**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 analyze a data set where the following descriptive statistics have been calculated for you in Excel:

*Please refer to the provided Salivary Amylase Datasheet to help you answer the following questions.*

*Begin on Sheet 1, "raw data files," in your Excel spreadsheet.*

1. Data for individuals from high-starch and low-starch diet populations are provided for you to analyze. These individuals are from the European-American, Hadza, and Japanese populations, and the # of AMY 1 gene copies each individual in this sample has is listed in column B. The individuals that comprise the low-starch diet populations are listed below the high-starch diet population. These individuals are from the Blaka, Mbuti, Yakut, and Datog populations, and the # of AMY 1 gene copies each individual in this sample has is listed in column B.

*Please refer to Sheet 2, titled "Calculations," in your Excel spreadsheet.*

2. Excel can be utilized to calculate descriptive statistics that can help you to understand the meaning of the data collected from each population. The sample size (n) has been calculated for each population by counting the number of individuals. Locate "n" and record the values below:

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

3. The central tendency of a data set can be determined by calculating the mean. The mean for each population has been calculated by summing all the data points in a data set (ΣX) and then dividing this number by the total number of data points. Locate the mean (average) number of AMY1 gene copies for each population and record the values below:

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.

4. 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.

The median number of AMY1 gene copies has been calculated for each population.

Record these values below:

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?

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, the standard deviation (s) of AMY1 gene copies for each population has been calculated as shown by the formula above as well as using specific Excel formulas. Locate and record the standard deviation for each population below:

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

Which population has more variation around the mean?

*Please refer to Sheet 3, titled "Generating Histogram," in your Excel spreadsheet.*

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, and allows you to visualize the distribution of the data for high-starch vs. low-starch populations.

a. For each population, the # of AMY1 Gene Copies have been sorted from minimum (low) to maximum (high). Record these values below:

1. High-starch population min= max=
2. Low-starch population min= max=

b. The "Bin" column next to this sorted data spans the range of AMY1 gene copies found in each population. The bins for each population are 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. The "Frequency" column next to the bin column indicates the number of individuals in each population 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. For each population, the proportion column next to the frequency column indicates the proportion of individuals in each bin. This was calculated by dividing the frequency by the sample size (n).

e. This data was used to create a histogram. Histograms are labeled to make the data represented clear and easy to analyze. Record the labels for the histogram provided below:

Title:

X-axis label (horizontal axis title):

Y-axis label (vertical axis title):

Legend:

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

1. How do the centers of each population compare with each other?
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