INTRODUCTION TO STATISTICS

Students will learn about population sampling, and they will measure and interpret average and variance of leaf sizes of different tree species.

LESSON LENGTH:

- 1.5-2 hours

GOALS:

- Gain an understanding of basic statistical techniques
- Analyze collected data to estimate a parameter of interest.

OBJECTIVES:

- Students will understand what a population sample is as well as what the average and variance of that sample means.
- Students will create and do basic analysis of a small dataset.
- Students will apply the idea of sample variance across different systems.

NATIONAL, STATE, LOCAL STANDARDS

NC Standard Course of Study

- NC.ECS.M1.SID.1: Given data, use technology to construct a simple graph (line, pie, bar, or picture) or table, and interpret the data.
- NC.ECS.M1.SID.2: Interpret general trends on a graph or chart. (more, less, increasing, decreasing). Given a graph, table, or word problem, calculate the average of a given data sets (when the number of data points is fewer than five) and compare the average.

STUDENT TAKEAWAYS FROM LESSON:

- Essential question / theme
  - Why do we need data/statistics?
  - What is the sample average and variation of the size of leaves on a tree?
- Key concepts and vocabulary (define them here)
  - Population – every observation that could possibly be made. For example, if you were trying to describe the height of trees in the forest, the population would be the height of every single tree in the area of interest. **This is different than the meaning of the word "population" in ecology, which means a group organisms of the same species living in the same place.**
  - Sample – a random subset of the population that is actually measured. If you are trying to estimate the height of trees in a very large, you may not have time to measure every tree, so you may randomly choose 50 trees to measure. That sample will give you an idea of what the rest of the population looks like. The assumption here is that the sample you measured represents the population.
  - Average – the central tendency of the data.
  - Variation – how spread out data is
  - Probability – measure of the likelihood that an event will occur.
  - Distribution – for each value of a parameter that can be measured, this function gives a relative frequency or probability with which that value will occur.

ASSESSMENTS:

- Estimate a population average and variation from a sample of n taken from that population.

DIVERSITY (REACHING STUDENTS OF ALL LEVELS/ABILITIES):

- This activity includes drawing and hands-on measuring.

MATERIALS & EQUIPMENT:

- Calculator (1-2)
- Ruler
- Pencil
- Paper/notebook

LOCATION:

- In the woods, preferably in close proximity to a tree.

BAD WEATHER ALTERNATIVE:

- Students could collect leaves and bring them back to a dry area to measure. If it's not possible to venture outside the tents/tarps, students could also measure themselves as a sample – height, foot length, etc.

LESSON:

Part 1: Height of Students

- *What the instructor should say is in italics.*

ENGAGE

- Ask the students the following questions:
  - *What comes to mind when you hear the word "statistics?"* Field some answers, and then emphasize that statistics is how we formally identify patterns, which helps us make predictions.
  - *What it we wanted to figure out a pattern about our heights? What statistical questions might we ask about our heights?* (Get students to questions like: What's the average height of a girl on this trip?)

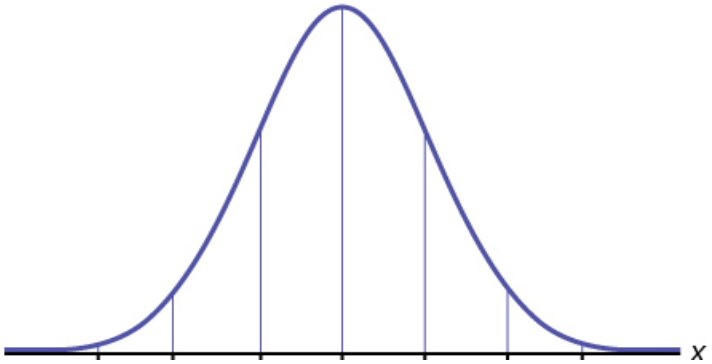
EXPLORE

- Tell students that we're going to do a statistical analysis of their heights!
  - Ask: *How could we figure out the average height of our group?*
  - Have a girl go to the whiteboard to ask and record everyone's height
  - Instructor should walk through how to translate the data into a T-chart (girl height and number of girls/frequency) on the whiteboard. Use ranges (5'0"-5'3"), called bins, instead of individual heights for the "girl height" column so that your histogram ends up looking more like a normal curve (girls copy T-chart into their notebooks).
  - Instructor should explain and draw a histogram on the whiteboard using that data, emphasizing axis labeling (y-axis = number of girls, or FREQUENCY, and x-axis = girl height) and what the axes mean (girls copy into their notebooks).
  - Instructor should draw a curve over the histogram (which hopefully looks a bit like a normal curve). Instructor can ask: *What do you notice about this histogram? Can we see any patterns about our heights?*
  - Ask: *What is the average height?* Walk girls through the meaning of average and the calculation and then find and label the average height on the histogram/curve.
  - Ask: *Is your average height a good representation of the average height of a high-school student in the U.S.? Why/Why not?* Get girls to think about the makeup of the sample and understand that this is not a good representation because we're all girls, there's too few of us, we're all from NC...
  - Ask: *How could we find a more accurate average height of a high school student in the U.S.?* (Collect more data.) *Could we get the data from every single high school student in the U.S.?* (No, that would take forever!)
    - *The population is every observation you could possibly make of the system you are testing. In this case, that would be every high school student in the U.S.* (Emphasize that the word "population" in statistics is different from the word "population" in ecology.) *Therefore, we chose a random subset of individuals, and that sample represent the population.*
    - *Theoretically, the properties of the sample should match the properties of the whole population. This is only true if the sample is truly* ***random*** *(unbiased) and a sufficiently* ***large*** *sample size.* (Write "sample = random and large" on the board).

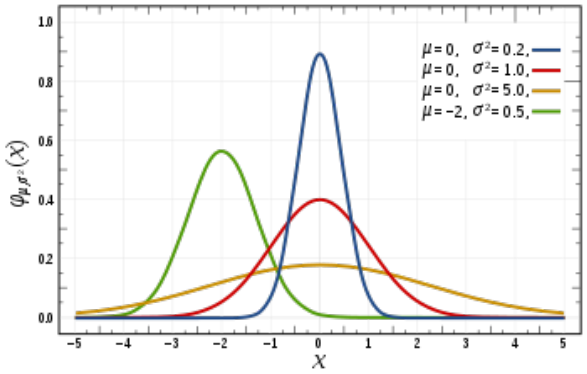
- Think-pair-share/Jigsaw: Have half of the pairs discuss why the sample needs to be random and the other half discuss why the sample size needs to be large. Sample share-outs/takeaway points below:
    - *What could happen if the sample wasn't random (was biased)?* Taking data just from girls, who are on average shorter than boys, would skew the data downwards
    - *What might happen if we don't have a large enough sample size?* Taking data from just the 8 of you isn't a good representation of the height of the whole population of U.S. high schoolers. A sample size is sufficiently large when taking more samples doesn't change your average much. There isn't an exact answer to the question of "how many samples should I take?" In general, the more the better – especially in ecology research, where our populations are often huge! With too small of a sample size, we could end up with a sample that does not represent the population (skewed upwards or downwards)

EXPLAIN

- Draw a normal distribution like the distribution below (without the vertical lines). Y-axis = frequency, x-axis = height. Ask: *Have you seen this before? Do you know what it is or what it's called?*



- *This is a normal distribution (aka a bell curve or normal curve). Most data, especially in nature, follow this pattern – if we were to use a random and large sample to collect data about high school student height, when we graphed it, it would probably look like this.*
- *Where is the average height?* (right in the middle – label it)
- *What is the most common height for high schoolers? (i.e. Which height has the most students that are that tall?)* (right in the middle too!)
- *As you can see from this graph, the most probable height – the one you will measure most often – is the average, but you will get height data on either side of that average.*
    - Explain that instead of the y-axis being "Frequency," we could also call the y-axis "Probability." (This is a difficult concept for students -- use the example of picking a random high school student and wanting to guess their height without looking at them -- what would give you the best chance of being right?)
- Draw a couple of example curves like the ones below (but *move the red one to the right.* The curves can represent the heights of girls aged 7-19 (green), gymnasts (blue), and woman aged 18-30 (red). Ask: *What do you notice about these bell curves? What are some differences you see?* (They have different averages. Some are skinny and tall, while some are wider and flat.)



- *We use a word called "variation" to describe the shape of the normal curve. What does "variation" sound like?* (sounds like vary, so variation is how different the values are from one another)
- *For the normal curve, the variation tells with whether the curve is tall and skinny (almost all the values you measure will likely be within a narrow range of the average) or if it is wide and flat (the values will likely be measured over a wide range of values).*
  Label the graphs as girls aged 7-19 (green), gymnasts (blue), and woman aged 18-30 (red)). Ask or talk about the connections between those labels and the shape and average of the bell curves.

Part 2: Leaf Area

- *What the instructor should say is in italics.*

ENGAGE

- Tell girls we're next going to look for patterns about leaf area (use a leaf to explain what "leaf area" is).
- *What might leaf area tell us about a forest?* (When thinking about forests and the way carbon and nutrients move through the environment, it's useful to have an idea of the leaf area of a forest or the sum of the area of every leaf in the forest. This can influence the amount of photosynthesis that occurs and therefore affects the uptake of carbon dioxide, use of nutrients, and uptake of water. It can also give you an idea of the health of a forest. Healthier forests generally grow more and bigger leaves.)
- *What kinds of factors do you think influence leaf area?* Field answers to this question and ask students HOW those factors would influence leaf area (i.e. more water = more or less leaf area?). Main things to think about:
  - tree species (show an example of a pine needle and another leaf)
  - how many of each species are in the forest
  - latitude (sunlight)
  - water availability
  - nutrient availability
  - Students might also mention seasons (more leaves in the summer than winter)

EXPLORE

- *We want to know: What is the average area of a leaf in this forest? We can use data and statistics to find these patterns and even predict what the leaf area might be across a whole forest.*
- *How would you propose we go about answering that question?* Field a few answers. *The first step in the process is to collect data from the population. Remember, the population is every observation you could possibly make of the system you are testing – so what would the population be here?* (That would be every leaf in the forest. That would be impossible to measure before they all fell off in the fall and you

would have to start all over again!) *Instead of collecting all of that data, what do we use?* A sample that needs to be **random and large!**

- Use a leaf to model how to measure and calculate the leaf area on the board (have girls copy the example into their notebook). Since these are often irregular shapes, advise them on how to estimate the area. They can assume it's approximately a circle or a square as long as they stay consistent (include a review of the area formulas for rectangle, triangle, and circle if necessary). Accuracy is not the goal here.
- Split the students into pairs of two. Two groups should measure two different trees of species 1 (red oak, loblolly pine, etc), and two groups should measure two trees of species 2 (so that 4 total trees are being measured, one by each group).
- *Each group will have 10 minutes to measure and record the area of 15-20 leaves. Record the leaf area of each leaf.*

EXPLAIN

- Gather everyone when they're finished measuring. *Now that we have our data, what do we want next? We need to find a way to describe the distribution of area.*
- *In your pairs, calculate the average of your data and draw a histogram. For the histogram, group the data into ranges, and graph how many measurements you took in that range. These ranges are called bins. Create about 5 bins for your data. The size of each bin can be calculated using the formula: (max leaf area measured-min leaf area measured)/5*
    - Draw a set of example ranges and a histogram on the board if necessary.
- When they're done, have groups that measured the same species of tree meet to compare results. *Did they get similar answers? Why might they be different?*
- Then bring the whole group together and compare between species. *Did different species have significantly different leaf areas? Why might that be?* (That's an adaptations/evolution question – cross-cutting concepts!)
- *What shape is your graph (skinny or wide)? What type of variation is in your sample? Is the variation different depending on the tree or species?*
- *How do you think you could improve your experimental design? Was your sample biased at all?*
- Concluding remarks: *We used the data to make simple observations about possible leaf areas for trees. Another way to approach a scientific problem is to start out with a specific hypothesis. For example, we could have hypothesized that species A had a higher average leaf area than species B. Using the distributions that were drawn and a couple more calculations, we could support or refute this hypothesis. This is likely how you will be using these concepts in your projects. Continue to think about how we can use data and distributions to test hypothesis.*

REFERENCE MATERIALS/RESOURCES

- Written by Kim Bourne for GALS

Based on: http://www.d.umn.edu/~tbates/curricularesources/ModelLessonPlan.pdf

NOTES:

- Wrote the lesson assuming a base knowledge of the scientific method.
- Assumes a basic understanding of probability.