OVERVIEW

Normal distribution, sometimes called the bell curve, is a common way to describe a continuous distribution in probability theory and statistics. In the natural sciences, scientists typically assume that a series of measurements of a population will be normally distributed, even though the actual distribution may be unknown. But even if you assume that measurements of a population should be normally distributed, a sample taken from that population will not necessarily be normally distributed. Why is that?

In this Click and Learn, you will explore what sample distribution looks like when samples are taken from an idealized population of a defined mean and standard deviation. Students will explore how standard deviation affects the distribution of measurements in a population. Next, they will explore how sample size affects the distribution of measurements and therefore the sample mean. Through this exploration, students will develop an understanding of how sample size affects the distribution of sample means drawn from the same population and how this phenomenon is modeled in an equation for calculating the standard error of the mean.

KEY CONCEPTS AND LEARNING OBJECTIVES

- The appearance of a histogram of measurements in a sample depends on the population from which the sample came.
- The appearance of the histogram also depends on the sample size.
- Small samples taken from a normally distributed population may not appear to be normally distributed. Larger samples start to approximate a normal distribution.
- When a population is sampled repeatedly, a mean can be calculated for each sample, to obtain many different means. If those means are plotted as a histogram, they will be approximately normally distributed.
- The standard deviation of such a distribution of means is called the standard error of the mean.

Students will be able to

- explain that standard deviation is a measure of the variation of the spread of the data around the mean.
- explain that larger sample sizes are desirable when collecting data about a population because they are more likely to reflect the distribution of measurements in a population.
- calculate Standard Error of the Mean ($SE_{\bar{x}}$, but also commonly referred to as SE, or SEM), using the equation $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$.
- explain that $SE_{\bar{x}}$ of the mean is a measure of the reliability of the mean of a sample as a reflection of the mean of the population from which the sample was drawn.

- use $SE_{\bar{x}}$ to determine the 95% Confidence Interval to add error bars to a graph and use these error bars to determine if there is a difference between the populations from which the sample came.

CURRICULUM CONNECTIONS

AP Biology (2012-2013) SP2, SP5 **NGSS (2013)** SEP4

KEY TERMS

measurement, sample, population, normal distribution, random sampling, mean, standard deviation, standard error of the mean, 95% Confidence Interval, error bar

TIME REQUIREMENTS

Completing all parts of this lesson will require up to three 50-minute class periods. However, some portions can be assigned for homework.

SUGGESTED AUDIENCE

Part 1 of this activity is appropriate for a first year and an advanced (honors, AP, or IB) high school biology course. Parts 2 and 3 are appropriate for an advanced (honors, AP, or IB) high school or introductory college biology course.

PRIOR KNOWLEDGE

Students should be familiar with

- statistical concept of mean as an average of a sample's measurements.
- histograms as a display of the frequency of measurements in a sample.

MATERIALS

- Sampling and Normal Distribution Click and Learn at http://www.hhmi.org/biointeractive/sampling-and-normal-distribution
- Distribution of Means grid (last page of this document; these can be laminated to be reused by multiple classes)

TEACHING TIPS

- This activity assumes no prior knowledge of Standard Deviation or Standard Error of the Mean. Therefore, it can be used to introduce the use of statistics to describe a data set. It is important that students can distinguish between the terms measurement, sample, and population. A sample is a collection of individual measurements drawn from a population. Prior to starting Part 1, students should understand that it is typically not possible to measure every individual in a large population. Therefore, a randomly selected sample of the population is measured and the data is used to represent the whole population.

- Students can often recognize that small sample sizes are not recommended when collecting data from a population. Doing a simple demonstration such as drawing only a few colored beads from a bag to determine the distribution of colors in the bag or measuring the height of only a few students to determine the mean height of the class can show students that small sample sizes can often lead to a misrepresentation of the population known as sampling error.
- The simulation in the Click and Learn is run by a program that calculates a random sample value from a normally distributed population of infinite size. In Part 1, the student can manipulate sample size as well as population mean and standard deviation. In Part 2, the student can manipulate sample size. Depending on the speed of your computer, resampling in Part 2 can take a few seconds and the calculations occurring in the background are complicated. (For those of you who are mathematically or statistically inclined, the program uses Box-Muller transform.)
- At the conclusion of Part 1 of the activity, students should be able to explain what standard deviation shows about the distribution of measurements in a population. They will also be able to explain, using evidence collected in the activity, why the means of larger sample sizes are more likely to be representative of a population's true mean.
- At the conclusion of Part 2 of the activity, students will understand why the equation for $SE_{\bar{x}}$ gives an estimate of the standard error of the mean based on a sample's size and standard deviation. They will also be able to use the equation to calculate $SE_{\bar{x}}$, 95% confidence intervals, and use the 95% CI to generate error bars on a bar graph. While this activity focuses on the effect of sample size (as sample size increases, $SE_{\bar{x}}$ decreases), students should be able to predict from the equation that there is a direct relationship between the standard deviation and $SE_{\bar{x}}$.
- Remind students that on page 1 of the Click and Learn "Sampling from a Normally Distributed Population," clicking "resample" is simulating collecting a new randomly selected set of measurements from the population. Therefore, sample means and standard deviations will likely be different from student to student. This will not affect the final outcome of the activity. Students should also be reminded that on page 2 of the Click and Learn "Standard Error of the Mean," "resample" represents repeating the sample collection 500 times, and each sample consists of a number of measurements equal to the sample size. This means that for a sample size of 100, the simulation took 50,000 measurements.
- In Part 2, the teacher may need to point out and discuss the difference between sample mean and standard deviation and the mean and standard deviation of 500 means. Sample mean and standard deviation is describing the data in the top graph, while the mean and standard deviation of 500 means is describing the data in the bottom graph.

SUGGESTED PROCEDURE

Depending on the skill level of the students in the course, this activity can be done independently or guided by the teacher. The procedure below is for a guided process, during which the instructor checks for student understanding at key points in the activity.

Introduction

1. Show students the graph below and ask them to interpret it. Ask them what the error bars mean. While it depends on an individual student's prior learning, most students will not be able to explain what the error bars mean. If this is the case, ask them to describe the error bar. Guide students to observations such as

- The error bar for dark does not overlap the error bar for light.
- The dark error bar is longer than the light error bar.
- The lengths of the error bar above and below the top of the bar are equal.

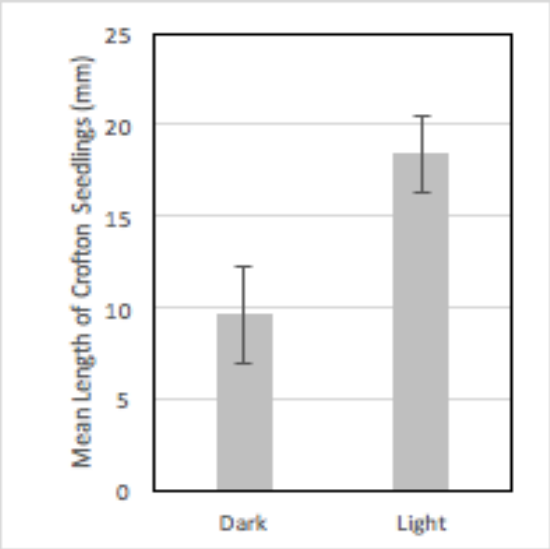


Figure 1. Mean Length of Crofton Seedlings after One Week in the Dark or in the Light. (From *Using BioInteractive Resources to Teach Mathematics and Statistics in Biology* http://www.hhmi.org/biointeractive/teacher-guide-math-and-statistics)

2. Instruct students to complete the Pre-assessment Question (which could be collected on note cards as a formative assessment). Then instruct students to access the Click and Learn at http://www.hhmi.org/biointeractive/sampling-and-normal-distribution and complete items 2 through 5. It is important at this point to ensure that students understand that an individual measurement is part of a sample taken from a larger population. Point out to students the characteristics of a normally distributed population by referencing the red line on the graph and that number of individual mass measurements are represented by the bars in the histogram. Note: This part along with Part 1 items 6 and 7 could be assigned to students for homework prior to completing the rest of Part 1 of the activity in class.

PART 1: SAMPLING FROM A NORMALLY DISTRIBUTED POPULATION

1. Students work through the task and complete items 6 and 7 to explore how modifying the standard deviation affects the distribution of measurements in the population. It should be pointed out to students that changing the parameters changes the simulation program. In a real data set, the standard deviation is determined by the actual measurements in the population or sample.

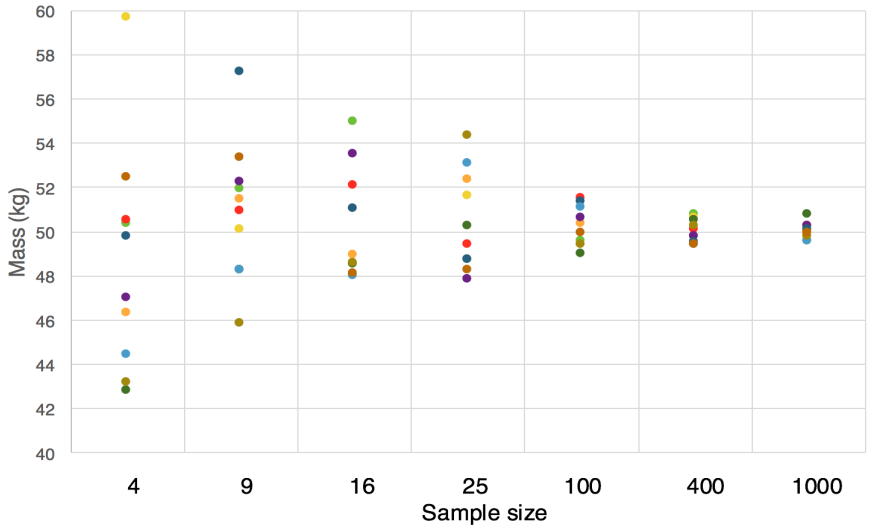
2. Have students read the summary description of standard deviation and discuss any questions they have about standard deviation and normal distribution.

3. In the rest of Part 1, students explore the effect of sample size on the sample mean compared to the true mean of the population. Remind students that they are setting parameters for the program running the simulation (population mean = 50 kg and standard deviation = 10 kg).

4. Item 8 can be used as a formative assessment to monitor student understanding of standard deviation. A correct student response would be: "For this population, 68% of the masses should be between 40 and 60 kg (1 standard deviation), while 95% of the masses should fall between 30 and 70 kg (2 standard deviations)."

5. Students complete items 9 and 10. After completing this task, students should recognize that a sample size of 1000 is more likely to give you a sample mean that reflects the true mean of the population because the larger number of measurements will reflect the normal distribution of the population. They should also recognize that collecting measurements from a sample of 1000 individuals could be time-consuming, expensive, or simply not practical.

6. Students complete "Selecting the appropriate sample size" by completing the task and items 11 through 13. Provide students with the "Distribution of Means" grid. This task can be completed in pairs or a small group. There should be at least one graph in the class for each sample size in the simulation (4, 9, 16, 25, 100, 400, 1000). Laminating the grids will allow them to be reused by several classes. Discuss item 13 as a whole class. Ask students to justify their answer to the question with evidence from the graphs. Students typically select 100 as an appropriate sample size. An example of the data generated from this task is shown below.

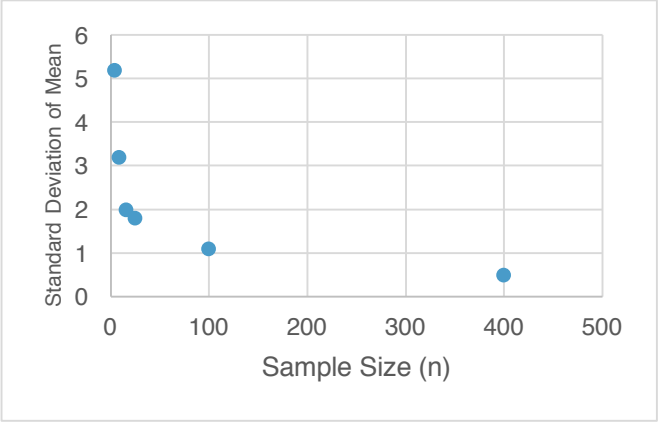


PART 2: STANDARD ERROR OF THE MEAN

1. Items 1 through 8 can be completed as a homework assignment. Students use the next page in the Click and Learn to extend their exploration of the effect of sample size on the distribution of sample means. In Part 2, resampling will generate a histogram of the means of 500 samples showing a normal distribution. Students should come to the conclusion that while the sample size does not affect the mean of 500 means, it does affect the standard deviation of the means. For smaller sample sizes, the sample mean could be quite different from the population mean, and this is reflected in the larger standard deviation of the means. This should reinforce the conclusion they came to at the end of Part 1 that larger sample sizes will provide a better representation of the population from which the sample was drawn.

2. Item 9 introduces students to the equation for standard error of the mean, $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$. Items 10 and 11 have students calculate the $SE_{\bar{x}}$ using the equation. They are then asked to compare the empirically measured $SE_{\bar{x}}$ (standard deviation of 500 means) to the calculated estimation of the $SE_{\bar{x}}$. Students should be reminded that the mathematical formula for $SE_{\bar{x}}$ allows them to estimate the real $SE_{\bar{x}}$ given a small sample, while repeating samples many times allows them to empirically measure the actual $SE_{\bar{x}}$. Students should find that, with the exception of very small sample sizes, using the equation $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$ is a reasonably accurate way to estimate the Standard Error of the Mean from the standard deviation of a sample and the sample size. Students can benefit from a discussion regarding how the equation models the distribution of means that they observed. Encourage students to discuss why sample size is in the denominator. As seen in the graph below, standard deviation of the means decreases as sample size increases. They should recall from their observations during Part 1 of the activity that larger sample sizes are more likely to be a truer reflection of the population from which the measurements are drawn and that very small sample sizes often result in inaccurate representations of the population.

It is helpful to refer students back to the distribution of 10 means they plotted in Part 1 of the activity. Emphasize that the $SE_{\bar{x}}$ equation allows one to estimate the spread of the means that would be expected from many samples drawn from the same population from the standard deviation and sample size of a single observed sample.



The effect of sample standard deviation on the $SE_{\bar{x}}$ is not explored in the simulation. This was done to avoid the misconception that sample size affects sample standard deviation. It still may be helpful to discuss the effect standard deviation of the sample has on the standard error of the mean. The standard deviation of the sample is in the numerator of the equation because a more varied population (larger sample standard deviation) will increase the likelihood that the sample measurements will not be a good representation of the population from which they are taken.

3. Have students read the summary and then complete items 12 and 13 to learn how to use the standard error of the mean to generate 95% Confidence Interval error bars. Reading the summary will show students how to interpret these error bars.

PART 3: APPLY WHAT YOU HAVE LEARNED

The data presented is authentic data collected by students conducting an experiment to test the effect of pectinase and cellulase on turning apple sauce into apple juice.

This part of the activity can be given as an assessment to determine students' level of understanding of the concept of standard error of the mean and how to use the statistic to analyze the experimental data.

RECOMMENDED FOLLOW-UP ACTIVITIES

Evolution in Action: Data Analysis (http://www.hhmi.org/biointeractive/evolution-action-data-analysis)
Rosemary and Peter Grant have provided morphological measurements, including wing length, body mass, and beak depth, taken from a sample of 100 medium ground finches (*Geospiza fortis*) living on the island of Daphne Major in the Galápagos archipelago. The complete data set is available in the accompanying Excel spreadsheet.

In one activity, entitled "Evolution in Action: Graphing and Statistics," students are guided through the analysis of this sample of the Grants' data by constructing and interpreting graphs, and calculating and interpreting descriptive statistics. The second activity, "Evolution in Action: Statistical Analysis," provides an example of how the data set can be analyzed using statistical tests, in particular the Student's t-test for independent samples, to help draw conclusions about the role of natural selection on morphological traits based on measurements.

Lizard Evolution Virtual Lab (http://www.hhmi.org/biointeractive/lizard-evolution-virtual-lab)
In the Lizard Evolution Virtual Lab, students explore the evolution of the anole lizards in the Caribbean by collecting and analyzing their own data.

The virtual lab includes four modules that investigate different concepts in evolutionary biology, including adaptation, convergent evolution, phylogenetic analysis, reproductive isolation, and speciation. Each module involves data collection, calculations, analysis, and answering questions.

AUTHOR

Valerie May, Woodstock Academy, Woodstock CT

Distribution of Means

Sample Size = _____