**Student Handout Module: Amylase Copy Number and Diet**

*Learning objectives:*

* Propose testable hypotheses to investigate correlation between diet and amylase gene copy number.
* Explain the function of salivary amylase in the hydrolysis of starch.
* Describe how genetic variation can provide a selective advantage to an organism in a particular environment.
* Create and/or analyze numerical summaries and graphical displays of data.
* For two groups on one continuous variable: Explain the meaning of key terms used in descriptive statistics such as: sample sizes, means, standard deviations, medians, minimums, and maximums.
* Analyze a histogram for two groups on a single continuous variable to identify critical components such as graph type, axis labels, and key to symbols.
* Compare and contrast measures of center and spread based on the shape of the histograms to compare and contrast high vs. low starch diet on the variable of interest, amylase gene copy number.

Starch, a plant polysaccharide composed of many building blocks of glucose, is a high-energy component of foods found in nature. As early humans transitioned from

hunting and gathering to more agrarian lifestyles, their diets changed to include more high starch foods. Some cultures incorporated more starch into their diets than others, and those cultural differences in human populations are still present in some cultures today.

Starch cannot be dissolved in water or in stomach acid, so an enzyme is needed to break it up. Starch digestion begins in the mouth with an enzyme called amylase. Salivary amylase breaks the glycosidic bonds between glucose units in starch by adding a water molecule. This chemical reaction is called hydrolysis. The gene that encodes salivary amylase (AMY1) is somewhat unique, as most humans have more than one diploid copy of the gene; in fact the number of copies ranges from 2 to 15. The authors of the paper from which this data was taken investigated whether there is a correlation between the number of AMY1 gene copies and the type of diet (high-starch or low-starch) of a population. A correlation would indicate that having more copies of the AMY1 gene provides a selective advantage, allowing individuals to break down starch more efficiently.

This activity will require you to use the Salivary Amylase data file to generate a histogram and calculate the following descriptive statistics in Excel:

* + n (sample size)
	+ Mean
	+ Median
	+ Minimum
	+ Maximum
	+ Variance and standard deviation

You will need the following to use as a reference as you create your spreadsheet and analyze your data:

[Spreadsheet Tutorial 1 -- Formulae, Functions, and Averages](http://www.hhmi.org/biointeractive/spreadsheet-tutorial-1-formulae-functions-and-averages)

[Spreadsheet Tutorial 2 -- Autofill Data, Cell References, and Standard Deviation](http://www.hhmi.org/biointeractive/spreadsheet-tutorial-2-autofill-data-cell-references-and-standard-deviation)

[Spreadsheet Tutorial 5: Histogram](http://www.hhmi.org/biointeractive/spreadsheet-tutorial-5-histogram)

*Please refer to spreadsheet tutorial 1 to help you answer questions 1-4.*

1. Begin on the "Calculations" tab. In the cell labeled "n", use the count function to calculate the sample size for each population. The sample size (n) is calculated by using the "count" formula.

1. High-starch population n=
2. Low-starch population n=

2. The central tendency of a data set can be determined by calculating the mean. The mean is calculated by summing all the data points in a data set (ΣX), and then dividing this number by the total number of data points.

In the cell labeled "mean," use the average function to calculate the mean (average) number of AMY1 gene copies for each population.

1. High-starch population mean=
2. Low-starch population mean=

What is the difference between these values in populations consuming high-starch vs. low-starch?

Using a complete sentence, write a statement that compares the high-starch and low-starch population, indicating in which population the average number of AMY1 gene copies is higher.

What question can this data help you to answer?

Keeping this question in mind, state a hypothesis that accounts for the difference in AMY1 gene copy number in populations having high-starch vs. low-starch diets.

3. When the data are ordered from the largest to the smallest, the median is the midpoint of the data. It is not distorted by extreme values, or even when the distribution is not normal. For this reason, it may be more useful for you to use the median as the main descriptive statistic for a sample of data in which some of the measurements are extremely large or extremely small.

In the cell labeled "median," use the median function to calculate the median number of AMY1 gene copies in for each population.

1. High-starch population median=
2. Low-starch population median=

Compare the mean and median for each population. Which value best describes the center of the data distribution (the central tendency) for each population?

*Please refer to spreadsheet tutorial 2 and spreadsheet tutorial 5 to help you answer question 5.*

5. Variability describes the extent to which numbers in a data set diverge from the central tendency. It is a measure of how “spread out” the data are. The standard deviation is the most widely used measure of variability. The sample standard deviation (s) is essentially the average of the deviation between each measurement in the sample and the sample mean (𝑥). The sample standard deviation estimates the standard deviation in the larger population.

The formula for calculating the sample standard deviation follows:

s = √(Σ (𝑥𝑖 − 𝑥)2)

 (𝑛 − 1)

To determine the variability in the data for each population, calculate the standard deviation (s) of AMY1 gene copies for each population. You may do this manually using the cells labeled (value-mean), (value-mean)squared, sum of squares, n-1, variance (sum of squares/(n-1), standard deviation as directed in the tutorial and/or you may use Excel functions to calculate variance and standard deviation. Record the standard deviation for each population.

1. High-starch population s=
2. Low-starch population s=

Which population has more variation around the mean?

*Please refer to spreadsheet tutorial 5 to help you answer question 6.*

6. A histogram is a graph of frequency distribution. It graphs how many data points fall between a range of values. It is a very useful graph for illustrating data distribution.

Visualize the distribution of the data for high-starch vs. low-starch populations by creating a histogram in the tab labeled "Generating Histogram."

a. Using the cell "sorted data" for each population, sort the # of AMY1 Gene Copies data from minimum (low) to maximum (high).

b. Use the "Bin" column next to this sorted data to create bins that span the range of AMY1 gene copies found in each population.

Your bins for each population will be 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, and 16 as the # of AMY1 gene copies in the two populations range from 2-15.

c. Use the "Frequency" column next to the bin column to record the number of individuals with that number of AMY1 gene copies. For example, there are 9 individuals with 4 AMY1 gene copies in the low-starch diet population, so the frequency associated with Bin "4" is 9.

d. Use the "Proportion" column next to the frequency column to calculate the proportion of individuals in each bin. This can be calculated by dividing the frequency by the sample size (n).

e. Use the clustered column graph to create your histogram. Label your histogram with the following labels:

Title: Amylase Copy Number and Diet

X-axis label (horizontal axis title): AMY1 diploid gene copy number

Y-axis label (vertical axis title): Population of Individuals

Legend: high-starch diet

 low-starch diet

7. Having studied your histogram, answer the following questions.

1. How do the centers of each population compare in each population?
2. How does the distribution of gene frequencies compare in each population?
3. Is there a relationship between the number of AMY1 gene copies and the type of diet (high-starch or low-starch) of a population? Does your conclusion support or nullify your hypothesis?

*Reflection*

8. State one thing you learned about using excel to calculate descriptive statistics in this exercise.

9. Write a specific statement of one way measuring mean, median, and standard deviation values can be used.

10. State one thing you learned about visualizing data using a histogram in this exercise.

11. For you, what was the most difficult thing about this exercise?

*References*

1. Amylase Copy Number and Diet, http://www.hhmi.org/biointeractive/amylase-copy-number-and-diet
2. Diet and the Evolution of Salivary Amylase, http://www.hhmi.org/biointeractive/diet-and-evolution-salivary-amylase
3. Perry, G. H. et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nature Genetics39:1256–1260.
4. Spreadsheet Data Analysis Tutorials, http://www.hhmi.org/biointeractive/spreadsheet-data-analysis-tutorials
5. Teacher Guide, Math and Statistics, http://www.hhmi.org/biointeractive/teacher-guide-math-and-statistics