direction of Skew**- Some distributions have many more observations on one side of graph than the other. Distributions with a tail on the right toward the higher values are said to be skewed right; and distributions with a tail on the left toward the lower values are said to be skewed left.
  - **Uniformity-** When observations in a set of data are equally spread across the range of the distribution, the distribution is called uniform distribution. A uniform distribution has no clear peaks.

- **Spread.** The spread of a distribution refers to the variability of the data. If the data cluster around a single central value, the spread is smaller. The further the observations fall from the center, the greater the spread or variability of the set. (range, interquartile range, Mean Absolute Deviation, and Standard Deviation measure the spread of data)

- **Standard Deviation. The square root of the variance.** $\sigma = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$

- **Statistics.** These are numerical values that describe the sample. The sample mean is symbolically represented by the statistic $\bar{x}$. The sample standard deviation is symbolically represented by the statistic s_x.

- **Variance.** The average of the squares of the deviations of the observations from their mean. $\sigma^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$

EVIDENCE OF LEARNING

By the conclusion of this unit, students should be able to demonstrate the following competencies:

- Construct appropriate graphical displays (dot plots, histogram, and box plot) to represent sets of data values.
- Describe a distribution using shape, center and spread and use the correct measure appropriate to the distribution
- Compare two or more different data sets using center and spread
- Recognize data that is described well by a normal distribution
- Estimate probabilities for normal distributions using area under the normal curve using calculators, spreadsheets and tables.
- Design a method to select a sample that represents a variable of interest from a population
- Design simulations of random sampling and explain the outcomes in context of population and know proportions or means
- Use sample means and proportions to estimate population values and calculate margins of error
- Read and explain in context data from real-world reports

FORMATIVE ASSESSMENT LESSONS (FAL)

Formative Assessment Lessons are intended to support teachers in formative assessment. They reveal and develop students' understanding of key mathematical ideas and applications. These lessons enable teachers and students to monitor in more detail their progress towards the targets of the standards. They assess students' understanding of important concepts and problem solving performance, and help teachers and their students to work effectively together to move each student's mathematical reasoning forward.

More information on Formative Assessment Lessons may be found in the Comprehensive Course Overview.

SPOTLIGHT TASKS

A Spotlight Task has been added to each GSE mathematics unit in the Georgia resources for middle and high school. The Spotlight Tasks serve as exemplars for the use of the Standards for Mathematical Practice, appropriate unit-level Georgia Standards of Excellence, and research-based pedagogical strategies for instruction and engagement. Each task includes teacher commentary and support for classroom implementation. Some of the Spotlight Tasks are revisions of existing Georgia tasks and some are newly created. Additionally, some of the Spotlight Tasks are 3-Act Tasks based on 3-Act Problems from Dan Meyer and Problem-Based Learning from Robert Kaplinsky.

TASKS

The following tasks represent the level of depth, rigor, and complexity expected of all Algebra II/Advanced Algebra students. These tasks, or tasks of similar depth and rigor, should be used to demonstrate evidence of learning. It is important that all elements of a task be addressed throughout the learning process so that students understand what is expected of them. While some tasks are identified as a performance task, they may also be used for teaching and learning (learning/scaffolding task).

Task Name	**Task Type** *Grouping Strategy*	**Content Addressed**
Math Award	Learning Task *Individual/Partner*	Calculate the mean absolute deviation, variance and standard deviation of two sets of data; compare center and spread of two or more different data sets; interpret differences to draw conclusions about data
What's Spread Got to Do With It? (FAL)	Formative Assessment Lesson *Partner/SmallGroup*	Summarize, represent, and interpret data on a single count or measurement variable; compare center and spread of two or more different data sets; make inferences and justify conclusions from sample surveys, experiments, and observational studies.
Empirical Rule	Group Learning Task *Partner/Small Group*	Identify data sets as approximately normal or not using the Empirical Rule; use the mean and standard deviation to fit data to a normal distribution where appropriate.
Let's Be Normal	Learning Task *Partner/Individual*	Use the mean and standard deviation to fit data to a normal distribution; use calculators or tables to estimate areas under the normal curve; interpret areas under a normal curve into context
And You Believe That?!	Learning Task *Individual*	Demonstrate understanding of the different kinds of sampling methods; discuss the appropriate way of choosing samples in context with limiting factors; recognize and understand bias in sampling methods
"Cost of Quality" in the Pulp & Paper Industry	Practice Task *Individual*	Analyze product quality data in the production of wood pulp (a major industry in the state of Georgia)
How Tall are Our Students	Practice Task *Partner/Individual*	Gather sample data, calculate statistical parameters and draw inferences about populations.
We're Watching You	Learning Task *Individual*	Explain how and why a sample represents the variable of interest from a population; demonstrate understanding of the different kinds of sampling methods; apply the steps involved in designing an observational study; apply the basic principles of experimental design

Color of Skittles	Group Learning Task *Small group*	Understand sample distributions of sample proportions through simulation; develop the formulas for the mean and standard deviation of the sampling distribution of a sample proportion; discover the Central Limit Theorem for a sample proportion
Pennies	Learning Task *Partner/Small Group*	Understand sample distributions of sample means through simulation; develop the formulas for the mean and standard deviation of the sampling distribution of a sample means; discover the Central Limit Theorem for a sample means; use sample means to estimate population values; conduct simulations of random sampling to gather sample means; explain what the results mean about variability in a population.
The Gettysburg's Address (Spotlight Task)	Learning Task *Partner/Individual*	Understand sample distributions of sample means through simulation; develop the formulas for the mean and standard deviation of the sampling distribution of a sample means; discover the Central Limit Theorem for sample means; use sample means to estimate population values; conduct simulations of random sampling to gather sample means; explain what the results mean about variability in a population.
How Confident Are You?	Learning Task *Partner/Individual*	Develop an understanding of margin of error through confidence intervals; calculate confidence intervals for sample proportions; calculate confidence intervals for sample means
Final Grades	Culminating Task	Understand how to choose summary statistics that are appropriate to the characteristics of the data distribution, such as the shape of the distribution or the existence of outliers; how the normal distribution uses area to make estimates of frequencies which can be expressed as probabilities and recognizing that only some data are well described by a normal distribution; understand how to make inferences about a population using a random sample
Draw You own Conclusions	Culminating Task	Understand how to choose summary statistics that are appropriate to the characteristics of the data distribution, such as the shape of the distribution or the existence of outliers; how the normal distribution uses area to make estimates of frequencies which can be expressed as probabilities and recognizing that only some data are well described by a normal distribution; understand how to make inferences about a population using a random sample; compare theoretical and empirical results to evaluate the effectiveness of a treatment; the way in which data is collected determines the scope and nature of the conclusions that

		can be drawn from the data; how to use statistics as a way of dealing with, but not eliminating, variability of results from experiments and inherent randomness.

Math Award Learning Task

Math Goals

- Calculate the population variance and population standard deviation by hand
- Compare variance and standard deviation to the mean absolute deviation (learned in Coordinate Algebra) as a measure of spread

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, ~~mean absolute deviation~~, standard deviation) of two or more different data sets.

Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them.
2. Attend to precision

Introduction

In this task students will learn how to calculate the population variance and population standard deviation by hand. They will compare it to the mean deviation as a measure of spread. They should have learned how to calculate the mean deviation in the sixth grade and reviewed the concept in Coordinate Algebra.

Materials

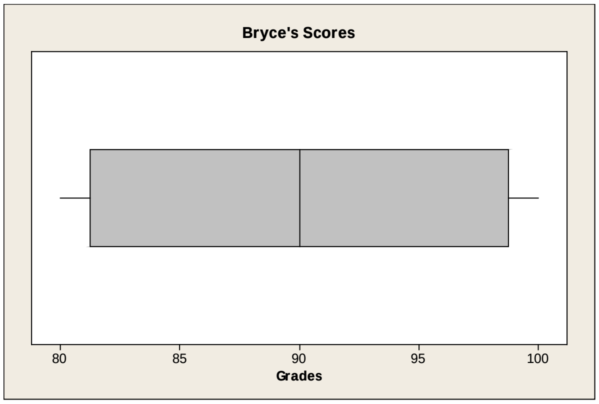
- Calculators
- Graph or centimeter grid paper optional

Your teacher has a problem and needs your input. She has to give one math award this year to a deserving student, but she can't make a decision. Here are the test grades for her two best students:

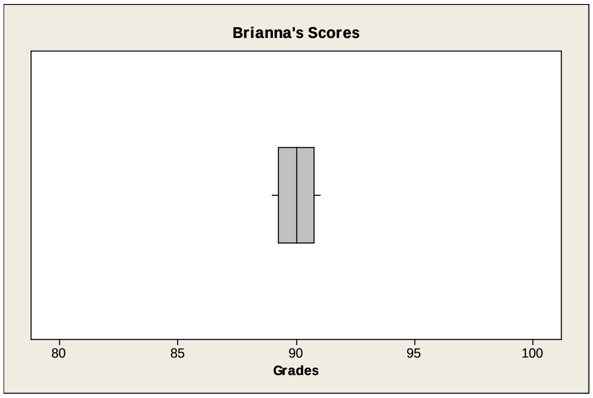
Bryce: 90, 90, 80, 100, 99, 81, 98, 82

Brianna: 90, 90, 91, 89, 91, 89, 90, 90

1) Create a boxplot for each student's grade distribution and record the five-number summary for each student.



Bryce's Scores



Brianna's Scores

Bryce: min = 80, Q1 = 81.5, median = 90, Q3 = 98.5, max = 100, IQR = 17

Briana: min = 89, Q1 = 89.5, median = 90, Q3 = 90.5, max = 91, IQR = 1

2) Based on your display, write down which of the two students should get the math award and discuss why they should be the one to receive it.

Your students will probably first calculate the average of each student. They will soon discover that both students have an average of 90. They will then either say that both deserve the award or go on to say that Bryce should get it because he had higher A's or Brianna should get it because she was more consistent. This should open up a discussion that it is very important to use a measure of spread (or variability) to describe a distribution. Many times we only look at a measure of center to describe a distribution.

3) Calculate the mean ($\bar{x}$) of Bryce's grade distribution. ***90***

Calculate the mean absolute deviation, variance, and standard deviation of Bryce's distribution.

In the first high school course, students calculated the mean absolute deviation. This is probably the first time for them to calculate the variance and the standard deviation. Point out that the sum of the column $X_i - \bar{X}$ ***will always be zero and that is why they have to take the absolute value or square those values before they average them to get the mean deviation or the variance. They will probably discover that there is a huge discrepancy between the mean deviation and the variance. That should lead into the discussion of why they take the square root of the variance to get the standard deviation.***

The formulas for mean absolute deviation, variance, and standard deviation are below.

mean absolute deviation: $MAD = \frac{1}{n}\sum|x_i - \bar{x}|$ **variance:** $\sigma^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$

standard deviation: $\sigma = \sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2}$ **which is the square root of the variance**

4) Fill out the table to help you calculate them by hand.

Scores for Bryce (x_i)	**Mean Deviation $x_i - \overline{x}$**	**Mean Absolute Deviation $\lvert x_i - \overline{x}\rvert$**	**Variance $(x_i - \overline{x})^2$**
90	*0*	*0*	*0*
90	*0*	*0*	*0*
80	*-10*	*10*	*100*
100	*10*	*10*	*100*
99	*9*	*9*	*81*
81	*-9*	*9*	*81*
98	*8*	*8*	*64*
82	*-8*	*8*	*64*
Total	*0*	*54*	*490*

MAD for Bryce: $\frac{54}{8} = 6.75$

Variance for Bryce: $\frac{490}{8} = 61.25$

Standard deviation for Bryce: $\sqrt{61.25} = 7.826$

5) What do these measures of spread tell you about Bryce's grades?

All of these values tell you how your data deviates from the mean. The variance is much larger than the mean deviation or the standard deviation because the deviations from the mean were squared. That's why you take the square root of the variance to get the standard deviation. The standard deviation and mean deviation are pretty close in value.

6) Calculate the mean of Brianna 's distribution. *90*

7) Calculate the mean deviation, variance, and standard deviation of Brianna's distribution.

Scores for Brianna (x_i)	**Mean Deviation $x_i - \overline{x}$**	**Mean Absolute Deviation $\lvert x_i - \overline{x} \rvert$**	**Variance $(x_i - \overline{x})^2$**
90	***0***	***0***	***0***
90	***0***	***0***	***0***
91	***1***	***1***	***1***
89	***-1***	***1***	***1***
91	***1***	***1***	***1***
89	***-1***	***1***	***1***
90	***0***	***0***	***0***
90	***0***	***0***	***0***
Total	***0***	***4***	***4***

MAD for Brianna: $\frac{4}{8} = 0.5$

Variance for Brianna: $\frac{4}{8} = 0.5$

Standard deviation for Brianna: $\sqrt{0.5} = 0.707$

8) What do these measures of spread tell you about Brianna's grades?

All of these values tell you how your data deviates from the mean. In this case, the mean deviation and the variance are the same. Unlike Bryce, and probably surprising to students, the standard deviation is larger than the variance.

9) Based on this information, write down which of the two students should get the math award and discuss why they should be the one to receive it.

Students will have different opinions of who should win. Students must be able to justify their answer using an analysis of the statistics.

Math Award Learning Task

Name ____ Date ____________________

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, ~~mean absolute deviation~~, standard deviation) of two or more different data sets.

Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them.
2. Attend to precision

Your teacher has a problem and needs your input. She has to give one math award this year to a deserving student, but she can’t make a decision. Here are the test grades for her two best students:

Bryce: 90, 90, 80, 100, 99, 81, 98, 82

Brianna: 90, 90, 91, 89, 91, 89, 90, 90

1) Create a boxplot for each student’s grade distribution and record the five-number summary for each student.

2) Based on your display, write down which of the two students should get the math award and discuss why they should be the one to receive it.

3) Calculate the mean ($\bar{x}$) of Bryce’s grade distribution.

Calculate the mean deviation, variance, and standard deviation of Bryce's distribution.

The formulas for mean absolute deviation, variance, and standard deviation are below.

mean absolute deviation: $MAD = \frac{1}{n}\sum|x_i - \bar{x}|$ **variance:** $s^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$

standard deviation: $s = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$ **which is the square root of the variance**

4) Fill out the table to help you calculate them by hand.

Scores for Bryce (x_i)	**Mean Deviation $x_i - \bar{x}$**	**Mean Absolute Deviation $|x_i - \bar{x}|$**	**Variance $(x_i - \bar{x})^2$**
90			
90			
80			
100			
99			
81			
98			
82			
Total			

MAD for Bryce: _______

Variance for Bryce: ______

Standard deviation for Bryce: _______

5) What do these measures of spread tell you about Bryce's grades?

6) Calculate the mean of Brianna 's distribution.

7) Calculate the mean deviation, variance, and standard deviation of Brianna's distribution.

Scores for Brianna (x_i)	**Mean Deviation $x_i - \overline{x}$**	**Mean Absolute Deviation $\lvert x_i - \overline{x} \rvert$**	**Variance $(x_i - \overline{x})^2$**
90			
90			
91			
89			
91			
89			
90			
90			
Total			

MAD for Brianna: _______

Variance for Brianna: ______

Standard deviation for Brianna: _______

8) What do these measures of spread tell you about Brianna's grades?

9) Based on this information, write down which of the two students should get the math award and discuss why they should be the one to receive it.

What's Spread Got to Do with It? (FAL)

Source: Georgia Mathematics Design Collaborative

This lesson is intended to help you assess how well students are able to:

- Interpret the meaning of standard deviation as a measure of the spread away from the mean of a set of data.
- Compare data sets using standard deviation

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID.2 **Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, ~~mean absolute deviation~~, standard deviation) of two or more different data sets.**

S.IC Make inferences and justify conclusions from sample surveys, experiments, and observational studies.

STANDARDS FOR MATHEMATICAL PRACTICE:

This lesson uses all of the practices with emphasis on:

2. Reason abstractly and quantitatively
3. Construct viable arguments and critique the reasoning of others
7. Look for and make use of structure

TASK DESCRIPTION, DEVELOPMENT AND DISCUSSION:

Tasks and lessons from the Georgia Mathematics Design Collaborative are specifically designed to help teachers effectively formatively assess their students. The way the tasks and lessons are designed gives the teacher a clear understanding of what the students are able to do and not do. Within the lesson, teachers will find suggestions and question prompts that will help guide students towards understanding.

The task, *What's Spread Got to Do with It?*, is a Formative Assessment Lesson (FAL) that can be found at: http://ccgpsmathematics9-10.wikispaces.com/Georgia+Mathematics+Design+Collaborative+Formative+Assessment+Lessons

Empirical Rule Learning Task

Mathematical Goals

- Identify data sets as approximately normal or not using the Empirical Rule
- Use the mean and standard deviation to fit data to a normal distribution where appropriate

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID. 4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets and tables to estimate areas under the normal curve

Standards for Mathematical Practice

1. **Make sense of problems and persevere in solving them.**
2. **Reason abstractly and quantitatively.**
3. **Use appropriate tools strategically**.

Introduction

This task provides students with an understanding of how standard deviation can be used to describe distributions. It also will cause them to realize that standard deviation is not resistant to strong skewness or outliers. The standard deviation is best used when the distribution is symmetric. The empirical rule is useful for Normal Distributions which are symmetric and bell-shaped. It will not apply to non-bell shaped curves. The instructions state “under certain conditions” the Empirical Rule can be used to make a good guess of the standard deviation. “Under certain conditions” should state “for a Normal Distribution.” Hopefully, by the end of this activity, students will realize that the Empirical rule applies only to normal distributions. They should also realize that the standard deviation is not a good measure of spread for non-symmetric data.

Materials

- pencil
- graphing calculator or statistical software package

Under certain conditions (those you will discover during this activity) the Empirical Rule can be used to help you make a good guess of the standard deviation of a distribution.

The Empirical Rule is as follows:

For certain conditions (which you will discover in this activity),

- 68% of the data will be located within one standard deviation symmetric to the mean
- 95% of the data will be located within two standard deviations symmetric to the mean
- 99.7% of the data will be located within three standard deviations symmetric to the mean

For example, suppose the data meets the conditions for which the empirical rule applies. If the mean of the distribution is 10, and the standard deviation of the distribution is 2, then about 68%

of the data will be between the numbers 8 and 12 since 10-2 =8 and 10+2 = 12. We would expect approximately 95% of the data to be located between the numbers 6 and 14 since 10-2(2) = 6 and 10 + 2(2) = 14. Finally, almost all of the data will be between the numbers 4 and 16 since 10 – 3(2) = 4 and 10 + 3(2) = 16.

For each of the dotplots below, use the Empirical Rule to estimate the mean and the standard deviation of each of the following distributions. Then, use your calculator to determine the mean and standard deviation of each of the distributions. Did the empirical rule give you a good estimate of the standard deviation?

You may need to tell the students a method for estimating the standard deviation if they cannot come up with a method. Since 95% of your data is within ±2 standard deviations of your mean, and 99.7% of your data is within ±3 standard deviations of your mean, then a way for you to estimate a standard deviation (if the empirical rule applies) is to find the range of your data and divide that number by 4 (for 2 standard deviations)because two standard deviations divides the curve into 4 sections and then divide the range by 6 (for 3 standard deviations) because three standard deviations divides the curves into 6 sections. Your standard deviation should be between these two numbers if the empirical rule applies.

In the Math Award Task students learned how to calculate the variance and standard deviation by hand using the formula. You may want to show them how to calculate the standard deviation on the calculator. To do this, follow the steps below: (these instructions are for a TI-83 or TI-84 calculator):

- ***Enter the data into your list by pressing "STAT" and then "EDIT." List 1 (L1) should pop up. Enter the data.***
- ***To calculate the standard deviation, press "STAT" "CALC" "1-Var Stats" "Enter". A list of summary statistics should appear. The mean is represented by the symbol $\overline{x}$. There are two standard deviations listed. S_x refers to the sample standard deviation and σ_x (referred to as sigma) is the population standard deviation.***

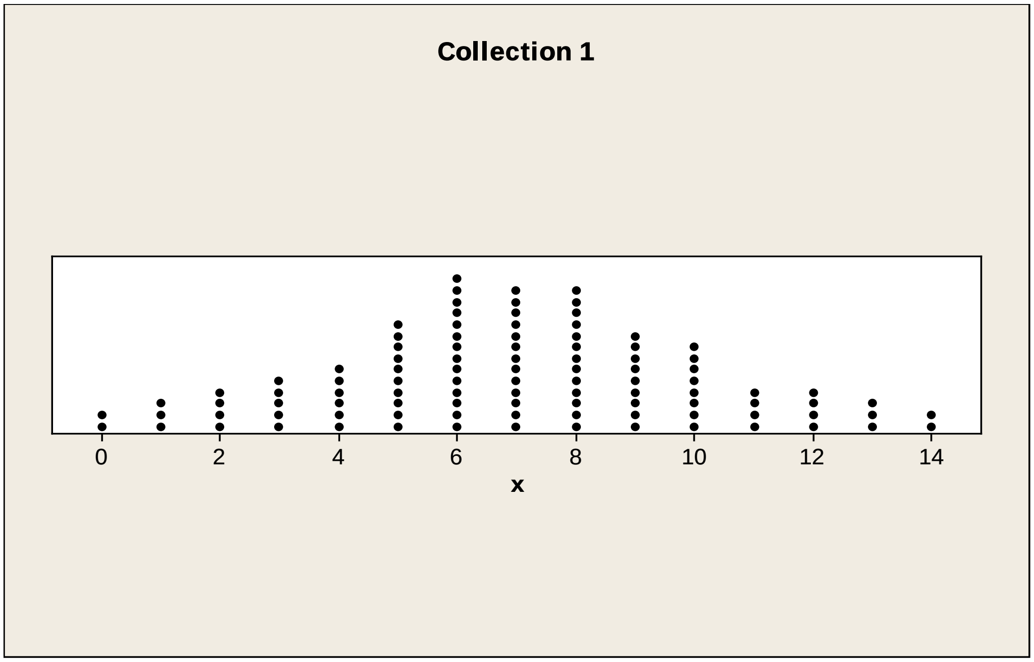
The formula that the students previously learned was for the population standard deviation. Today, have them use the population standard deviation as well.

An easier method for finding the exact standard deviation, rather than typing 100 numbers into L1, is as follows:

- ***Enter the values of x into list 1 on your calculator, but do not enter repeated values.***
- ***Enter the frequencies of these values into list 2. For example if you have the data 0,0,0,0,1,1,1,1,1, then only 0 and 1 would appear in list 1 and the numbers 4, and 5 would appear in List 2 adjacent to 0 and 1.***
- ***On the TI-83 or TI-84 calculator, press "Stat" "CALC" "1-Var Stats"….do not press enter yet***

- ***On your home screen, you will see "1-Var Stats" Type L1, L2, so you should see on your home screen 1-Var Stats L1, L2 (the comma is located over the 7 button)***
- ***Now press "Enter." The mean and standard deviation will appear in a list of summary statistics.***

For your convenience, there are 100 data points for each dotplot.



Estimated mean: *7* Estimated standard deviation*: between* $\frac{14}{6} = 2.3$ *and* $\frac{14}{4} = 3.5$

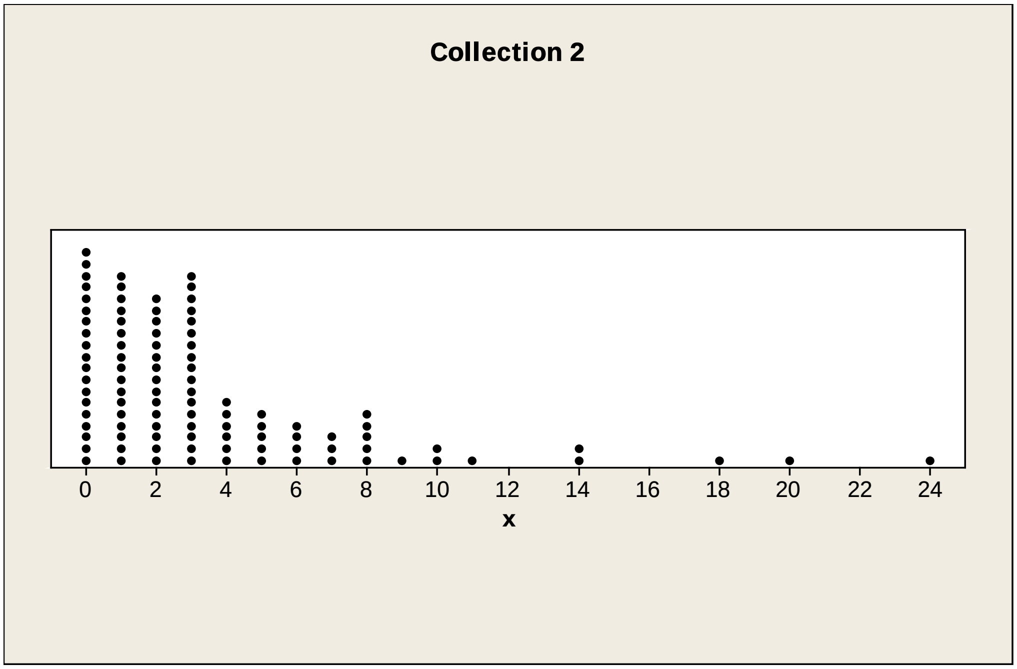
Actual mean: *6.79* Actual standard deviation: *3.14*

Did the empirical rule help give you a good estimate of the standard deviation?

Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ *3.65* $\mu + \sigma =$ ___ *9.93*

Locate these numbers on the dotplot above. How many dots are between these numbers?__*65*__ Is this close to 68%?___*yes*___ Do you think that the empirical rule should apply to this distribution?__*yes*____



Estimated mean*: 3.5* Estimated standard deviation*: between* $\frac{24}{6} = 4$ *and* $\frac{24}{4} = 6$

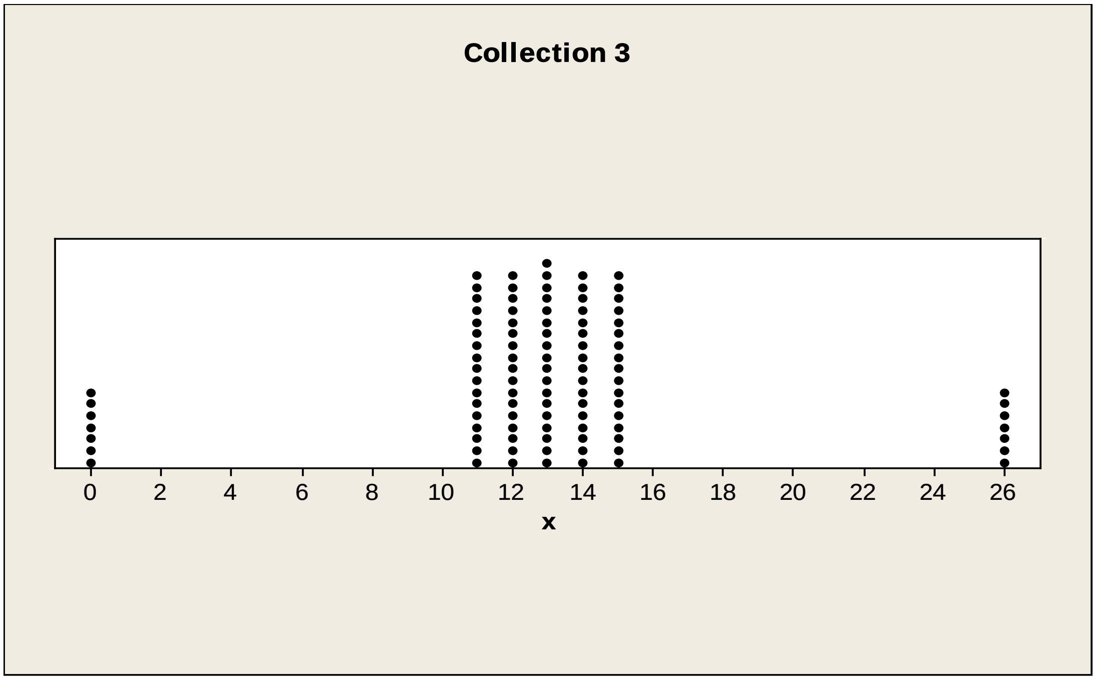
Actual mean*:3.62* Actual standard deviation: *4.30*

Did the empirical rule help give you a good estimate of the standard deviation?

Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ *-0.68* $\mu + \sigma =$ ___ *7.92*

Locate these numbers on the dotplot above. How many dots are between these numbers?___*86* ___ Is this close to 68%?____*no*__ Do you think that the empirical rule should apply to this distribution?__*no*____



Estimated mean:***13*** Estimated standard deviation: *Between* $\frac{26}{6} = 4.3\ and\ \frac{26}{4} = 6.5$

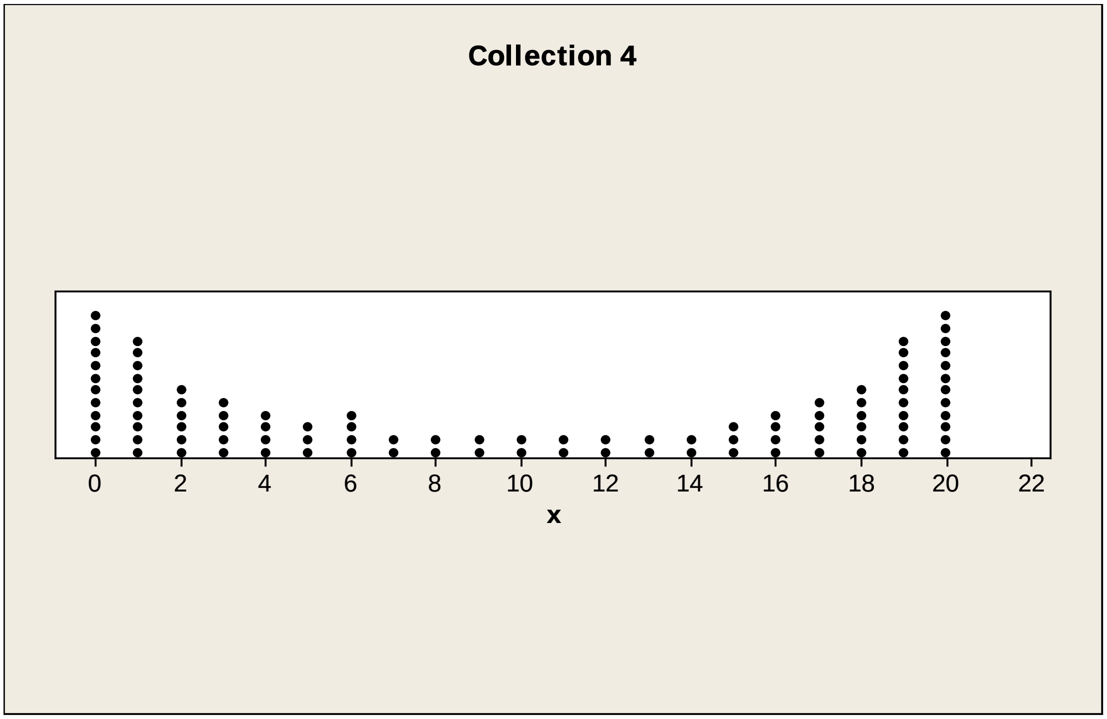
Actual mean: ***13*** Actual standard deviation: ***5.04***

Did the empirical rule help give you a good estimate of the standard deviation?

Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ ***7.96*** $\mu + \sigma =$ ___ ***18.04***

Locate these numbers on the dotplot above. How many dots are between these numbers?__***86***____ Is this close to 68%?__***no***____ Do you think that the empirical rule should apply to this distribution?__***no***____


Collection 4

Estimated mean: *10*

Estimated standard deviation: *between* $\frac{20}{6} = 3.3$ *and* $\frac{20}{4} = 5$

Actual mean: *9.92*

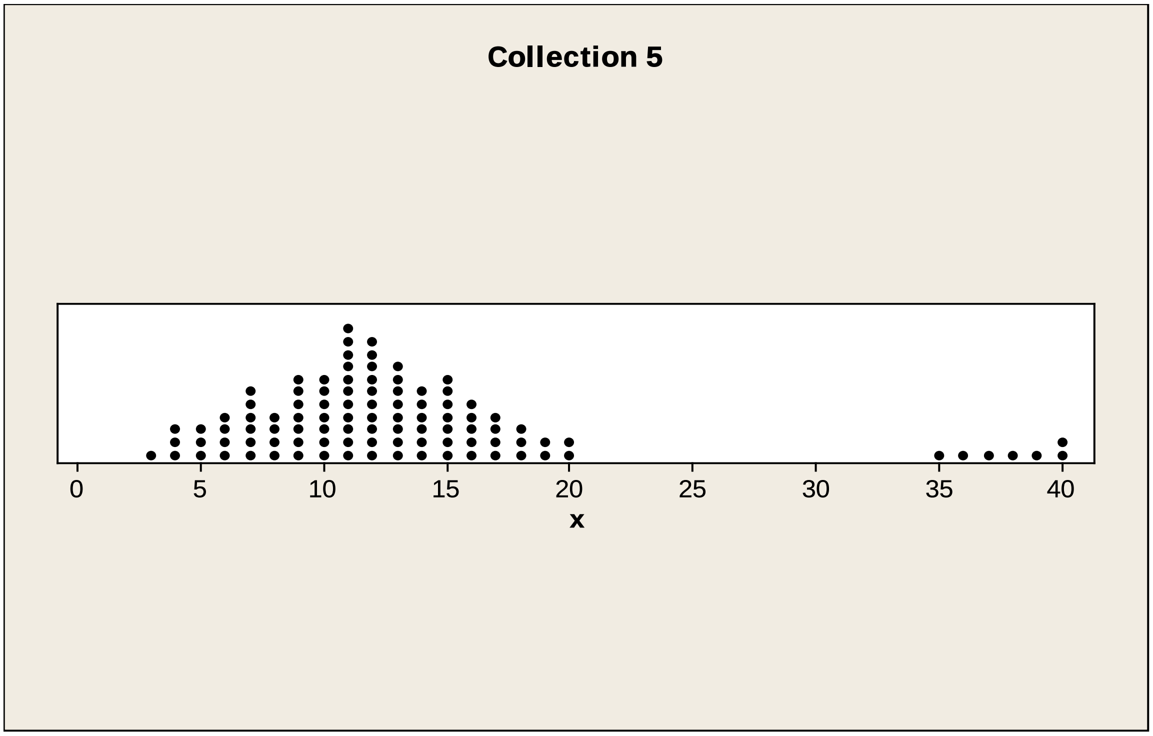
Actual standard deviation:*7.66*

Did the empirical rule help give you a good estimate of the standard deviation?

Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ *2.26* $\mu + \sigma =$ ___ *17.58*

Locate these numbers on the dotplot above. How many dots are between these numbers?___*44*___ Is this close to 68%?__*no*____ Do you think that the empirical rule should apply to this distribution?___*no*___



Estimated mean: ***12*** Estimated standard deviation: ***between*** $\frac{37}{6} = 6.167$ ***and*** $\frac{37}{4} = 9.25$

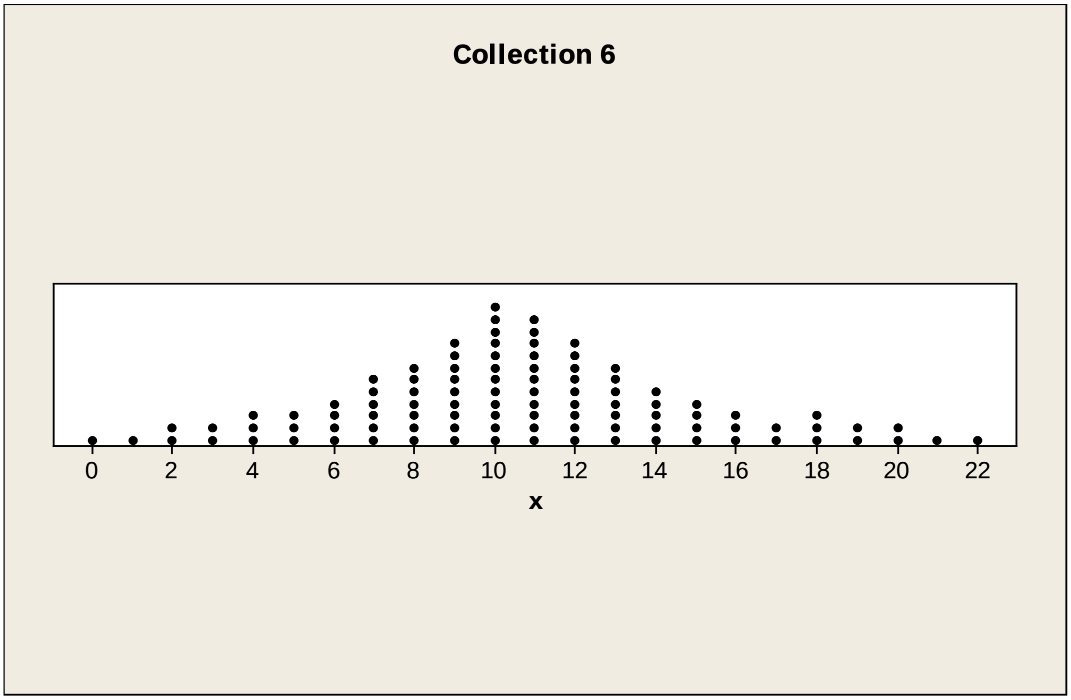
Actual mean: ***13.4*** Actual standard deviation: ***7.74***

Did the empirical rule help give you a good estimate of the standard deviation? ***yes***

Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ ***5.66*** $\mu + \sigma =$ ___ ***21.14***

Locate these numbers on the dotplot above. How many dots are between these numbers?__***86***____ Is this close to 68%?___***no***___ Do you think that the empirical rule should apply to this distribution?___***no***___



Estimated mean: ***11*** Estimated standard deviation: $\boldsymbol{between\ \frac{22}{6} = 3.6\ and\ \frac{22}{4} = 5.5}$

Actual mean:***10.68*** Actual standard deviation***: 4.49***

Did the empirical rule help give you a good estimate of the standard deviation? ***yes***

Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ ***6.19*** $\mu + \sigma =$ ___ ***15.17***

Locate these numbers on the dotplot above. How many dots are between these numbers?___***70***___ Is this close to 68%?___***yes***___ Do you think that the empirical rule should apply to this distribution?___***yes***___

For which distributions did you give a good estimate of the standard deviation based on the empirical rule?

The empirical rule only applies to normal distributions. All normal distributions are bell shaped. The first and last distributions are the only bell-shaped distributions without outliers. It is important to note, however, that all bell-shaped distributions are not normal.

Which distributions did not give a good estimate of the standard deviation based on the empirical rule? ***See student's work***

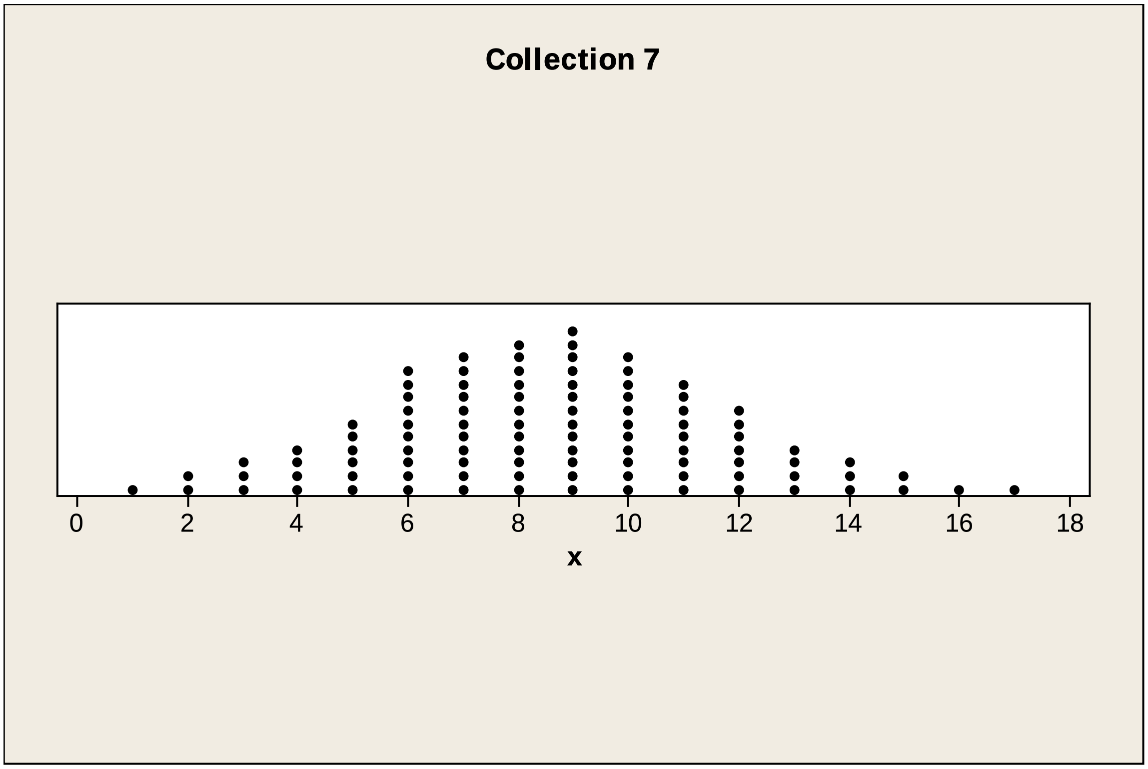
Which distributions had close to 68% of the data within one standard deviation of the mean? What do they have in common? ***The first and the last as discussed above***

For which type of distributions do you think the Empirical rule applies? ***Student's should say something like "bell shaped…mound-shaped…symmetrical"***

As you discovered, the empirical rule does not work unless your data is bell-shaped. However, not all bell-shaped graphs are normal. The next two dotplots are bell-shaped graphs. You will apply the empirical rule to determine if the bell-shaped graph is normal or not.

Using the dotplot below, calculate the mean and the standard deviation of the distribution.

Mean = 8.6 ***Standard deviation = 3.21***



Mark the mean on your dotplot above.

Calculate one standard deviation above and below the mean.

$\mu_X - \sigma_X$ = ____*5.39* and $\mu_X + \sigma_X$ = _____*11.81*. Mark these points on the x-axis of the dotplot. How many data points are between these values? ___*66*___

Calculate two standard deviations below and above the mean.

$\mu_X + 2\sigma_X =$ _____*15.02* and $\mu_X - 2\sigma_X =$ _____*2.18*. Mark these points on the x-axis of the dotplot. How many data points are between these values? ___*95*_____

Calculate three standard deviations below and above the mean.

$\mu_X - 3\sigma_X =$ _____*-1.03* and $\mu_X + 3\sigma_X =$ _____*18.23*. Mark these points on the x-axis of the dotplot. How many data points are between these values? __*100*_____

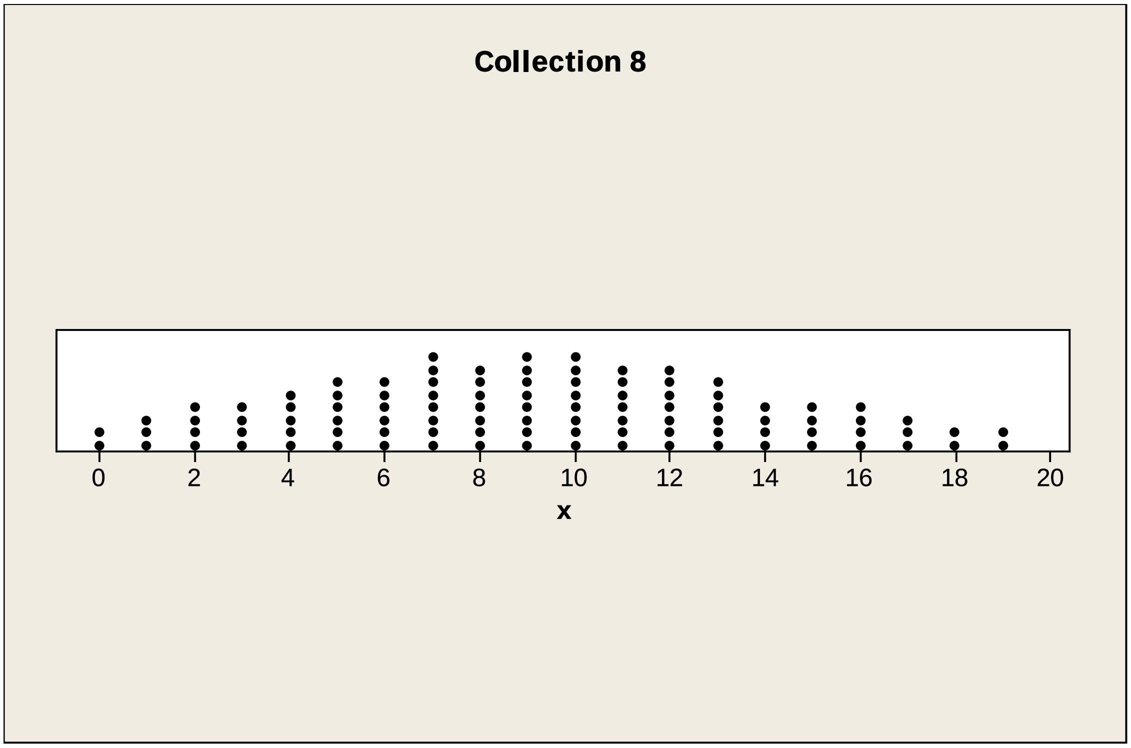
Is it likely that this sample is from a normal population? Explain your thinking. ***Yes. The percents are very close to the empirical rule…68%, 95%, 99.7%***

Outliers are values that are beyond two standard deviations from the mean in either direction.

Which values from the data would be considered to be outliers? _______*1, 2, 2, 16, 17*________

Using the dotplot below, calculate the mean and the standard deviation of the distribution.

Mean 9.17 and Standard deviation = 4.68



Mark the mean on your dotplot above.

Calculate one standard deviation above and below the mean.

$\mu_X - \sigma_X$ = ______***4.49*** and $\mu_X + \sigma_X$ = _______***13.85***. Mark these points on the x-axis of the dotplot. How many data points are between these values? ___***63***_____

Calculate two standard deviations below and above the mean.

$\mu_X + 2\sigma_X$ =_______***-0.19*** and $\mu_X - 2\sigma_X$ = _______***18.53***. Mark these points on the x-axis of the dotplot. How many data points are between these values? ____***98***______

Calculate three standard deviations below and above the mean.

$\mu_X - 3\sigma_X$ = _______***-4.87*** and $\mu_X + 3\sigma_X$ = _______***23.21***. Mark these points on the x-axis of the dotplot. How many data points are between these values? ___***100***______

Is it likely that this sample is from a normal population? Explain your thinking. ***This distribution is not as likely to be from a normal population as the other distribution because the values were not as close to the empirical rule; however, there could be sampling error.***

One definition of an outlier is a value that is beyond two standard deviations from the mean in either direction.

Which values from the data would be considered to be outliers? ________***19, 19***_______

Based on your observations when is it appropriate to conclude that a data set is approximately normal?

When the distribution is bell shaped, symmetric with no gaps or outliers.

Empirical Rule Learning Task

Name ____ Date______________________

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID. 4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets and tables to estimate areas under the normal curve

Standards for Mathematical Practice

1. **Make sense of problems and persevere in solving them.**
2. **Reason abstractly and quantitatively.**
3. **Use appropriate tools strategically**.

Under certain conditions (those you will discover during this activity) the Empirical Rule can be used to help you make a good guess of the standard deviation of a distribution.

The Empirical Rule is as follows:

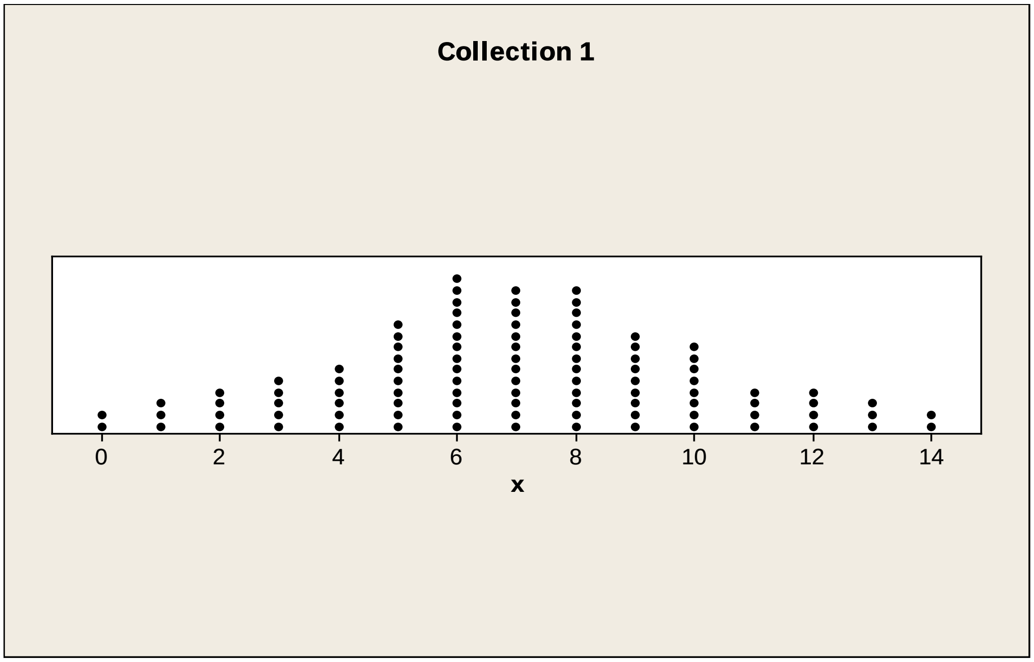
For certain conditions (which you will discover in this activity),

- 68% of the data will be located within one standard deviation symmetric to the mean
- 95% of the data will be located within two standard deviations symmetric to the mean
- 99.7% of the data will be located within three standard deviations symmetric to the mean

For example, suppose the data meets the conditions for which the empirical rule applies. If the mean of the distribution is 10, and the standard deviation of the distribution is 2, then about 68% of the data will be between the numbers 8 and 12 since 10-2 =8 and 10+2 = 12. We would expect approximately 95% of the data to be located between the numbers 6 and 14 since 10-2(2) = 6 and 10 + 2(2) = 14. Finally, almost all of the data will be between the numbers 4 and 16 since 10 – 3(2) = 4 and 10 + 3(2) = 16.

For each of the dotplots below, use the Empirical Rule to estimate the mean and the standard deviation of each of the following distributions. Then, use your calculator to determine the mean and standard deviation of each of the distributions. Did the empirical rule give you a good estimate of the standard deviation?

For your convenience, there are 100 data points for each dotplot.



Estimated mean:______________ Estimated standard deviation:___________

Actual mean:_____________ Actual standard deviation:__________

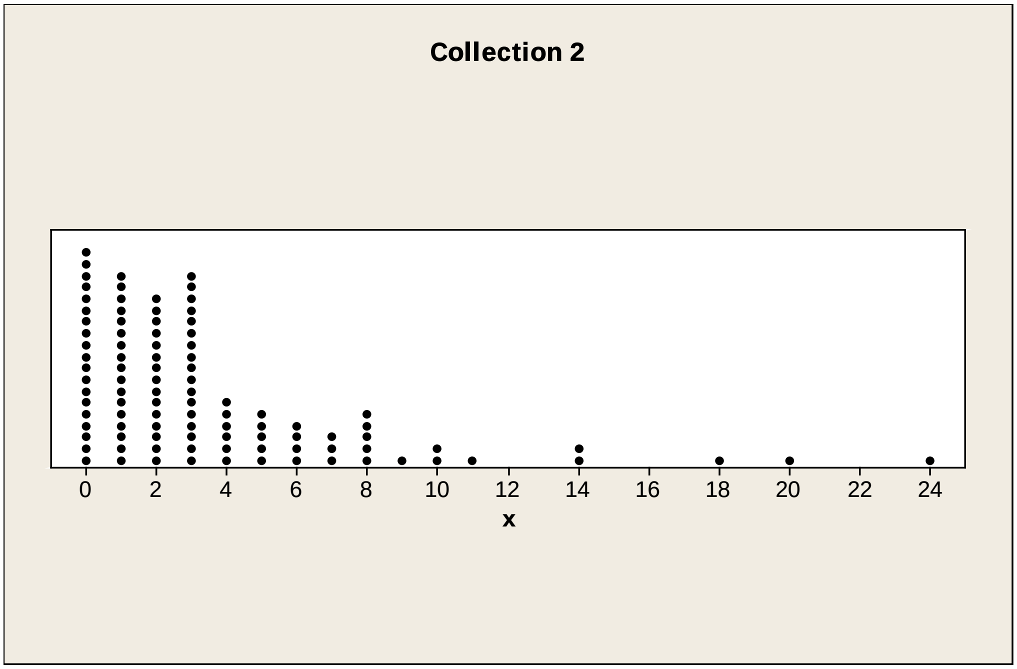
Did the empirical rule help give you a good estimate of the standard deviation?

Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ $\mu + \sigma =$ ___

Locate these numbers on the dotplot above. How many dots are between these numbers?______

Is this close to 68%?______ Do you think that the empirical rule should apply to this distribution?______



Estimated mean:_____________ Estimated standard deviation:__________

Actual mean:____________ Actual standard deviation:_________

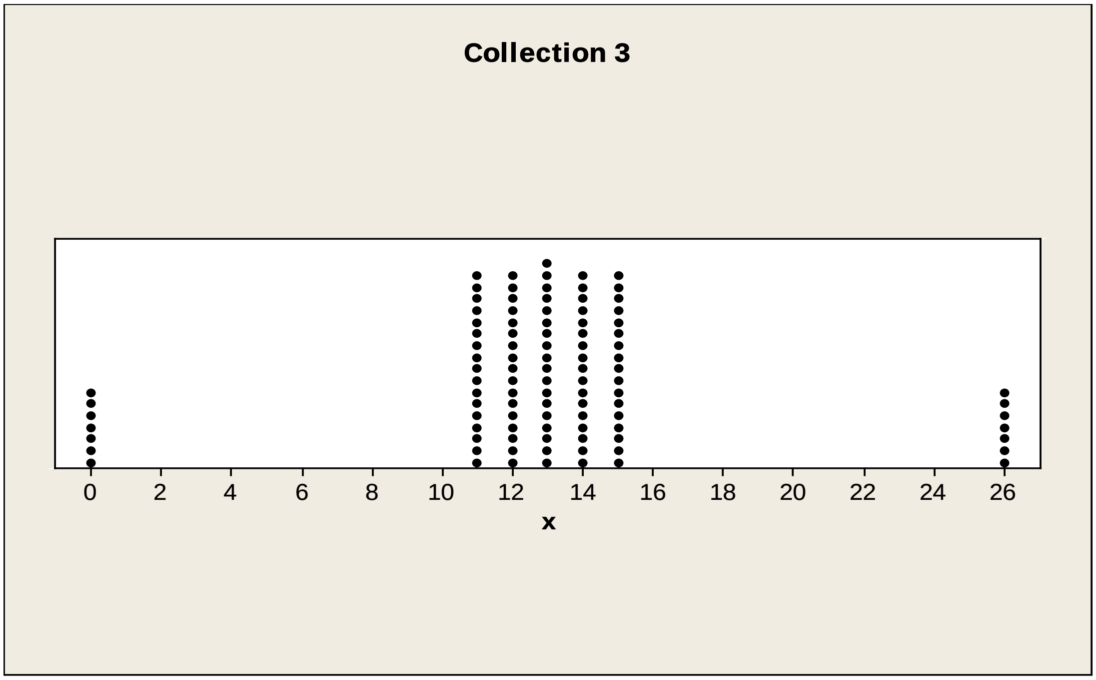
Did the empirical rule help give you a good estimate of the standard deviation?

Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ $\mu + \sigma =$ ___

Locate these numbers on the dotplot above. How many dots are between these numbers?______

Is this close to 68%?______ Do you think that the empirical rule should apply to this distribution?______



Estimated mean:______________ Estimated standard deviation:___________

Actual mean:_____________ Actual standard deviation:__________

Did the empirical rule help give you a good estimate of the standard deviation?

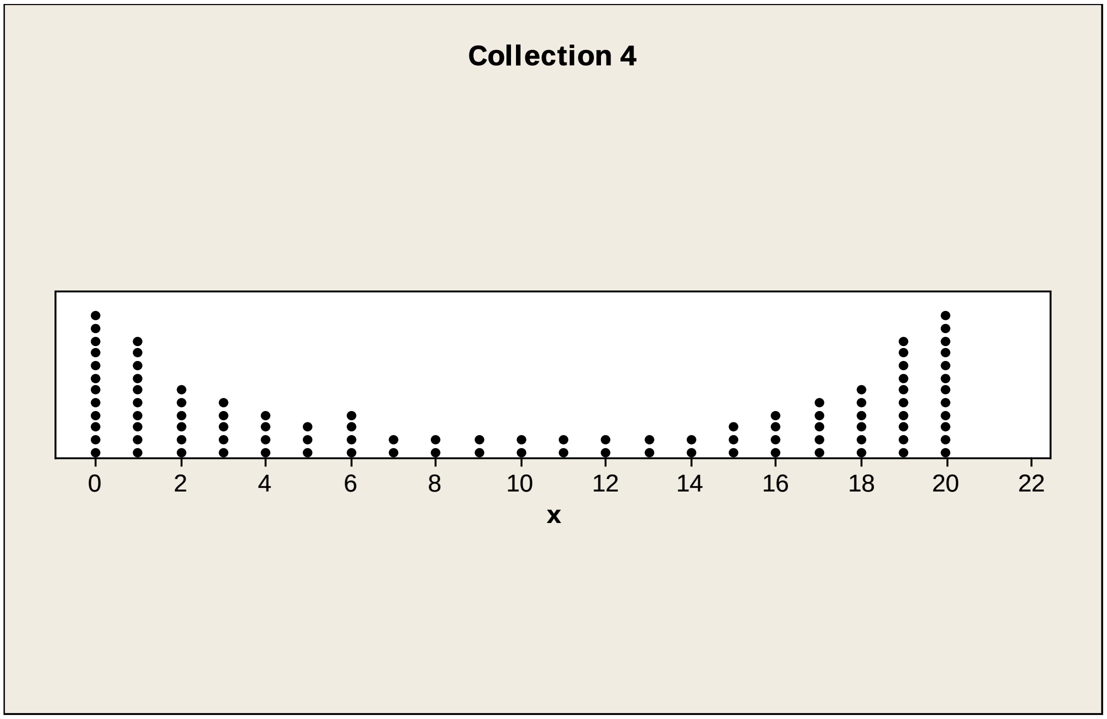
Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ $\mu + \sigma =$ ___

Locate these numbers on the dotplot above. How many dots are between these numbers?______

Is this close to 68%?______ Do you think that the empirical rule should apply to this

distribution?______



Estimated mean:______________ Estimated standard deviation:___________

Actual mean:_____________ Actual standard deviation:__________

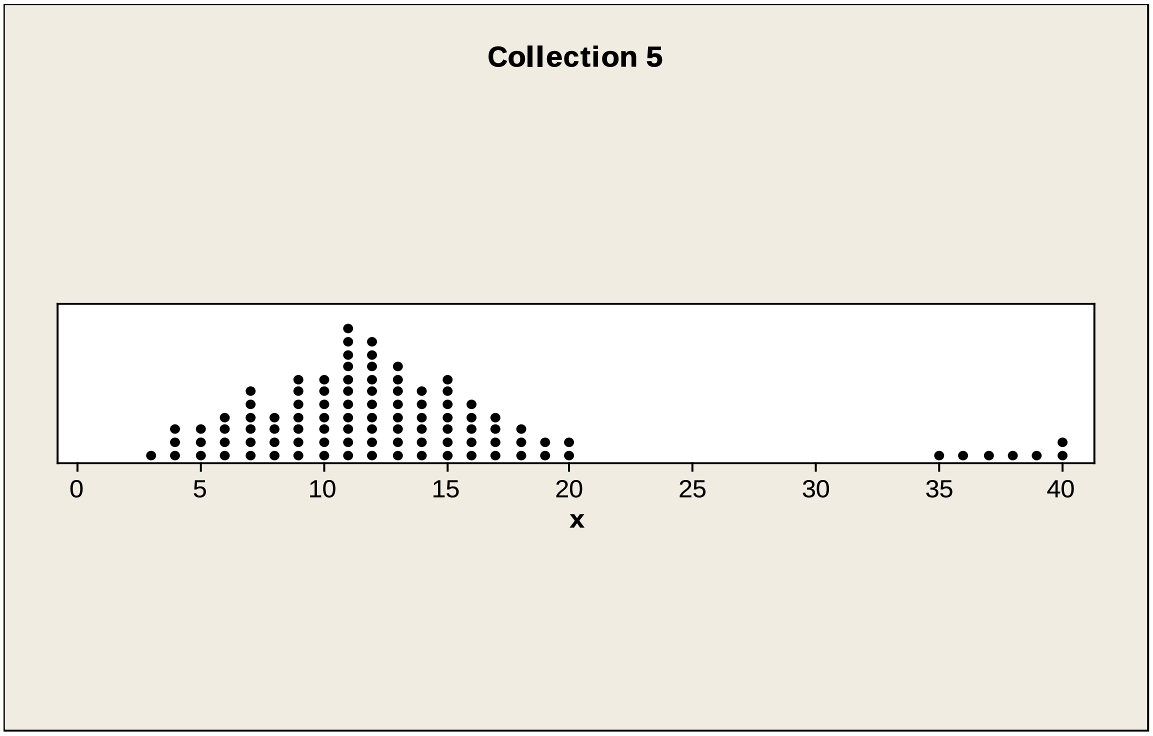
Did the empirical rule help give you a good estimate of the standard deviation?

Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ $\mu + \sigma =$ ___

Locate these numbers on the dotplot above. How many dots are between these numbers?______

Is this close to 68%?______ Do you think that the empirical rule should apply to this distribution?______



Estimated mean:______________ Estimated standard deviation:___________

Actual mean:_____________ Actual standard deviation:__________

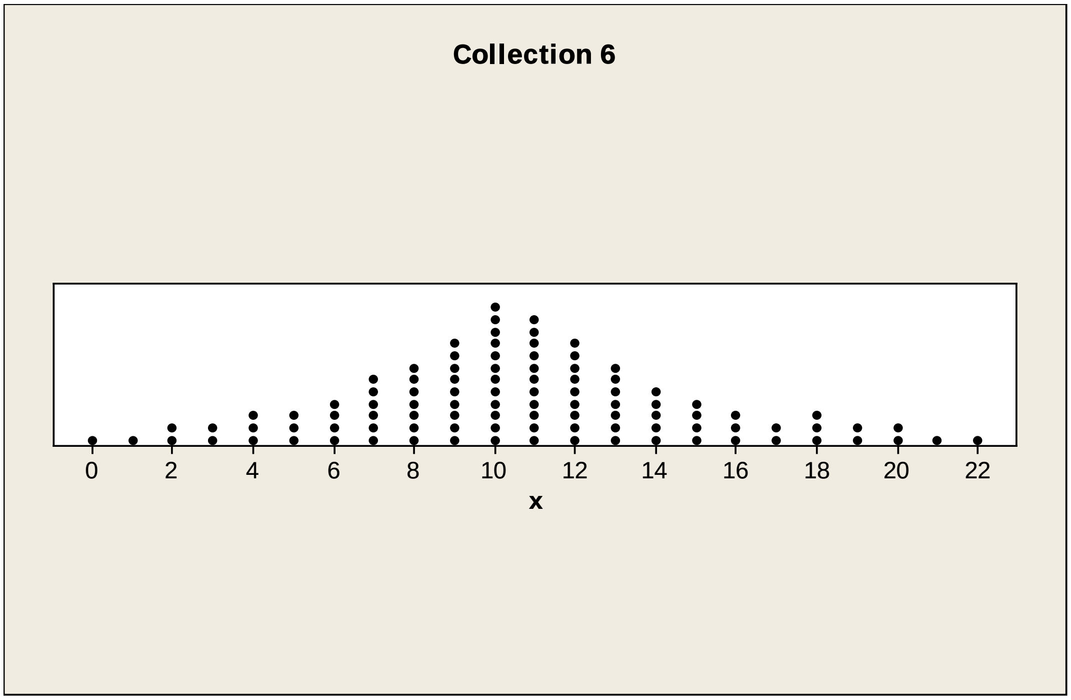
Did the empirical rule help give you a good estimate of the standard deviation?

Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ $\mu + \sigma =$ ___

Locate these numbers on the dotplot above. How many dots are between these numbers?______

Is this close to 68%?______ Do you think that the empirical rule should apply to this distribution?______



Estimated mean:_____________ Estimated standard deviation:__________

Actual mean:____________ Actual standard deviation:_________

Did the empirical rule help give you a good estimate of the standard deviation?

Now that you know what the actual mean and standard deviation, calculate one standard deviation below the mean and one standard deviation above the mean.

$\mu - \sigma =$ ___ $\mu + \sigma =$ ___

Locate these numbers on the dotplot above. How many dots are between these numbers?______

Is this close to 68%?______ Do you think that the empirical rule should apply to this distribution?______

Describe the distributions that gave you a good estimate of the standard deviation based on the empirical rule?

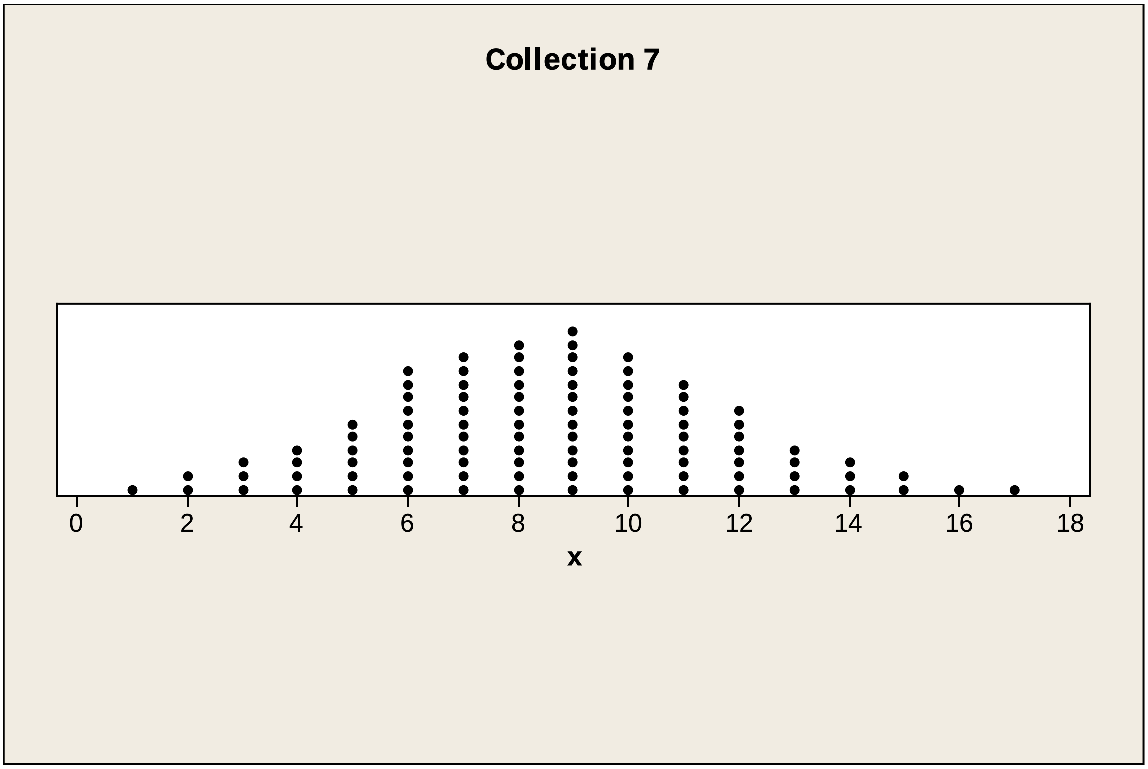
Describe the distributions that did not give you a good estimate of the standard deviation based on the empirical rule?

Which distributions had close to 68% of the data within one standard deviation of the mean? What do they have in common?

For which type of distributions do you think the Empirical rule applies?

As you discovered, the empirical rule does not work unless your data is bell-shaped. However, not all bell-shaped graphs are normal. The next two dotplots are bell-shaped graphs. You will apply the empirical rule to determine if the bell-shaped graph is normal or not.

Using the dotplot below, calculate the mean and the standard deviation of the distribution.



Mark the mean on your dotplot above.

Calculate one standard deviation above and below the mean.

$\mu_X - \sigma_X$ = ____ and $\mu_X + \sigma_X$ = _____. Mark these points on the x-axis of the dotplot. How many data points are between these values? ______

Calculate two standard deviations below and above the mean.

$\mu_X + 2\sigma_X$ =_____ and $\mu_X - 2\sigma_X$ = _____. Mark these points on the x-axis of the dotplot. How many data points are between these values? ________

Calculate three standard deviations below and above the mean.

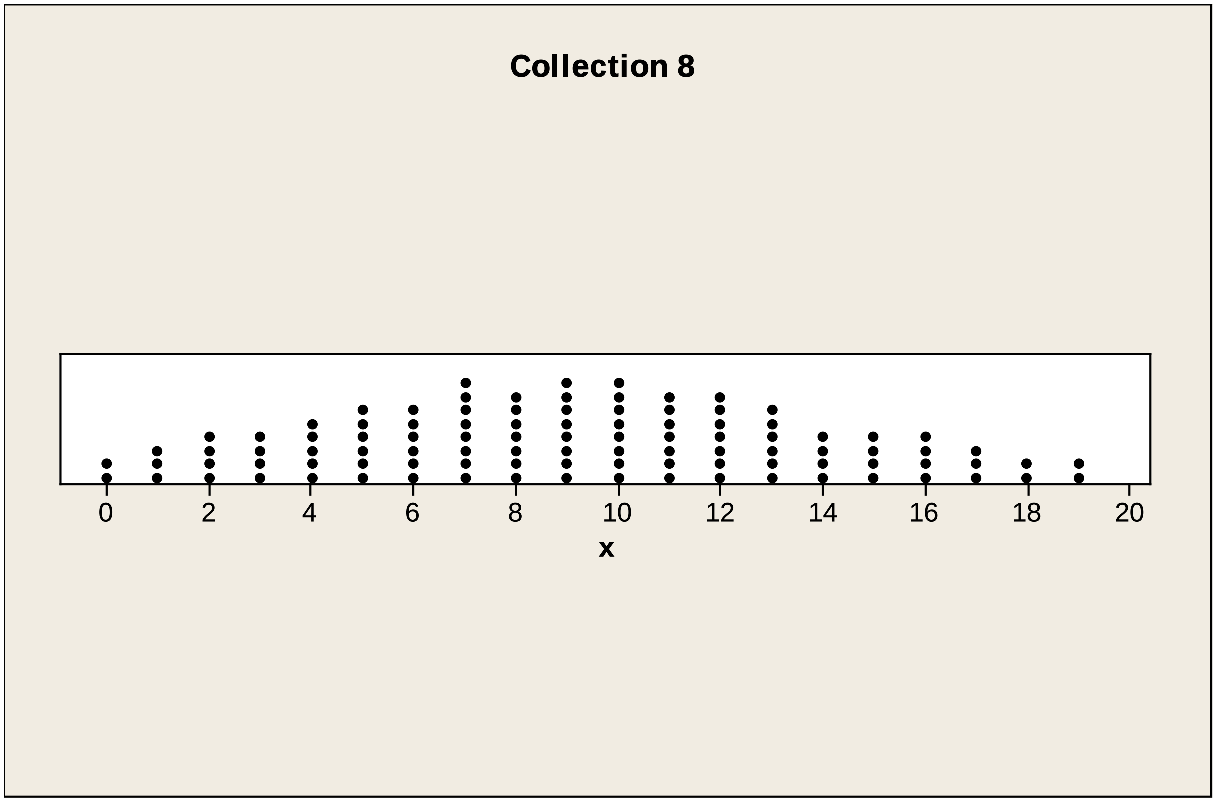
$\mu_X - 3\sigma_X$ = _____ and $\mu_X + 3\sigma_X$ = _____. Mark these points on the x-axis of the dotplot. How many data points are between these values? _______

Is it likely that this sample is from a normal population? Explain your thinking.

Outliers are values that are beyond two standard deviations from the mean in either direction.

Which values from the data would be considered to be outliers? _______________

Using the dotplot below, calculate the mean and the standard deviation of the distribution.



Mark the mean on your dotplot above.

Calculate one standard deviation above and below the mean.

$\mu_X - \sigma_X$ = ______ and $\mu_X + \sigma_X$ = _______. Mark these points on the x-axis of the dotplot.
How many data points are between these values? ________

Calculate two standard deviations below and above the mean.

$\mu_X + 2\sigma_X$ =_______ and $\mu_X - 2\sigma_X$ = _______. Mark these points on the x-axis of the dotplot.
How many data points are between these values? __________

Calculate three standard deviations below and above the mean.

$\mu_X - 3\sigma_X$ = _______ and $\mu_X + 3\sigma_X$ = _______. Mark these points on the x-axis of the dotplot.
How many data points are between these values? _________

Is it likely that this sample is from a normal population? Explain your thinking.

One definition of an outlier is a value that is beyond two standard deviations from the mean in either direction.

Which values from the data would be considered to be outliers? _______________

Based on your observations when is it appropriate to conclude that a data set is approximately normal?

Let's Be Normal Learning Task

Mathematical Goals

- Use the mean and standard deviation to fit data to a normal distribution
- Use calculators or tables to estimate areas under the normal curve
- Interpret areas under a normal curve in context

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID. 4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets and tables to estimate areas under the normal curve

Standards for Mathematical Practice

1. **Make sense of problems and persevere in solving them.**
2. **Reason abstractly and quantitatively.**
3. **Use appropriate tools strategically**.

Introduction

The purpose of this task is to have students understand how to estimate probabilities under the standard normal curve. To do this, students will be asked to standardize scores (find z-scores). Then using that information, they will be asked to find probabilities under the standard normal curve.

Materials

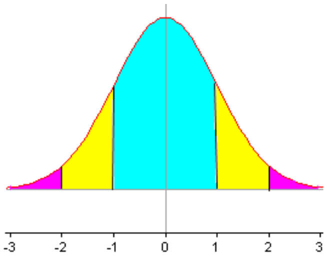
- Pencil
- paper/graph paper
- graphing calculator
- copy of the Table of Standard normal probabilities

Until this task, we have focused on distributions of discrete data. We will now direct our attention to continuous data. Where a **discrete variable** has a **finite number of possible values**, a continuous variable can assume all values in a given interval of values. Therefore, a **continuous random variab**le can assume an **infinite number of values**.

We will focus our attention specifically on continuous random variables with distributions that are approximately normal. Remember that normal distributions are symmetric, bell-shaped curves that follow the Empirical Rule.

The Empirical Rule for Normal Distributions states that

- 68% of the data values will fall within one standard deviation of the mean,
- 95% of the data values will fall within two standard deviations of the mean, and
- 99.7% of the data values will fall within three standard deviations of the mean.



In the last task, dot plots were used to explore this type of distribution and you spent time determining whether or not a given distribution was approximately normal. In this task, we will use probability histograms and approximate these histograms to a smooth curve that displays the shape of the distribution without the boxiness of the histogram. *We will also assume that all of the data we use is approximately normally distributed.*

Review:

1**)** The distribution of heights of adult American women is approximately normal with a mean of 65.5 inches and a standard deviation of 2.5 inches. Draw a normal curve and label the mean and points one, two, and three standard deviations above and below the mean.

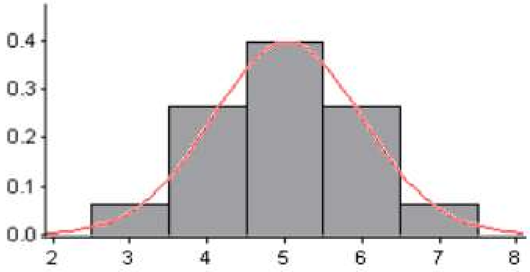
a) What percent of women are taller than 70.5 inches? ***2.5%***

b) Between what heights do the middle 95% of women fall? ***60.5 inches and 70.5 inches***

c) What percent of women are shorter than 63 inches? ***16%***

d) A height of 68 inches corresponds to what percentile of adult female American heights? ***84th percentile***

2. This is an example of a probability histogram of a continuous random variable with an approximate mean of five ($\mu = 5$) and standard deviation of one ($\sigma = 1$). A normal curve with the same mean and standard deviation has been overlaid on the histogram.

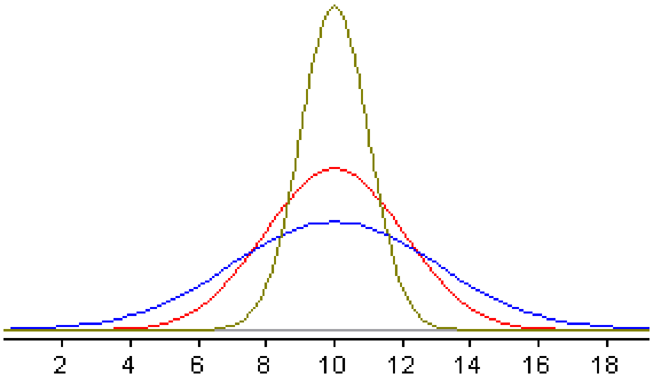


a) What do you notice about these two graphs?

Answers may vary. Sample response include: The two graphs have the same peak at 5. The normal curve cuts through the midpoint of the top of each rectangle of the histogram. You could also have students estimate the number of observations in certain intervals on this histogram when provided a sample size.

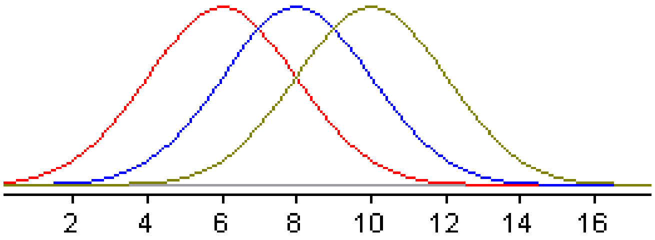
3. First, you must realize that normally distributed data may have any value for its mean and standard deviation. Below are two graphs with three sets of normal curves.

a) In the first set, all three curves have the same mean, 10, but different standard deviations. Approximate the standard deviation of each of the three curves.



Tall curve: $\sigma = 1$ Middle curve: $\sigma = 2$ Shortest curve: $\sigma = 3$

b) In this set, each curve has the same standard deviation. Determine the standard deviation and then each mean.



$\sigma = 2$ for all three curves

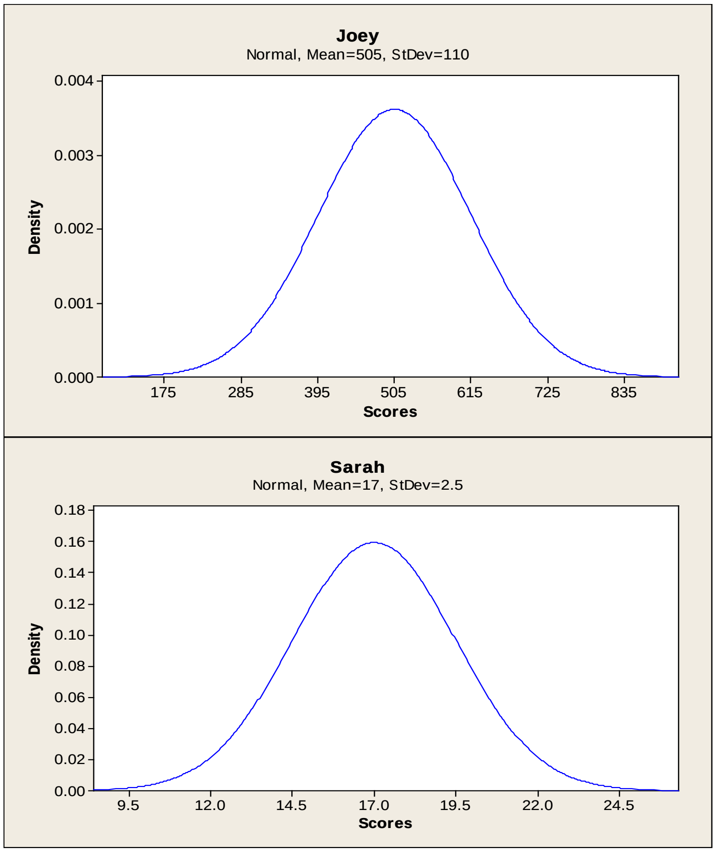
Left curve: $\mu = 6$ Middle curve: $\mu = 8$ Right curve: $\mu = 10$

c) Based on these two examples, make a conjecture regarding the effect changing the mean has on a normal distribution. Make a conjecture regarding the effect changing the standard deviation has on a normal distribution. ***The mean moves the curve left and right. Smaller values for the mean move the curve to the left. Standard deviation is a measure of variability. Standard***

deviation affects the width, and therefore the height, of the curve. The more narrow a curve is the smaller its standard deviation will be. Similarly, wider normal curves have larger standard deviations.

4. The SAT Verbal Test in recent years follow approximately a normal distribution with a mean of 505 and a standard deviation of 110. Joey took this test made a score of 600. The scores on the ACT English Test are approximately a normal distribution with a mean of 17 and a standard deviation of 2.5. Sarah took this test and made a score of 18. Which student made the better score? How do you know?

Answer: Joey. Have students sketch a normal curve for each student and place their score on the curve. Students should be able to visually recognize that Sarah's score is closer to the mean than Joey's based on the standard deviations



The ***Standard Normal Distribution*** is a normal distribution with a mean of 0 and a standard deviation of 1. To more easily compute the probability of a particular observation given a normally distributed variable, we can transform any normal distribution to this standard normal distribution using the following formula:

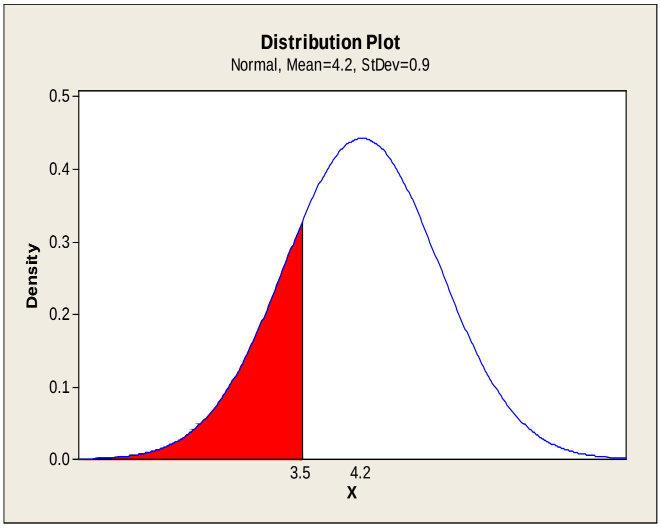
$$z = \frac{X - \mu}{\sigma}$$

When you find this value for a given value, it is referred to as the z-score. The **z-score** is a standard score for a data value that indicates **the number of standard deviations** that the data value is away from its respective mean.

5. We can use the z-score to find the probability of many other events. Let's explore those now.

a) Suppose that the mean time a typical American teenager spends doing homework each week is 4.2 hours. Assume the standard deviation is 0.9 hour. Assuming the variable is normally distributed, find the percentage of American teenagers who spend less than 3.5 hours doing homework each week.

- First, sketch a normal curve for this situation and shade the probability in which you are interested.



- Next, find the z-score for X = 3.5.

$$z = \frac{3.5 - 4.2}{0.9} = -0.78$$

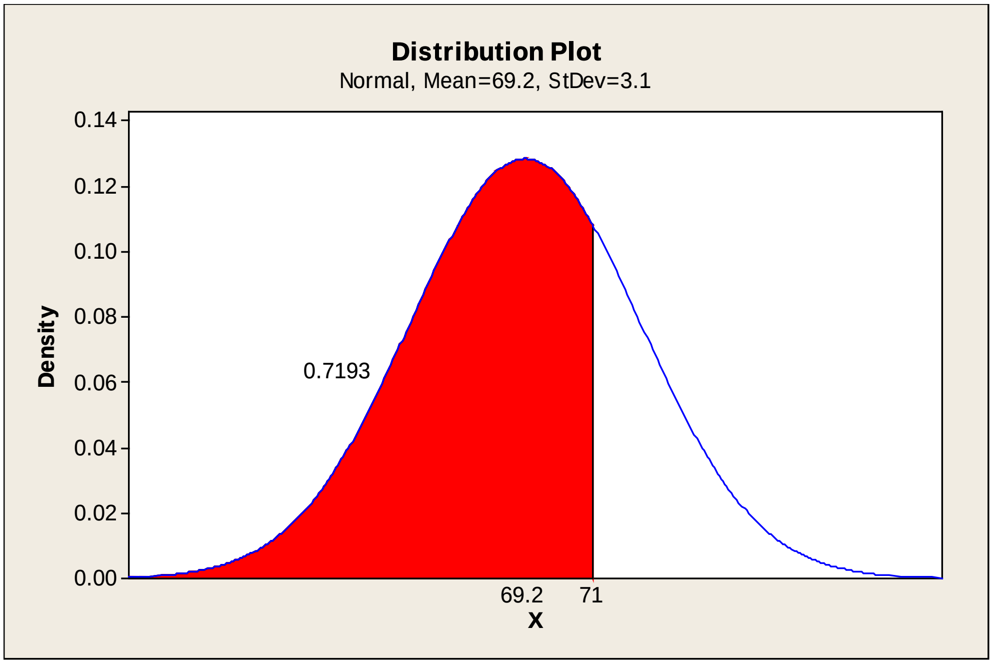
- Now, use the table of standard normal probabilities or your graphing calculator to determine the probability of $P(X < 3.5)$.

At this point, you will need to show students how to use either a z-score table or a graphing calculator to calculate the following probabilities

$P(X < 3.5) = P(z < -0.78) = \mathbf{0.217}$

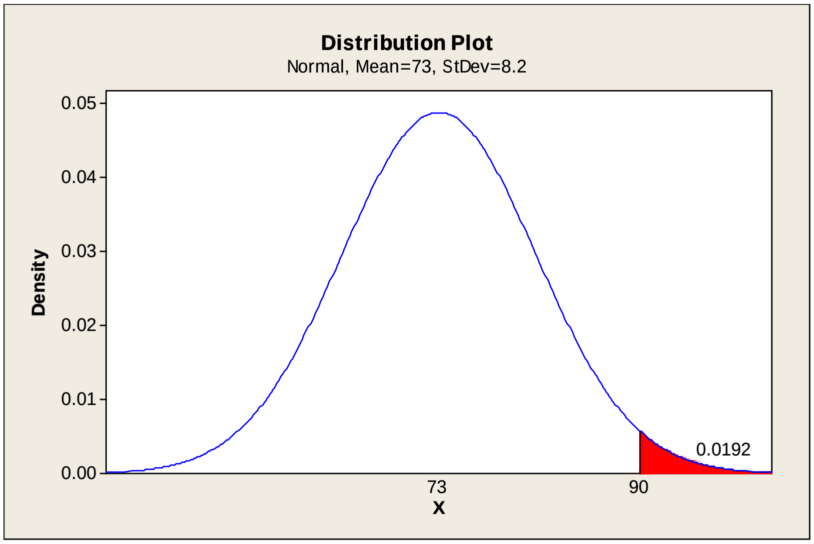
b) The average height of adult American males is 69.2 inches. If the standard deviation is 3.1 inches, determine the probability that a randomly selected adult American male will be at most 71 inches tall. Assume a normal distribution.

$$P(X \leq 71) = P(z \leq 0.5806) = 0.7193$$



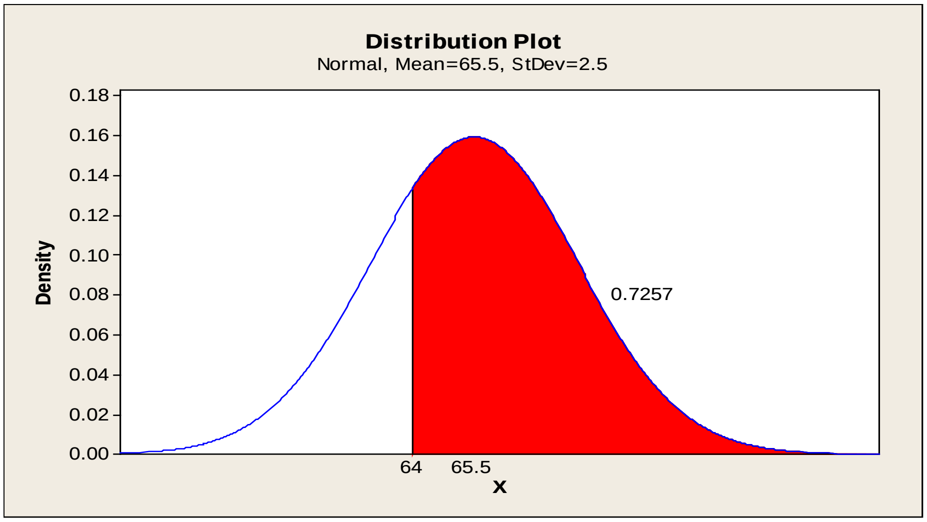
6. We can also use z-scores to find the percentage or probability of events above a given observation.

a) The average on the most recent test Ms. Cox gave her French students was 73 with a standard deviation of 8.2. Assume the test scores were normally distributed. Determine the probability that a student in Ms. Cox's class scored 90 or more on the test.



$$P(X \geq 90) = 1 - P\left(z \leq \frac{90-73}{8.2}\right) = 1 - P(z \leq 2.073) = 1 - 0.9808 = 0.0192$$

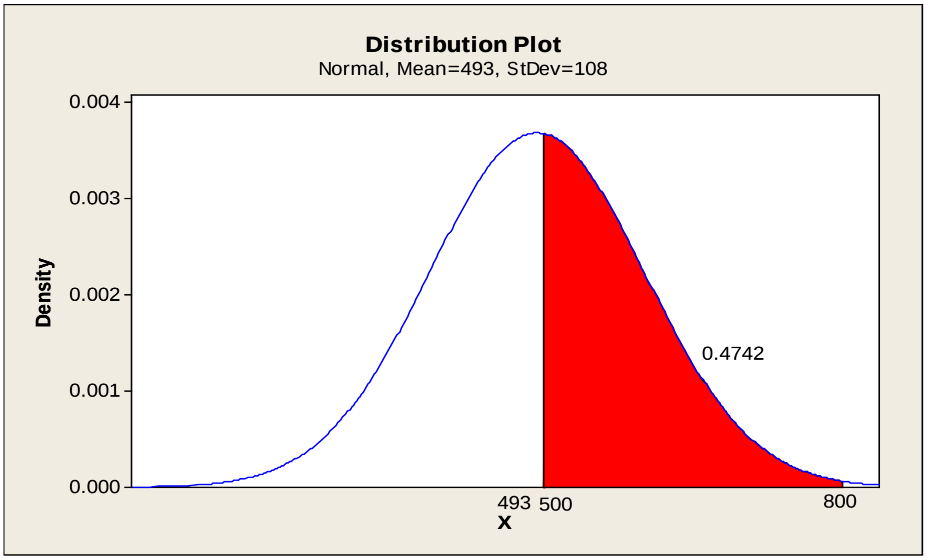
b) Women's heights are approximately normally distributed with μ = 65.5 inches and σ = 2.5 inches. Determine the probability of a randomly selected woman having a height of at least 64 inches.



$$P(X \geq 64) = 1 - P\left(z \leq \frac{64-65.5}{2.5}\right) = 1 - P(z \leq -0.60) = 1 - 0.2743 = 0.7257$$

7. We can also determine the probability between two values of a random variable.

a) According to the College Board, **Georgia** seniors graduating in 2008 had a mean Math SAT score of 493 with a standard deviation of 108. Assuming the distribution of these scores is normal, find the probability of a member of the 2008 graduating class in Georgia scoring between 500 and 800 on the Math portion of the SAT Reasoning Test.



$P(500 \le X \le 800) = P(X \le 800) - P(X \le 500)$

$P(X \le 800) - P(X \le 500)$

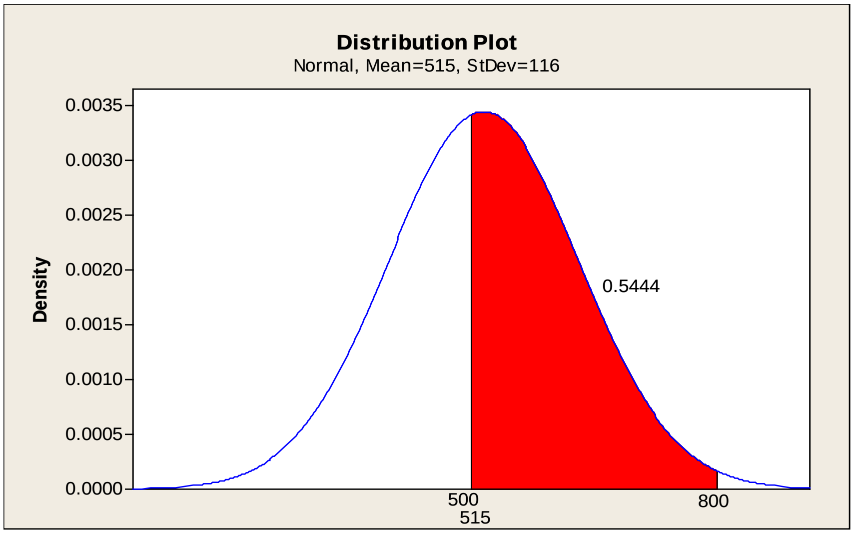
$1 - P(z \le 0.0648)$

$1 - 0.5258 = 0.4742$

Comment:

Calculating the probability for $P(X \le 800)$ ***is somewhat redundant for this problem. Hopefully all students know that all valid SAT scores are 800 or lower, which makes the probability of this event 1. However, students may choose to calculate this probability using the formula and will find the probability to be*** $P(z \le 2.8426) = 0.9978$. ***This will change the final answer just slightly.***

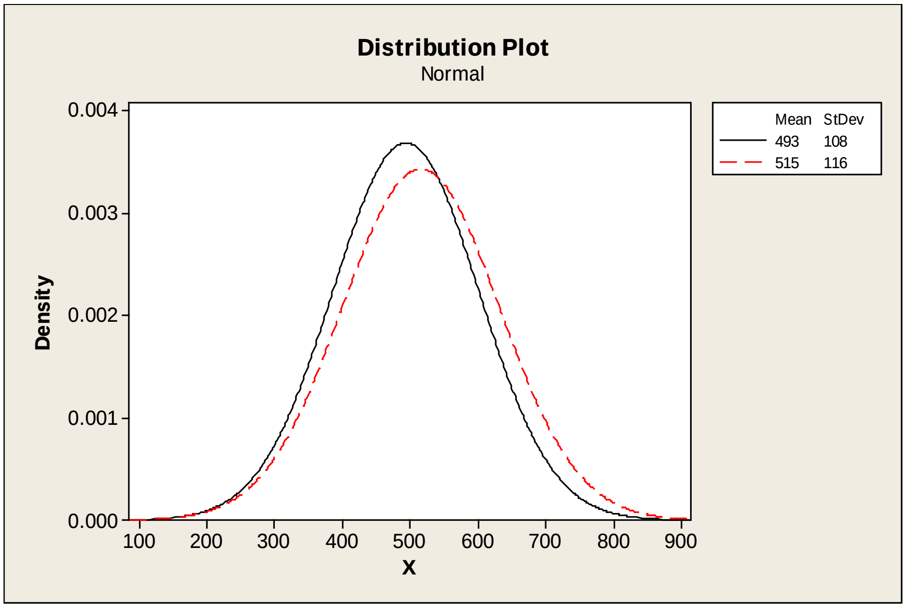
b) According the same College Board report, the population of **American** 2008 high school graduates had a mean Math SAT score of 515 with $\sigma = 116$. What is the probability that of a randomly selected senior from this population scoring between 500 and 800 on the Math portion of the SAT Reasoning Test?



$$P(500 \le X \le 800) = P(X \le 800) - P(X \le 500)$$

$P(X \le 800) - P(X \le 500)$		$P(X \le 800) - P(X \le 500)$
$1 - P(z \le -0.1293)$	or	$P(z \le 2.4569) - P(z \le -0.1293)$
$1 - 0.4486 = 0.5514$		$0.9930 - 0.4486 = 0.5444$

c) Compare the probabilities from parts a) and b). Explain the differences in the two probabilities.



Comment:

Overlapping the two distributions gives a visual representation of how close the probabilities (shaded regions) really are to each other.**

Solution:

For the population of all U.S. graduating seniors, the score of 500 is below (to the left) the mean. For the population of Georgia graduating seniors, the score of 493 is above (to the right) of the mean. This means that a greater percentage of the seniors in the US population should score above 500 than in the Georgia population.

Students could also mention that the variability is a little more in the US population than in the Georgia population as indicated by the somewhat higher standard deviation. Asking your students why this is true can potentially generate a nice conversation about population size and variability.

Table A: Standard Normal Probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Table A: Standard Normal Probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Let's Be Normal Learning Task

Name____Date _________________________

STANDARDS ADDRESSED IN THIS TASK:
MGSE9-12.S.ID. 4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets and tables to estimate areas under the normal curve

Standards for Mathematical Practice

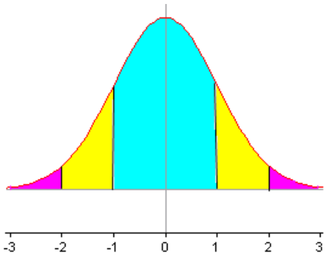
1. **Make sense of problems and persevere in solving them.**
2. **Reason abstractly and quantitatively.**
3. **Use appropriate tools strategically**.

Until this task, we have focused on distributions of discrete data. We will now direct our attention to continuous data. Where a **discrete variable** has a **finite number of possible values**, a continuous variable can assume all values in a given interval of values. Therefore, a **continuous random variab**le can assume an **infinite number of values**.

We will focus our attention specifically on continuous random variables with distributions that are approximately normal. Remember that normal distributions are symmetric, bell-shaped curves that follow the Empirical Rule.

The Empirical Rule for Normal Distributions states that

- 68% of the data values will fall within one standard deviation of the mean,
- 95% of the data values will fall within two standard deviations of the mean, and
- 99.7% of the data values will fall within three standard deviations of the mean.



In the last task, dot plots were used to explore this type of distribution and you spent time determining whether or not a given distribution was approximately normal. In this task, we will use probability histograms and approximate these histograms to a smooth curve that displays the shape of the distribution without the boxiness of the histogram. *We will also assume that all of the data we use is approximately normally distributed.*

Review:
1**)** The distribution of heights of adult American women is approximately normal with a mean of 65.5 inches ($\mu = 65.5$) and a standard deviation of 2.5 inches ($\sigma = 2.5$).

Draw a normal curve and label the mean and points one, two, and three standard deviations above and below the mean.

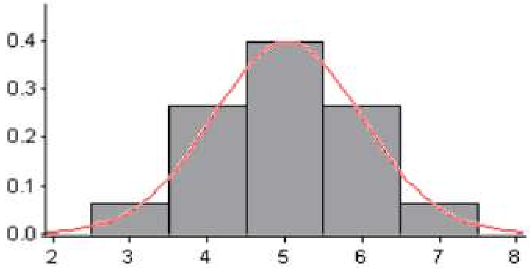
a) What percent of women are taller than 70.5 inches?

b) Between what heights do the middle 95% of women fall?

c) What percent of women are shorter than 63 inches?

d) A height of 68 inches corresponds to what percentile of adult female American heights?

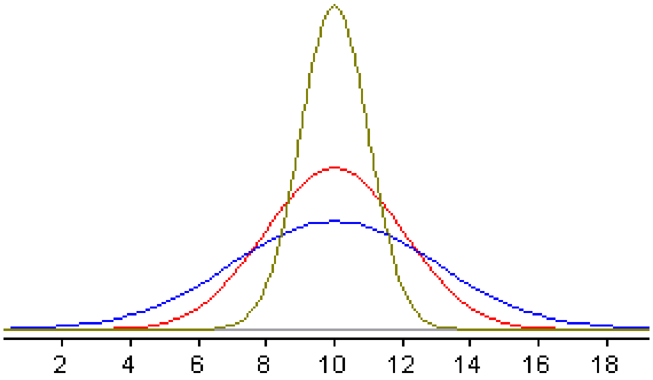
2. This is an example of a probability histogram of a continuous random variable with an approximate mean of five ($\mu = 5$) and standard deviation of one ($\sigma = 1$). A normal curve with the same mean and standard deviation has been overlaid on the histogram.



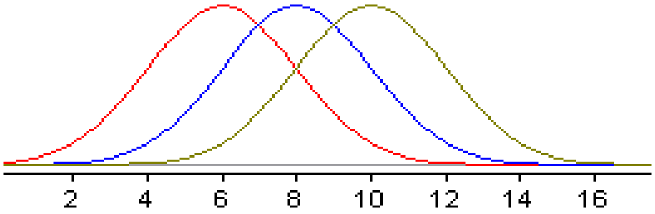
a) What do you notice about these two graphs?

3. First, you must realize that normally distributed data may have any value for its mean and standard deviation. Below are two graphs with three sets of normal curves.

a) In the first set, all three curves have the same mean, 10, but different standard deviations. Approximate the standard deviation of each of the three curves.



b) In this set, each curve has the same standard deviation. Determine the standard deviation and then each mean.



c) Based on these two examples, make a conjecture regarding the effect changing the mean has on a normal distribution. Make a conjecture regarding the effect changing the standard deviation has on a normal distribution.

4. The SAT Verbal Test in recent years follow approximately a normal distribution with a mean of 505 ($\mu = 505$) and a standard deviation of 110 ($\sigma = 110$). Joey took this test made a score of 600. The scores on the ACT English Test are approximately a normal distribution with a mean of 17 ($\mu = 17$) and a standard deviation of 2.5 ($\sigma = 2.5$). Sarah took this test and made a score of 18. Which student made the better score? How do you know?

The ***Standard Normal Distribution*** is a normal distribution with a mean of 0 and a standard deviation of 1. To more easily compute the probability of a particular observation given a normally distributed variable, we can transform any normal distribution to this standard normal distribution using the following formula:

$$z = \frac{X - \mu}{\sigma}$$

When you find this value for a given value, it is referred to as the z-score. The **z-score** is a standard score for a data value that indicates **the number of standard deviations** that the data value is away from its respective mean.

5. We can use the z-score to find the probability of many other events. Let's explore those now.

a) Suppose that the mean time a typical American teenager spends doing homework each week is 4.2 hours. Assume the standard deviation is 0.9 hour. Assuming the variable is normally distributed, find the percentage of American teenagers who spend less than 3.5 hours doing homework each week.

- First, sketch a normal curve for this situation and shade the probability in which you are interested.

- Next, find the z-score for X = 3.5.

- Now, use the table of standard normal probabilities to determine the probability of $P(X < 3.5)$.

b) The average height of adult American males is 69.2 inches. If the standard deviation is 3.1 inches, determine the probability that a randomly selected adult American male will be at most 71 inches tall. Assume a normal distribution.

6. We can also use z-scores to find the percentage or probability of events above a given observation.

a) The average on the most recent test Ms. Cox gave her French students was 73 with a standard deviation of 8.2. Assume the test scores were normally distributed. Determine the probability that a student in Ms. Cox's class scored a 90 or more on the test.

b) Women's heights are approximately normally distributed with $\mu = 65.5$ inches and $\sigma = 2.5$ inches. Determine the probability of a randomly selected woman having a height of at least 64 inches.

7. We can also determine the probability between two values of a random variable.

a) According to the College Board, **Georgia** seniors graduating in 2008 had a mean Math SAT score of 493 with a standard deviation of 108. Assuming the distribution of these scores is normal, find the probability of a member of the 2008 graduating class in Georgia scoring between 500 and 800 on the Math portion of the SAT Reasoning Test.

b) According the same College Board report, the population of **American** 2008 high school graduates had a mean Math SAT score of 515 with $\sigma = 116$. What is the probability that of a randomly selected senior from this population scoring between 500 and 800 on the Math portion of the SAT Reasoning Test?

c) Compare the probabilities from parts a) and b). Explain the differences in the two probabilities.

Table A: Standard Normal Probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Table A: Standard Normal Probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

And You Believed That?! Learning Task

Mathematical Goals

- Demonstrate understanding of the different kinds of sampling methods
- Discuss the appropriate way of choosing samples in context with limiting factors
- Recognize and understand bias in sampling methods

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population

MGSE9-12.S.IC. 3 Recognize the purposed of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

MGSE9-12.S.IC.6 **Evaluate reports based on data.** ***For example, determining quantitative or categorical data; collection methods; biases or flaws in data.***

Standards for Mathematical Practice

1. **Construct viable arguments and critique the reasoning of others**
2. **Model with mathematics**

Introduction

The purpose of this task is to have students understand the most appropriate way to gather and analyze data using different types of sampling methods. Students will also have the opportunity to look for bias in the sampling methods and understand the effect of bias in results.

Materials

- Pencil

A solid knowledge of statistical procedures will help you be an educated consumer of information. Every day, we are confronted by a news report citing a new relationship researchers have discovered, a claim being made in advertising, or other data-driven statements. Beginning with this task, we will examine the questions you should keep in mind when analyzing such claims.

1. Read the article below.

Facebook use linked to less textbook time

By Mary Beth Marklein, USA TODAY (April 13, 2009)

Does Facebook lead to lower grades? Or do college students with lower grades use Facebook more than their higher-achieving peers?

A study of 219 students at Ohio State University being presented at a conference this week doesn't answer those questions definitively. But it suggests a link between the social networking site and academic performance.

Students who said they used Facebook reported grade-point averages between 3.0 and 3.5; those who don't use it said they average 3.5 to 4.0. Also, Facebook users said they studied one to five hours a week, vs. non-users' 11 hours or more.

Ohio State doctoral student Aryn Karpinski, who conducted the research with graduate student Adam Duberstein, says the study is too narrow to conclude that Facebook and academics don't mix.

"It cannot be stated (that) Facebook use causes a student to study less or get lower grades," she says. "I'm just saying that they're related somehow, and we need to look into it further." Of the 68% of students who said they used Facebook, 65% accessed the site daily or multiple times daily.

Karpinski says 79% of Facebook users believe it has no impact on their academics; some say it helps them form study groups.

She says faculty ought to consider harnessing it as a learning tool. Yet a preliminary peek at a second survey suggests "a lot of faculty … didn't even know what Facebook is," she says.

1. What claim is being made? What is special about the claim? ***The claim being made is that college students that use Facebook with have lower grades or that college students with lower grades use Facebook.***

Bias in Research and Studies

The first question raised when evaluating the believability of a claim is whether or not the questions and procedures were designed in such a way as to eliminate bias. It is critical for statisticians and researchers to avoid leading questions and questions that are vague or contain confusing wording. For example, asking someone each of the following questions may illicit different responses even though all three questions address the same topic.

- "Is it really possible for a person to still believe that wearing a seat belt is not completely necessary?"
- "Is wearing a seat belt necessary for the complete safety of all passengers?"
- "Wearing a seat belt is currently required by state law. Do you agree with this law?"

2. How would you answer each of these questions? Did the wording of the questions influence your responses? ***Answers will vary. Students should understand how the wording of the questions can affect how a person answers the question.***

3. Refer to the article at the beginning of the task. What questions could the researchers have asked? Can you write two unbiased questions related to the article that researchers might have asked the subjects of the study? ***Answers will vary***

Another possible source of bias in studies is in the *sampling* technique. Remember that a sample is a subgroup of the population. It is important that researchers use unbiased samples. In order to have an unbiased sample, the sample must be selected at random. There are many types of random samples. The most common are the following.

- A *simple random sample*, in which every possible sample of the same size has the same chance of being selected. This can be accomplished by assigning every member of the population a distinct number and then using a random number generator or table to select members of the sample.
- A *systematic sample*, in which every member of population is assigned a number or put in order and then members of the sample are selected at set intervals, for example every tenth member is selected for the sample.
- A *stratified random sample*, in which members of the population are grouped by a specific characteristic and then members from each group, or strata, are selected using a simple random sample procedure.
- A *cluster sample*, in which the researcher identifies pre-existing groups, or clusters, within the population and then randomly selects a set numbers of these clusters as the sample. In this case, every member of the selected cluster is a part of the sample.

There are also sampling methods that create bias in the study. Types of these methods are the following.

- *convenience sampling* – choosing individuals who are easiest to reach (asking the first ten people who walk by)
- *voluntary response sampling*- consists of people who choose themselves by responding to a general appeal (asking radio listeners to call in to share responses or vote on a particular issue or asking subjects to return a survey by mail or email).

Both convenience sampling and voluntary response lack the critical element of randomization.

4. Determine whether each study below has a source of bias. If there is a source of bias, describe the bias and why this bias makes the sample unrepresentative. ***Sample responses are given***

a) A medical company uses sick patients to test their competitors' drugs for side effects
There is bias because the side effects could be masked by the patient's illness

b) A medical company uses healthy patients to test their competitors' drugs for side effects
There is bias because certain side effects will only occur in patients that are not healthy

c) A newspaper polls 9th grade students to measure if students are going into the Armed Forces after high school
There is bias because 9th graders are not representative of the entire population that will graduate from high school

d) The Department of Education conducts and online poll that asks "Do you have internet service at home?" ***There is bias because people who are online already have internet service***

e) A survey is mailed to voters in Augusta asking "Will you vote for the one cent sales tax increase in Augusta?" ***There is no bias in this study***

f) A survey is mailed to voters in Augusta who make more than $90,000 a year asking "Will you vote for the one cent sales tax increase in Augusta?" ***There is bias because the survey leaves out the part of the population that makes less than $90,000***

5. For each experiment, determine which sampling technique would be most appropriate. Then explain how you would obtain a sample and why the technique you chose was appropriate.

a) A company wants to decide who likes wheat bread more, men or women ***stratified random sampling – divide the population into two groups, men and women, and randomly select from each group***

b) A newspaper wants to determine which areas of town have the least number of subscriptions ***stratified random or cluster – divide the population into groups according to their geographic location (north, south, east, west) then randomly select from each group***

c) You want to estimate the number of people in your school who are vegans.
Simple random or systematic sampling – assign each student a number then randomly generate a sample of students

d) You want to determine whether 9th graders or 12th graders are more likely to be vegans.
Stratified random sampling – divide population into two groups, 9th and 12th graders, then randomly select from each group

e) The Department of Children Services wants to count the number of homeless children in a city. They only have enough counters to cover one-sixth of the city. ***cluster sampling- divide town into six group then select one of the groups***

f) A manufacturer wants to test the taste of their frozen vegetables as the bags of vegetables come out of a freezer. ***Simple random sample – randomly choose bags of vegetables from different locations in the freezer***

6. Referring to the article "Facebook use linked to less textbook time," what type of sampling technique do you believe the researchers may have used? Why?
Answers will vary. Students need to be able to justify their answers

7. Consider the student body of your high school to be the population for a study being conducted by the school newspaper. One of the newspaper students, Emma, is writing an article on study habits. She has carefully designed a survey of five questions. Is it reasonable to think that she can survey the entire student body? ***Answers will vary***

8. Emma has decided to survey 50 students. She is trying to decide which type of random sample will be the most appropriate and easiest for her to complete successfully. For the following, explain how she could select each type of sample from the students at your school. ***Answers will vary***

Simple:

Systematic:

Stratified:

Cluster:

If you were gathering information for this article, which of these samples would you use? Explain.

9. Another student on the staff thinks that Emma is making the assignment too difficult and suggested that she simply survey the students in her first period class. Would this be an appropriate sampling method? Explain. ***Answers will vary***

10. Emma's friend, Marcus, is on the yearbook staff and is currently involved in the staff's effort to design this year's cover. The group wants to create a cover design that depicts the "typical" student from the school. In order to determine the typical student, they have decided to design a fifteen question survey focusing on physical characteristics, classes and extracurricular activities. The problem is that the students are having difficulty writing unbiased questions to gather the data they need. They have agreed that seven questions should address physical characteristics, four should address classes taken, and four should address extracurricular activities. Pretend you are a member of this yearbook staff and write fifteen unbiased questions for this survey. ***Answers will vary***

And You Believed That?! Learning Task

Name ____
Date________________________

STANDARDS ADDRESSED IN THIS TASK:
MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population

MGSE9-12.S.IC. 3 Recognize the purposed of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

MGSE9-12.S.IC.6 Evaluate reports based on data. ***For example, determining quantitative or categorical data; collection methods; biases or flaws in data.***

Standards for Mathematical Practice
1. **Construct viable arguments and critique the reasoning of others**
2. **Model with mathematics**

A solid knowledge of statistical procedures will help you be an educated consumer of information. Every day, we are confronted by a news report citing a new relationship researchers have discovered, a claim being made in advertising, or other data-driven statement. Beginning with this task, we will examine the questions you should keep in mind when analyzing such claims.

1. Read the article below.

Facebook use linked to less textbook time

By Mary Beth Marklein, USA TODAY (April 13, 2009)

Does Facebook lead to lower grades? Or do college students with lower grades use Facebook more than their higher-achieving peers?

A study of 219 students at Ohio State University being presented at a conference this week doesn't answer those questions definitively. But it suggests a link between the social networking site and academic performance.

Students who said they used Facebook reported grade-point averages between 3.0 and 3.5; those who don't use it said they average 3.5 to 4.0. Also, Facebook users said they studied one to five hours a week, vs. non-users' 11 hours or more.

Ohio State doctoral student Aryn Karpinski, who conducted the research with graduate student Adam Duberstein, says the study is too narrow to conclude that Facebook and academics don't mix.

"It cannot be stated (that) Facebook use causes a student to study less or get lower grades," she says. "I'm just saying that they're related somehow, and we need to look into it further." Of the 68% of students who said they used Facebook, 65% accessed the site daily or multiple times daily.

Karpinski says 79% of Facebook users believe it has no impact on their academics; some say it helps them form study groups.

She says faculty ought to consider harnessing it as a learning tool. Yet a preliminary peek at a second survey suggests "a lot of faculty … didn't even know what Facebook is," she says.

1. What claim is being made? What is special about the claim?

Bias in Research and Studies

The first question raised when evaluating the believability of a claim is whether or not the questions and procedures were designed in such a way as to eliminate bias. It is critical for statisticians and researchers to avoid leading questions and questions that are vague or contain confusing wording. For example, asking someone each of the following questions may illicit different responses even though all three questions address the same topic.

- "Is it really possible for a person to still believe that wearing a seat belt is not completely necessary?"
- "Is wearing a seat belt necessary for the complete safety of all passengers?"
- "Wearing a seat belt is currently required by state law. Do you agree with this law?"

2. How would you answer each of these questions? Did the wording of the questions influence your responses?

3. Refer to the article at the beginning of the task. What questions could the researchers have asked? Can you write two unbiased questions related to the article that researchers might have asked the subjects of the study?

Another possible source of bias in studies is in the *sampling* technique. Remember that a sample is a subgroup of the population. It is important that researchers use unbiased samples. In order to have an unbiased sample, the sample must be selected at random. There are many types of random samples. The most common are the following.

- A *simple random sample*, in which every possible sample of the same size has the same chance of being selected. This can be accomplished by assigning every member of the population a distinct number and then using a random number generator or table to select members of the sample.
- A *systematic sample*, in which every member of population is assigned a number or put in order and then members of the sample are selected at set intervals, for example every tenth member is selected for the sample.
- A *stratified random sample*, in which members of the population are grouped by a specific characteristic and then members from each group, or strata, are selected using a simple random sample procedure.
- A *cluster sample*, in which the researcher identifies pre-existing groups, or clusters, within the population and then randomly selects a set numbers of these clusters as the sample. In this case, every member of the selected cluster is a part of the sample.

There are also sampling methods that create bias in the study. Types of these methods are the following.

- *convenience sampling* – choosing individuals who are easiest to reach (asking the first ten people who walk by)
- *voluntary response sampling*- consists of people who choose themselves by responding to a general appeal (asking radio listeners to call in to share responses or vote on a particular issue or asking subjects to return a survey by mail or email).

Both convenience sampling and voluntary response lack the critical element of randomization.

4. Determine whether each study below has a source of bias. If there is a source of bias, describe the bias and why this bias makes the sample unrepresentative.

a) A medical company uses sick patients to test their competitors’ drugs for side effects

b) A medical company uses healthy patients to test their competitors’ drugs for side effects

c) A newspaper polls 9th grade students to measure if students are going into the Armed Forces after high school

d) The Department of Education conducts and online poll that asks “Do you have internet service at home?”

e) A survey is mailed to voters in Augusta asking “Will you vote for the one cent sales tax increase in Augusta?”
f) A survey is mailed to voters in Augusta who make more than $150,000 a year asking “Will you vote for the one cent sales tax increase in Augusta?”

5. For each experiment, determine which sampling technique would be most appropriate. Then explain how you would obtain a sample and why the technique you chose was appropriate.

a) A company wants to decide who likes wheat bread more, men or women

b) A newspaper wants to determine which areas of town have the least number of subscriptions

c) You want to estimate the number of people in your school who are vegans.

d) You want to determine whether 9th graders or 12th graders are more likely to be vegans.

e) The Department of Children Services wants to count the number of homeless children in a city. They only have enough counters to cover one-sixth of the city.

f) A manufacturer wants to test the taste of their frozen vegetables as the bags of vegetables come out of a freezer.

6. Referring to the article “Facebook use linked to less textbook time,” what type of sampling technique do you believe the researchers may have used? Why?

7. Consider the student body of your high school to be the population for a study being conducted by the school newspaper. One of the newspaper students, Emma, is writing an article on study habits. She has carefully designed a survey of five questions. Is it reasonable to think that she can survey the entire student body?

8. Emma has decided to survey 50 students. She is trying to decide which type of random sample will be the most appropriate and easiest for her to complete successfully. For the following, explain how she could select each type of sample from the students at your school.

Simple:

Systematic:

Stratified:

Cluster:

If you were gathering information for this article, which of these samples would you use? Explain.

9. Another student on the staff thinks that Emma is making the assignment too difficult and suggested that she simply survey the students in her first period class. Would this be an appropriate sampling method? Explain.

10. Emma's friend, Marcus, is on the yearbook staff and is currently involved in the staff's effort to design this year's cover. The group wants to create a cover design that depicts the "typical" student from the school. In order to determine the typical student, they have decided to design a fifteen question survey focusing on physical characteristics, classes and extracurricular activities. The problem is that the students are having difficulty writing unbiased questions to gather the data they need. They have agreed that seven questions should address physical characteristics, four should address classes taken, and four should address extracurricular activities. Pretend you are a member of this yearbook staff and write fifteen unbiased questions for this survey.

“Cost of Quality” in the Pulp & Paper Industry

In this task, students will analyze product quality data in the production of wood pulp (a major industry in the state of Georgia). Utilizing data analysis tools, the students will understand the basics of an industrial product quality measurement system and the impact of product quality on a business’ profitability.

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, ~~mean absolute deviation~~, standard deviation) of two or more different data sets.

MGSE9-12.S.ID.4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

MGSE9-12.S.IC.1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

MGSE9-12.S.IC.6 Evaluate reports based on data. ***For example, determining quantitative or categorical data; collection methods; biases or flaws in data.***

STANDARDS FOR MATHEMATICAL PRACTICE

1. Make sense of problems and persevere in solving them.
4. Model with mathematics.
5. Use appropriate tools strategically.

BACKGROUND KNOWLEDGE In order for students to be successful, the following skills and concepts need to be maintained:

- Students will need to be familiar with histograms, normal distributionsand with statistical measures of central tendency (mean) and dispersal (standard deviation)
- Students must be able to read and create tabular data to be analyzed.
- Students must have a rudimentary understanding of microeconomics.

ESSENTIAL QUESTIONS

- How do I analyze product quality data to determine whether a particular batch meets a customer’s specifications? What data measures, tabulating and graphical representations will assist me?
- How do I apply my analysis of product quality data to determine impacts on the viability/profitability of the business?
- How can I determine the impact of product improvement initiatives in the business?

MATERIALS: Scientific calculator

GROUPING Individual or Partners

TASK DESCRIPTION, DEVELOPMENT AND DISCUSSION

Georgia manufacturers of wood pulp products utilize statistical data to manage their product quality process. Several product qualities measurements are taken on each batch (***jumbo roll***) of product produced.

These quality measurements include:

- **Intrinsic Viscosity (IV)** – A measure of the relative degradation of cellulose fiber during the pulping /bleaching process. (Unit of Measure = mPa.s)
- **Brightness** - A measure of how much light is reflected by the pulp - an indicator of cleanliness and purity. (Unit of Measure = % ISO)
- **Caliper** - The thickness of a pulp sheet. (Unit of Measure = thousandths of an inch or "mils")
- **Dirt Levels** – The amount of dirt contamination allowed. (Unit of Measure = mm^2/kg)
- **Burst Index** - The resistance of a pulp sheet to rupture. (Unit of Measure = $kPa.m^2/g$)

The manufacturers' customers require that the quality properties of the wood pulp they purchase fall within specific ranges. Each customer has a different sales price for pulps that meet their quality specifications. For example:

Customer	Quality	IV	Brightness	Dirt Levels	Burst Index	Caliper	Sales Price ($/ton)
A	High	21-25	> 90	≤ 1.2	2.0 – 4.0	40 - 50	$1,200
B	Good	19-28	> 86	< 2.0	1.5 – 7.5	40 - 50	$900
Off Grade	Poor	None	> 70	< 5.0	None	> 30	$400

Over the course of a production run, the following samples were taken from 10 batches (jumbo rolls). Circle the ones that do **NOT** meet Customer A's specification.

Jumbo Roll #	IV	Brightness	Dirt Levels	Burst Index	Caliper
10010	22.6	88.5	1.41	3.1	45.0
10011	19.8	91.0	1.20	2.8	51.0
10012	24.5	92.5	1.04	3.9	48.3
10013	26.1	89.3	0.93	3.6	47.6
10014	22.3	94.0	0.95	2.9	45.3
10015	15.3	85.7	2.02	1.9	48.2
10016	21.0	90.9	1.08	3.7	47.9
10017	21.1	91.0	1.13	2.8	48.3
10018	19.9	92.0	1.02	3.4	49.4
10019	14.3	85.9	1.48	4.2	47.9

The goal of the manufacturer was to sell all of this production to Customer A, as they have the highest sales price. Please list which jumbo roll numbers may be sold to Customer A, which ones may be sold to Customer B and which ones will have to be sold as “Off Grade” (pulp meets few specs and must be sold at a loss to low grade users).

Customer	**Jumbo Roll #**			
A	*10012*	*10014*	*10016*	*10017*
B	*10010*	*10013*	*10018*	
Off Grade	*10011*	*10015*	*10019*	

If the cost of production is $850/ton and all batches weigh 10 tons, calculate the profit/loss for sales to each customer, and the opportunity cost of low quality product (not being able to sell lower grade product for the highest price, i.e. to Customer A).

Opportunity Cost = (Highest price per ton – Actual Price per ton)*# Tons

Opportunity Cost of product sold to Customer B = **$300**/ton ($1200-$900)* # Tons

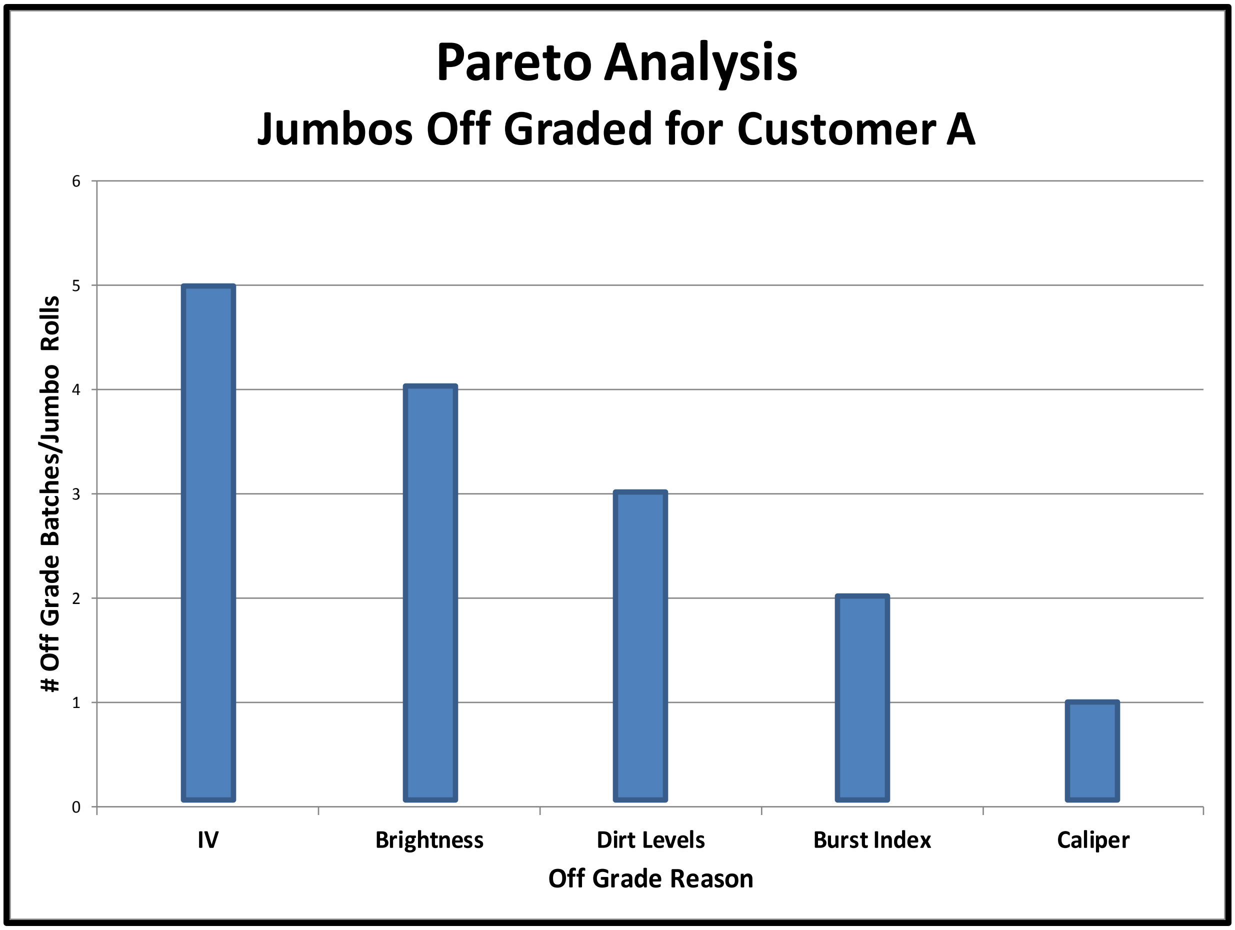
Opportunity Cost of product sold for Off Grade = **$800**/ton ($1200-$400)* # Tons

Customer	**# Batches**	**# Tons**	**Price/Ton**	**Cost/Ton**	**Profit/Loss per Ton**	**Total Profit/(Loss$^+$)**	**Opportunity Cost**
A	*4*	*40*	$1200	$850	$350	*$14,000*	*$0*
B	*3*	*30*	$900	$850	$50	*$1,500*	*$9,000*
Off Grade	*3*	*30*	$400	$850	-$450	*($13,500)*	*$24,000*
					TOTAL FINANCIAL IMPACT	*$2,000*	*$33,000*

$^+$ A loss is indicated by the dollar figure being inside parenthesis.

The manufacturer is unhappy with the profit that it did not realize due to the opportunity costs of low grade product. They wish to perform a data analysis of the reasons for the low grade pulp. Determine how many off grade measurements were recorded for each of these reasons and then graph the results. Use the list of jumbo rolls above, with the Offgrade specifications circled.

# Jumbos Off Grade for Customer A	**IV**	**Brightness**	**Dirt Levels**	**Burst Index**	**Caliper**
	5	*4*	*3*	*2*	*1*



* Pareto analysis (sometimes referred to as the 80/20 rule and as ABC analysis) is a method of classifying items, events, or activities according to magnitude of their occurrence. Pareto analysis is used to prioritize the most important items or factors.

Intrinsic Viscosity Metric

The manufacturer determines that the Intrinsic Viscosity (IV) is the biggest area of opportunity for improving quality. They record the **IV measurements** from 50 batches/Jumbo Rolls:

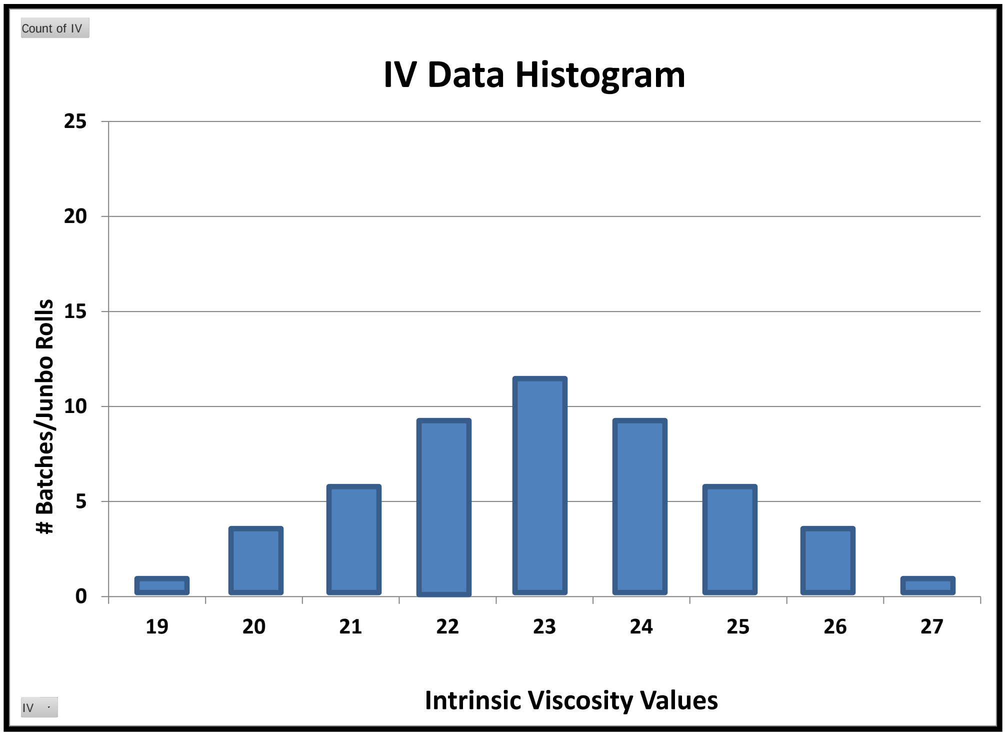
Jumbo #	IV	Jumbo #	IV	Jumbo #	IV	Jumbo #	IV
12101	24	12114	22	12127	21	12139	23
12102	22	12115	23	12128	20	12140	25
12103	23	12116	25	12129	22	12141	22
12104	27	12117	24	12130	23	12142	23
12105	24	12118	23	12131	24	12143	21
12106	25	12119	23	12132	24	12144	23
12107	26	12120	25	12133	25	12145	21
12108	21	12121	26	12134	21	12146	24
12109	20	12122	25	12135	20	12147	24
12110	19	12123	22	12136	22	12148	26
12111	22	12124	23	12137	23	12149	24
12112	22	12125	23	12138	24	12150	21
12113	23	12126	22	**Whole #'s used to simplify calculations.**			

Create a probability distribution chart and calculate the mean and sample standard deviation of the data.

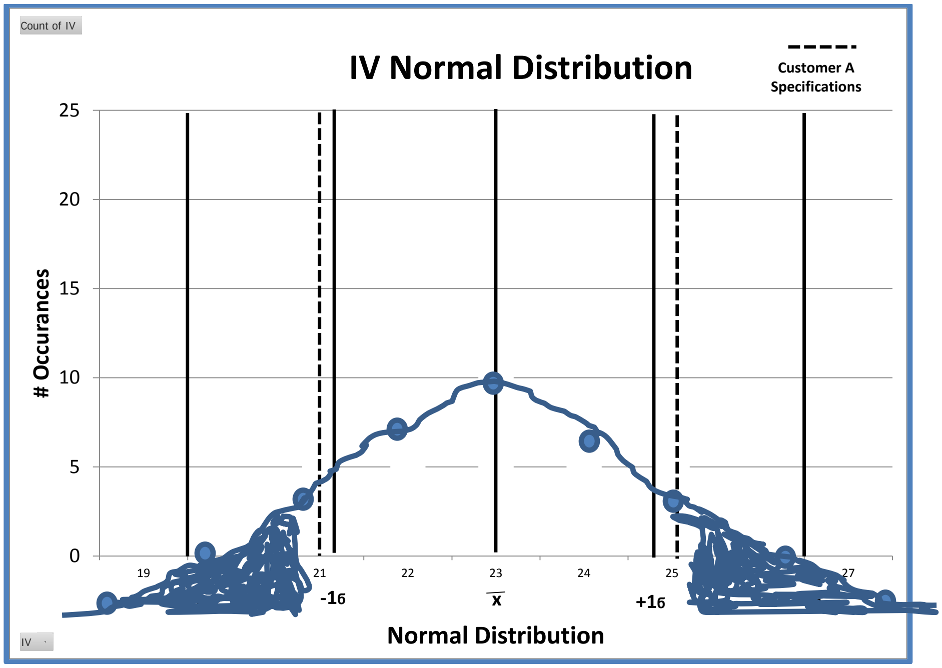
IV Metric	# Batches	Probability	Mean
19	1	0.02	23
20	2	0.06	
21	3	0.12	
22	9	0.18	
23	12	0.24	**Std Dev**
24	9	0.18	**1.76**
25	6	0.12	
26	3	0.06	
27	1	0.02	
Grand Total	**50**	**1.00**	

Sample Standard Deviation (S)

Create a Probability Histogram of the Batch IV's.



Convert the histogram to a Normal Distribution (estimate the bell curve/shade in the poor quality area).



Compute the "**Cost of Poor Quality**" by determining the **Opportunity Cost** of not being able to sell the pulp that does **not** meet Customer A's specifications for the highest price (having to sell to Customer B instead).

Customer A Spec	# Batches Below Spec	# Batches Above Spec	Total Batches "out of spec"	Total Tons (10 tons/batch)	Opportunity cost/ton	Cost Of Poor Quality
21 – 25	*4*	*4*	*8*	*80*	$300	*$24,000*

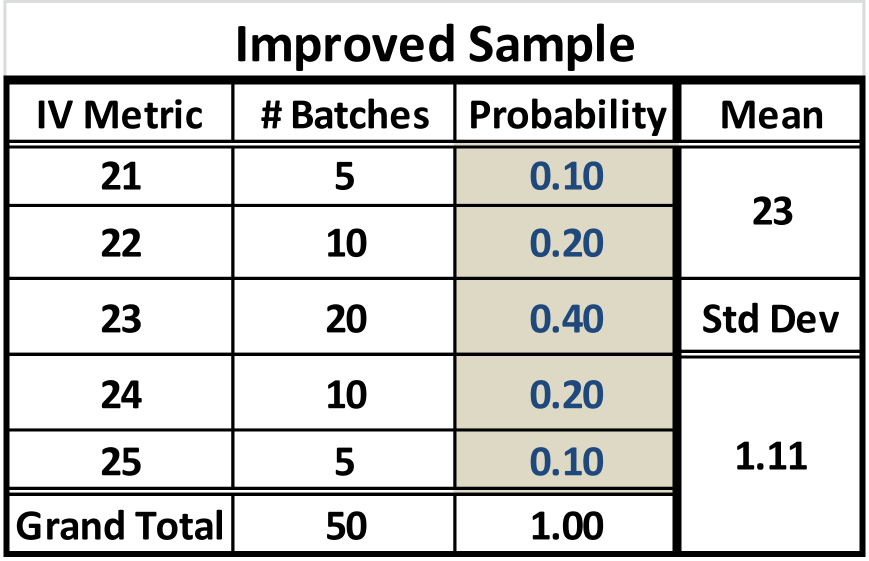
If the manufacturer produces 100,000 tons, of this grade, every year, what is the total annual "Cost of Poor Quality".

% of Bad Pulp Produced*	# Tons Produced Annually	Cost of Poor Quality/Ton**	Annual Cost of Poor Quality
16%	100,000	$1,200 - $900 = $300	*$4,800,000*

***Probability of pulp being out of spec.** ****Assumes product meets Customer B Specs – Cost even higher if "off-graded".**

(2% + 6% + 6% + 2%)

With 6 months of work and large capital expenditures ($9.6 million), the manufacturer improves the manufacturing process and is now able to more closely control the wood pulp's Intrinsic Viscosity. They produce 50 batches/jumbo rolls with the IV parameters below.

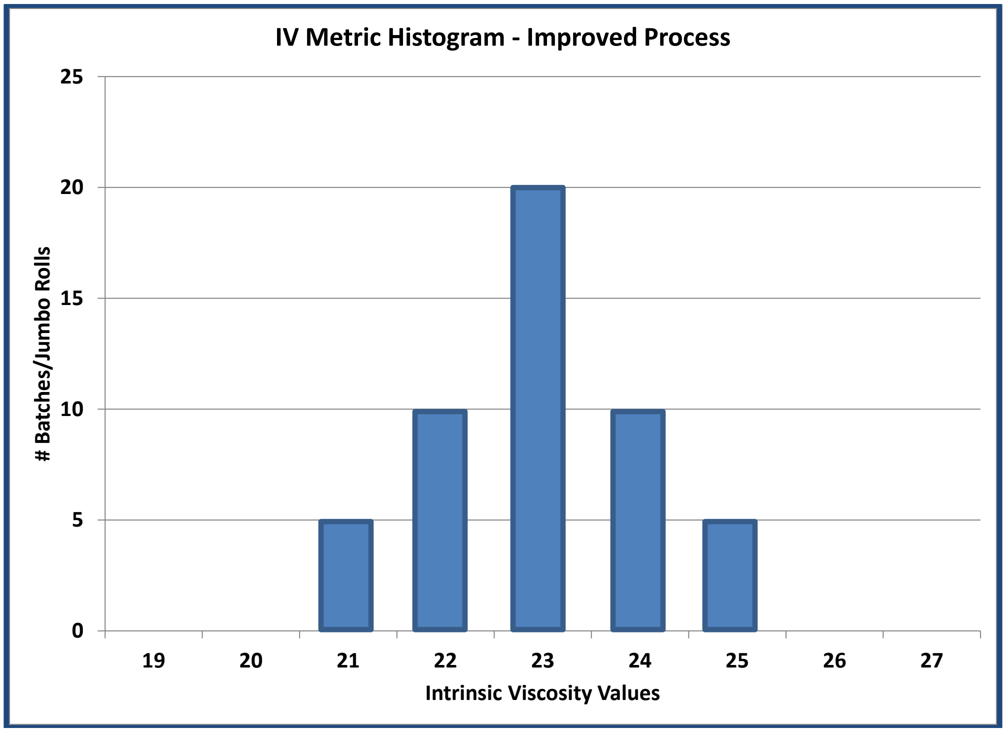


Improved Sample

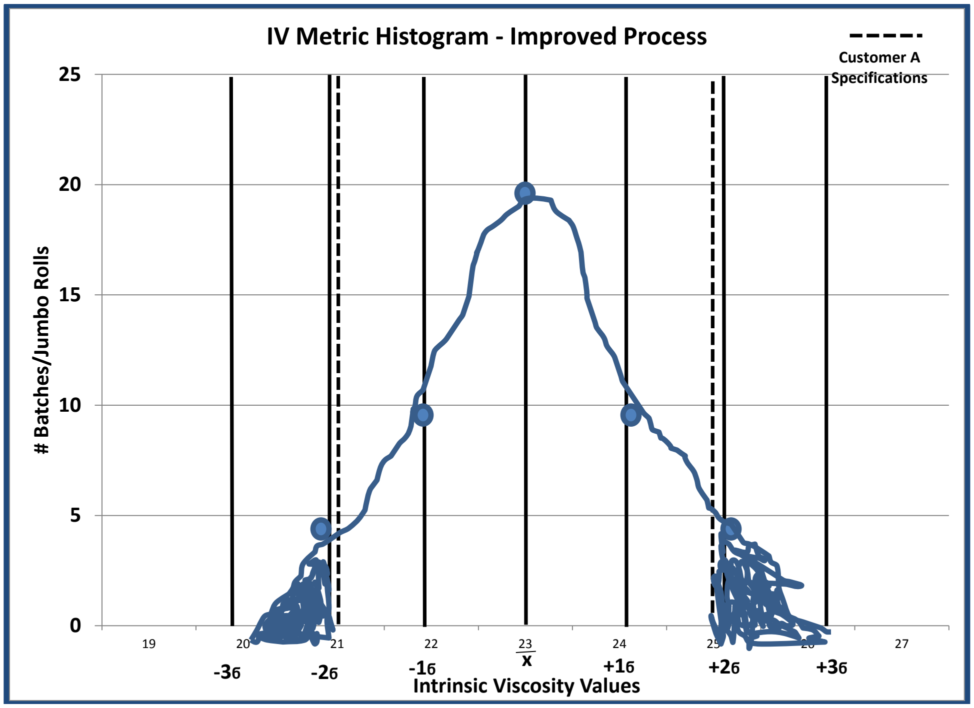
IV Metric	# Batches	Probability	Mean
21	5	0.10	23
22	10	0.20	
23	20	0.40	Std Dev
24	10	0.20	1.11
25	5	0.10	
Grand Total	50	1.00	

Sample Standard Deviation (S)

Create a Probability Histogram of the Batch IV's.



Convert the histogram to a Normal Distribution (estimate the bell curve).



FORMATIVE ASSESSMENT

- How do the normal distribution curves of the original pulp samples and the improved pulp samples compare?

The normal distribution curve for the Improved Sample is higher and narrower. There is less area outside (lower and higher than) Customer A's specifications, i.e. less "poor quality" product.

- Why would the manufacturer prefer to make the improved pulp?

The Improved Pulp is preferable because the manufacturer can sell it for a higher price, realizing a greater profit → approximately $4.8 million per year.

- What is the "payback" period for the manufacturer? i.e. how many years does it take for the manufacturer to recoup its capital investment ($9.6 million)?

Formula – Capital Investment/Annual Cost of "Poor Quality" avoided.

$$\frac{Capital\ Investment}{Annual\ Cost\ of\ Poor\ Quality} = \frac{\$9.6\ \cancel{million}}{\$4.8\ \cancel{million}/year}$$

$$9.6 \div 4.8 = \boxed{2\text{ years}}$$

- Do you think this is a good investment to make?

Yes. The cost of the process improvement is return in two years and all of the improvement contributes to the manufacturer's profit after that.

“Cost of Quality” in the Pulp & Paper Industry

In this task, students will analyze product quality data in the production of wood pulp (a major industry in the state of Georgia). Utilizing data analysis tools, the students will understand the basics of an industrial product quality measurement system and the impact of product quality on a business’ profitability.

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, ~~mean absolute deviation~~, standard deviation) of two or more different data sets.

MGSE9-12.S.ID.4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

MGSE9-12.S.IC.1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

MGSE9-12.S.IC.6 Evaluate reports based on data. *For example, determining quantitative or categorical data; collection methods; biases or flaws in data.*

STANDARDS FOR MATHEMATICAL PRACTICE

1. Make sense of problems and persevere in solving them.
4. Model with mathematics.
5. Use appropriate tools strategically.

TASK DESCRIPTION, DEVELOPMENT AND DISCUSSION

This scenario from a Georgia manufacturing process will give students insight into the use of data analysis in a real world situation. The use of product quality data will demonstrate the importance of data analysis techniques to determine product profitability and the “cost of poor quality”.

Georgia manufacturers of wood pulp products utilize statistical data to manage their product quality process. Several product qualities measurements are taken on each batch (***jumbo roll***) of product produced.

These quality measurements include:

- **Intrinsic Viscosity (IV)** – A measure of the relative degradation of cellulose fiber during the pulping /bleaching process. (Unit of Measure = mPa.s)
- **Brightness** - A measure of how much light is reflected by the pulp - an indicator of cleanliness and purity. (Unit of Measure = % ISO)

- **Caliper** - The thickness of a pulp sheet. (Unit of Measure = thousandths of an inch or "mils")
- **Dirt Levels** – The amount of dirt contamination allowed. (Unit of Measure = mm^2/kg)
- **Burst Index** - The resistance of a pulp sheet to rupture. (Unit of Measure = $kPa.m^2/g$)

The manufacturers' customers require that the quality properties of the wood pulp they purchase fall within specific ranges. Each customer has a different sales price for pulps that meet their quality specifications. For example:

Customer	Quality	IV	Brightness	Dirt Levels	Burst Index	Caliper	Sales Price ($/ton)
A	**High**	**21-25**	**> 90**	**≤ 1.2**	**2.0 – 4.0**	**40 - 50**	**$1,200**
B	**Good**	**19-28**	**> 86**	**< 2.0**	**1.5 – 7.5**	**40 - 50**	**$900**
Off Grade	**Poor**	**None**	**> 70**	**< 5.0**	**None**	**> 30**	**$400**

Over the course of a production run, the following samples were taken from 10 batches (jumbo rolls). Circle the ones that do **NOT** meet Customer A's specification.

Jumbo Roll #	IV	Brightness	Dirt Levels	Burst Index	Caliper
10010	22.6	88.5	1.41	3.1	45.0
10011	19.8	91.0	1.20	2.8	51.0
10012	24.5	92.5	1.04	3.9	48.3
10013	26.1	89.3	0.93	3.6	47.6
10014	22.3	94.0	0.95	2.9	45.3
10015	15.3	85.7	2.02	1.9	48.2
10016	21.0	90.9	1.08	3.7	47.9
10017	21.1	91.0	1.13	2.8	48.3
10018	19.9	92.0	1.02	3.4	49.4
10019	14.3	85.9	1.48	4.2	47.9

The goal of the manufacturer was to sell all of this production to Customer A, as they have the highest sales price. Please list which jumbo roll numbers may be sold to Customer A, which ones may be sold to Customer B and which ones will have to be sold as "Off Grade" (pulp meets few specs and must be sold at a loss to low grade users).

Customer	Jumbo Roll #			
A				
B				
Off Grade				

If the cost of production is $850/ton and all batches weigh 10 tons, calculate the profit/loss for sales to each customer, and the opportunity cost of low quality product (not being able to sell lower grade product for the highest price, i.e. to Customer A).

Opportunity Cost = (Highest price per ton – Actual Price per ton)*# Tons

Opportunity Cost of product sold to Customer B = **$300**/ton ($1200-$900)* # Tons

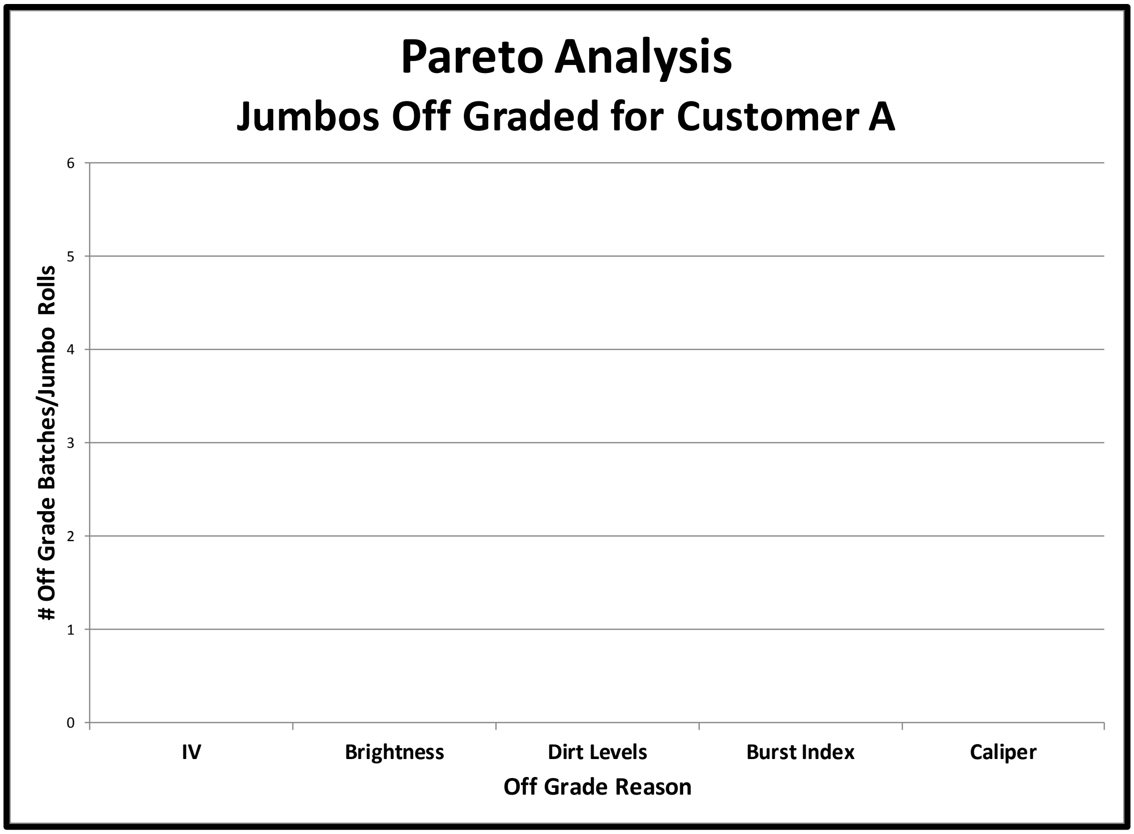
Opportunity Cost of product sold for Off Grade = **$800**/ton ($1200-$400)* # Tons

Customer	# Batches	# Tons	Price/Ton	Cost/Ton	Profit/Loss per Ton	Total Profit/(Loss$^+$)	Opportunity Cost
A			$1200	$850	$350	**$**	**$ ---------**
B			$900	$850	$50	**$**	**$**
Off Grade			$400	$850	-$450	**$**	**$**
					TOTAL FINANCIAL IMPACT	**$**	**$**

$^+$ A loss is indicated by the dollar figure being inside parenthesis.

The manufacturer is unhappy with the profit that they did not realize due to the opportunity costs of low quality product. They wish to perform a data analysis of the reasons for the low grade pulp. Determine how many off grade measurements were recorded for each of these reasons and then graph the results, in a Pareto* analysis histogram. Use the list of jumbo rolls above, with the Offgrade specifications circled.

# Jumbos Off Grade for Customer A	**IV**	**Brightness**	**Dirt Levels**	**Burst Index**	**Caliper**



* Pareto analysis (sometimes referred to as the 80/20 rule and as ABC analysis) is a method of classifying items, events, or activities according to magnitude of their occurrence. Pareto analysis is used to prioritize the most important items or factors.

Intrinsic Viscosity Metric

The manufacturer determines that the Intrinsic Viscosity (IV) is the biggest area of opportunity for improving quality. They record the **IV measurements** from 50 batches/Jumbo Rolls:

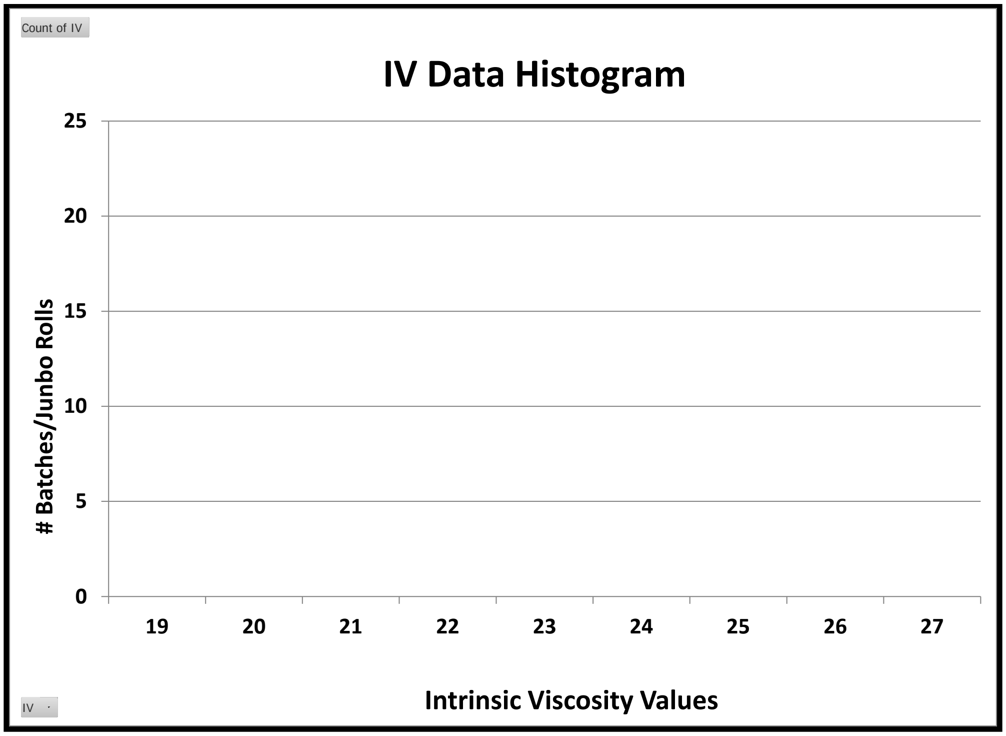
Jumbo #	IV	Jumbo #	IV	Jumbo #	IV	Jumbo #	IV
12101	24	12114	22	12127	21	12139	23
12102	22	12115	23	12128	20	12140	25
12103	23	12116	25	12129	22	12141	22
12104	27	12117	24	12130	23	12142	23
12105	24	12118	23	12131	24	12143	21
12106	25	12119	23	12132	24	12144	23
12107	26	12120	25	12133	25	12145	21
12108	21	12121	26	12134	21	12146	24
12109	20	12122	25	12135	20	12147	24
12110	19	12123	22	12136	22	12148	26
12111	22	12124	23	12137	23	12149	24
12112	22	12125	23	12138	24	12150	21
12113	23	12126	22	**Whole #'s used to simplify calculations.**			

Create a probability distribution chart and calculate the <u>mean</u> and sample <u>standard deviation</u> of the data.

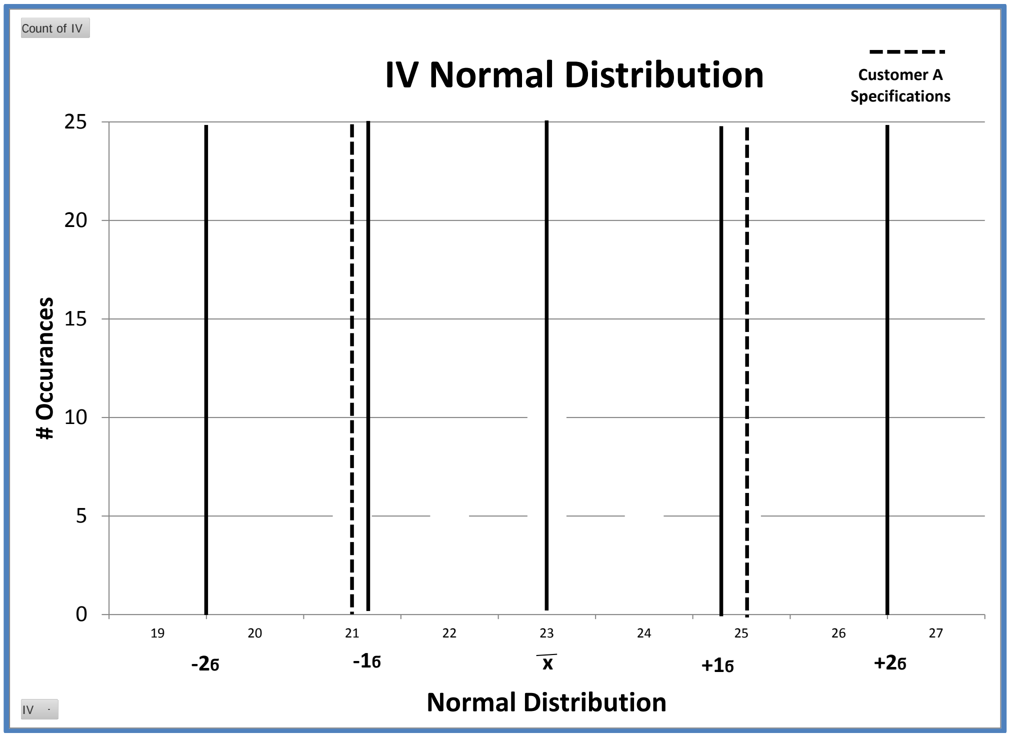
IV Metric	# Batches	Probability	Mean
19			
20			
21			
22			
23			Std Dev
24			
25			
26			
27			
Grand Total	50	1.00	

Sample Standard Deviation (S)

Create a Probability Histogram of the Batch IV's.



Convert the histogram to a Normal Distribution (estimate the bell curve/shade in the poor quality area).



Compute the "**Cost of Poor Quality**" by determining the **Opportunity Cost** of not being able to sell the pulp that does **not** meet Customer A's specifications for the highest price (having to sell to Customer B instead).

Customer A Spec	# Batches Below Spec	# Batches Above Spec	Total Batches "out of spec"	Total Tons (10 tons/batch)	Opportunity cost/ton	Cost Of Poor Quality
21 – 25					$300	$

If the manufacturer produces 100,000 tons, of this grade, every year, what is the total annual "Cost of Poor Quality".

% of Bad Pulp Produced*	# Tons Produced Annually	Cost of Poor Quality/Ton**	Annual Cost of Poor Quality
%	100,000	$1,200 - $900 = $300	$

*Probability of pulp being out of spec.

**Assumes product meets Customer B Specs – Cost even higher if "off-graded".

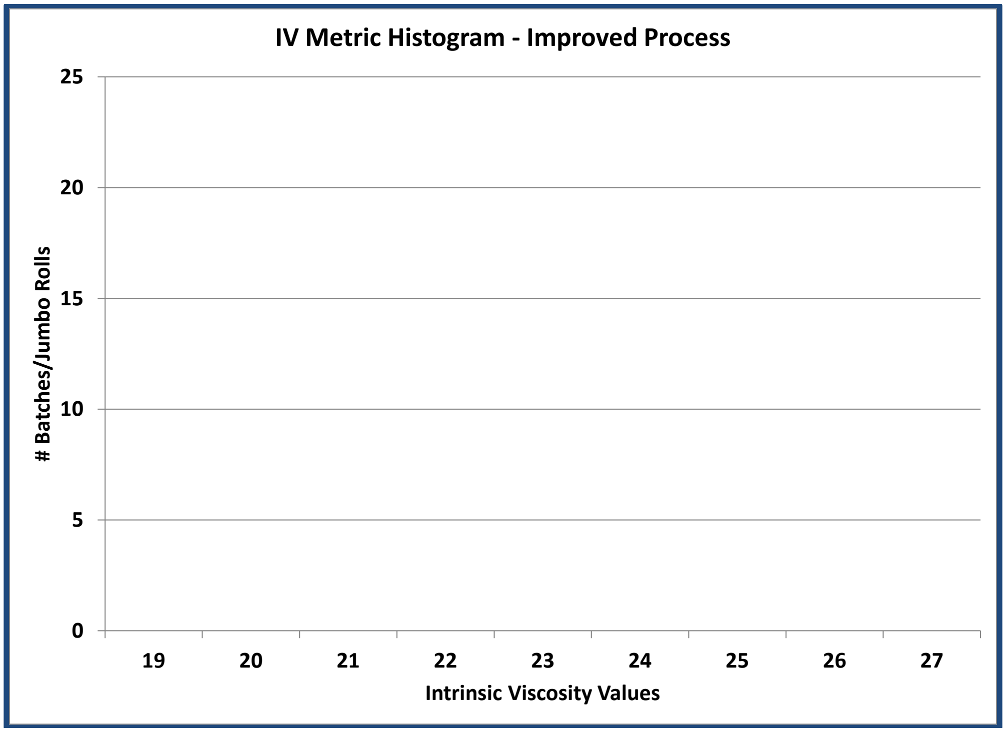
With 6 months of work and large capital expenditures ($9.6 million), the manufacturer improves the manufacturing process and is now able to more closely control the wood pulp's Intrinsic Viscosity. They produce 50 batches/jumbo rolls with the IV parameters below.

Improved Sample

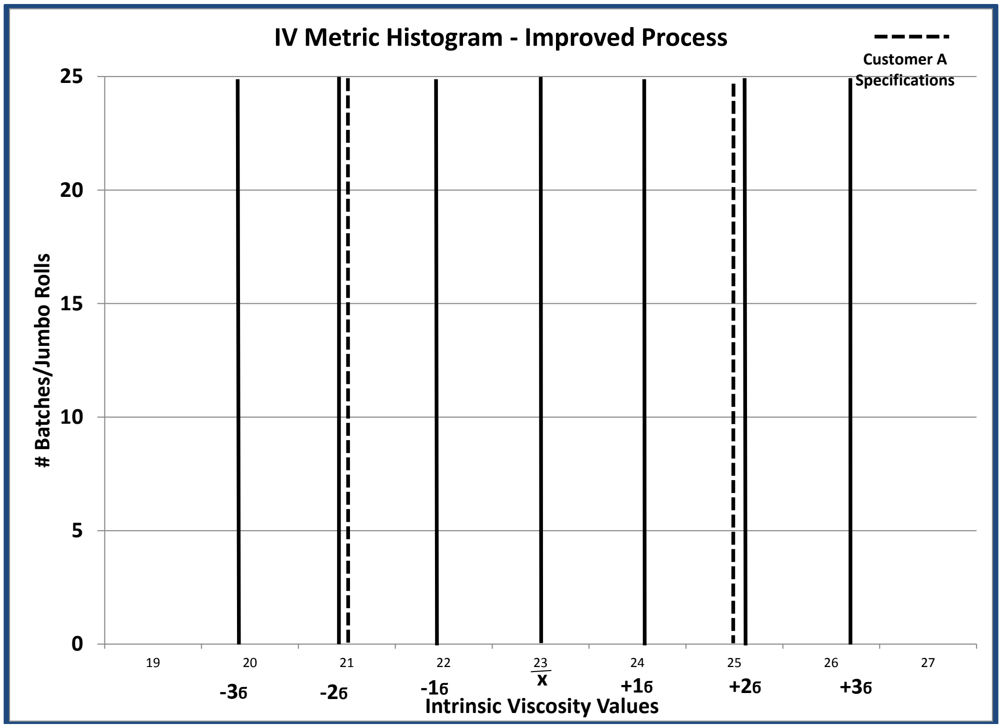
IV Metric	# Batches	Probability	Mean
21	5		23
22	10		
23	20		Std Dev
24	10		1.11
25	5		
Grand Total	50	1.00	

Sample Standard Deviation (S)

Create a Probability Histogram of the Batch IV's.



Convert the histogram to a Normal Distribution (estimate the bell curve).



FORMATIVE ASSESSMENT

- How do the normal distribution curves of the original pulp samples and the improved pulp samples compare?

- Why would the manufacturer prefer to make the improved pulp?

- What is the “payback” period for the manufacturer? i.e. how many years does it take for the manufacturer to recoup its capital investment ($9.6 million)?

Formula – Capital Investment/Annual Cost of “Poor Quality” avoided.

- Do you think this is a good investment to make?

How Tall are Our Students?

In this task, students will gather sample data, calculate statistical parameters and draw inferences about populations. Students will learn to take samples, understand different types of sampling (random, convenience, self-selected, and systematic) and sampling bias, extrapolate from a sample to a population and use statistical methodologies to make predictions regarding the full population.

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID.2 **Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, ~~mean absolute deviation~~, standard deviation) of two or more different data sets.**

MGSE9-12.S.ID.4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

MGSE9-12.S.IC.1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

STANDARDS FOR MATHEMATICAL PRACTICE

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.

BACKGROUND KNOWLEDGE

In order for students to be successful, the following skills and concepts need to be maintained:

- Students will need to be familiar with histograms, normal distributions and with statistical measures of central tendency (mean) and dispersal (standard deviation)
- Students must understand the use of z-scores to make predictions about a population from smaller samples.
- Students must understand sampling techniques and sampling types, including random, convenience, self-selected, and systematic and sampling bias
- Students must be able to read and create tabular data to be analyzed.

ESSENTIAL QUESTIONS

- How do I conduct a sampling operation to gather data about a population?
- How do I develop statistical parameters of the data that was gathered?
- How do I make predictions about the total population, based on the sample?

MATERIALS:

- Scientific calculator.
- Tape Measure or other measurement device ("height pipes" can be created from thin wall $1\frac{1}{2}$ inch PVC pipe).

GROUPING

Teams of two to four students.

TASK DESCRIPTION, DEVELOPMENT AND DISCUSSION

I. Purpose

To give students the experience of gathering sample data, calculating statistical parameters and drawing inferences about populations.

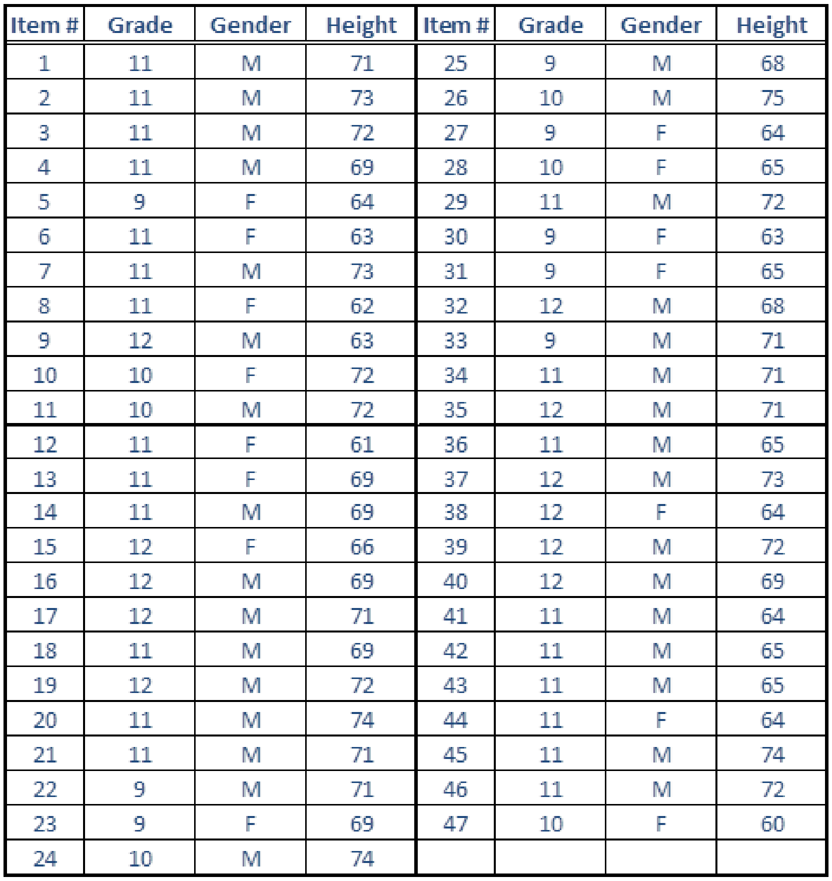
II. Data Sampling (Height Measurement)

a. Each team will utilize a measuring tape or "height pipe" and the "Height Sampling Data" worksheet. They will measure other students, recording Grade, Gender and Height (to the nearest inch), and record their observations on the worksheet. (This works best when a wide range of students are available, e.g. during lunch)

b. Teams should NOT duplicate students measured, i.e. ask the subject if they have already been measured by another team. If they answer yes, then find another student to measure.

Answers will vary depending on the sample data. This answer key is representative of how the response should be structured.

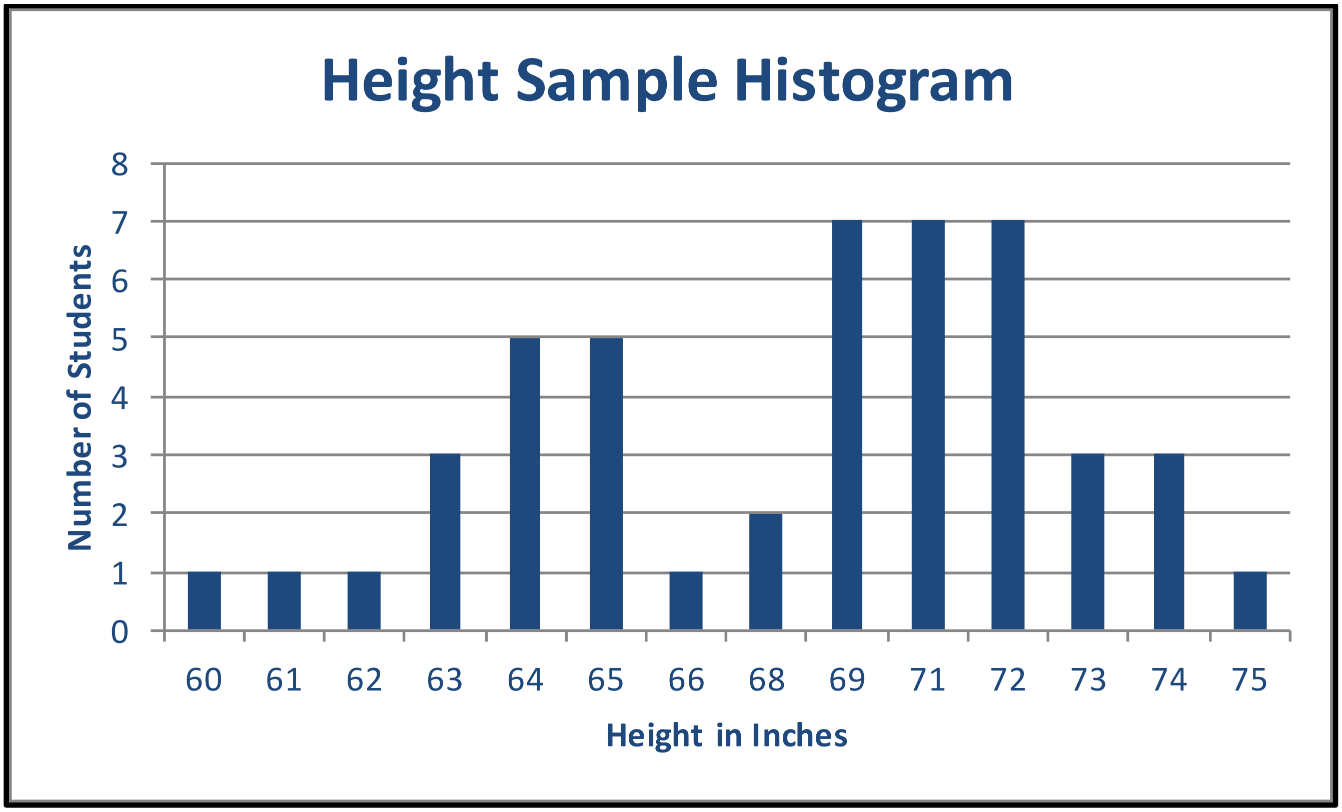
II. Data Sampling (Height Measurement)



Item #	Grade	Gender	Height	Item #	Grade	Gender	Height
1	11	M	71	25	9	M	68
2	11	M	73	26	10	M	75
3	11	M	72	27	9	F	64
4	11	M	69	28	10	F	65
5	9	F	64	29	11	M	72
6	11	F	63	30	9	F	63
7	11	M	73	31	9	F	65
8	11	F	62	32	12	M	68
9	12	M	63	33	9	M	71
10	10	F	72	34	11	M	71
11	10	M	72	35	12	M	71
12	11	F	61	36	11	M	65
13	11	F	69	37	12	M	73
14	11	M	69	38	12	F	64
15	12	F	66	39	12	M	72
16	12	M	69	40	12	M	69
17	12	M	71	41	11	M	64
18	11	M	69	42	11	M	65
19	12	M	72	43	11	M	65
20	11	M	74	44	11	F	64
21	11	M	71	45	11	M	74
22	9	M	71	46	11	M	72
23	9	F	69	47	10	F	60
24	10	M	74				

III. Develop Statistical Parameters of the Sample

a. Draw a Histogram Chart for your sample data.

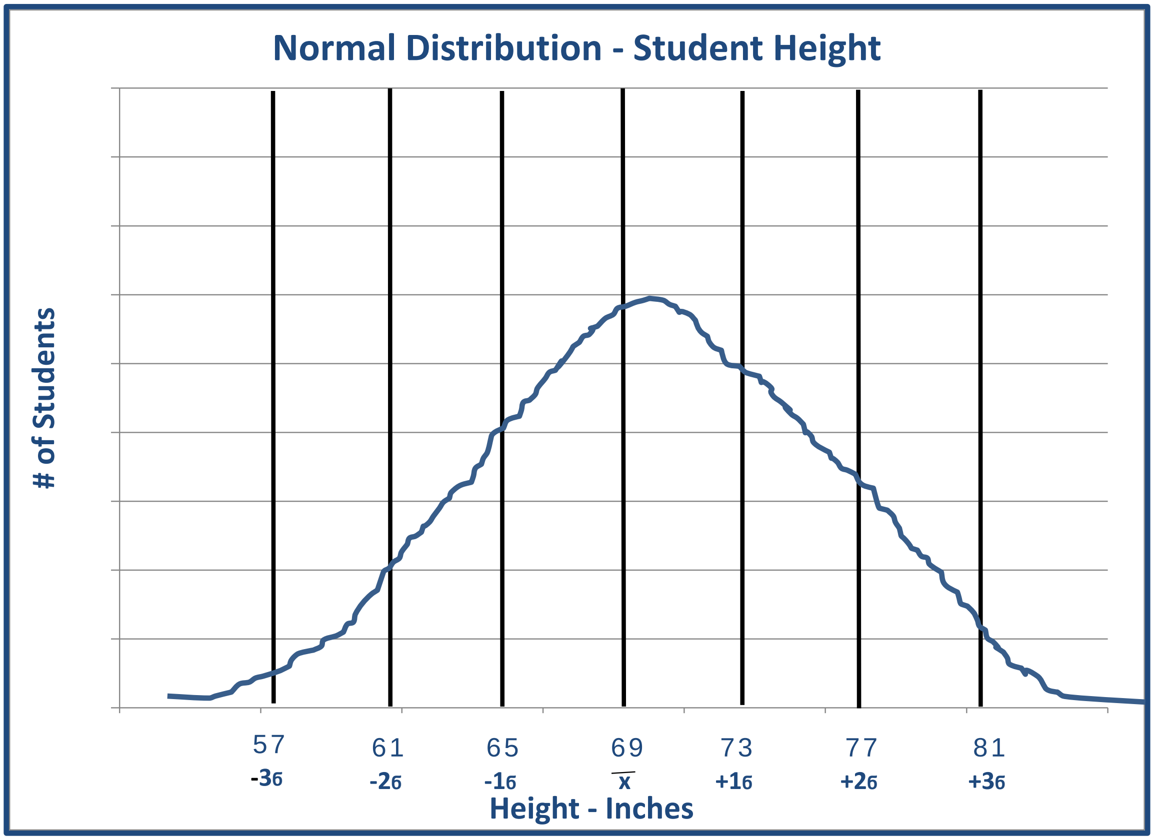


b. Calculate the mean, standard deviation and % of the total population (total number of students in the population*) for your sample. (Use the Sample Standard Deviation S on your calculator).

Sample Size (n)	% Total Population*	Mean ($\overline{x}$)	Standard Deviation (S)
47	3.4%	69.0	4.0

* Need to obtain total population numbers for comparison.
Total School population of 1,374

c. Draw a Normal Distribution Chart for your sample data.



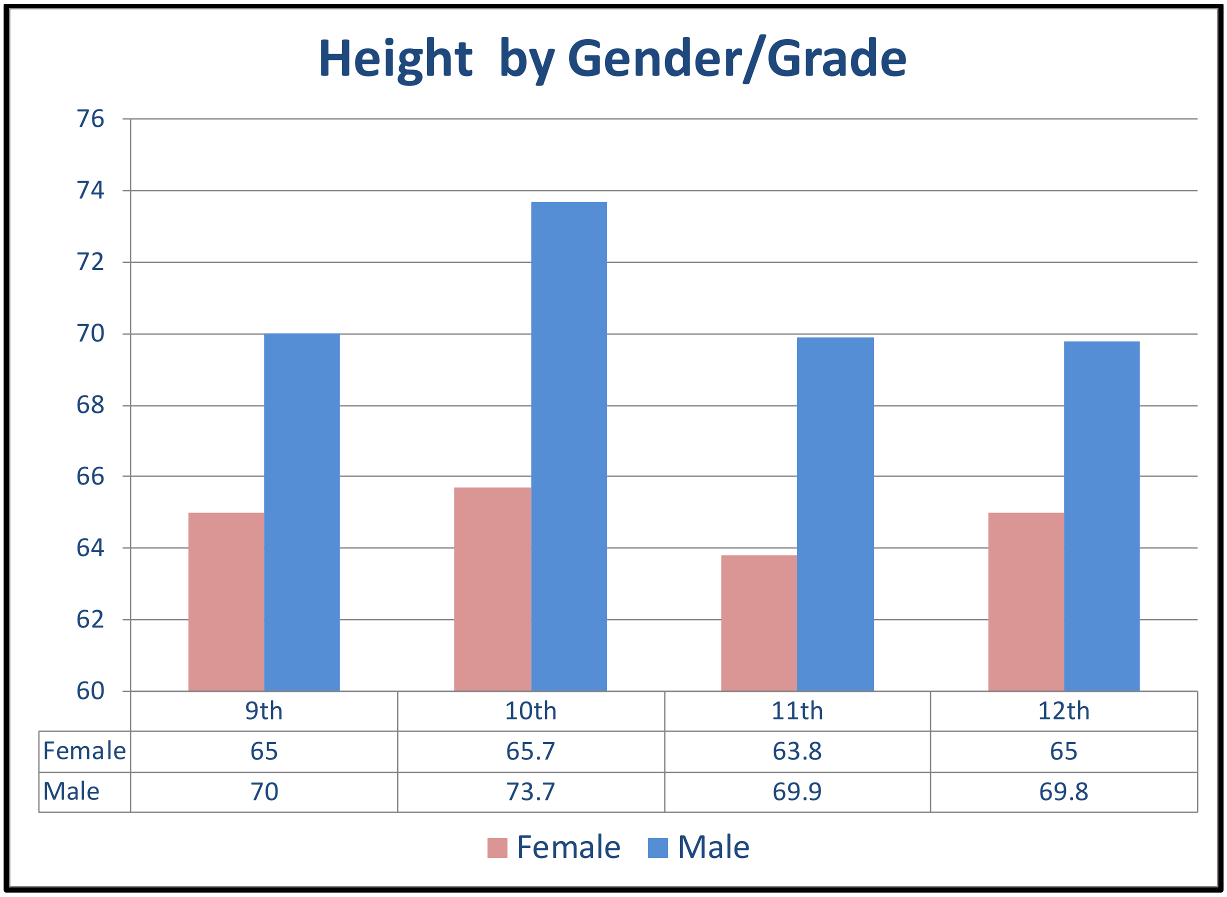
IV. In-depth Sample Analysis

a. "Disaggregate" your data into gender and grade and compare your sample size to the total school population* breakdown.

Grade	Female				Male				Total School*
	Sample	Mean	Total School*	%	Sample	Mean	Total School*	%	
9th	5	65.0	189	2.6%	3	70.0	230	1.3%	419
10th	3	65.7	151	2.0%	3	73.7	184	1.6%	335
11th	5	63.8	141	3.5%	17	69.9	174	9.8%	315
12th	2	65.0	160	1.3%	9	69.8	145	6.2%	305
TOTAL	15	64.7	641	2.3%	32	70.3	733	4.4%	1374

* Need to obtain total population numbers for comparison.
Total School population of 1,374

b. Draw a Histogram of your Means by Gender and Grade.



	9th	10th	11th	12th
Female	65	65.7	63.8	65
Male	70	73.7	69.9	69.8

V. Draw Inferences about the Entire Population

$$z = \frac{x - \bar{x}}{S}$$

Using your normal distribution and the z-score table determine what percent of students are predicted to be:

$$z = \frac{60 - 69}{4} = 2.25$$

100% minus % less than 5 feet + % greater than 6 feet

$$z = \frac{72 - 69}{4} = 0.75$$

Less than 5 feet (60 inches)	Between 5 feet and 6 feet	Greater than 6 feet (72 inches)
1.2%	76.1%	22.7%

VI. Representative Sample?

a. How would you describe your sampling methods?

__ Self-Selected [✓ __ Convenience] __ Systematic __ Random

Describe your method:

The samples were obtained by measuring student heights during 1st lunch (one lunch out of three). Students were asked if we could measure their height and we collected the data.

b. Do you think that your sample is representative of the total population? Why?

NO.

Students were only measured during one lunch. Any possible sample data from the other two lunches was missed.

Looking at the data, juniors and seniors are over-represented and freshmen and sophomores are under-represented. Also more male than female measurements were taken.

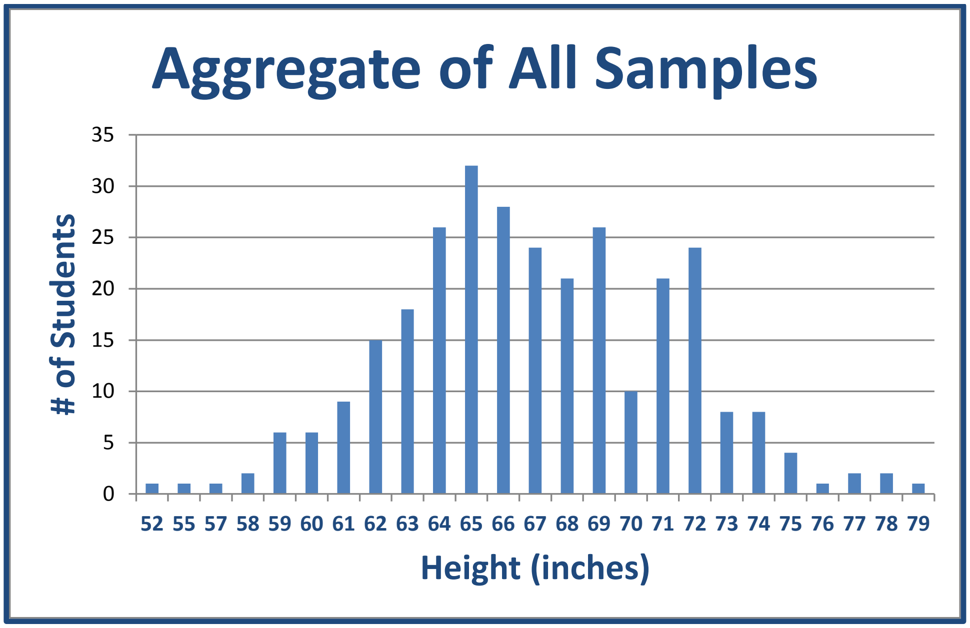
- ***The students taking the measurements were juniors and seniors. They were more easily able to approach fellow juniors and seniors to be sampled.***
- ***Males may have been more willing to volunteer than females.***
- ***The lunch period sampled MAY have contained more males and/or more upper classmen.***

This was definitely a CONVENIENCE sample.

The following is if all teams aggregate the data to have a larger sample. Answers will vary depending on the sample data. This answer key is representative of how the response should be structured.

III. Develop Statistical Parameters of the Sample

a. Draw a Histogram Chart for your sample data.

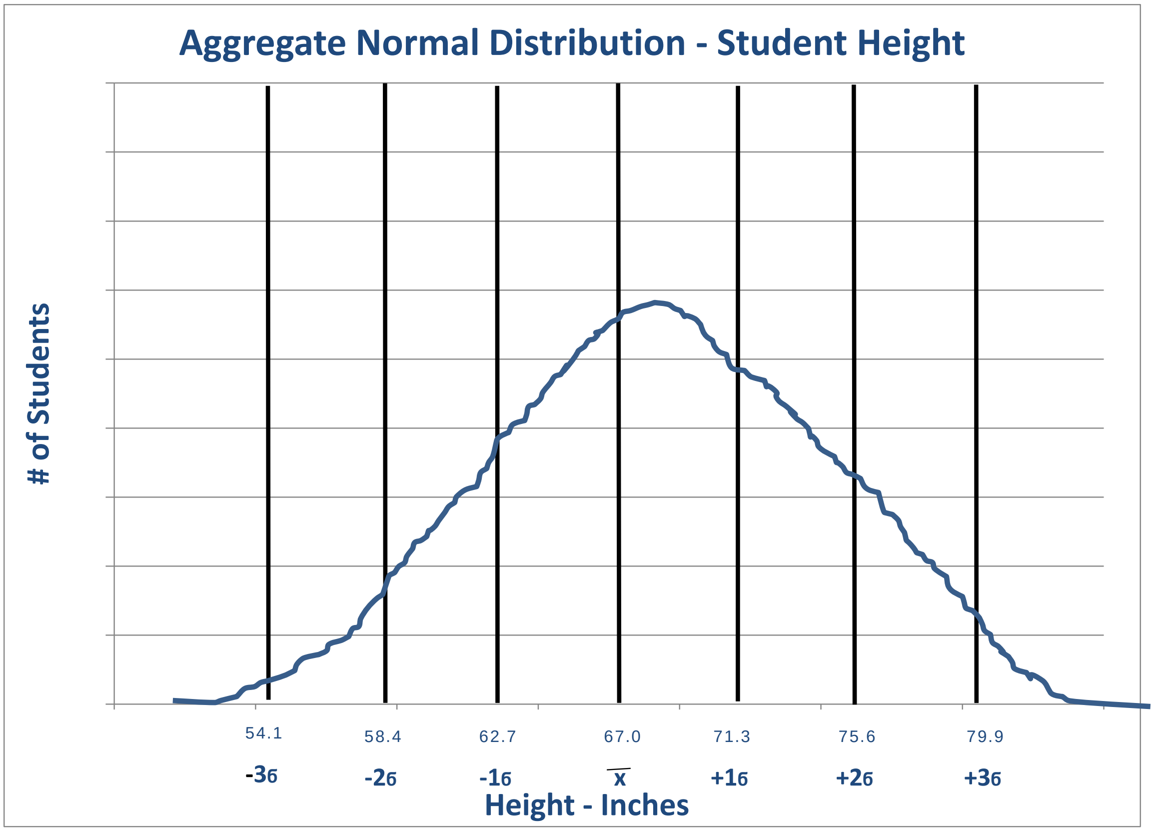


b. Calculate the mean, standard deviation and % of the total population (total number of students in the population*) for your sample.

Team #	Sample Size (n)	% Total Population	Mean ($\bar{x}$)	Std Deviation S
1	40	2.9	66.0	3.8
2	40	2.9	67.9	3.6
3	31	2.2	65.0	3.0
4	38	2.8	65.4	4.1
5	33	2.4	65.8	4.0
6	40	2.9	68.4	5.4
7	29	2.0	67.4	4.2
8	47	3.4	69.0	4.0
TOTAL	**297**	**21.6**	**67.0**	**4.3**

* Need to obtain total population numbers for comparison.
Total School population of 1,374

c. Draw a Normal Distribution Chart for your sample data.



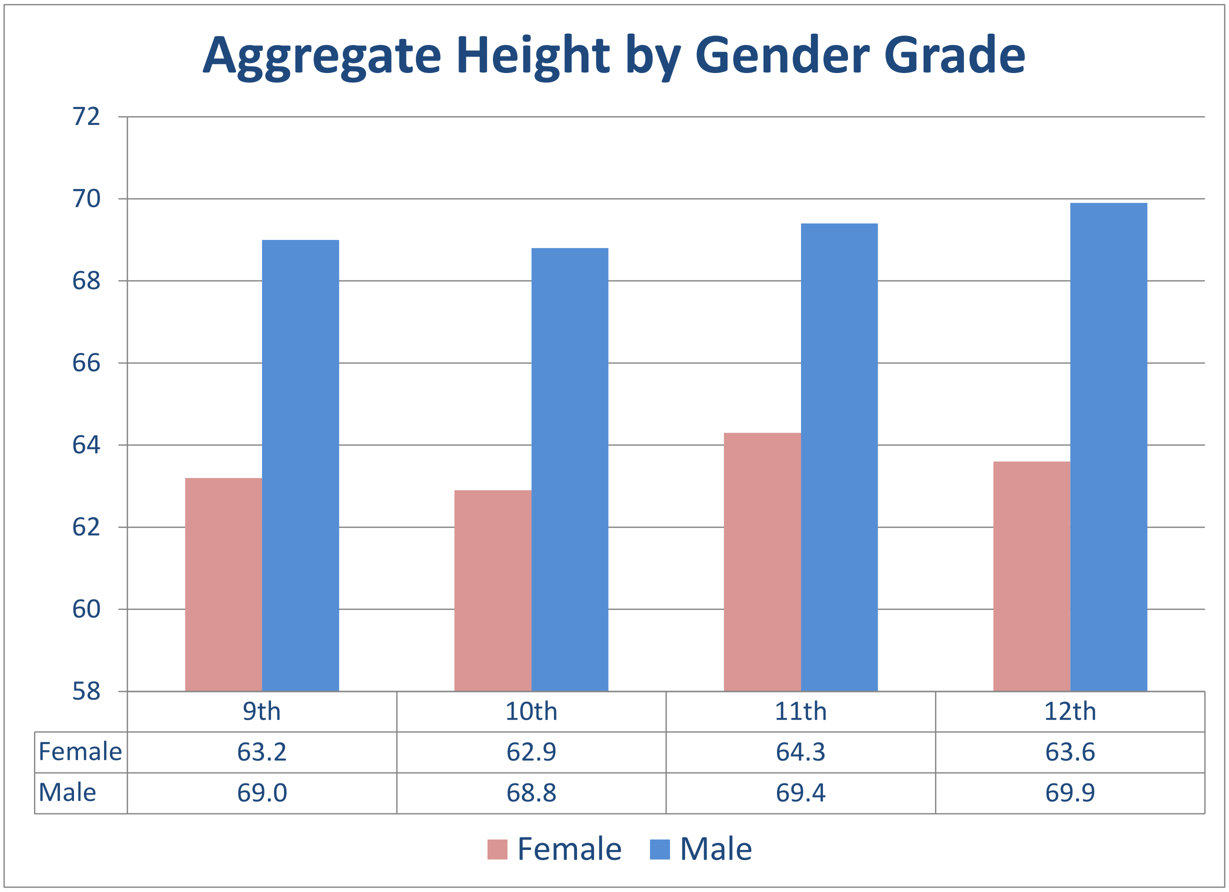
IV. In-depth Sample Analysis

a. “Disaggregate” your data into gender and grade and compare your sample size to the total school population* breakdown.

	Female				Male				Total School *
	Sample	Mean	Total School *	%	Sample	Mean	Total School *	%	
9th	19	63.2	189	10.0%	22	69.0	230	9.6%	419
10th	23	62.9	151	15.2%	25	68.8	184	13.6%	335
11th	49	64.3	141	34.8%	59	69.4	174	33.9%	315
12th	36	63.6	160	22.5%	64	69.9	145	44.1%	305
TOTAL	127	63.7	641	19.8%	169	69.4	733	23.1%	1374

* Need to obtain total population numbers for comparison.
Total School population of 1,374

b. Draw a Histogram of your Means by Gender and Grade.



	9th	10th	11th	12th
Female	63.2	62.9	64.3	63.6
Male	69.0	68.8	69.4	69.9

V. Draw Inferences about the Entire Population

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

Using your normal distribution and the z-score table determine what percent of students are predicted to be:

$$z = \frac{60 - 67}{4.3} = 1.63$$

100% minus % less than 5 feet + % greater than 6 feet

$$z = \frac{72 - 67}{4.3} = 1.16$$

Less than 5 feet (60 inches)	Between 5 feet and 6 feet	Greater than 6 feet (72 inches)
5.3%	82.5%	12.2%

VI. Representative Sample?

b. How would you describe your sampling methods?

__ Self-Selected [✓ __ Convenience] __ Systematic __ Random

Describe your method:

The samples were obtained by measuring student heights during 1st lunch (one lunch out of three). Students were asked if we could measure their height and we collected the data.

c. Do you think that your sample is representative of the total population? Why?

NO.

Students were only measured during one lunch. Any possible sample data from the other two lunches was missed.

Looking at the data, juniors and seniors are over-represented and freshmen and sophomores are under-represented. Also more male than female measurements were taken.

- ***The students taking the measurements were juniors and seniors. They were more easily able to approach fellow juniors and seniors to be sampled.***
- ***Males may have been more willing to volunteer than females.***
- ***The lunch period sampled MAY have contained more males and/or more upper classmen.***

This was definitely a CONVENIENCE sample.

How Tall are Our Students?

In this task, students will gather sample data, calculate statistical parameters and draw inferences about populations. Students will learn to take samples, understand different types of sampling (random, convenience, self-selected, and systematic) and sampling bias, extrapolate from a sample to a population and use statistical methodologies to make predictions regarding the full population.

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, ~~mean absolute deviation~~, standard deviation) of two or more different data sets.

MGSE9-12.S.ID.4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

MGSE9-12.S.IC.1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

STANDARDS FOR MATHEMATICAL PRACTICE

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.

TASK DESCRIPTION, DEVELOPMENT AND DISCUSSION

I. Purpose

To give students the experience of gathering sample data, calculating statistical parameters and drawing inferences about populations.

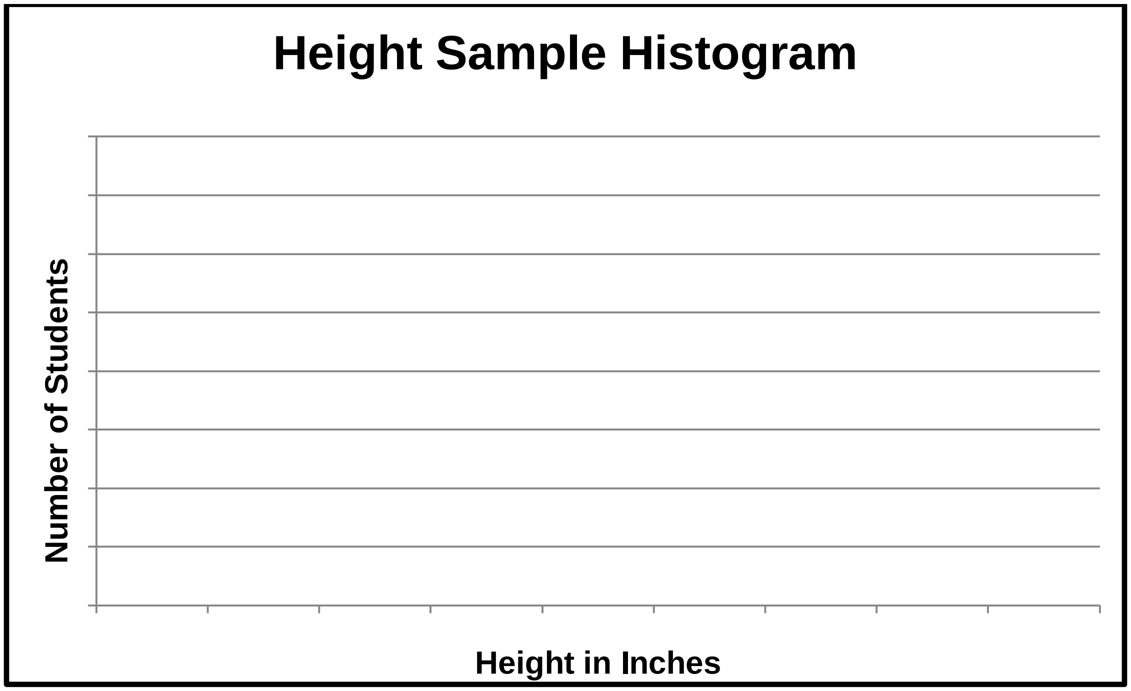
II. Data Sampling (Height Measurement)

a. Each team will utilize a measuring tape or "height pipe" and the "Height Sampling Data" worksheet. They will measure other students, recording Grade, Gender and Height (to the nearest inch), and record their observations on the worksheet. (This works best when a wide range of students are available, e.g. during lunch)

III. Teams should NOT duplicate students measured, i.e. ask the subject if they have already been measured by another team. If they answer yes, then find another student to measure.

IV. Develop Statistical Parameters of the Sample

a. Draw a Histogram Chart for your sample data.

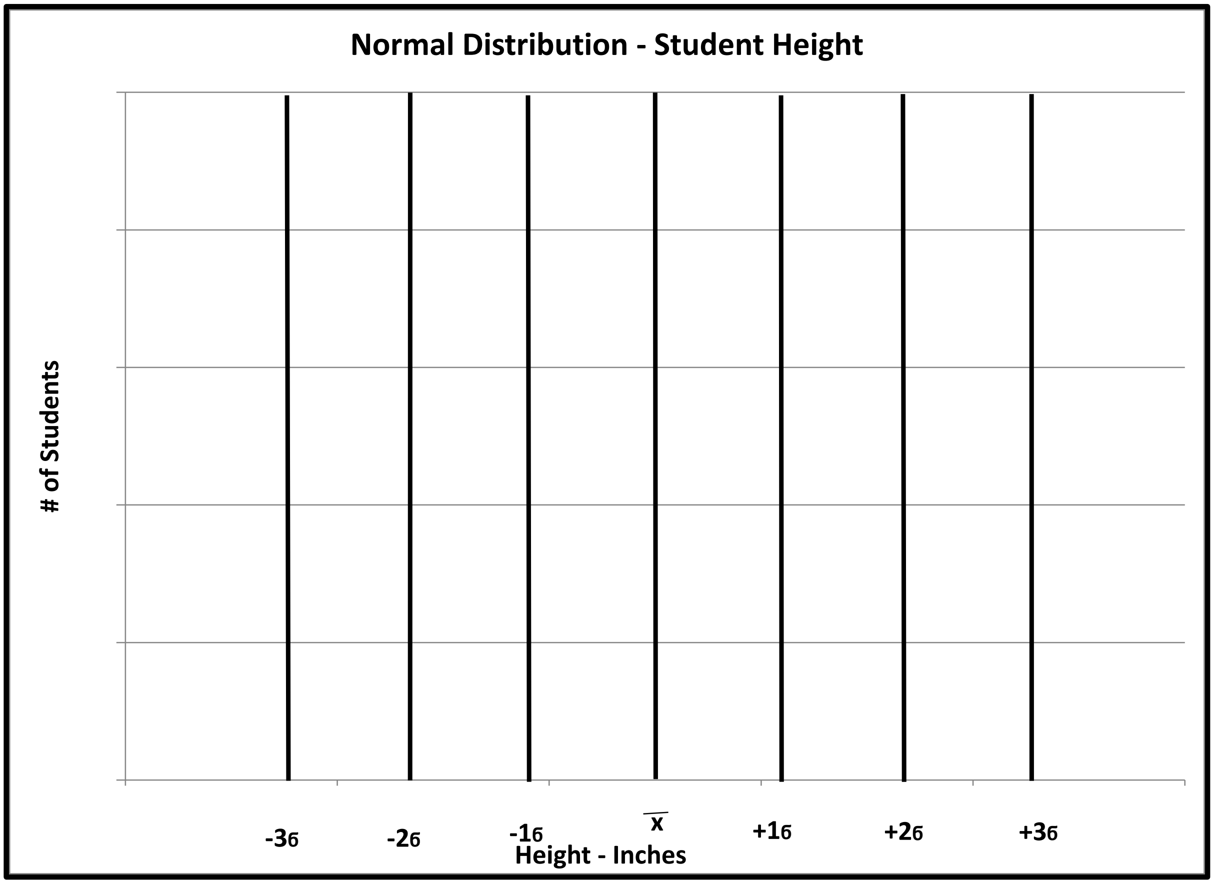


b. Calculate the mean, standard deviation and % of the total population (total number of students in the population*) for your sample. (Use the Sample Standard Deviation S on your calculator).

Sample Size (n)	% Total Population*	Mean ($\overline{x}$)	Standard Deviation (S)

* Need to obtain total population numbers for comparison.

c. Draw a Normal Distribution Chart for your sample data.



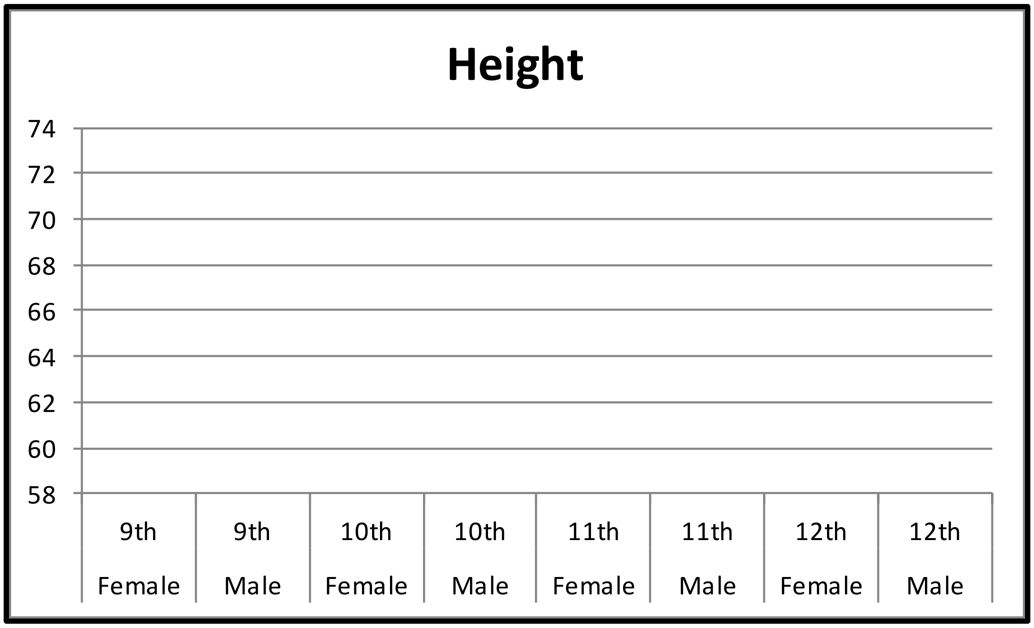
V. In-depth Sample Analysis

a. “Disaggregate” your data into gender and grade and compare your sample size to the total school population* breakdown.

Grade	Female				Male				Total School*
	Sample	Mean	Total School*	%	Sample	Mean	Total School*	%	
9th									
10th									
11th									
12th									
TOTAL									

* Need to obtain total population numbers for comparison.

b. Draw a Histogram of your Means by Gender and Grade.



V. Draw Inferences about the Entire Population

$$z = \frac{x - \bar{x}}{S}$$

Using your normal distribution and the z-score table determine what percent of students are predicted to be:

Less than 5 feet	Between 5 feet and 6 feet	Greater than 6 feet

VI. Representative Sample?

c. How would you describe your sampling methods?

__ Self-Selected __ Convenience __ Systematic __ Random

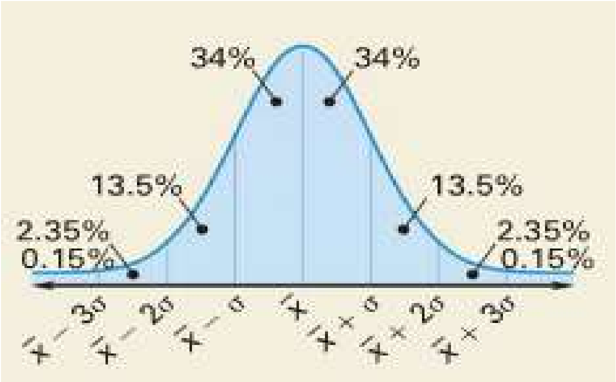
Describe your method:

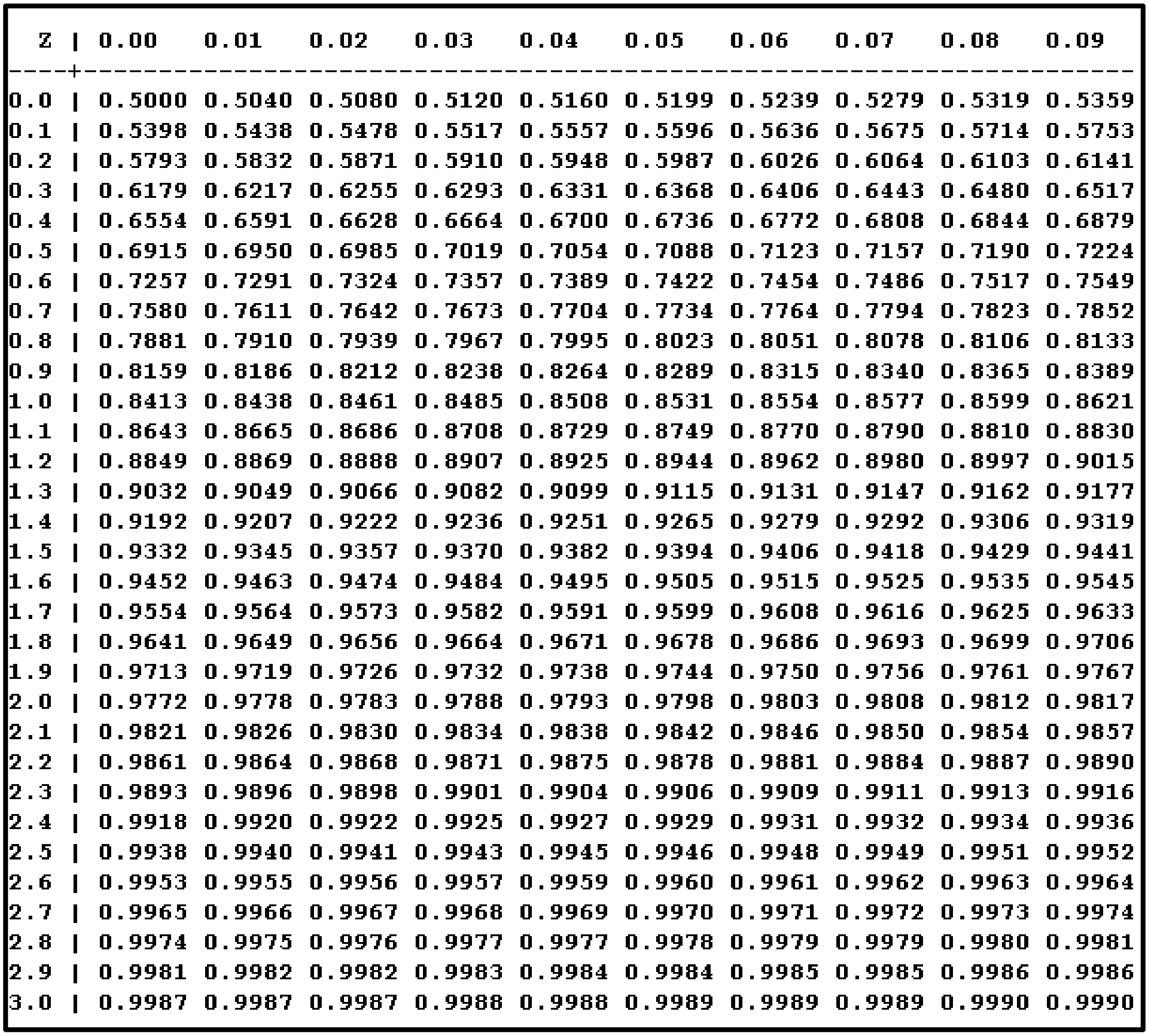
d. Do you think that your sample is representative of the total population? Why?

Height Sampling Data Worksheet

#	Grade	Gender	Height (inches)	#	Grade	Gender	Height (inches)
1				21			
2				22			
3				23			
4				24			
5				25			
6				26			
7				27			
8				28			
9				29			
10				30			
11				31			
12				32			
13				33			
14				34			
15				35			
16				36			
17				37			
18				38			
19				39			
20				40			

Z-Score Table





z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

We're Watching You Learning Task

Mathematical Goals

- Explain how well and why a sample represent the variable of interest from a population
- Demonstrate understanding of the different kinds of sampling methods
- Apply the steps involved in designing an observational study
- Apply the basic principles of experimental design

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

MGSE9-12.S.IC. 2 Decide if a specified model is consistent with results from a given data-generating process, e.g. using simulation.

MGSE9-12.S.IC. 3 Recognize the purpose of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

Standards for Mathematical Practice

1. **Construct viable arguments and critique the reasoning of others**
2. **Model with mathematics**

Introduction

The purpose of this task is for students to design an observational study or experiment using the basic principles of experimental design.

Materials

- Pencil

There are two approaches to collecting data in statistics – observational studies and experiments. In observational studies, researchers observe characteristics from samples of an existing population and use the information collected to make inferences about the population. In an observational study, the researcher gathers data without trying to influence responses or imposing any controls on the situation. In an experiment, researchers gather data by imposing a treatment and observing responses.

There are several key steps involved in designing an observational study.
1. Determine the focus of the study. What is the variable of interest? What information is needed to answer the main question of interest?

2. Develop a plan to collect data. How will subjects be observed?

3. Determine the most appropriate sampling method and select the sample.

4. Collect the data.

5. Describe and interpret the data using appropriate statistical procedures and graphs.

6. Report the findings of the study.

The basic principles of experimental design are
1. **Randomization –** Experimental units/subjects should be randomly assigned to treatment groups;

2. **Control** - Experimenters need to control any lurking variables, generally by comparing multiple treatment groups;

3. **Replication –** The experiment should involve many experimental units/subjects.

Use the information above as well as what you have learned in class to explore the following situations.

1. A local community has just installed red light cameras at its busiest intersection. The police department hopes that the cameras will encourage drivers to be more careful and that incidents of drivers running red lights at this intersection will decrease. Design an observational study that the police department could use to determine if the installation of the traffic light has had the deserved effect.

a. What is the focus of the study? What is the variable of interest? ***Do red light cameras reduce the number of drivers who run red lights? Number of cars running a red light at the city's busiest intersection***

b. Determine the data collection plan. ***Answers will vary. One response might be to collect data for one month prior to the installation of the camera and then for one month following the installation of the cameras. A student might include sampling in his/her response by saying that data might be collected only on certain days of the week or at given times of the day.***

c. Funds are limited and there are only a few days to conduct the study. What is the most appropriate sampling method? ***Students should incorporate a simple random sample into the response. However, their specific responses will vary***

d. The police chief also wonders if there is a difference in driver behavior at different times of day. Can you incorporate this concern into your sampling method? ***A stratified random sample should be used. The strata students use may vary. Possible strata are morning, afternoon and evening. Another method might be heavy commute times vs. times when most people are at work.***

***** Question c and d provide good opportunities for class discussion once groups have determined their sampling methods.*****

2. A few years ago, a study was conducted at Johns Hopkins hospital in Boston to see if exposure to ultrasound could affect the birth weight of a baby. Investigators followed unborn babies and their mothers until their birth and notes their birth weight. A comparison was made between the birth weight of those babies exposed to ultrasound and those babies not exposed to ultrasound. Whether an ultrasound was used on the baby was a decision made by the mother's doctor, based on medical justification. Was this study an experiment or an observational study? Explain. List any possible confounding variables in this study. *(source: Chris Franklin, University of Georgia)*

Observational study – While the study does compare two groups, it is not an experiment because the group the babies went into was determined by the mother's doctor, based on medical justification. There was no randomization.

Confounding variables – Answers will vary. Make sure that students justify how the variable they suggest would result in a group difference being observed. One possible confounding variable is the mother's age and/or health.

3. Suppose the faculty of a Statistics department at a large university wanted to look at how students in the introductory Statistics courses might perform on exams under different environmental conditions. They decided to consider the effect of the size of the classroom (a smaller classroom where there are just enough seats for the students versus a large classroom where the students can spread out with an empty seat between each student). When the next exam is given in one section of the introductory Statistics course, 60 students will be randomly assigned to one of the treatments. The scores on the exam will then be compared. *(source: Chris Franklin, University of Georgia)*

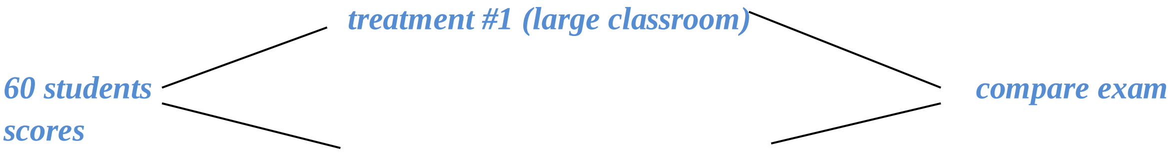
a. Is this study an experiment or observational study. Explain. ***experiment – subjects are randomly assigned to treatment groups. There are a significant number of subjects for replication and there are two treatment groups to control for lurking variables.***

b. Name the explanatory variable. ***size of the classroom***

c. Name the response variable. ***exam scores.***

d. How many treatments will this study compare? Name the treatments. ***two treatments (classroom size) – smaller classrooms and larger classrooms***

e. Diagram a completely randomized design for this study.



treatment #2 (small classroom)

****30 students will be randomly assigned into each treatment group***

4. You want know if talking on a hands-free cell phone distracts drivers. Forty college students "drove" in a simulator equipped with a hands free cell phone. The car ahead brakes: how quickly does the subject respond?

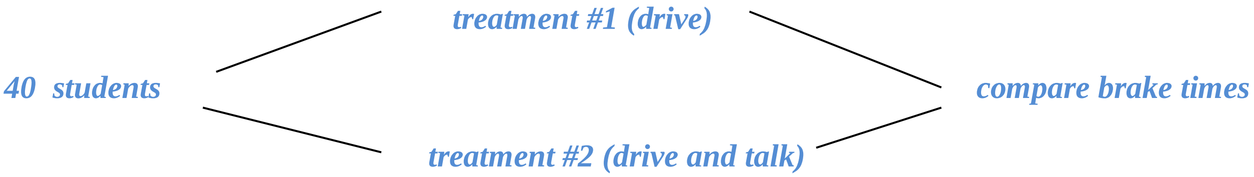
a. What are the experimental units? ***40 college students***

b. What is the explanatory variable? ***use of hands free cell phone***

c. What are the treatments? ***two treatments - drive and drive while talking***

d. What is the response variable? ***brake time***

e. Outline the design of the *above experiment.*



****20 students will be randomly assigned into each treatment group***

5. You want to determine the best color for attracting cereal leaf beetles to boards on which they will be trapped. You will compare four colors: blue, green, white and yellow. You plan to count the number of beetles trapped. You will mount one board on each of 16 poles evenly spaced in a square field.

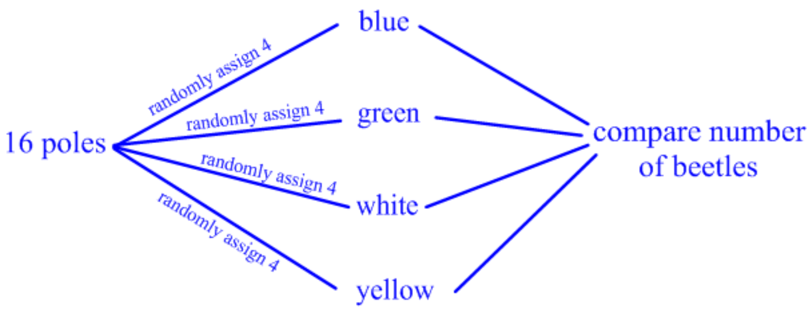
a. What are the experimental units? ***sixteen poles***

b. What is the explanatory variable? ***board color***

c. What are the treatments? ***blue, green, white and yellow***

d. What is the response variable? ***number of beetles trapped***

e. Outline the design of the above experiment.


16 poles
randomly assign 4
randomly assign 4
randomly assign 4
randomly assign 4
blue
green
white
yellow
compare number of beetles

We're Watching You Learning Task

Name ____ Date___________________

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

MGSE9-12.S.IC. 2 Decide if a specified model is consistent with results from a given data-generating process, e.g. using simulation.

MGSE9-12.S.IC. 3 Recognize the purpose of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

Standards for Mathematical Practice

1. **Construct viable arguments and critique the reasoning of others**
2. **Model with mathematics**

There are two approaches to collecting data in statistics – observational studies and experiments. In observational studies, researchers observe characteristics from samples of an existing population and use the information collected to make inferences about the population. In an observational study, the researcher gathers data without trying to influence responses or imposing any controls on the situation. In an experiment, researchers gather data by imposing a treatment and observing responses.

- There are several key steps involved in designing an observational study.

  1. Determine the focus of the study. What is the variable of interest? What information is needed to answer the main question of interest?

  2. Develop a plan to collect data. How will subjects be observed?

  3. Determine the most appropriate sampling method and select the sample.

  4. Collect the data.

  5. Describe and interpret the data using appropriate statistical procedures and graphs.

  6. Report the findings of the study.

The basic principles of experimental design are

1. **Randomization** – Experimental units/subjects should be randomly assigned to treatment groups;

2. **Control** - Experimenters need to control any lurking variables, generally by comparing multiple treatment groups;

3. **Replication** – The experiment should involve many experimental units/subjects.

Use the information above as well as what you have learned in class to explore the following situations.

1. A local community has just installed red light cameras at its busiest intersection. The police department hopes that the cameras will encourage drivers to be more careful and that incidents of drivers running red lights at this intersection will decrease. Design an observational study that the police department could use to determine if the installation of the traffic light has had the deserved effect.

a. What is the focus of the study? What is the variable of interest?

b. Determine the data collection plan.

c. Funds are limited and there are only a few days to conduct the study. What is the most appropriate sampling method?

d. The police chief also wonders if there is a difference in driver behavior at different times of day. Can you incorporate this concern into your sampling method?

2. A few years ago, a study was conducted at Johns Hopkins hospital in Boston to see if exposure to ultrasound could affect the birth weight of a baby. Investigators followed unborn babies and their mothers until their birth and notes their birth weight. A comparison was made between the birth weight of those babies exposed to ultrasound and those babies not exposed to ultrasound. Whether an ultrasound was used on the baby was a decision made by the mother’s doctor, based on medical justification. Was this study an experiment or an observational study? Explain. List any possible confounding variables in this study. *(source: Chris Franklin, University of Georgia)*

3. Suppose the faculty of a Statistics department at a large university wanted to look at how students in the introductory Statistics courses might perform on exams under different environmental conditions. They decided to consider the effect of the size of the classroom (a

smaller classroom where there are just enough seats for the students versus a large classroom where the students can spread out with an empty seat between each student). When the next exam is given in one section of the introductory Statistics course, 60 students will be randomly assigned to one of the treatments. The scores on the exam will then be compared. *(source: Chris Franklin, University of Georgia)*

a. Is this study an experiment or observational study. Explain.

b. Name the explanatory variable.

c. Name the response variable.

d. How many treatments will this study compare? Name the treatments

e. Diagram a completely randomized design for this study.

4. You want know if talking on a hands-free cell phone distracts drivers. Forty college students “drove” in a simulator equipped with a hands free cell phone. The car ahead brakes: how quickly does the subject respond?

a. What are the experimental units?

b. What is the explanatory variable?

c. What are the treatments?

d. What is the response variable?

e. Outline the design of the *above experiment.*

5. You want to determine the best color for attracting cereal leaf beetles to boards on which they will be trapped. You will compare four colors: blue, green, white and yellow. You plan to count the number of beetles trapped. You will mount one board on each of 16 poles evenly spaced in a square field.

a. What are the experimental units?

b. What is the explanatory variable?

c. What are the treatments?

d. What is the response variable?

e. Outline the design of the above experiment.

Colors of Skittles Learning Task

Mathematical Goals

- Understand sample distributions of sample proportions through simulation
- Develop the formulas for the mean and standard deviation of the sampling distribution of a sample proportion
- Discover the Central Limit Theorem for a sample proportion

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

MGSE9-12.S.IC. 2 Decide if a specified model is consistent with results from a given data-generating process, e.g. using simulation.

MGSE9-12.S.IC. 4 Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling

Standards for Mathematical Practice

1. Reason abstractly and quantitatively
2. Construct viable arguments and critique the reasoning of others
3. Model with mathematics

Introduction

The purpose of this task is to review ideas about sampling distributions of sample proportions and develop this knowledge into the Central Limit Theorem (CLT) for Proportions. In addition to collecting data with actual candy, students will use simulation to fully develop the CLT. The task concludes with a problem intended to synthesize the ideas from the simulation.

Modification Ideas:

- The task has been modified to use Skittles instead of Reese's Pieces. Part three can be left off; however, the class should review the ideas of sampling variability and the impact of the sample size on the variability and the usefulness of the statistic approximating the parameter.
- This task can be modified to address any question of interest that can be modeled with a binomial distribution. For example, if you prefer M&M's, you could look at the percent of blue M&M's out of a bag.
- The students could conduct surveys (or simulate surveys on their calculators). They could ask if their fellow students like country music, use a social networking site, or listen to a digital music player. These options require very small modifications to the activity below.

Materials

- One regular bag of Skittles for each group

- Graphing calculator or statistical software package
- Graph paper

Part 1

1. Reviewing some basics.

a) Think about a single bag of Skittles. Does this single bag represent a ***sample*** of Skittles candies or the ***population*** of Skittles candies? ***Solution: The single bag represents a sample. The population is all of the Skittles made. Or it could be considered to be all of the Skittles made in the same batch.***

b) We use the term ***statistic*** to refer to measures based on samples and the term ***parameter*** to refer to measures of the entire population. If there are 50 Skittles in your bag, is 50 a statistic or a parameter? If Mars claims that 20% of all Skittles are yellow, is 20% a statistic or a parameter?

Solution: 50 is a statistic. 20% is a parameter.

c) What is **sampling variability**? ***The idea that the value of a statistic varies as we take another sample***

2. How many orange candies should I expect in a bag of Skittles?

a) From your bag of Skittles, take a random sample of 10 candies. Record the count and proportion of each color in your sample. ***Answers will vary.***

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion ($\hat{p}$)					

b) Do you think that every student in the class obtained the same proportion of orange candies in his or her sample? Why or why not? ***Solution: No. Samples have variability.***

c) Combine your results with the rest of the class and produce a *dot plot* for the distribution of sample proportions of **orange** candies (out of a sample of **10 candies**) obtained by the class members.

Make sure you label the axes of your dot plot correctly.

Answers will vary. Be sure that the students are labeling the horizontal axis of the dotplot. Other dotplots will also be created and the students will need to be able to distinguish them.

d) What is the average of the sample proportions obtained by your class?

Answers will vary.

e) Put the Skittles back in the bag and take a random sample of 25 candies. Record the count and proportion of each color in your sample.

Answers will vary.

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion ($\hat{p}$)					

f) Combine your results with the rest of the class and produce a dot plot for the distribution of sample proportions of **orange** candies (out of a sample of **25 candies**) obtained by the class members. Is there more or less **variability** than when you sampled 10 candies? Is this what you expected? Explain.

Answers will vary. Be sure that the students are labeling the horizontal axis of the dotplot. There should be less variability in this plot. That is, the data should be clustered together more. If the students remember discussing sampling variability, this result should be expected.

g) What is the average of the sample proportions (from the **samples of 25**) obtained by your class? Do you think this is closer or farther from the true proportion of **oranges** than the value you found in part *d*? Explain.

Answers will vary. The proportion should be closer to the true proportion because the sample size is larger.

h) This time, take a random sample of 40 candies. Record the count and proportion of each color in your sample.

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion ($\hat{p}$)					

Answers will vary.

i) Combine your results with the rest of the class and produce a dot plot for the distribution of sample proportions of **orange** candies (out of a sample of **40 candies**) obtained by the class members. Is there more or less **variability** than the previous two samples? Is this what you expected? Explain.

Answers will vary. Be sure that the students are labeling the horizontal axis of the dotplot. There should be less variability in this plot than in the previous two. If the students remember discussing sampling variability this result should be expected.

j) What is the average of the sample proportions (from the **samples of 40**) obtained by your class? Do you think this is closer or farther from the true proportion of **oranges** than the values you found in parts *d* and g? Explain.

Answers will vary. The proportion should be closer to the true proportion because the sample size is larger.

Part 2

We have been looking a number of different ***sampling distributions*** of $\hat{p}$ (the distribution of the statistic for *all possible samples* of a given size), but we have seen that there is **variability** in the distributions.

We would like to know that $\hat{p}$ is a good estimate for the true proportion of orange Skittles. However, there are guidelines for when we can use the statistic to estimate the parameter (in other words using sample data to make assumption about the entire population).

First, however, we need to understand the center, shape, and spread of the sampling distribution of $\hat{p}$.

We know that if we are counting the number of Skittles that are orange and comparing with those that are not orange, then the counts of orange follow a binomial distribution (given that the population is much larger than our sample size). There are two outcomes of Skittles, either orange or not-orange and each sample selected of Skittles was independent from the others.

Conditions for Binomial Distribution are:

- The experiment consists of *n* repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success, probability *p* (ranging from 0 to 1), and the other, a failure, probability, (1 – *p*).
- The *n* trials are independent; that is, the outcome on one trial does not affect the outcome on other trials. (Think of replacement with your skittles.)

The formulas for the mean and standard deviation of a **binomial distribution:**

$$\mu_X = np \qquad \sigma_X = \sqrt{np(1-p)}$$

where *n* = numbers of trials and *p* = proportion

a) Given that $\hat{p} = \frac{x}{n}$, where *x* is the count of oranges and *n* is the total in the sample, we find $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$ by dividing each by *n* also. Find the formulas for each statistic.

Solution: $\mu_{\hat{p}} = \dfrac{1}{n}np = p$

$$\sigma_{\hat{p}} = \frac{1}{n}\sqrt{np(1-p)} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}}.$$

b) This leads us to the statement of the characteristics of the sampling distribution of a sample proportion.

The Sampling Distribution of a Sample Proportion:

Choose a simple random sample of size *n* from a large population with population parameter *p* having some characteristic of interest. Let $\hat{p}$ be the proportion of the sample having that characteristic. Then

- The mean of the sampling distribution $\left(\mu_{\hat{p}}\right)$ is ____***p***_______; and
- The standard deviation of the sampling distribution ($\sigma_{\hat{p}}$) is ______$\sqrt{\frac{\boldsymbol{p(1-p)}}{\boldsymbol{n}}}$___________.

c) Let's look at the standard deviation a bit more. What happens to the standard deviation as the sample size increases? Try a few examples to verify your conclusion. Then use the formula to explain why your conjecture is true.

Sample Size (*n*)	Let *p* = 0.2	Let *p* = 0.7
1	***0.4***	***0.46***
5	***0.18***	***0.20***
10	***0.13***	***0.15***
25	***0.08***	***0.09***
50	***0.06***	***0.065***
100	***0.04***	***0.046***
1000	***0.013***	***0.014***

Solution: As the sample size increases, the standard deviation decreases. If we rewrite the standard deviation as $\sigma_{\hat{p}} = \sqrt{\frac{1}{n}}\sqrt{p(1-p)}$***, we can think only about what happens as n increases. Students have previously studied the graph of y = 1/x, so they should know that as x***

increases, the value of y decreases. The square roots of such numbers also decrease as x increases.

If we wanted to cut the standard deviation in **half**, thus decreasing the **variability** of $\hat{p}$, what would we need to do in terms of our sample size? (Hint: multiply the formula for standard deviation by ½).

Solution: We would need to increase the sample size by a factor of 4:

$$\sigma_{\hat{p}} = \frac{1}{2}\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{p(1-p)}{(2^2)n}} = \sqrt{\frac{p(1-p)}{4n}}$$

Caution: We can only use the formula for the standard deviation of $\hat{p}$ when the population is at least **10** times as large as the sample.

d) For each of the samples taken in Part 1 of the Skittles Task, determine what the population of Skittles must be for us to use the standard deviation formula derived above. Is it safe to assume that the population is at least as large as these amounts? Explain.

Solution: The population would need to at least 100, 250, and 400, respectively. It is reasonable to believe the population is larger than these amounts. Explanations will vary.

Part 3

Simulating the Selection of Orange Skittles

As we have seen, there is variation in the distributions depending on the size of your sample and which sample is chosen. To better investigate the distribution of the sample proportions, we need more samples and we need samples of **larger size**. We will turn to technology to help with this sampling. For this simulation, we need to assume a value for the true proportion of orange candies. **Let's assume *p* = *0.20***.

Teaching Note: The following simulation is explained using the TI calculator. However, other brands as well as software programs can be substituted. Teachers will need to consult their calculator and software manuals to determine how to conduct these simulations. Also, these simulations only call for 100 samples. If you are using a computer or high powered calculators, you can use more samples.

a) First, let's imagine that there are 100 students in the class and each takes a sample of 50 Skittles. We can simulate this situation with your calculator.

Type *randBin(50,0.20)* in your calculator. The command *randBin* is found in the following way.

- TI-83/84: Math → PROB → 7:randBin(

What number did you get? Compare with a neighbor. What do you think this command does?

Answers will vary. The command considers a binomial distribution with a sample size of 50 and a success probability of 0.20. It runs a simulation and counts the number of "successes" there are out of 50 trials.

How could you obtain the proportion that are orange?

Solution: Divide the output by 50.

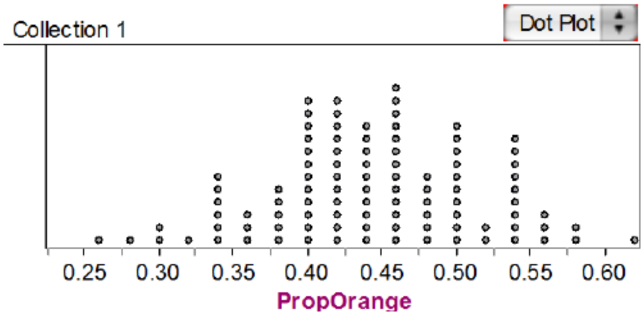
b) Now, we want to generate 100 samples of size 50. This time, input *randBin(50,0.20,100)/50→*L_1. [The → is found by using the STO button on the bottom left of the calculator (TI-83/84).] The latter part (store in L_1) puts all of the outputs into List 1. **(Be patient, this takes a while.)**

Using your Stat Plots, create a histogram or stem-and-leaf plot of the proportions of orange candies. (Plot 1 → On → Type: third selection, top row will give you a histogram. Zoom → 9: ZoomStat → Graph)

Sketch the graph below.

Do you notice a pattern in the distribution of the sample proportions? Explain.

Dotplots will vary. Here is one example: The pattern should be that the plot is close to a normal distribution. Students may need to be reminded what a normal distribution is and why it is important.



c) Find the sample mean and sample standard deviation of the theoretical sample proportion of orange Skittles with sample size 50 and proportion 0.20. Next, find the mean and standard deviation of the output using 1-Var Stats. How do these compare with the theoretical mean and standard deviation for a sampling distribution of a sample proportion?

Theoretical Sample: Mean: ____*p=0.20*_______ Standard Deviation: $\sqrt{\frac{0.2(0.8)}{50}} \approx$ 0.057

Your Sample: Mean: _______________ Standard Deviation: _______________

d) Use the TRACE button on the calculator to count how many of the 100 sample proportions are within ± 0.057 of 0.20. Note: 0.057 is close to the standard deviation you found above, so we are going about one standard deviation on each side of the mean. Then repeat for within ± 0.114 and for within ± 0.171. Record the results below:

	Number of the 100 Sample Proportions	Percentage of the 100 Sample Proportions
Within ± 0.057 of 0.20		
Within ± 0.114 of 0.20		
Within ± 0.171 of 0.20		

Answers will vary

e) If each of the 100 students from your theoretical sample who sampled Skittles were to estimate the population proportion of orange candies by going a distance of 0.114 on either side of his or her sample proportion, what percentage of the 100 students would capture the actual proportion (0.20) within this interval? ***Answers will vary***

f) Simulate drawing out 200 Skittles 100 times [randBin(200, .20, 100)/200→L1]. Find the mean and standard deviation of the set of sample proportions in this simulation. Compare with the theoretical mean and standard deviation of the sampling distribution with sample size 200. Create and study the dot plot for this data set.

Theoretical Sample: Mean: ____***p=0.20***_______ Standard Deviation: $\sqrt{\frac{0.2(0.8)}{50}} \approx$

0.057

Your Sample: Mean: ______________ Standard Deviation: ______________

Answers will vary

g) How is the plot of the sampling distribution from part f different from the plot in part b? How do the mean and standard deviation compare?

Answers will vary

h) What percentage of the 200 sample proportions fall within 0.114 of 0.20 (or approximately 2 standard deviations)? How does this compare with the answer to part e?

Answers will vary

i) You should notice that these distributions follow an approximately normal distribution. The Empirical Rule states how much of the data will fall within 1, 2, and 3 standard deviations of the mean. Restate the rule:

Empirical Rule: In a normal distribution with mean μ and standard deviation σ,

- ____% of the observations fall within 1 standard deviation (1σ) of the mean (μ),
- ____% of the observations fall within 2 standard deviations (2σ) of the mean (μ), and
- ____% of the observations fall within 3 standard deviation (3σ) of the mean (μ).

Solutions: 68, 95, 99.7

j) Do your answers to parts e and h agree with the Empirical Rule? Explain.

Answers will vary. For the simulations above, the answers do generally agree with the Empirical Rule.

This leads us to an important result in statistics: the **Central Limit Theorem (CLT) for a Sample Proportion**:

Choose a simple random sample of size n from a large population with population parameter p having some characteristic of interest. Then the sampling distribution of the sample proportion $\hat{p}$ is approximately normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$. This approximation becomes more and more accurate as the sample size n increases, and it is generally considered valid if the population is much larger than the sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$.

k) How might this theorem be helpful? What advantage does this theorem provide in determining the likelihood of events?

Answers will vary. Students should be aware that this theorem allows us to take random variables $\hat{p}$***, whose original distributions are not normal, and apply normal distribution calculations. At this point, students should be reminded how to find standardized scores (z-scores), the characteristics of the standard normal curve, and how to calculate probabilities with the standard normal.***

Colors of Skittles Learning Task

Name ____

Date________________________

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

MGSE9-12.S.IC. 2 Decide if a specified model is consistent with results from a given data-generating process, e.g. using simulation.

MGSE9-12.S.IC. 4 Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling

Standards for Mathematical Practice

1. **Reason abstractly and quantitatively**
2. **Construct viable arguments and critique the reasoning of others**
3. **Model with mathematics**

Part 1:

1. Reviewing some basics.

a) Think about a single bag of Skittles. Does this single bag represent a ***sample*** of Skittles candies or the ***population*** of Skittles candies?

b) We use the term ***statistic*** to refer to measures based on samples and the term ***parameter*** to refer to measures of the entire population. If there are 50 Skittles in your bag, is 50 a statistic or a parameter? If Mars claims that 20% of all Skittles are yellow, is 20% a statistic or a parameter?

c) What is **sampling variability**?

2. How many orange candies should I expect in a bag of Skittles?

a) From your bag of Skittles, take a random sample of 10 candies. Record the count and proportion of each color in your sample.

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion ($\hat{p}$)					

b) Do you think that every student in the class obtained the same proportion of orange candies in his or her sample? Why or why not?

c) Combine your results with the rest of the class and produce a *dot plot* for the distribution of sample proportions of **orange** candies (out of a sample of **10 candies**) obtained by the class members.

Make sure you label the axes of your dot plot correctly.

d) What is the average of the sample proportions obtained by your class?

e) Put the Skittles back in the bag and take a random sample of 25 candies. Record the count and proportion of each color in your sample.

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion ($\hat{p}$ **)**					

f) Combine your results with the rest of the class and produce a dot plot for the distribution of sample proportions of **orange** candies (out of a sample of **25 candies**) obtained by the class members. Is there more or less **variability** than when you sampled 10 candies? Is this what you expected? Explain.

g) What is the average of the sample proportions (from the **samples of 25**) obtained by your class? Do you think this is closer or farther from the true proportion of **oranges** than the value you found in part *d*? Explain.

h) This time, take a random sample of 40 candies. Record the count and proportion of each color in your sample.

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion ($\hat{p}$ **)**					

i) Combine your results with the rest of the class and produce a dot plot for the distribution of sample proportions of **orange** candies (out of a sample of **40 candies**) obtained by the class members. Is there more or less **variability** than the previous two samples? Is this what you expected? Explain.

j) What is the average of the sample proportions (from the **samples of 40**) obtained by your class? Do you think this is closer or farther from the true proportion of **oranges** than the values you found in parts *d* and g? Explain.

Names of group
members:____

COLORS OF SKITTLES CANDIES LEARNING TASK – PART 1 ANSWER SHEET

1a. ________________________

1b. 50: ________________________ 20%: ________________________

1c.
____
_

2a. Random Sample of 10 Skittles

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion ($\hat{p}$)					

2b.
____

2c.

2d. ________________________

2e. Random Sample of 25 Skittles

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion ($\hat{p}$)					

2f.

Variability? _________________________ What you expected? ____________________

Explain: ____

____

2g. Class average: _________________________ True proportion? __________________

Explain:____

____
_

2h. Random Sample of 40 Skittles

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion ($\hat{p}$)					

2i.

Variability? _________________________ What you expected? _________________________

Explain:____

____
__

2j. Class average: ________________________ True proportion? _________________

Explain:____
____
__

Part 2

We have been looking a number of different ***sampling distributions*** of $\hat{p}$ (the distribution of the statistic for *all possible samples* of a given size), but we have seen that there is **variability** in the distributions.

We would like to know that $\hat{p}$ is a good estimate for the true proportion of orange Skittles. However, there are guidelines for when we can use the statistic to estimate the parameter (in other words using sample data to make assumption about the entire population).

First, however, we need to understand the center, shape, and spread of the sampling distribution of $\hat{p}$.

We know that if we are counting the number of Skittles that are orange and comparing with those that are not orange, then the counts of orange follow a binomial distribution (given that the population is much larger than our sample size). There are two outcomes of Skittles, either orange or not-orange and each sample selected of Skittles was independent from the others.

Conditions for Binomial Distribution are:

- The experiment consists of n repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success, probability p (ranging from 0 to 1), and the other, a failure, probability, $(1 - p)$.
- The n trials are independent; that is, the outcome on one trial does not affect the outcome on other trials. (Think of replacement with your skittles.)

The formulas for the mean and standard deviation of a **binomial distribution:**

$$\mu_X = np \qquad \sigma_X = \sqrt{np(1-p)}$$

where n = numbers of trials and p = proportion

a) Given that $\hat{p} = \frac{x}{n}$, where x is the count of oranges and n is the total in the sample, we find $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$ by dividing each by n also. Find the formulas for each statistic.

b) This leads us to the statement of the characteristics of the sampling distribution of a sample proportion.

The Sampling Distribution of a Sample Proportion:

Choose a simple random sample of size *n* from a large population with population parameter *p* having some characteristic of interest. Let $\hat{p}$ be the proportion of the sample having that characteristic. Then

- The mean of the sampling distribution $\left(\mu_{\hat{p}}\right)$ is ____ _______; and
- The standard deviation of the sampling distribution ($\sigma_{\hat{p}}$) is ________________.

c) Let's look at the standard deviation a bit more. What happens to the standard deviation as the sample size increases? Try a few examples to verify your conclusion. Then use the formula to explain why your conjecture is true.

Sample Size (*n*)	Let $p = 0.2$	Let $p = 0.7$
1		
5		
10		
25		
50		
100		
1000		

If we wanted to cut the standard deviation in **half**, thus decreasing the **variability** of $\hat{p}$, what would we need to do in terms of our sample size? (<u>Hint</u>: multiply the formula for standard deviation by ½).

Caution: We can only use the formula for the standard deviation of $\hat{p}$ when the population is at least **10** times as large as the sample.

d) For each of the samples taken in Part 1 of the Skittles Task, determine what the population of Skittles must be for us to use the standard deviation formula derived above. Is it safe to assume that the population is at least as large as these amounts? Explain.

Part 3

Simulating the Selection of Orange Skittles

As we have seen, there is variation in the distributions depending on the size of your sample and which sample is chosen. To better investigate the distribution of the sample proportions, we need more samples and we need samples of **larger size**. We will turn to technology to help with this sampling. For this simulation, we need to assume a value for the true proportion of orange candies. **Let's assume $p = 0.20$**.

a) First, let's imagine that there are 100 students in the class and each takes a sample of 50 Skittles. We can simulate this situation with your calculator.

> Type *randBin(50,0.20)* in your calculator. The command *randBin* is found in the following way.
>
> - TI-83/84: Math → PROB → 7:randBin(

What number did you get? Compare with a neighbor. What do you think this command does?

How could you obtain the proportion that are orange?

b) Now, we want to generate 100 samples of size 50. This time, input *randBin(50,0.20,100)/50*$\rightarrow L_1$. [The → is found by using the STO button on the bottom left of the calculator (TI-83/84).] The latter part (store in L_1) puts all of the outputs into List 1. **(Be patient, this takes a while.)**

Using your Stat Plots, create a histogram or stem-and-leaf plot of the proportions of orange candies. (Plot 1 → On → Type: third selection, top row will give you a histogram. Zoom → 9: ZoomStat → Graph)

Sketch the graph below.

Do you notice a pattern in the distribution of the sample proportions? Explain.

c) Find the sample mean and sample standard deviation of the theoretical sample proportion of orange Skittles with sample size 50 and proportion 0.20. Next, find the mean and standard deviation of the output using 1-Var Stats. How do these compare with the theoretical mean and standard deviation for a sampling distribution of a sample proportion?

Theoretical Sample: Mean: ____________ Standard Deviation: _____________

Your Sample: Mean: ______________ Standard Deviation:
____

d) Use the TRACE button on the calculator to count how many of the 100 sample proportions are within ± 0.057 of 0.20. Note: 0.057 is close to the standard deviation you found above, so we are going about one standard deviation on each side of the mean. Then repeat for within ± 0.114 and for within ± 0.171. Record the results below:

	Number of the 100 Sample Proportions	Percentage of the 100 Sample Proportions
Within ± 0.057 of 0.20		
Within ± 0.114 of 0.20		
Within ± 0.171 of 0.20		

e) If each of the 100 students from your theoretical sample who sampled Skittles were to estimate the population proportion of orange candies by going a distance of 0.114 on either side of his or her sample proportion, what percentage of the 100 students would capture the actual proportion (0.20) within this interval?

f) Simulate drawing out 200 Skittles 100 times [randBin(200, .20, 100)/200→L1]. Find the mean and standard deviation of the set of sample proportions in this simulation. Compare with the theoretical mean and standard deviation of the sampling distribution with sample size 200. Create and study the dot plot for this data set.

Theoretical Sample: Mean: ______________ Standard Deviation: _____________

Your Sample: Mean: ______________ Standard Deviation: ______________

g) How is the plot of the sampling distribution from part f different from the plot in part b? How do the mean and standard deviation compare?

h) What percentage of the 200 sample proportions fall within 0.114 of 0.20 (or approximately 2 standard deviations)? How does this compare with the answer to part e?

i) You should notice that these distributions follow an approximately normal distribution. The Empirical Rule states how much of the data will fall within 1, 2, and 3 standard deviations of the mean. Restate the rule:

Empirical Rule: In a normal distribution with mean μ and standard deviation σ

- ____% of the observations fall within 1 standard deviation (1σ) of the mean (μ),
- ____% of the observations fall within 2 standard deviations (2σ) of the mean (μ), and
- ____% of the observations fall within 3 standard deviation (3σ) of the mean (μ).

j) Do your answers to parts e and h agree with the Empirical Rule? Explain.

This leads us to an important result in statistics: the **Central Limit Theorem (CLT) for a Sample Proportion**:

Choose a simple random sample of size n from a large population with population parameter p having some characteristic of interest. Then the sampling distribution of the sample proportion $\hat{p}$ is approximately normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$. This approximation becomes more and more accurate as the sample size n increases, and it is generally considered valid if the population is much larger than the sample, i.e. $np \geq 10$ and $n(1 - p) \geq 10$.

k) How might this theorem be helpful? What advantage does this theorem provide in determining the likelihood of events?

Pennies Learning Task

Mathematical Goals

- Understand sample distributions of sample means through simulation
- Develop the formulas for the mean and standard deviation of the sampling distribution of a sample means
- Discover the Central Limit Theorem for a sample means
- Use sample means to estimate population values
- Conduct simulations of random sampling to gather sample means.
- Explain what the results mean about variability in a population.

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

MGSE9-12.S.IC. 2 Decide if a specified model is consistent with results from a given data-generating process, e.g. using simulation.

MGSE9-12.S.IC. 4 Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling

Standards for Mathematical Practice

1. Reason abstractly and quantitatively
2. Construct viable arguments and critique the reasoning of others
3. Model with mathematics

Introduction

This task is similar to last except that it focuses on the Central Limit Theorem as it applies to sample means. The activity is done quite often in AP Statistics courses and some version of it is found in many statistics texts. The task presented here is largely based on the version found in Yates, D. S., Moore, D. S., & Starnes, D. S. (2003). The Practice of Statistics: TI-83/89 Graphing Calculator Enhanced (2^{nd} ed.). New York: W.H. Freeman. Because Part 2 is completely based on student data, it is not possible to provide solutions; instead, general ideas of what is to be expected are provided.

Materials

- Graphing calculator or statistical software package
- Masking tape
- Small Post-It notes (5 per student)
- Pennies (students will need to bring these in before the task begins)

Part 1:
Sampling Distribution of a Sample Mean from a Normal Population

1) The scores of individual students on the ACT entrance exam have a normal distribution with mean 18.6 and standard deviation 5.9.

a) Use your calculator to simulate the scores of 25 randomly selected students who took the ACT. Record the mean and standard deviations of these 25 people in the table below. Repeat, simulating the scores of 100 people. (To do this, use the following command: $randNorm(\mu,\sigma, n) \rightarrow L_1$.)

	Mean	**Standard Deviation**
Population	18.6	5.9
25 people		
100 people		

Solution: Actual answers will vary.

b) As a class, compile the means for the sample of 25 people.

- Determine the mean and standard deviation of this set of means. That is, calculate $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$.

Answers will vary.

- How does the mean of the sample means compare with the population mean?

The mean of the sample means should be close to the population mean 18.6.

- How does the standard deviation of the sample means compare with the population standard deviation?

The standard deviation of the sample means should be approximately $\frac{5.9}{\sqrt{25}} = 1.18$.

c) Describe the plot of this set of means. How does the plot compare with the normal distribution?

Solution: The plot should be approximately a normal distribution.

d) As a class, compile the means for the sample of 100 people.

- Determine the mean and standard deviation of this set of means. That is, calculate $\mu_{\overline{x}}$ and $\sigma_{\overline{x}}$.

Answers will vary.

- How does the mean of the sample means compare with the population mean?

The mean of the sample means should be close to the population mean 18.6.

- How does the standard deviation of the sample means compare with the population standard deviation?

The standard deviation of the sample means should be approximately $\frac{5.9}{\sqrt{100}} = 0.59$

e) Describe the plot of this set of means. How does the plot compare with the normal distribution?

Solution: The plot should be approximately a normal distribution.

f) Determine formulas for the mean of the sample means, $\mu_{\overline{x}}$, and the standard deviation of the sample means, $\sigma_{\overline{x}}$. Compare with a neighbor.

Solution: Students may engage in much conjecturing and discussion here; however, this is a fact they learned previously2. $\mu_{\overline{x}} = \mu$ ***and*** $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$.

g) Just as we saw with proportions, the sample mean is an *unbiased estimator* of the population mean.

The Sampling Distribution of a Sample Mean: Choose a simple random sample of size *n* from a large population with mean μ and standard deviation σ. Then:

- o The mean of the sampling distribution of $\overline{x}$ is ____.
- o The standard deviation of the sampling distribution of $\overline{x}$ is ___________.

$$\mu_{\overline{x}} = \mu \text{ and } \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}.$$

h) Again, we must be cautious about when we use the formula for the standard deviation of $\bar{x}$. What was the rule when we looked at proportions? It is the same here.

Solution: The population must be at least 10 times the sample size.

i) Put these latter facts together with your response to part b to complete the following statement:

> Choose a simple random sample of size *n* from a population that has a **normal distribution** with mean μ and standard deviation σ. Then the sample mean $\bar{x}$ has a ____________distribution with mean ________ and standard deviation ____________.

Solution: normal; $\mu_{\bar{x}} = \mu$; $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

This problem centered on a population that was known to be normally distributed. What about populations that are not normally distributed? Can we still use the facts above? Let's investigate!

Part 2:

How old are your pennies?[1]

Year	*Age*	*Frequency*
2012	0	3
2011	1	6
2010	2	3
...	...	...

a) Make a frequency table of the year and the age of the 25 pennies you brought to class like the one on the right. Do this on your own paper. Find the average age of the 25 pennies. Record the mean age as $\bar{x}(25)$.

$\bar{x}(25)$ = ___________

b) Put your 25 pennies in a cup or bag and randomly select 5 pennies. Find the average age of the 5 pennies in your sample, and record the mean age as $\bar{x}(5)$. Replace the pennies in the cup, and repeat.

$\bar{x}_1(5)=$ ________________

$\bar{x}_2(5)=$ ________________

c) Repeat the process two more times, this time removing 10 pennies at a time. Calculate the average age of the sample of 10 pennies and record as $\bar{x}(10)$.

$\bar{x}_1(10)=$ ________________

$\bar{x}_2(10)=$ ________________

---

[1] The *Workshop Statistics* books provide an alternatives to actually working with pennies; however, it is not possible to recreate the methods here. It requires assigning three-digit numbers to penny ages and using a random number generator to "sample" pennies.

Teaching Note: The masking tape number lines should be set up before class or while the students are sampling and finding averages. As soon as students finishing calculating their averages, they should begin making their dotplots. If some students finish before others, they can help others with plotting their coins.

d) Make a penny dotplot of your 25 pennies.

Look at the shape of the final dotplot. Describe the distribution of the pennies' ages.

What to Expect: The dotplot will be quite right skewed, with many more young pennies than middle or old pennies. Of course, the range will be from 0 to the age of the oldest penny brought in.

e) Make a second penny dotplot of the means for the sample size 5.

What is the shape of the dotplot for the distribution of $\bar{x}(5)$? How does it compare with the original distribution of pennies' ages?

What to Expect: This plot shows less skewing and less

f) Make a third penny dotplot of the means for the sample size 10.

What is the shape of the dotplot for the distribution of $\bar{x}(10)$? How does it compare with the previous plots?

What to Expect: Again, this plot shows less skewing and less variability. It may begin looking somewhat approximately normally distributed.

g) Finally, make a fourth penny dotplot. Use this to record the means for the sample size of 25.

Describe the shape of the dotplot.

What to Expect: The means are clustered together rather closely and the plot looks approximately normally distributed.

h) Now, find the mean and standard deviation of the sample means. Record your results in the chart below.

	Mean	**Standard Deviation**	**Shape of the Distribution**
“Population”			
Samples of 5			
Samples of 10			
Samples of 25			

i) Previously, we stated that if samples were taken from a normal distribution, that the mean and standard deviation of the sampling distribution of sample means was also normal with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. In this activity, we did not begin with a normal distribution. However, compare the means for the samples of 5, 10, and 25 with the overall mean of the pennies. Then compare the standard deviations with the standard deviation of all the pennies. Do these formulas appear to hold despite the population of penny ages being obviously non-normal? Explain.

Answers may vary. The calculations of the standard deviation for each sampling distribution should be relatively close to the calculation of the “population” standard deviation divided by the square root of n. The formulas should appear to hold.

j) Suppose that the U.S. Department of Treasury estimated that the average age of pennies presently in circulation is 12.25 years with a standard deviation of 9.5. Determine the theoretical means and standard deviations for the sampling distributions of sample size 5, 10, and 25.

	Mean	**Standard Deviation**
Population	12.25	9.5
Samples of 5	*12.25*	$\frac{9.5}{\sqrt{5}} \approx 4.2485$
Samples of 10	*12.25*	$\frac{9.5}{\sqrt{10}} \approx 3.0042$
Samples of 25	*12.25*	$\frac{9.5}{\sqrt{25}} = 1.9$

k) This brings us to the **Central Limit Theorem (CLT) for Sample Means:**

> Choose a simple random sample of size n from any population, regardless of the original shape of the distribution, with mean μ and finite standard deviation σ. When n is large,

the sampling distribution of the sample mean $\overline{x}$ is approximately normal with mean ______ and standard deviation ________.

Note: The statement “when *n* is large” seems a bit ambiguous. A good rule of thumb is that the sample size should be at least 30, as we can see in the dotplots above. When the sample sizes were sample, i.e. 5 and 10, the plots were still quite right skewed.

Solution: $\mu_{\overline{x}} = \mu$ ***and*** $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

Pennies Learning Task

Name ____

Date____________________

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

MGSE9-12.S.IC. 2 Decide if a specified model is consistent with results from a given data-generating process, e.g. using simulation.

MGSE9-12.S.IC. 4 Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling

Standards for Mathematical Practice

1. Reason abstractly and quantitatively

2. Construct viable arguments and critique the reasoning of others

3. Model with mathematics

Part 1:
Sampling Distribution of a Sample Mean from a Normal Population

1) The scores of individual students on the ACT entrance exam have a normal distribution with mean 18.6 and standard deviation 5.9.

a) Use your calculator to simulate the scores of 25 randomly selected students who took the ACT. Record the mean and standard deviations of these 25 people in the table below. Repeat, simulating the scores of 100 people. (To do this, use the following command: *randNorm*$(\mu,\sigma, n) \rightarrow L_1$.)

	Mean	**Standard Deviation**
Population	18.6	5.9
25 people		
100 people		

b) As a class, compile the means for the sample of 25 people.

- Determine the mean and standard deviation of this set of means. That is, calculate $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$.

- How does the mean of the sample means compare with the population mean?

- How does the standard deviation of the sample means compare with the population standard deviation?

c) Describe the plot of this set of means. How does the plot compare with the normal distribution?

d) As a class, compile the means for the sample of 100 people.

- Determine the mean and standard deviation of this set of means. That is, calculate $\mu_{\overline{x}}$ and $\sigma_{\overline{x}}$.

- How does the mean of the sample means compare with the population mean?

- How does the standard deviation of the sample means compare with the population standard deviation?

e) Describe the plot of this set of means. How does the plot compare with the normal distribution?

f) Determine formulas for the mean of the sample means, $\mu_{\overline{x}}$, and the standard deviation of the sample means, $\sigma_{\overline{x}}$. Compare with a neighbor.

g) Just as we saw with proportions, the sample mean is an *unbiased estimator* of the population mean.

The Sampling Distribution of a Sample Mean: Choose a simple random sample of size *n* from a large population with mean μ and standard deviation σ. Then

- The mean of the sampling distribution of $\overline{x}$ is ____ and

- The standard deviation of the sampling distribution of $\overline{x}$ is ___________.

h) Again, we must be cautious about when we use the formula for the standard deviation of $\overline{x}$. What was the rule when we looked at proportions? It is the same here.

i) Put these latter facts together with your response to part b to complete the following statement:

> Choose a simple random sample of size n from a population that has a **normal distribution** with mean μ and standard deviation σ. Then the sample mean $\overline{x}$ has a _____________distribution with mean _________ and standard deviation _____________.

This problem centered on a population that was known to be normally distributed. What about populations that are not normally distributed? Can we still use the facts above? Let's investigate!

Part 2:
How Old are Your Pennies?[2]

a) Make a frequency table of the year and the age of the 25 pennies you brought to class like the one on the right. Do this on your own paper. Find the average age of the 25 pennies. Record the mean age as $\overline{x}(25)$.

$\overline{x}(25)$ = ___________

Year	***Age***	***Frequency***
2012	0	3
2011	1	6
2010	2	3
...	...	...

b) Put your 25 pennies in a cup or bag and randomly select 5 pennies. Find the average age of the 5 pennies in your sample, and record the mean age as $\overline{x}(5)$. Replace the pennies in the cup, and repeat.

$\overline{x}_1(5)=$ ________________

$\overline{x}_2(5)=$ ________________

c) Repeat the process two more times, this time removing 10 pennies at a time. Calculate the average age of the sample of 10 pennies and record as $\overline{x}(10)$.

$\overline{x}_1(10)=$ ________________

$\overline{x}_2(10)=$ ________________

d) Make a penny dotplot of your 25 pennies.

Look at the shape of the final dotplot. Describe the distribution of the pennies' ages.

e) Make a second penny dotplot of the means for the sample size 5.

---

[2] The *Workshop Statistics* books provide an alternatives to actually working with pennies; however, it is not possible to recreate the methods here. It requires assigning three-digit numbers to penny ages and using a random number generator to "sample" pennies.

What is the shape of the dotplot for the distribution of $\bar{x}(5)$? How does it compare with the original distribution of pennies' ages?

f) Make a third penny dotplot of the means for the sample size 10.

What is the shape of the dotplot for the distribution of $\bar{x}(10)$? How does it compare with the previous plots?

g) Finally, make a fourth penny dotplot. Use this to record the means for the sample size of 25.

Describe the shape of the dotplot.

h) Now, find the mean and standard deviation of the sample means. Record your results in the chart below.

	Mean	**Standard Deviation**	**Shape of the Distribution**
"Population"			
Samples of 5			
Samples of 10			
Samples of 25			

i) Previously, we stated that if samples were taken from a normal distribution, that the mean and standard deviation of the sampling distribution of sample means was also normal with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. In this activity, we did not begin with a normal distribution. However, compare the means for the samples of 5, 10, and 25 with the overall mean of the pennies. Then compare the standard deviations with the standard deviation of all the pennies. Do these formulas appear to hold despite the population of penny ages being obviously non-normal? Explain.

j) Suppose that the U.S. Department of Treasury estimated that the average age of pennies presently in circulation is 12.25 years with a standard deviation of 9.5. Determine the theoretical means and standard deviations for the sampling distributions of sample size 5, 10, and 25.

	Mean	**Standard Deviation**
Population	12.25	9.5
Samples of 5		
Samples of 10		
Samples of 25		

k) This brings us to the **Central Limit Theorem (CLT) for Sample Means:**

Choose a simple random sample of size n from any population, regardless of the original shape of the distribution, with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normal with mean ______ and standard deviation ________.

Note: The statement "when n is large" seems a bit ambiguous. A good rule of thumb is that the sample size should be at least 30, as we can see in the dotplots above. When the sample sizes were sample, i.e. 5 and 10, the plots were still quite right skewed.

Gettysburg Address (Spotlight Task)

Mathematical Goals

- Distinguish between a population distribution, a sample data distribution, and a sampling distribution of a statistic.
- Understand sampling distributions of sample means through simulation.
- Use sample means to estimate population values and understand the role of variability
- Conduct simulations of self-selection sampling to gather sample means.
- Conduct simulations of random sampling to gather sample means.
- Understand bias, in particular, sampling bias.
- Understand the role of randomness in sampling.
- Understand the role of sample size in sampling.

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population

MGSE9-12.S.IC. 4 Use data from a population mean or proportion;

Standards for Mathematical Practice

1. **Make sense of problems and persevere in solving them**
2. **Construct viable arguments and critique the reasoning of others**
3. **Model with mathematics**
4. **Attend to precision**

Introduction
This task is adapted from a sampling distributions activity developed by Christine Franklin and Gary Kader. The vocabulary in bold at the beginning of the task is important. Emphasize to students that should always define the words in bold for the scenario they are considering. The concept of a sampling distribution is difficult for students. Part of this difficulty is not being comfortable with distinguishing the three type of distributions: population, sample data distribution (distribution on one sample), and the sampling distribution (distribution of a statistics from repeated samples). The students first sample by using their eyes and will see from the simulated sampling distribution that their results are biased; i.e, by using their eyes, they tend to overestimate the actual population mean. By taking random samples, they will see that bias is minimized. That is the role of randomization – to minimize bias. The role of sample size is to minimize variability allowing more precision in predicting the actual population mean.

Technology Note: If you have access to Fathom or Mini-Tab and can input the Gettysburg Address word list prior to class, this activity can go quite quickly.

Materials

- Graphing calculator or statistical software package
- Copies of the Gettysburg Address
- Word list from the Gettysburg Address (at end of student edition)

Authorship of literary works is often a topic for debate. One method researchers use to decide who was the author is to look at word patterns from known writing of the author and compare these findings to an unknown work. To help us understand this process we will analyze the length of the words in the Gettysburg Address, authored by Abraham Lincoln.

The Gettysburg Address

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we cannot dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under god, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.

Statistical Question: How long are the words in the Gettysburg address?

In this problem, the **variable** of interest is the length of a word in the Gettysburg address, which is a discrete, quantitative variable. Note that the word lengths vary and that the population of all word lengths has a distribution, a mean and a standard deviation. We desire to estimate the <u>mean</u> word length. This is our **parameter** of interest. A parameter is a numerical summary of the population. In statistics, we select a sample and hope that the **distribution of the sample** is similar to the **distribution of the population.**

We could examine each and every word in the Gettysburg Address but to make the most efficient use of our time, we will instead take a subset of the words. We are considering this passage a **population** of words, and the 10 words you selected are considered a **sample** from this population. In most studies, we do not have access to the entire population and can only consider results for a sample from that population. The goal is to learn something about a very large population (e.g., all American adults, all American registered voters) by studying a sample. The key is in carefully selecting the sample so that the results in the sample are **representative** of the larger population (i.e., has the same characteristics). The **population** is the entire collection of observational units that we are interested in examining. A **sample** is a subset of observational units from the population. Keep in mind that these are objects or people, and then we need to determine what variable we want to measure about these entities and then the parameter of interest. In this scenario, we will use the sample mean, referred to as a **statistic,** to predict the population mean (the parameter).

1. Circle 10 words you think are “representative” of the word length using your eyes. Record the words and word lengths below.

Word Number	Word	Length
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

2. Summarize your data on "Word Length" in a dotplot, a graphical representation of the distribution of sample data. Compare your sample data distribution to least 2 classmates' distributions. Are they the same or different?

1 2 3 4 5 6 7 8 9 10 11

Dotplot for Sample of Word Lengths

3. Determine the mean for your sample of words.

4. Let's examine the distribution of the sample means. That is, each of you has a sample and each sample has a mean. Did you each get the same mean?

5. The sample means vary from one sample to another. This illustrates one of the most important ideas in statistics – the concept that a sample statistic (in this case, the sample mean) varies from one sample to another. Let's summarize the variation in our sample means in a dotplot.

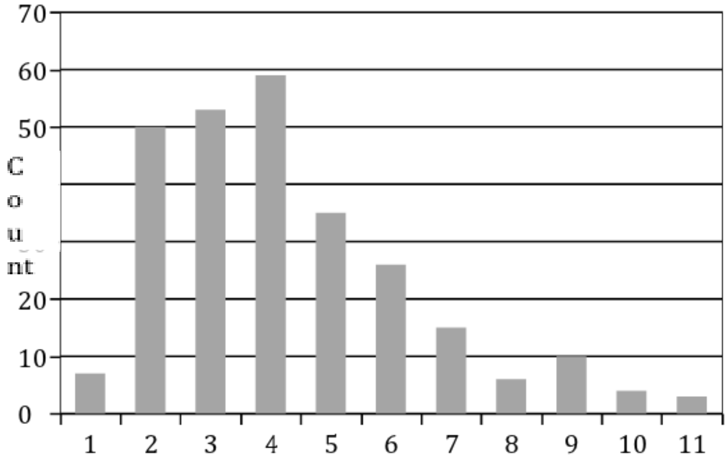
Record class sample means here:						

1 2 3 4 5 6 7 8 9 10 11

Mean word length/samples of size 10

6. The above plot summarizes the sample-to-sample variation in our sample means. It represents a simulated **sampling distribution of the sample mean.** Estimate the average (mean) of this distribution of sample means. Estimate the variation from this average as measured by the standard deviation.

7. How do these sample means compare with the actual population mean?
Below is the distribution of word lengths for the entire population as represented by a histogram.



Word Lengths

Estimate the mean and standard deviation of this population distribution. Also, comment on the distribution shape.

8. Using the histogram for the population of 268 words in the Gettysburg Address, it was calculated that the population Mean Word Length is 4.3. That is, the population mean is $\mu = 4.3$. The population standard deviation is 2.12 and the distribution shape is right skewed. How do the sample means in the dotplot in part 5 compare to 4.3? Is your estimated average for the simulated sampling distribution close or noticeably higher or lower to the population mean?

Samples that are "self-selected," tend to produce **biased** results. In this case, in our self-selected samples, the means from the samples tend to overestimate the population means. Your eyes are drawn to the larger words. That is, the sampling method produces samples with means generally larger than the population mean. This is called **sampling bias.**

Self-selected samples tend to produce sample distributions that are not representative of the population. In statistics, randomness is introduced into the sampling procedure in order to produce samples that tend to be representative of the population. In simple random sampling, each sample of a given size has the same chance (probability) of being selected. This "fairness" in selection tends to produce unbiased sample results. We want to select random samples of size n and to examine the behavior of the sample means from sample to sample.

How do we select a simple random sample of 10 words from the Gettysburg address?

On the last page of this task is a list of the words from the Gettysburg address. Note that there are 268 words, and each word is assigned a number from 1 (001) to 268. Many calculators will produce random integers; however, they are not guaranteed to all be different. To be safe, we will generate 20 random integers between 1 and 268 and use the first 10 distinct integers. To generate 20 "random numbers" between 1 and 268 on a TI-84, enter the following commands:

```
MATH → PRB → randInt(→ ENTER
randInt(1,268,20) → STO → L1 → ENTER
```

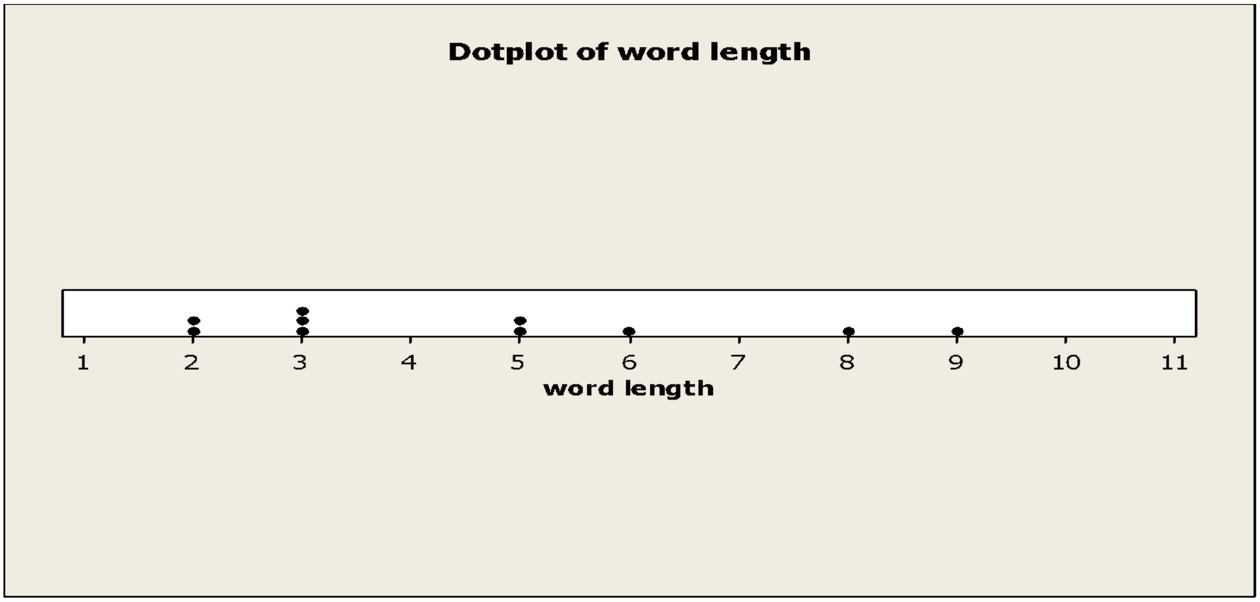
- You can find the list of numbers STAT → EDIT → L1

Suppose the above sequence of commands produced the following random integers:
{33 152 114 93 248 170 233 98 114 22 224 37 88 214 7 45 25 118 25 4}

Then our sample would consist of the following words and associated word lengths:

Word Number	33	152	114	93	248	170	233	98	22	224
Word	are	but	we	is	birth	dedicated	shall	proper	to	devotion
Word Length	3	3	3	2	5	9	5	6	2	8

The dotplot for these data follows. Also, the sample mean word length is 5.



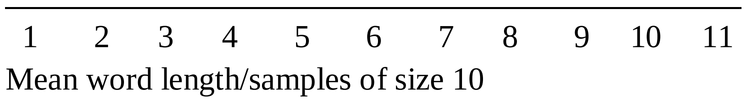
9. Use your calculator to randomly generate 20 integers between 1 and 268 and use these to select a Simple Random Sample (SRS) of 10 words. Record these below and find the sample mean.

Word #	Random Integer	Word	Length		Word #	Random Integer	Word	Length
1					6			
2					7			
3					8			
4					9			
5					10			

The sample Mean word length is ____________

10. Summarize the variation in the sample means by creating a dotplot displaying the sample means from the different simple random samples we have generated.

Record the class sample means here



11. Based on the dotplot, estimate the average of this simulated sampling distribution of sample means. How do the means from our samples compare with the population mean of 4.3? Based on the dotplot, do simple random samples appear to produce unbiased results? Explain.

    We expect the average of the simulated sampling distribution to be close to the population mean of 4.3. Yes, random sampling appears to produce unbiased results since not overestimating or underestimating population mean on average.

12. Based on the dotplot, estimate the standard deviation of this simulated sampling distribution of sample means. How does this standard deviation of the sample means compare with the population standard deviation of 2.12? Is it similar, smaller, or larger?
    Expect the standard deviation of the simulate sampling distribution to be around 0.67, thus smaller than the population standard deviation of 2.12.
13. What distribution shape do you observe emerging for the simulated sampling distribution of the sample mean? How does this shape compare to the shape of the population distribution?

    Expect to see unimodal symmetric bell shaped distribution, compared to the right skewed population distribution.

14. What would happen to the behavior of the sampling distribution of the sample mean is the sample size was increase to 20? Make your prediction about shape, mean, and standard deviation.

    Expect the shape and mean to be similar to sampling distribution for sample size of 10. However, expect the standard to become smaller than 0.67. We believe that by increasing the sample size to 20, we will decrease the variability.

15. Repeat parts 9-13. In part 9, you should generate 30-40 integers to guarantee 20 unique integers. Do your results confirm your predictions in part 14?

Answers will vary.

Gettysburg address word list (page 1)

Number	Word	Length	Number	Word	Length	Number	Word	Length
001	Four	4	046	Nation	6	091	Live.	4
002	Score	5	047	So	2	092	It	2
003	And	3	048	Conceived	9	093	Is	2
004	Seven	5	049	And	3	094	Altogether	10
005	Years	5	050	So	2	095	Fitting	7
006	Ago.	3	051	Dedicated,	9	096	And	3
007	Our	3	052	Can	3	097	Proper	6
008	Fathers	7	053	Long	4	098	That	4
009	Brought	7	054	Endure.	5	099	We	2
010	Forth	5	055	We	2	100	Should	6
011	Upon	4	056	Are	3	101	Do	2
012	This	4	057	Met	3	102	This.	4
013	Continent	9	058	On	2	103	But	3
014	A	1	059	A	1	104	In	2
015	New	3	060	Great	5	105	A	1
016	Nation:	6	061	Battlefield	11	106	Larger	6
017	Conceived	9	062	Of	2	107	Sense,	5
018	In	2	063	That	4	108	We	2
019	Liberty,	7	064	War.	3	109	Cannot	6
020	And	3	065	We	2	110	Dedicate,	8
021	Dedicated	9	066	Have	4	111	We	2
022	To	2	067	Come	4	112	Cannot	6
023	The	3	068	To	2	113	Consecrate,	10
024	Proposition	11	069	Dedicate	8	114	We	2
025	That	4	070	A	1	115	Cannot	6

026	All	3		071	Portion	7		116	Hallow	6
027	Men	3		072	Of	2		117	This	4
028	Are	3		073	That	4		118	Ground.	6
029	Created	7		074	Field	5		119	The	3
030	Equal.	5		075	As	2		120	Brave	5
031	Now	3		076	A	1		121	Men,	3
032	We	2		077	Final	5		122	Living	6
033	Are	3		078	Resting	7		123	And	3
034	Engaged	7		079	Place	5		124	Dead,	4
035	In	2		080	For	3		125	Who	3
036	A	1		081	Those	5		126	Struggled	9
037	Great	5		082	Who	3		127	Here	4
038	Civil	5		083	Here	4		128	Have	4
039	War,	3		084	Gave	4		129	Consecrated	11
040	Testing	7		085	Their	5		130	It,	2
041	Whether	7		086	Lives	5		131	Far	3
042	That	4		087	That	4		132	Above	5
043	Nation,	6		088	That	4		133	Our	3
044	Or	2		089	Nation	6		134	Poor	4
045	Any	3		090	Might	5		135	Power	5

Gettysburg address word list (page 2)

Number	Word	Length		Number	Word	Length		Number	Word	Length
136	To	2		181	Have	4		226	We	2
137	Add	3		182	Thus	4		227	Here	4
138	Or	2		183	Far	3		228	Highly	6
139	Detract.	7		184	So	2		229	Resolve	7
140	The	3		185	Nobly	5		230	That	4
141	World	5		186	Advanced.	8		231	These	5
142	Will	4		187	It	2		232	Dead	4
143	Little	6		188	Is	2		233	Shall	5
144	Note,	4		189	Rather	6		234	Not	3
145	Nor	3		190	For	3		235	Have	4
146	Long	4		191	Us	2		236	Died	4
147	Remember	8		192	Here	4		237	In	2
148	What	4		193	To	2		238	Vain,	4
149	We	2		194	Be	2		239	That	4
150	Say	3		195	Dedicated	9		240	This	4
151	Here,	4		196	To	2		241	Nation,	6
152	But	3		197	The	3		242	Under	5

153	It	2		198	Great	5		243	God,	3
154	Can	3		199	Task	4		244	Shall	5
155	Never	5		200	Remaining	9		245	Have	4
156	Forget	6		201	Before	6		246	A	1
157	What	4		202	Us,	2		247	New	3
158	They	4		203	That	4		248	Birth	5
159	Did	3		204	From	4		249	Of	2
160	Here.	4		205	These	5		250	Freedom,	7
161	It	2		206	Honored	7		251	And	3
162	Is	2		207	Dead	4		252	That	4
163	For	3		208	We	2		253	Government	10
164	Us	2		209	Take	4		254	Of	2
165	The	3		210	Increased	9		255	The	3
166	Living,	6		211	Devotion	8		256	People,	6
167	Rather,	6		212	To	2		257	By	2
168	To	2		213	That	4		258	The	3
169	Be	2		214	Cause	5		259	People,	6
170	Dedicated	9		215	To	2		260	For	3
171	Here	4		216	Which	5		261	The	3
172	To	2		217	They	4		262	People,	6
173	The	3		218	Gave	4		263	Shall	5
174	Unfinished	10		219	The	3		264	Not	3
175	Work	4		220	Last	4		265	Perish	6
176	Which	5		221	Full	4		266	From	4
177	They	4		222	Measure	7		267	The	3
178	Who	3		223	Of	2		268	Earth.	5
179	Fought	6		224	Devotion,	8				
180	Here	4		225	That	4				

Gettysburg Address Spotlight Task

Authorship of literary works is often a topic for debate. One method researchers use to decide who was the author is to look at word patterns from known writing of the author and compare these findings to an unknown work. To help us understand this process we will analyze the length of the words in the Gettysburg Address, authored by Abraham Lincoln.

The Gettysburg Address

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we cannot dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under god, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.

Statistical Question: How long are the words in the Gettysburg address?

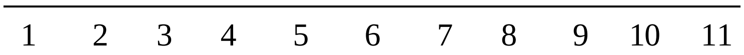
In this problem, the **variable** of interest is the length of a word in the Gettysburg address, which is a discrete, quantitative variable. Note that the word lengths vary and that the population of all word lengths has a distribution, a mean and a standard deviation. We desire to estimate the mean word length. This is our **parameter** of interest. A parameter is a numerical summary of the population. In statistics, we select a sample and hope that the **distribution of the sample** is similar to the **distribution of the population.**

We could examine each and every word in the Gettysburg Address but to make the most efficient use of our time, we will instead take a subset of the words. We are considering this passage a **population** of words, and the 10 words you selected are considered a **sample** from this population. In most studies, we do not have access to the entire population and can only consider results for a sample from that population. The goal is to learn something about a very large population (e.g., all American adults, all American registered voters) by studying a sample. The key is in carefully selecting the sample so that the results in the sample are **representative** of the larger population (i.e., has the same characteristics). The **population** is the entire collection of observational units that we are interested in examining. A **sample** is a subset of observational units from the population. Keep in mind that these are objects or people, and then we need to determine what variable we want to measure about these entities and then the parameter of interest. In this scenario, we will use the sample mean, referred to as a **statistic,** to predict the population mean (the parameter).

1. Circle 10 words you think are "representative" of the word length using your eyes. Record the words and word lengths below.

Word Number	Word	Length
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

2. Summarize your data on "Word Length" in a dotplot, a graphical representation of the distribution of sample data. Compare your sample data distribution to least 2 classmates' distributions. Are they the same or different?



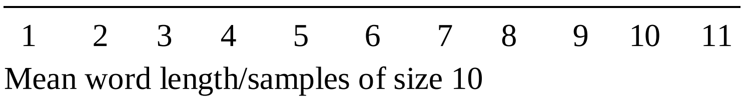
Dotplot for Sample of Word Lengths

3. Determine the mean for your sample of words.

4. Let’s examine the distribution of the sample means. That is, each of you has a sample and each sample has a mean. Did you each get the same mean?

5. The sample means vary from one sample to another. This illustrates one of the most important ideas in statistics – the concept that a sample statistic (in this case, the sample mean) varies from one sample to another. Let’s summarize the variation in our sample means in a dotplot.

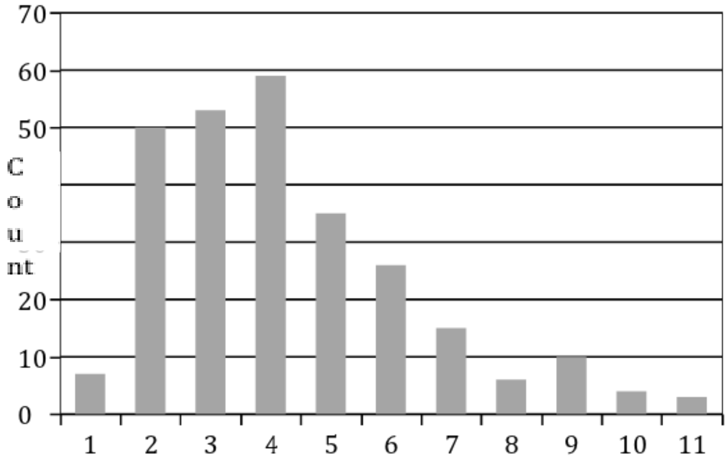
<table>
<tr><th colspan="6">Record the class sample means here</th></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>



6. The above plot summarizes the sample-to-sample variation in our sample means. It represents a simulated **sampling distribution of the sample mean.** Estimate the average (mean) of this distribution of sample means. Estimate the variation from this average as measured by the standard deviation.

7. How do these sample means compare with the actual population mean?

Below is the distribution of word lengths for the entire population as represented by a histogram.



Word Lengths

Estimate the mean and standard deviation of this population distribution. Also, comment on the distribution shape.

8. Using the histogram for the population of 268 words in the Gettysburg Address, it was calculated that the population Mean Word Length is 4.3. That is, the population mean is $\mu = 4.3$. The population standard deviation is 2.12 and the distribution shape is right skewed. How do the sample means in the dotplot in part 5 compare to 4.3? Is your estimated average for the simulated sampling distribution close or noticeably higher or lower to the population mean?

Samples that are "self-selected," tend to produce **biased** results. In this case, in our self-selected samples, the means from the samples tend to overestimate the population means. Your eyes are drawn to the larger words. That is, the sampling method produces samples with means generally larger than the population mean. This is called **sampling bias.**

Self-selected samples tend to produce sample distributions that are not representative of the population. In statistics, randomness is introduced into the sampling procedure in order to produce samples that tend to be representative of the population. In simple random sampling, each sample of a given size has the same chance (probability) of being selected. This "fairness" in selection tends to produce unbiased sample results. We want to select random samples of size n and to examine the behavior of the sample means from sample to sample.

How do we select a simple random sample of 10 words from the Gettysburg address?

On the last page of this task is a list of the words from the Gettysburg address. Note that there are 268 words, and each word is assigned a number from 1 (001) to 268. Many calculators will produce random integers; however, they are not guaranteed to all be different. To be safe, we will generate 20 random integers between 1 and 268 and use the first 10 distinct integers. To generate 20 "random numbers" between 1 and 268 on a TI-84, enter the following commands:

```
MATH → PRB → randInt(→ ENTER
randInt(1,268,20) → STO → L1 → ENTER
```

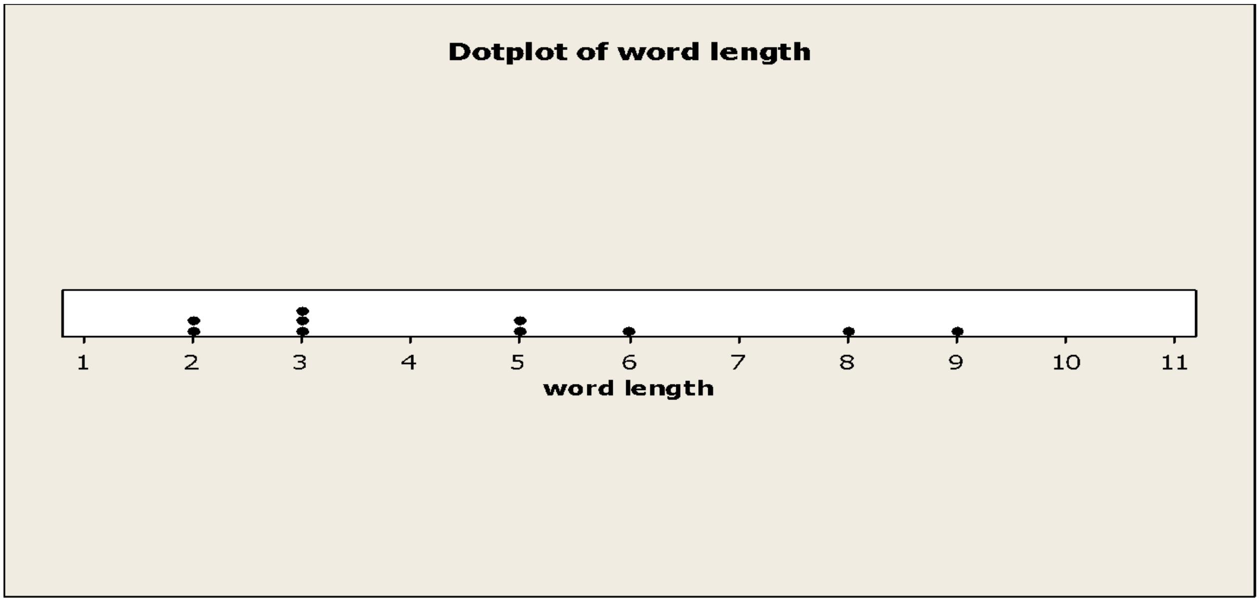
➢ You can find the list of numbers STAT → EDIT → L1

Suppose the above sequence of commands produced the following random integers:
{33 152 114 93 248 170 233 98 114 22 224 37 88 214 7 45 25 118 25 4}

Then our sample would consist of the following words and associated word lengths:

Word Number	33	152	114	93	248	170	233	98	22	224
Word	are	but	we	is	birth	dedicated	shall	proper	to	devotion
Word Length	3	3	3	2	5	9	5	6	2	8

The dotplot for these data follows. Also, the sample mean word length is 5.



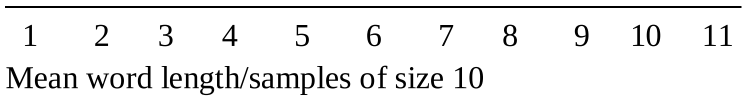
9. Use your calculator to randomly generate 20 integers between 1 and 268 and use these to select a Simple Random Sample (SRS) of 10 words. Record these below and find the sample mean.

Word #	**Random Integer**	**Word**	**Length**		**Word #**	**Random Integer**	**Word**	**Length**
1					**6**			
2					**7**			
3					**8**			
4					**9**			
5					**10**			

The sample Mean word length is ____________

10. Summarize the variation in the sample means by creating a dotplot displaying the sample means from the different simple random samples we have generated.

Record the class sample means here					



11. Based on the dotplot, estimate the average of this simulated sampling distribution of sample means. How do the means from our samples compare with the population mean of 4.3? Based on the dotplot, do simple random samples appear to produce unbiased results? Explain.

12. Based on the dotplot, estimate the standard deviation of this simulated sampling distribution of sample means. How does this standard deviation of the sample means compare with the population standard deviation of 2.12? Is it similar, smaller, or larger?

13. What distribution shape do you observe emerging for the simulated sampling distribution of the sample mean? How does this shape compare to the shape of the population distribution?

14. What would happen to the behavior of the sampling distribution of the sample mean is the sample size was increase to 20? Make your prediction about shape, mean, and standard deviation.

15. Repeat parts 9-13. In part 9, you should generate 30-40 integers to guarantee 20 unique integers. Do your results confirm your predictions in part 14?

Gettysburg address word list (page 1)

Number	Word	Length	Number	Word	Length	Number	Word	Length
001	Four	4	046	Nation	6	091	Live.	4
002	Score	5	047	So	2	092	It	2
003	And	3	048	Conceived	9	093	Is	2
004	Seven	5	049	And	3	094	Altogether	10
005	Years	5	050	So	2	095	Fitting	7
006	Ago.	3	051	Dedicated,	9	096	And	3
007	Our	3	052	Can	3	097	Proper	6
008	Fathers	7	053	Long	4	098	That	4
009	Brought	7	054	Endure.	5	099	We	2
010	Forth	5	055	We	2	100	Should	6
011	Upon	4	056	Are	3	101	Do	2
012	This	4	057	Met	3	102	This.	4
013	Continent	9	058	On	2	103	But	3
014	A	1	059	A	1	104	In	2
015	New	3	060	Great	5	105	A	1
016	Nation:	6	061	Battlefield	11	106	Larger	6
017	Conceived	9	062	Of	2	107	Sense,	5
018	In	2	063	That	4	108	We	2
019	Liberty,	7	064	War.	3	109	Cannot	6
020	And	3	065	We	2	110	Dedicate,	8
021	Dedicated	9	066	Have	4	111	We	2
022	To	2	067	Come	4	112	Cannot	6
023	The	3	068	To	2	113	Consecrate,	10
024	Proposition	11	069	Dedicate	8	114	We	2
025	That	4	070	A	1	115	Cannot	6
026	All	3	071	Portion	7	116	Hallow	6
027	Men	3	072	Of	2	117	This	4
028	Are	3	073	That	4	118	Ground.	6
029	Created	7	074	Field	5	119	The	3
030	Equal.	5	075	As	2	120	Brave	5
031	Now	3	076	A	1	121	Men,	3
032	We	2	077	Final	5	122	Living	6
033	Are	3	078	Resting	7	123	And	3
034	Engaged	7	079	Place	5	124	Dead,	4
035	In	2	080	For	3	125	Who	3
036	A	1	081	Those	5	126	Struggled	9
037	Great	5	082	Who	3	127	Here	4
038	Civil	5	083	Here	4	128	Have	4
039	War,	3	084	Gave	4	129	Consecrated	11

040	Testing	7		085	Their	5		130	It,	2
041	Whether	7		086	Lives	5		131	Far	3
042	That	4		087	That	4		132	Above	5
043	Nation,	6		088	That	4		133	Our	3
044	Or	2		089	Nation	6		134	Poor	4
045	Any	3		090	Might	5		135	Power	5

Gettysburg address word list (page 2)

Number	Word	Length		Number	Word	Length		Number	Word	Length
136	To	2		181	Have	4		226	We	2
137	Add	3		182	Thus	4		227	Here	4
138	Or	2		183	Far	3		228	Highly	6
139	Detract.	7		184	So	2		229	Resolve	7
140	The	3		185	Nobly	5		230	That	4
141	World	5		186	Advanced.	8		231	These	5
142	Will	4		187	It	2		232	Dead	4
143	Little	6		188	Is	2		233	Shall	5
144	Note,	4		189	Rather	6		234	Not	3
145	Nor	3		190	For	3		235	Have	4
146	Long	4		191	Us	2		236	Died	4
147	Remember	8		192	Here	4		237	In	2
148	What	4		193	To	2		238	Vain,	4
149	We	2		194	Be	2		239	That	4
150	Say	3		195	Dedicated	9		240	This	4
151	Here,	4		196	To	2		241	Nation,	6
152	But	3		197	The	3		242	Under	5
153	It	2		198	Great	5		243	God,	3
154	Can	3		199	Task	4		244	Shall	5
155	Never	5		200	Remaining	9		245	Have	4
156	Forget	6		201	Before	6		246	A	1
157	What	4		202	Us,	2		247	New	3
158	They	4		203	That	4		248	Birth	5
159	Did	3		204	From	4		249	Of	2
160	Here.	4		205	These	5		250	Freedom,	7
161	It	2		206	Honored	7		251	And	3
162	Is	2		207	Dead	4		252	That	4
163	For	3		208	We	2		253	Government	10
164	Us	2		209	Take	4		254	Of	2
165	The	3		210	Increased	9		255	The	3
166	Living,	6		211	Devotion	8		256	People,	6

167	Rather,	6		212	To	2		257	By	2
168	To	2		213	That	4		258	The	3
169	Be	2		214	Cause	5		259	People,	6
170	Dedicated	9		215	To	2		260	For	3
171	Here	4		216	Which	5		261	The	3
172	To	2		217	They	4		262	People,	6
173	The	3		218	Gave	4		263	Shall	5
174	Unfinished	10		219	The	3		264	Not	3
175	Work	4		220	Last	4		265	Perish	6
176	Which	5		221	Full	4		266	From	4
177	They	4		222	Measure	7		267	The	3
178	Who	3		223	Of	2		268	Earth.	5
179	Fought	6		224	Devotion,	8				
180	Here	4		225	That	4				

How Confident Are You? Learning Task

Mathematical Goals

- Develop an understanding of margin of error through confidence intervals
- Calculate confidence intervals for sample proportions
- Calculate confidence intervals for sample means

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population

MGSE9-12.S.IC. 4 Use data from a population mean or proportion; develop a margin of error through the use of simulation models for random sampling

MGSE9-12.S.IC.6 Evaluate reports based on data. ***For example, determining quantitative or categorical data; collection methods; biases or flaws in data.***

Standards for Mathematical Practice

1. **Make sense of problems and persevere in solving them**
2. **Construct viable arguments and critique the reasoning of others**
3. **Model with mathematics**
4. **Attend to precision**

Introduction

This task develops students’ ability to use statistical inference in the form of confidence intervals. The standard does not address student understanding of confidence intervals; however to deepen the students’ understanding of margin of error it seems necessary to have students investigate confidence intervals since margins of error are integral to finding confidence intervals. Part one reviews the empirical rule, a topic first learned in the beginning of the unit. Building from their knowledge of the empirical rule, the basics of confidence intervals are introduced. The task then engages students with confidence intervals for proportions and means. The students will also use simulations to generate data and use the calculator to construct confidence intervals.

Materials

- Calculator with statistical capabilities
- Standard normal tables (optional)

Sampling Distribution & Confidence Intervals

- ***Statistical Inference** provides methods for drawing conclusions about a population from sample data.*
- The two major types of statistical inference are **confidence intervals** and tests of significance. We will focus only on understanding and using confidence intervals. Tests of significance are a major focus of Statistics courses.
- We cannot be certain that our conclusions are correct – a different sample might lead to different conclusions. Statistical inference uses the language of probability to express the strength of our conclusions. Probability allows us to take chance variation into account and to correct our judgment by calculation.
- We will use confidence intervals to help us estimate the population mean and population proportion from a sample.
- A confidence level gives the probability that the interval will capture the true parameter value in repeated samples. That is, the confidence level is the success rate for the method.

Example 1:

a. Suppose that the mean SAT Math score for seniors in Georgia was 550 with a standard deviation of 50 points. Consider a simple random sample of 100 Georgia seniors who take the SAT. Describe the distribution of the sample mean scores.

Solution: The distribution, given that there are more than 1000 seniors who take the SAT, should be approximately normal.

b. What are the mean and standard deviation of this sampling distribution?

Solutions: The mean is 550 and the standard deviation is 50/(sqrt 100) = 5.

c. Use the Empirical Rule to determine between what two scores 68% of the data falls, 95% of the data falls, and 99.7% of the data falls.

- 68% of the data is between ______ and ______ and is ______ standard deviations away from the mean.
- 95% of the data is between ______ and ______ and is ______ standard deviations away from the mean.
- 99.7% of the data is between ______ and ______ and is ______ standard deviations away from the mean.

Solutions:

- ***68% of the data would be one standard deviation on either side of the mean, so (550 – 5, 550 + 5) = (545, 555)***
- ***95% of the data would be within two standard deviations of the mean or (540, 560).***
- ***99.7% of the data is within three standard deviations or (535, 565).***

For the 95% interval, this means that in 95% of all samples of 100 students from this population, the mean score for the sample will fall within _______ standard deviations of the true population mean or _________ points from the mean. ***2 and 10***

In the above problem, we took the mean and added/subtracted a certain number of standard deviations. That is, we calculated $\mu_{\bar{x}} \pm 2\sigma_{\bar{x}} = \mu_{\bar{x}} \pm 2\frac{\sigma}{\sqrt{n}}$ for the 95% interval and

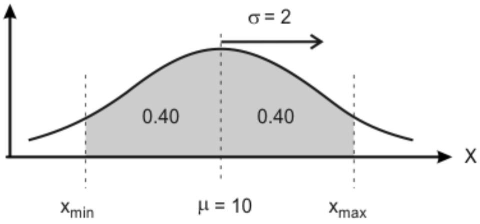
$\mu_{\bar{x}} \pm 3\sigma_{\bar{x}} = \mu_{\bar{x}} \pm 3\frac{\sigma}{\sqrt{n}}$ for the 99.7% interval.

The Basics of Confidence Intervals

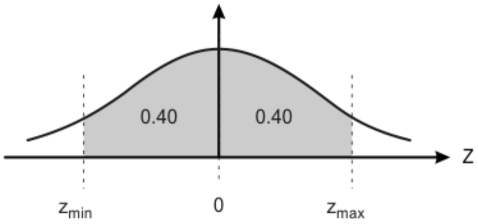
Example 2:

You are told that a population is distributed normally with a mean, μ= 10, and a standard deviation, σ = 2. In this case, 80% of the data fall between what two values?

From our discussions about normal curves, you should recognize the situation to be as follows:



The area on each side of the mean = 0.80/2 = 0.40, and you are asked to find x_{min} and x_{max}. You also know that this can be represented as follows using the standard normal curve:



Using your Standard Normal Chart you can find the values of z:

z_{min} = ______ **-1.28** z_{max} = ______ **1.28**

*You can also find these values by using the invNorm() function on your calculator. This is found on the TI-83/84: 2nd VARS [DISTR], ARROW DOWN to select 3:**invNormal(**enter z and then press ENTER.

Using the rearranged form of the z-score equation, we find that:

$$\pm z = \frac{x-\mu}{\sigma}$$
$$\pm z\sigma = x - \mu$$
$$x = \mu \pm z\sigma$$

$x_{max} = \mu + 1.28\sigma = 10 + 1.28(2) = 12.56$
$x_{min} = \mu - 1.28\sigma = 10 - 1.28(2) = 7.44$

In other words, for a population distributed normally with μ= 10 and σ= 2:
80% of the data fall in the range between _*7.44*__ and ____*12.56*____. You just figured out your first confidence interval!

In the language of confidence intervals, you could say that for the situation described above:

- You are 80% confident that any point chosen at random will fall between ___*7.44*________ and ___*12.56*________.
- Your 80% confidence interval is [___*7.44*______, _____*12.56*_____]

Confidence intervals are expressed in percentages, such as the 80% confidence interval or the 95% confidence interval. The percentage values 80% and 95% are known as the confidence level.

Above, we knew the population mean, but in practice, we often do not. So we take samples and create confidence intervals as a method of estimating the true value of the parameter. When we find a 95% confidence interval, we believe with 95% of all samples the true parameter falls within our interval. However, we must accept that 5% of all samples will give intervals which do not include the parameter. Every confidence interval takes the same shape:

estimate ± margin of error.

The margin of error has two main components: the number of standard deviations from the mean (i.e. the *z*-score) and the standard deviation. (**Margin of error =** $z \times \sigma$.)

Since we do not usually know the details of a population parameter (e.g. mean and standard deviation), we must use estimates of these values. So our margin of error becomes $m = z(\sigma_{estimate})$. Therefore, the confidence interval becomes

$$\textit{estimate} \pm \textit{margin of error} \rightarrow \textit{estimate} \pm z \times \sigma_{estimate}$$

Example 3:
What is the 80% confidence interval for a population for which $\mu_{\text{estimate}} = 36.20$ and $\sigma_{\text{estimate}} = 12.30$?

We know from the last example that for an 80% confidence level, the values of z are:
$z_{\text{min}} = -1.28$, $z_{\text{max}} = 1.28$

And using the rearranged form of the z-score equation again, we find that:

$x_{\text{max}} = \mu_{\text{estimate}} + 1.28 * \sigma_{\text{estimate}} = 36.2 + 1.28\ (12.3) = 51.944$

$x_{\text{min}} = \mu_{\text{estimate}} - 1.28 * \sigma_{\text{estimate}} = 36.2 – 1.28(12.3) = 20.456$

In other words, for a population distributed normally with $\mu_{\text{estimate}} = 36.2$ and $\sigma_{\text{estimate}} = 12.3$:
The 80% confidence interval is [_*20.456*________, ___*51.944*___].

What are the Common Confidence Intervals?

The *z*-score used in the confidence interval depends on how *confident* one wants to be. There are a few common levels of confidence used in practice: 90%, 95%, and 99%.

Confidence Level	Corresponding z-score	Corresponding Interval $\mu \pm z \times \sigma$
90%	*1.645*	$\mu \pm 1.645\sigma$
95%	*1.96*	$\mu \pm 1.96\sigma$
99%	*2.576*	$\mu \pm 2.576\sigma$

Example 4:
The admissions director at Big City University proposed using the IQ scores of current students as a marketing tool. The university agrees to provide him with enough money to administer the IQ tests to 50 students. The director gives the test to a simple random sample of 50 of the university's 5000 freshman. The mean IQ score for the sample is $\bar{x}$ =112. The IQ test that he administered is known to have a standard deviation of σ = 15. What is the 95% confidence interval for the population mean μ? That is what can the director say about the mean score μ of the population of all 5000 freshman?

$$x = 112 \pm 1.96(2.12)$$

$$x = 107.84, 116.16$$

Let's make sure you recognize what's going on here:

- We were told we knew σ,but not μ → So we estimated μ by sampling and getting $\bar{x}$

The 95% confidence interval is [___*107.84*_______ , _____*116.16*______]
Meaning this interval "captures" the true μ in about 95% of all possible samples.

Example 5:
President's Approval Ratings: Part 1 (Confidence Intervals for Proportions)

Read the article below:

Excerpt from http://www.cnn.com/2005/POLITICS/12/19/bush.poll/

Poll: Iraq speeches, election don't help Bush

Tuesday, December 20, 2005; Posted: 12:56 a.m. EST (05:56 GMT)

CNN -- President Bush's approval ratings do not appear to have changed significantly, despite a number of recent speeches he's given to shore up public support for the war in Iraq and its historic elections on Thursday.

A CNN/USA Today Gallup poll conducted over the weekend found his approval rating stood at 41 percent, while more than half, or 56 percent, disapprove of how the president is handling his job. A majority, or 52 percent, say it was a mistake to send troops to Iraq, and 61 percent say they disapprove of how he is handling Iraq specifically. The margin of error was plus or minus 3 percentage points.

...

The poll was nearly split, 49 percent to 47 percent, between those who thought the U.S. will either "definitely" or "probably" win, and those who said the U.S. will lose. That said, 69 percent of those polled expressed optimism that the U.S. can win the war. The margin of error for how respondents assessed the war was plus or minus 4.5 percentage points.

...

Although half those polled said that a stable government in Iraq was likely within a year, 62 percent said Iraqi forces were unlikely to ensure security without U.S. assistance. And 63 percent said Iraq was unlikely to prevent terrorists from using Iraq as a base. The margin of error on questions pertaining to troop duration in Iraq, as well as the country's future, was plus or minus 3 percentage points.

The poll interviewed 1,003 adult Americans and found that the public has also grown more skeptical about Bush's key arguments in favor of the war. Compared with two years ago, when 57 percent considered Iraq a part of the war on terrorism, 43 percent think so now. In the weekend poll, 55 percent said they view the war in Iraq as separate from the war on terror. The margin of error on this line of questioning was plus or minus 3 percentage points.

On the domestic front, 56 percent of those polled say they disapprove of how Bush is handling the economy; by contrast, 41 percent approve. The margin of error was plus or minus 3 percentage points.

The president may find support for his call to renew the Patriot Act. Forty-four percent said they felt the Patriot Act is about right, and 18 percent said it doesn't go far enough. A third of respondents say they believe the Patriot Act has gone too far in restricting people's civil liberties to investigate suspected terrorism.

Nearly two-thirds said they are not willing to sacrifice civil liberties to prevent terrorism, as compared to 49 percent saying so in 2002. The margin of error was plus or minus 4.5 percentage points for those questions.

a. In the article, Bush’s approval rating was 41 percent with a margin of error of plus or minus 3 points. Write this statement as a confidence interval.

(0.38, 0.44)

b. The article does not state the confidence level for the ratings. Use the margin of error to determine the standard error of the estimate (standard deviation) if a 95% confidence level was used. What is the standard error if a 99% confidence level was used?

$\boldsymbol{ME = z \times \sigma_{est}}$

95%: $\boldsymbol{0.03 = 1.96 \times \sigma_{est}}$ $\boldsymbol{\sigma_{est} = 0.0153}$

99%: $\boldsymbol{0.03 = 2.576 \times \sigma_{est}}$ $\boldsymbol{\sigma_{est} = 0.0116}$

c. What is the formula for calculating the standard deviation of sample proportions? In practice, we do not know the true population parameter. Instead, we must estimate using the sample proportion. Rewrite this equation using the symbol for the sample proportion. This is called the standard error of the sample proportion.

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ SE = $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

d. If the Gallup poll in the article used a 95% confidence level, what was the sample size? What if the poll used a 99% confidence level? (Use the answers to b and c. Recall that the sample proportion was 41%.) Refer to the article. How many people were polled?

95%	99%
n=1034	***N=1798***

What does that tell you about the confidence level employed? ***The article says that 1003 people were surveyed so the confidence interval was likely 95%***

e. Assume the pollsters were undecided about how many citizens to include in their poll. However, they knew that they wanted the margin of error to be 3 percent or less and they wanted to be 95% confident in their results.

That is, they wanted $z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le m$ or $1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le .03$.

What value of *p* will always give the greatest margin of error? Make a conjecture. Explain your reasoning. Compare with a neighbor. Call this value *p**.

$\boldsymbol{p(1-p) gives\ the\ greatest\ value}$

$\mathbf{0.5(0.50 = \ 0.25}$

$\mathbf{0.4(0.6) = \ 0.24}$

$\mathbf{0.51(0.49) = 0.2499}$

So p= 0.5 will give the greatest value

Use the value of *p** you found to determine the sample size needed by the pollsters.

95%	99%
1068 people	***1858 people***

f. Increase the sample sizes you found in part *e* by 500. Still using your p^*, determine the margin of error in both the 95% and 99% cases for intervals with these new sample sizes. How different is the margin of error?
(**Margin of error** = $z \times SE_{estimate}$.)

95%	99%
0.0247 Since the original ME was 0.03, it is half a percent less	***0.0265 Since the original ME was 0.03 it is now less than half a percent less***

What do you notice? ***The margin of error does not change much when the sample size is increased by 500; however, it does decrease.***

How Confident Are You? Learning Task

Name ____Date_________________________

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population

MGSE9-12.S.IC. 4 Use data from a population mean or proportion; develop a margin of error through the use of simulation models for random sampling

MGSE9-12.S.IC.6 Evaluate reports based on data. *For example, determining quantitative or categorical data; collection methods; biases or flaws in data.*

Standards for Mathematical Practice

1. **Make sense of problems and persevere in solving them**
2. **Construct viable arguments and critique the reasoning of others**
3. **Model with mathematics**
4. **Attend to precision**

Sampling Distribution & Confidence Intervals

- ***Statistical Inference** provides methods for drawing conclusions about a population from sample data.*
- The two major types of statistical inference are **confidence intervals** and tests of significance. We will focus only on understanding and using confidence intervals. Tests of significance are a major focus of Statistics courses.
- We cannot be certain that our conclusions are correct – a different sample might lead to different conclusions. Statistical inference uses the language of probability to express the strength of our conclusions. Probability allows us to take chance variation into account and to correct our judgment by calculation.
- We will use confidence intervals to help us estimate the population mean and population proportion from a sample.
- A confidence level gives the probability that the interval will capture the true parameter value in repeated samples. That is, the confidence level is the success rate for the method.

Example 1:

a. Suppose that the mean SAT Math score for seniors in Georgia was 550 with a standard deviation of 50 points. Consider a simple random sample of 100 Georgia seniors who take the SAT. Describe the distribution of the sample mean scores.

b. What are the mean and standard deviation of this sampling distribution?

c. Use the Empirical Rule to determine between what two scores 68% of the data falls, 95% of the data falls, and 99.7% of the data falls.

- 68% of the data is between ______ and ______ and is ______ standard deviations away from the mean.
- 95% of the data is between ______ and ______ and is ______ standard deviations away from the mean.
- 99.7% of the data is between ______ and ______ and is ______ standard deviations away from the mean.

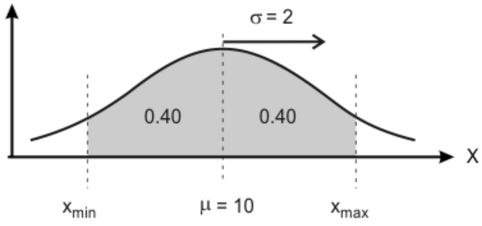
For the 95% interval, this means that in 95% of all samples of 100 students from this population, the mean score for the sample will fall within _______ standard deviations of the true population mean or _________ points from the mean.

In the above problem, we took the mean and added/subtracted a certain number of standard deviations. That is, we calculated $\mu_{\bar{x}} \pm 2\sigma_{\bar{x}} = \mu_{\bar{x}} \pm 2\frac{\sigma}{\sqrt{n}}$ for the 95% interval and

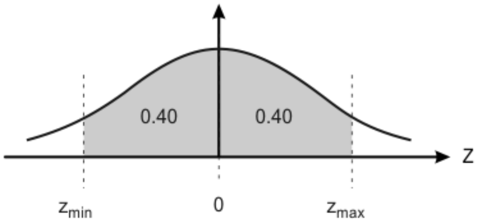
$\mu_{\bar{x}} \pm 3\sigma_{\bar{x}} = \mu_{\bar{x}} \pm 3\frac{\sigma}{\sqrt{n}}$ for the 99.7% interval.

The Basics of Confidence Intervals

Example 2:
You are told that a population is distributed normally with a mean, μ= 10, and a standard deviation, σ = 2. In this case, 80% of the data fall between what two values?
From our discussions about normal curves, you should recognize the situation to be as follows:



The area on each side of the mean = 0.80/2 = 0.40, and you are asked to find x_{min} and x_{max}. You also know that this can be represented as follows using the standard normal curve:



Using your Standard Normal Chart you can find the values of z:

z_{min} = ______ z_{max} = ______

*You can also find these values by using the invNorm() function on your calculator. This is found on the TI-83/84: 2nd VARS [DISTR], ARROW DOWN to select 3:**invNormal(**enter z and then press ENTER.

Using the rearranged form of the z-score equation, we find that:

$$\pm z = \frac{x-\mu}{\sigma}$$
$$\pm z\sigma = x - \mu$$
$$x = \mu \pm z\sigma$$

$x_{max} = \mu + 1.28\sigma = 10 + 1.28(2) = 12.56$
$x_{min} = \mu - 1.28\sigma = 10 - 1.28(2) = 7.44$

In other words, for a population distributed normally with μ= 10 and σ= 2:
80% of the data fall in the range between ________ and _______. You just figured out your first confidence interval!

In the language of confidence intervals, you could say that for the situation described above:

- You are 80% confident that any point chosen at random will fall between ___________ and ___________.
- Your 80% confidence interval is [_________, __________]

Confidence intervals are expressed in percentages, such as the 80% confidence interval or the 95% confidence interval. The percentage values 80% and 95% are known as the confidence level.

Above, we knew the population mean, but in practice, we often do not. So we take samples and create confidence intervals as a method of estimating the true value of the parameter. When we find a 95% confidence interval, we believe with 95% of all samples the true parameter falls within our interval. However, we must accept that 5% of all samples will give intervals which do not include the parameter. Every confidence interval takes the same shape:

$$\textbf{\textit{estimate}} \pm \textbf{\textit{margin of error}}.$$

The margin of error has two main components: the number of standard deviations from the mean (i.e. the z-score) and the standard deviation. (**Margin of error = $z \times \sigma$.**)

Since we do not usually know the details of a population parameter (e.g. mean and standard deviation), we must use estimates of these values. So our margin of error becomes $m = z(\sigma_{estimate})$. Therefore, the confidence interval becomes

$$\textit{estimate} \pm \textit{margin of error} \rightarrow \textit{estimate} \pm z \times \sigma_{estimate}$$

Example 3:
What is the 80% confidence interval for a population for which $\mu_{\text{estimate}} = 36.20$ and $\sigma_{\text{estimate}} = 12.30$?

We know from the last example that for an 80% confidence level, the values of z are:
$z_{\text{min}} = -1.28$, $z_{\text{max}} = 1.28$

And using the rearranged form of the z-score equation again, we find that:

$x_{\text{max}} = \mu_{\text{estimate}} + 1.28^* \sigma_{\text{estimate}} =$

$x_{\text{min}} = \mu_{\text{estimate}} - 1.28^* \sigma_{\text{estimate}} =$

In other words, for a population distributed normally with $\mu_{\text{estimate}} = 36.2$ and $\sigma_{\text{estimate}} = 12.3$:
The 80% confidence interval is [________, ______].

What are the Common Confidence Intervals?

The *z*-score used in the confidence interval depends on how *confident* one wants to be. There are a few common levels of confidence used in practice: 90%, 95%, and 99%.

Confidence Level	**Corresponding z-score**	**Corresponding Interval** $\mu \pm z \times \sigma$
90%		
95%		
99%		

Example 4:
The admissions director at Big City University proposed using the IQ scores of current students as a marketing tool. The university agrees to provide him with enough money to administer the IQ tests to 50 students. The director gives the test to a simple random sample of 50 of the university's 5000 freshman. The mean IQ score for the sample is $\bar{x}$ =112. The IQ test that he administered is known to have a standard deviation of σ = 15. What is the 95% confidence interval for the population mean μ? That is what can the director say about the mean score μ of the population of all 5000 freshman?

Let's make sure you recognize what's going on here:

- We were told we knew σ,but not μ → So we estimated μ by sampling and getting $\bar{x}$

The 95% confidence interval is [__________ , ___________]
Meaning this interval "captures" the true μ in about 95% of all possible samples.

Example 5:
President's Approval Ratings: Part 1 (Confidence Intervals for Proportions)

Read the article below:

Excerpt from http://www.cnn.com/2005/POLITICS/12/19/bush.poll/

Poll: Iraq speeches, election don't help Bush

Tuesday, December 20, 2005; Posted: 12:56 a.m. EST (05:56 GMT)

CNN -- President Bush's approval ratings do not appear to have changed significantly, despite a number of recent speeches he's given to shore up public support for the war in Iraq and its historic elections on Thursday.

A CNN/USA Today Gallup poll conducted over the weekend found his approval rating stood at 41 percent, while more than half, or 56 percent, disapprove of how the president is handling his job. A majority, or 52 percent, say it was a mistake to send troops to Iraq, and 61 percent say they disapprove of how he is handling Iraq specifically. The margin of error was plus or minus 3 percentage points.

...

The poll was nearly split, 49 percent to 47 percent, between those who thought the U.S. will either "definitely" or "probably" win, and those who said the U.S. will lose. That said, 69 percent of those polled expressed optimism that the U.S. can win the war. The margin of error for how respondents assessed the war was plus or minus 4.5 percentage points.

...

Although half those polled said that a stable government in Iraq was likely within a year, 62 percent said Iraqi forces were unlikely to ensure security without U.S. assistance. And 63 percent said Iraq was unlikely to prevent terrorists from using Iraq as a base. The margin of error on questions pertaining to troop duration in Iraq, as well as the country's future, was plus or minus 3 percentage points.

The poll interviewed 1,003 adult Americans and found that the public has also grown more skeptical about Bush's key arguments in favor of the war. Compared with two years ago, when 57 percent considered Iraq a part of the war on terrorism, 43 percent think so now. In the weekend poll, 55 percent said they view the war in Iraq as separate from the war on terror. The margin of error on this line of questioning was plus or minus 3 percentage points.

On the domestic front, 56 percent of those polled say they disapprove of how Bush is handling the economy; by contrast, 41 percent approve. The margin of error was plus or minus 3 percentage points.

The president may find support for his call to renew the Patriot Act. Forty-four percent said they felt the Patriot Act is about right, and 18 percent said it doesn't go far enough. A third of respondents say they believe the Patriot Act has gone too far in restricting people's civil liberties to investigate suspected terrorism.

> Nearly two-thirds said they are not willing to sacrifice civil liberties to prevent terrorism, as compared to 49 percent saying so in 2002. The margin of error was plus or minus 4.5 percentage points for those questions.

a. In the article, Bush's approval rating was 41 percent with a margin of error of plus or minus 3 points. Write this statement as a confidence interval.

b. The article does not state the confidence level for the ratings. Use the margin of error to determine the standard error of the estimate (standard deviation) if a 95% confidence level was used. What is the standard error if a 99% confidence level was used?

c. What is the formula for calculating the standard deviation of sample proportions? In practice, we do not know the true population parameter. Instead, we must estimate using the sample proportion. Rewrite this equation using the symbol for the sample proportion. This is called the standard error of the sample proportion.

$\sigma_{\hat{p}} =$ SE =

d. If the Gallup poll in the article used a 95% confidence level, what was the sample size? What if the poll used a 99% confidence level? (Use the answers to b and c. Recall that the sample proportion was 41%.) Refer to the article. How many people were polled?

95%	99%

What does that tell you about the confidence level employed?

e. Assume the pollsters were undecided about how many citizens to include in their poll. However, they knew that they wanted the margin of error to be 3 percent or less and they wanted to be 95% confident in their results.

That is, they wanted $z*\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} \le m$ or $1.96\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} \le .03$.

What value of p will always give the greatest margin of error? Make a conjecture. Explain your reasoning. Compare with a neighbor. Call this value $p*$.

Use the value of $p*$ you found to determine the sample size needed by the pollsters.

95%	99%

g. Increase the sample sizes you found in part *e* by 500. Still using your $p*$, determine the margin of error in both the 95% and 99% cases for intervals with these new sample sizes. How different is the margin of error?
(**Margin of error =** $z \times SE_{estimate}$.)

95%	99%

What do you notice?

Culminating Task: Final Grades

Math Goals:

- Understand how to choose summary statistics that are appropriate to the characteristics of the data distribution, such as the shape of the distribution or the existence of outliers.

- Understand how the normal distribution uses area to make estimates of frequencies which can be expressed as probabilities and recognizing that only some data are well described by a normal distribution.

- Understand how to make inferences about a population using a random sample

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, ~~mean absolute deviation~~, standard deviation) of two or more different data sets.

MGSE9-12.S.ID.4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

MGSE9-12.S.IC.1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population**.**

MGSE9-12.S.IC.2 Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation. *For example, a model says a spinning coin falls heads up with probability 0.5. Would a result of 5 tails in a row cause you to question the model?*

Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them
2. Reason abstractly and quantitatively
3. Construct viable arguments and critique the reasoning of others
4. Model with mathematics
5. Use appropriate tools strategically
6. Attend to precision
7. Look for and make use of structure
8. Look for and express regularity in repeated reasoning

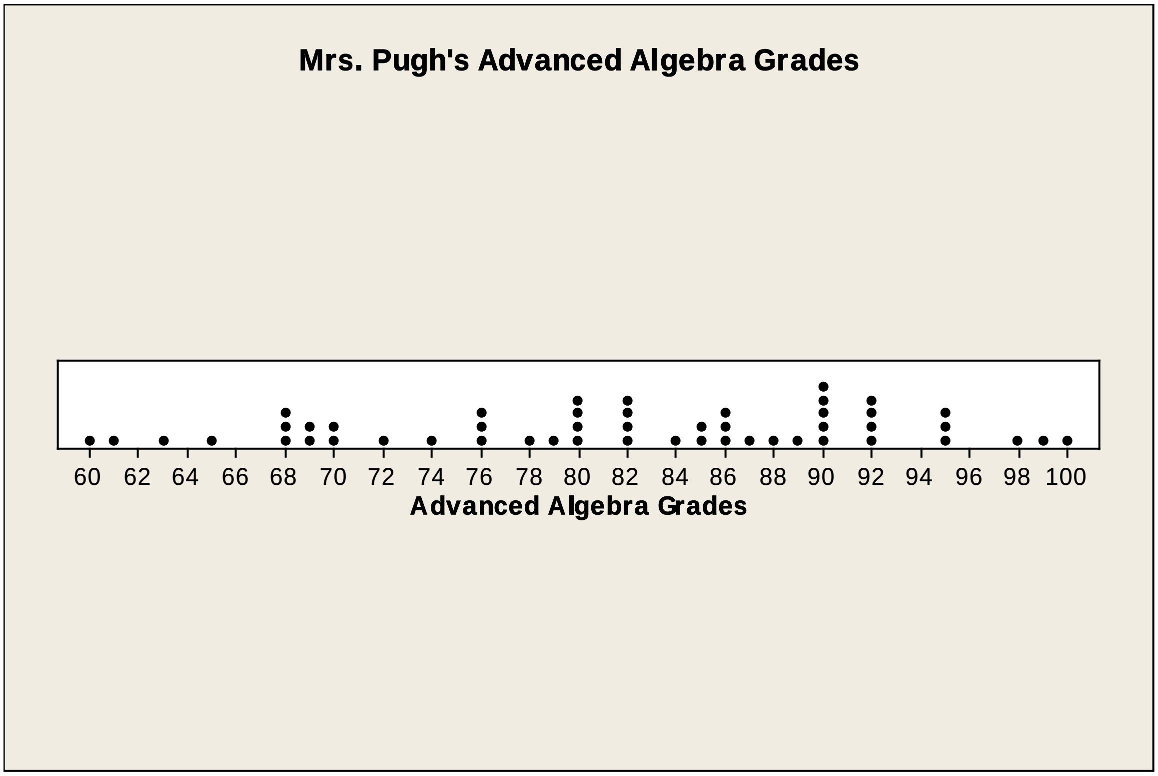
Introduction

Students will be given data and asked to describe the data. They will then be asked to estimate percentages using the data.

Materials

- Graphing calculator or statistical software
- Table of normal probabilities (optional)

I. Mrs. Pugh recorded all of the semester grades (on a scale of 100) of her 50 Advanced Algebra students in the following dotplot:



What is the mean and standard deviation of the distribution of individual grades?

1. $\bar{x}$ = ________________ 2. σ = ________________

Mean is 81.72 (81.7) and the standard deviation is 10.37 (10.4).

Identify the following **and** label the dotplot above

3. $\mu \pm 1\sigma$ = ________________ and ________________ ***71.3 and 92.1***

4. $\mu \pm 2\sigma =$ _______________ and _______________ ***60.9 and 102.5***

5. $\mu \pm 3\sigma =$ _______________ and _______________ ***50.5 and 112.9***

6. Does the data indicate a normal distribution? ______________ ***yes***

7. Why or why not? ***66% of the data fell within one standard deviation of the mean, 98% fell within two standard deviations of the mean and 100% fell within 3 standard deviations of the mean. This is very close to the Empirical Rule for a normal distribution***

8. We want to know the probability that a student selected randomly from her class would have an "A" (90 or above) in her class. Find the probability using two different methods.

Method 1: 15 dots out of the 50 are 90 or better so fifteen divided by fifty is 0.3

Method 2: Since the Empirical Rule applies, find a z-score of 0.7981. Use the chart you get a probability of 1 - .7881 = 0.2119

II. Mrs. Pugh went on vacation and could not be reached. Before she left, she turned in her individual student grades to her principal. The parent of the student who made a 68 in the class called and insisted to know the class average of her child's class by the end of the day. Unfortunately, the principal could not retrieve the exact class average because he only had individual student scores, but he told the parent that he could give her a range of scores that the class average would most likely be located within by the end of the day.

He first took a random sample of five students and calculated the average of the five students.

Simulate what the principal did below:

Seed your calculator by typing "1" "STO" "rand" on your calculator. To make sure you did this correctly, type randint(1,50) and you should get the number 38.

9. Simulate what the principal did by typing randint(1,50,5) to get 5 numbers. What are the numbers? ***43, 12, 24, 35, 5***

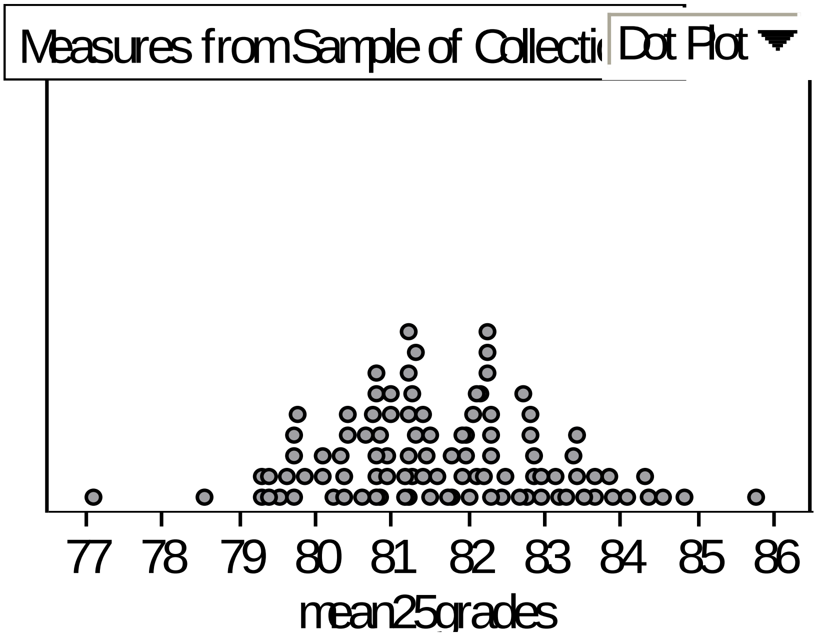
10. Locate these values on the dotplot. 1 corresponds to the lowest test score, and 50 corresponds to the highest test score. What are the associated test scores? ***92, 72, 82, 89, 68***

11. Find the average of these 5 test scores. ***80.6***

12. Is this sample mean the same as the population mean? ***no***

13. Why or why not? ***It was close, but the small sample size, 5, we would not expect it to be exactly accurate***

III. Although using a sample size of 5 gave him a good idea what the class average was, he wanted to give the parent a smaller range in which the average could be located. He figured that Mrs. Pugh had at least 25 students in each of her classes. He took 50 random samples of size 25, calculated their means, and recorded them in a dotplot.



Estimate the mean and standard deviation of this dotplot

14. $\bar{x}$ = ***around 81.5*** 15. σ = ***between 1.43 and 2.15***

The principal knows that it's unlikely a value is beyond two standard deviations away from the mean. Using the mean and standard deviation that you just approximated above for the class average of 25 students, find the values:

16. Mean – 2*standard deviation = ***about 77.5***

17. Mean + 2*standard deviation = ***about 85.5***

18. Identify the range of scores that would fall between 2 standard deviations **on the dotplot**.

19. He called the parent and told her that he was very confident that the class average was between

________________ and ________________. ***77.5 and 85.5***

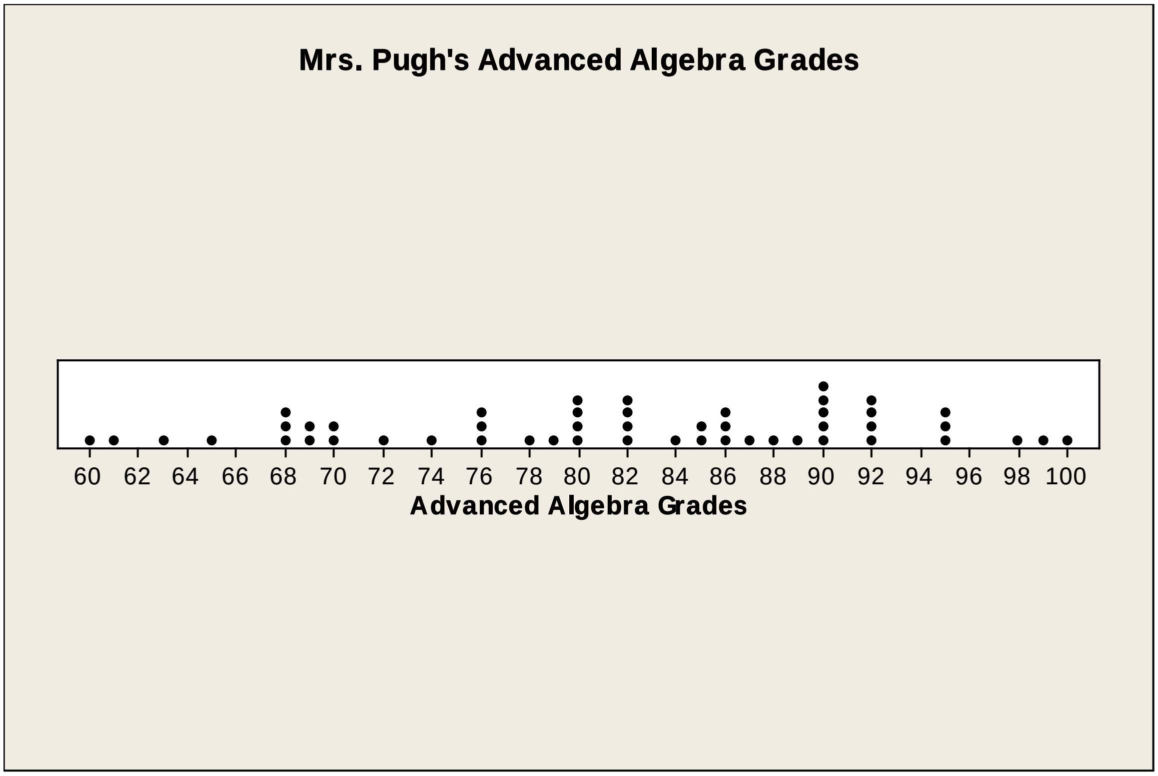
The parent replied that the principal’s response was not possible because she knew of 2 other students in Ms. Pugh’s class that made a 60 and 61.

20. What did the principal explain to the parent? ***Individual results vary. The parent asked for the class average and not the distribution of individual scores. It would be impossible for the class average to be a failing average if there were 25 students in each of Ms. Pugh’s classes because there were only 11 failing scores. Also, the distribution of the sample means has less variance and a smaller standard deviation than the distribution of individual scores. That is why a failing score is not located on the interval for the class average.***

Culminating Task: Final Grades

Name ____
Date_______________________

I. Mrs. Pugh recorded all of the semester grades (on a scale of 100) of her 50 Advanced Algebra students in the following dotplot:



What is the mean and standard deviation of the distribution of individual grades?

1. $\bar{x}$ = _______________ 2. σ = _______________

Identify the following **and** label the dotplot above

3. $\mu \pm 1\sigma$ = _______________ and _______________

4. $\mu \pm 2\sigma$ = _______________ and _______________

5. $\mu \pm 3\sigma$ = _______________ and _______________

6. Does the data indicate a normal distribution? ______________

7. Why or why not?

____

____

_

____

_

8. We want to know the probability that a student selected randomly from her class would have an "A" (90 or above) in her class. Find the probability using two different methods.

II. Mrs. Pugh went on vacation and could not be reached. Before she left, she turned in her individual student grades to her principal. The parent of the student who made a 68 in the class called and insisted to know the class average of her child's class by the end of the day. Unfortunately, the principal could not retrieve the exact class average because he only had individual student scores, but he told the parent that he could give her a range of scores that the class average would most likely be located within by the end of the day.

He first took a random sample of five students and calculated the average of the five students.

Simulate what the principal did below:

Seed your calculator by typing "1" "STO" "rand" on your calculator. To make sure you did this correctly, type randint(1,50) and you should get the number 38.

9. Simulate what the principal did by typing randint(1,50,5) to get 5 numbers. What are the

numbers?____

10. Locate these values on the dotplot. 1 corresponds to the lowest test score, and 50 corresponds to the highest test score. What are the associated test scores?

____

11. Find the average of these 5 test scores. ____

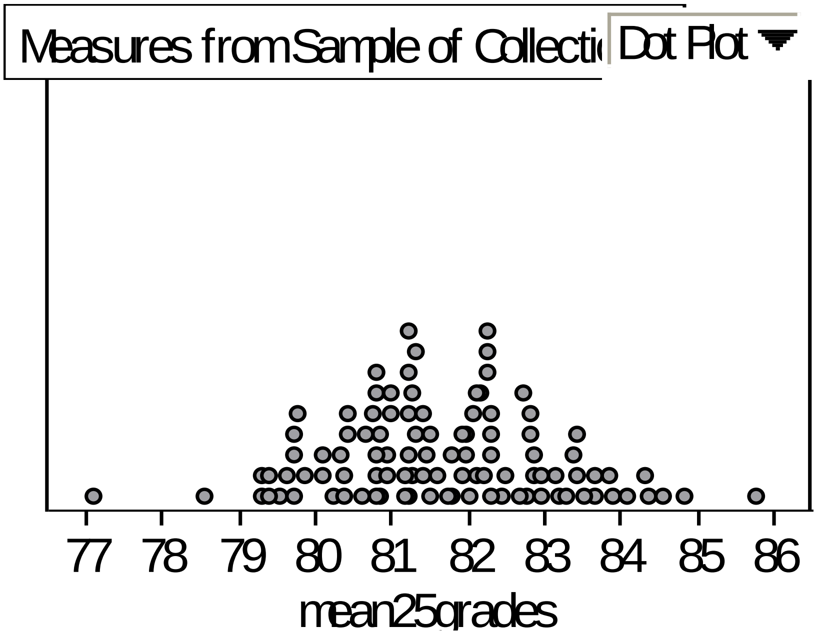
12. Is this sample mean the same as the population mean? _________

13. Why or why not?

____

____
_

III. Although using a sample size of 5 gave him a good idea what the class average was, he wanted to give the parent a smaller range in which the average could be located. He figured that Mrs. Pugh had at least 25 students in each of her classes. He took 50 random samples of size 25, calculated their means, and recorded them in a dotplot.



Estimate the mean and standard deviation of this dotplot

14. $\bar{x}$ = _______________ 15. σ = _______________

The principal knows that it's unlikely a value is beyond two standard deviations away from the mean. Using the mean and standard deviation that you just approximated above for the class average of 25 students, find the values:

16. Mean – 2*standard deviation = __________ 17. Mean + 2*standard deviation =
____

18. Identify the range of scores that would fall between 2 standard deviations **on the dotplot**.

19. He called the parent and told her that he was very confident that the class average was between

________________ and ________________.

The parent replied that the principal's response was not possible because she knew of 2 other students in Ms. Pugh's class that made a 60 and 61.

20. What did the principal explain to the parent? ____

____

____

Culminating Task: Draw You Own Conclusions

Math Goals:

- Understand how to choose summary statistics that are appropriate to the characteristics of the data distribution, such as the shape of the distribution or the existence of outliers.
- How the normal distribution uses area to make estimates of frequencies which can be expressed as probabilities and recognizing that only some data are well described by a normal distribution.
- Understand how to make inferences about a population using a random sample
- Comparing theoretical and empirical results for evaluate the effectiveness of a treatment.
- The way in which data is collected determines the scope and nature of the conclusions that can be drawn from the data.
- How to use statistics as a way of dealing with, but not eliminating, variability of results from experiments and inherent randomness.

STANDARDS ADDRESSED IN THIS TASK:

MGSE9-12.S.ID.2 **Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, ~~mean absolute deviation~~, standard deviation) of two or more different data sets.**

MGSE9-12.S.ID.4 Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

MGSE9-12.S.IC.1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population**.**

MGSE9-12.S.IC.2 Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation. *For example, a model says a spinning coin falls heads up with probability 0.5. Would a result of 5 tails in a row cause you to question the model?*

MGSE9-12.S.IC.3 Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

MGSE9-12.S.IC.4 Use data from a sample survey to estimate a population mean or proportion develop a margin of error through the use of simulation models for random sampling.

MGSE9-12.S.IC.5 Use data from a randomized experiment to compare two treatments; use simulations to decide if differences between parameters are significant.

Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them
2. Reason abstractly and quantitatively
3. Construct viable arguments and critique the reasoning of others
4. Model with mathematics
5. Use appropriate tools strategically
6. Attend to precision
7. Look for and make use of structure
8. Look for and express regularity in repeated reasoning

Introduction

In this task, students will be asked to combine what they have learned about statistics throughout this unit. Students, individually or with a group, will formulate a question for which they can collect data from a random sample of their identified population. Using the sample standard deviation to estimate the population standard deviation, they will be asked to determine the sample size needed to ensure a specific margin of error. The students will calculate margins of error and confidence intervals and interpret their meanings. Finally, students will use technology to simulate additional samples and illustrate the meaning of a confidence interval.

Complete this task individually or with a group.

a. Formulate a question that you can answer with a simple random sample or an experiment. The answer can be in the form of a proportion or a mean.

b. Describe your population and how you will ensure you collect a simple random sample or a completely randomized experiment. (Your population needs to be at least 10 times the size of your sample.)

c. Collect your data.

d. Calculate summary statistics for your data.

e. Describe the sampling distribution of the sample means (or sample proportions). Is it safe to use the formula for the standard deviation? Explain. Can you use the normal approximation with this distribution? Explain.

f. Calculate the 95% confidence interval for the true parameter you are investigating. Show the formula and the final confidence interval.

g. Suppose you were to expand your sample / experiment. If you plan to construct a 99% confidence interval and you wanted your margin of error to do less than 2 units from the mean (or proportion), how large would your sample / experiment need to be? Show your work. (If you are working with means, use the sample standard deviation from your data. If you are working with proportions, use $p^* = .5$.)

h. Design a simulation to generate additional samples to answer your question. Explain your simulation. Use your simulation 100 times and construct 100 95% confidence intervals. Explain your findings.

Present your work either on a poster board, technology or in the form of a paper.

Answers will vary. Students should be encouraged to be creative but be reasonable. The topics addressed in this unit can be adapted into this project.

Culminating Task: Draw You Own Conclusions

Name____
Date__________________

Complete this task individually or with a group.

a. Formulate a question that you can answer with a simple random sample or an experiment. The answer can be in the form of a proportion or a mean.

b. Describe your population and how you will ensure you collect a simple random sample or a completely randomized experiment. (Your population needs to be at least 10 times the size of your sample.)

c. Collect your data.

d. Calculate summary statistics for your data.

e. Describe the sampling distribution of the sample means (or sample proportions). Is it safe to use the formula for the standard deviation? Explain. Can you use the normal approximation with this distribution? Explain.

f. Calculate the 95% confidence interval for the true parameter you are investigating. Show the formula and the final confidence interval.

g. Suppose you were to expand your sample / experiment. If you plan to construct a 99% confidence interval and you wanted your margin of error to do less than 2 units from the mean (or proportion), how large would your sample / experiment need to be? Show your work. (If you are working with means, use the sample standard deviation from your data. If you are working with proportions, use $p^* = .5$.)

h. Design a simulation to generate additional samples to answer your question. Explain your simulation. Use your simulation 100 times and construct 100 95% confidence intervals. Explain your findings.

Present your work either on a poster board, technology or in the form of a paper.