# Week 1

Goal: In these assignments, students select a simple cluster sample of residents of The Islands, collect their own data, and use StatKey to describe the data from their sample. Students develop an understanding of where data comes from, learn best practices in preparing data files for statistical software, and reason about data both in terms of an individual case as well as the aggregate when examining visualizations and summaries.

## Part 1

#### INTRODUCTION

This week we are going to practice collecting data, creating a dataset, and summarizing data. To do this, we will be using the Island. Find more information about [The Islands](http://www.islandsinschools.com.au/) here. Log in to the Islands by going to [this website](https://theislands.umn.edu). Log in using your GWU email address. Click on 'forgot password' to set up your password for the first time.

Explore the islands by exploring what variables you can collect from Islanders. [Watch this video](https://youtu.be/v9Dp7Ac-xek) to help you plan ahead for the type of data collection you will be doing. Note: the video is from an old semester, and the assignment has since changed. Follow the instructions posted here, not the instructions I say in the video.

This week, we will visit the Islands to collect data. We will be collecting the following information from a series of Islanders: age, gender, height, weight, blood cholesterol, blood glucose. We will focus our investigation on the city of Vardo, the northernmost city in the Islands. Complete the following steps to answering your question. Submit your results following this same outline:

#### SAMPLE SELECTION

(1) Even though we are interested in all residents of the city, we will select a sample in order to make data collection manageable in terms of time. Visit Vardo and scroll down to see how many total houses there are in the city. Use [this webtool](https://www.randomizer.org) to generate a list of 30 houses in your city by setting the following inputs: How many sets = 1, How many numbers = 30, Number range = 1 to # of houses in Vardo, Remain Unique = yes, Sort the numbers = No. This list will be the houses you will include in your sample. Include all residents for the houses in your list.  Note: some houses may be empty, that is okay. Note: this type of sample is called a simple random cluster sample.

As you collect data from your sample, store the data in a table in the following format, using Microsoft Excel. You have free access to Microsoft Office 365 through your GWU student account.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| House Number | Name | Age | Gender | Height (in inches) | Weight (in lbs) | Blood Cholesterol (in mg/dL) | Blood Glucose (in mg/dL) |
| 17 | Alex Vargas | 25 | M | 68 | 172 | 186 | 101 |
| 49 | Michelle Dawson | 52 | F | 66 | 143 | 172 | 84 |

(2) Collect data from your sample of Islanders. For each sampled house, click on the house number, and then open a tab for everyone that lives in the house. You will then see their name, age, and gender. Click on 'tasks' and obtain consent. If the person refuses, write down their house number, name, age, and gender, and leave the other columns blank.  Click on 'physiology' to measure each Islander's height and weight. Click on 'blood tests' to measure each Islander's cholesterol and glucose levels. View the results by clicking on 'History'. Even if people refuse, keep them on your data file. Note: The Islanders sleep! Unfortunately this is a setting I cannot change. You will have to collect data during day time. Be sure to start your data collection early.

Save your excel file as a '.csv' file.

**due Sunday:**Submit your csv data file.

## Part 2

Complete the following instructions using the data you collected for in Part 1.

#### SUMMARY STATISTICS

Create the following summary statistics.

(1a) Summarize the gender distribution of participants in your sample with a bar chart. Go to StatKey, and under 'Descriptive Statistics and Graphs', click on 'One Categorical Variable'. Click 'upload file', and upload your saved csv data file. Select the gender variable.

(1b) Estimate the true proportion of females in the city by creating a confidence interval for a proportion. Go to StatKey, and under 'Bootstrap Confidence Intervals', click on 'CI for a single proportion'. Click on 'edit data' and enter in the correct summary statistics from the summary table in (1). Generate at least 10,000 fictitious bootstrap samples and write down the standard error. Use the following equation to create an interval estimate: Lower Limit = Original Sample Proportion - 2\*Standard Error, and, Upper Limit = Original Sample Proportion + 2\*Standard Error.

(2a) Summarize the glucose level distribution of participants in your sample with a dotplot.  Go to StatKey, and under 'Descriptive Statistics and Graphs', click on 'One Quantitative Variable'. Click 'upload file', and upload your saved csv data file. Select the glucose variable.

(2b)  Estimate the true mean glucose level of all Islanders in the city by creating a confidence interval for a mean. Go to StatKey, and under 'Bootstrap Confidence Intervals', click on 'CI for a single mean'. Click on 'upload file' and upload your saved csv data file. Select the glucose variable. Generate at least 10,000 fictitious bootstrap samples and write down the standard error. Use the following equation to create an interval estimate: Lower Limit = Original Sample Mean - 2\*Standard Error, and, Upper Limit = Original Sample Mean + 2\*Standard Error.

(3) Summarize the age distribution of participants in your sample with a summary table.  Go to StatKey, and under 'Descriptive Statistics and Graphs', click on 'One Quantitative Variable'. Click 'upload file', and upload your saved csv data file. Select the age variable.

(4) Summarize the relationship between weight and cholesterol level in participants in your sample with a scatterplot and summary table.  Go to StatKey, and under 'Descriptive Statistics and Graphs', click on 'Two Quantitative Variables'. Click 'upload file', and upload your saved csv data file. Select the weight variable first, and then the cholesterol variable.

(5) Summarize the relationship between gender and height in participants in your sample with a side-by-side boxplot and summary table. Go to StatKey, and under 'Descriptive Statistics and Graphs', click on 'One Quantitative and one Categorical Variable'. Click 'upload file', and upload your saved csv data file. Select the height variable first, and then the gender variable.

**due Sunday:** Submit a single .docx or .pdf file with your completed work, including screenshots of StatKey, written out equations, and explanations as appropriate.