**ESS 116 Lab 5**

**Descriptive Statistics**

General Instructions: To download necessary files for the lab assignment, please go to the Assignments folder on Canvas. Include your answers in a single PDF report containing your name. Please also name your file as “Lab1\_LastName.pdf”, e.g., “Lab1\_Pritchard.pdf”. You are welcome to use your favorite document editor (e.g., MS Word), but convert it to PDF before submitting it. The report is due in Canvas in roughly one week (**May 9th, by 10am**). Late reports will not be accepted; if you have a problem submitting your work on time, please e-mail the TA. Lastly, while you are welcome to talk with your fellow classmates about how to go about completing the assignment, you must do your own work. I should not see duplicate answers.

Goals

This laboratory exercise will help you familiarize yourself with MATLAB’s functionalities for descriptive statistics (such as calculating a mean, standard deviation, plotting a histogram, etc).

In 1991, an alien species from Planet X sends two UFOs to study planet Earth’s atmosphere. One UFO is dispatched to a location near the equator (0˚N, 180˚E) and the other UFO is sent to the North Pacific Ocean at a midlatitude location (42˚N, 180˚E). For five years, from 1991-1995, both UFOs hovered in the middle of the atmosphere, sitting on the 500 hPa pressure surface (about 5-6 km above the ocean), taking daily measurements of the atmospheric temperature (T) and the altitude (Z) as the pressure surface bobbed up and down with weather fluctuations.

Part 1: Descriptive Statistics and Histograms

Before beginning the process of warping back to their home planet, the alien UFO pilots calculate and transmit summary statistics of the time series measurements they have just taken back to their home planet, Planet X. Since they are only able to transmit 8 numbers from such a long distance, they calculate and send the bare minimum statistical information needed to quantitatively describe the location and the dispersion characteristics of each data variable’s distribution.

1. Describe which aspects of a data distribution the mean and standard deviation measure and how they relate to a histogram (1 sentence)

The mean measures the average value of a set of numbers and it is related to a histogram because the mean may be the central tendency –the average position among the whole distribution which may be the mean, median, or mode.

The standard deviation measures the extent of deviation from the average of a set of values and it is related to a histogram because the standard deviation may be the dispersion –the extent to which the values are spread around the central tendency.

1. The UFO pilot who had been sitting over the midlatitude ocean sends its 4 numbers first (a mean and standard deviation for both temperature and altitude). Write a MATLAB script called “EarthInfo\_LASTNAME.m” that loads “*aliendata.mat*” and displays the 4 numbers received on Planet X (replace “UNIT” by the correct unit). You don’t need to store the output of *load* in a variable because *aliendata.mat* is storing 5 variables (use *whos* to see what they are).

Earth Properties @ Midlatitude

Mean Temperature: XX.XX UNIT

Mean Altitude: XXXX.XX UNIT

Std Dev Temperature: X.XX UNIT

Std Dev Altitude: XXX.XX UNIT

1. The UFO pilot who had been sitting over the equator sends its data next. Continue to edit Earthinfo, skip one line and print the equator properties:

Earth Properties @ Equator

Mean Temperature: X.XX UNIT

Mean Altitude: XXXX.XX UNIT

Std Dev Temperature: X.XX UNIT

Std Dev Altitude: XX.XX UNIT

1. Back on Planet X, the alien scientists receiving this information notice a striking difference indicating a likely difference in the dispersion characteristics of the data collected near the equator compared to the midlatitudes. Explain which aspect of the statistical values calculated in 2) and 3) they are referring to. What do they predict will be different about the data distributions of the variables in the two locations? (2 sentences)

The aspect of the statistical values that are referring to is the standard deviation for the equator versus the midlatitudes because the standard deviation is much greater in the midlatitudes than at the equator for both the average temperature and altitude. What they can predict about the data distributions of the variables in the two locations is that the midlatitude data will have a greater range and spread than the equator data.

The scientists back on Planet X request additional information from the alien UFO pilots to test their prediction about differences in the data distribution in the equator vs. midlatitudes. They ask the UFO pilots for detailed histograms of each variable, in each location, and request a 4-panel figure containing, on the top row, a plot of T and Z histograms for the midlatitude location, and on the bottom row, the same thing for the equatorial location. The UFO pilots work on this while in hyperspace by using a copy of MATLAB that they pirated from the humans’ fledgling Internet.

1. Edit a new MATLAB script called “EarthHist\_LASTNAME.m” that shows the figure, making sure to satisfy all the rules below (Note: you’ll need to load the data again so that this script can run independently from ‘EarthInfo\_LASTNAME.m’):
   * Use an appropriate number of histogram bins for the number of data values in each sample.
   * Allow MATLAB to auto-distribute the bin centers according to the range of the different datasets.
   * Enforce common x- and y- axis limits for each variable so you can visually compare the distribution between different locations. For temperature, a range of -60 C to 0 C will work. For altitude, a range of 4800 m to 5900 m is best. For the count/frequency axis, a maximum limit of 150 should work, if you have chosen your bins appropriately.
   * Add a title for each subplot that indicates the location (either “Midlat” or “Equator”).
   * Add a label to the x axis that is either “T (UNIT)” or “Z (UNIT)” with the appropriate unit.
2. Is the shape of the histograms consistent with your prediction based on its descriptive statistics from Part 1? Explain with reference to specific features of the histograms and how they relate to the specific statistical measure values from Part 1. (2 sentences)

The shapes of the 4 histograms are consistent with my predictions and measured standard deviations from part one because the data collected at the equator has a narrower spread than that collected at the midlatitudes for both the temperature and altitude. The standard deviation calculated in part one is therefore visually plotted via histogram when applied within a 150-frequency limit and appropriate x-limits for temperature and altitude, and in addition the 4 means calculated are also displayed because they are the bins with the higher frequencies.

1. Based on the shape of the histograms, do you think the data distributions have a lot of skewness? Test your prediction by calculating the skewness of the data in EarthHist.m and display the output as:

Equator Z skewness: XX.XX

Equator T skewness: XX.XX

Midlat. Z skewness: XX.XX

Midlat. T skewness: XX.XX

Based on the shape of the histograms I predict that the data distributions that the skewness of all 4 plots will be small because none of them appear to have a significant tail, and I additionally predict that the skewness of the equator temperature and altitude data will be less than that of the midlatitudes because the standard deviation is smaller.

1. Is the data distribution unimodal, bimodal, or polymodal

The data distributions for all 4 data sets are unimodal because they each have a single bin with a highest occurring frequency that is relatively higher than other bins.

1. Do you think the data is approximately normally distributed? That is, do you think it could be fairly described with a Gaussian distribution?

I think the data is approximately normally distributed because the skewness is small and therefore the majority of the would likely data fall within plus or minus 3σ of the

Part 2: Sodium in volcanic ash

We now explore a dataset of 51 microprobe analyses on glass shards hand-picked from a volcanic ash. The volcanic glass has been affected by chemical weathering at an initial stage and the shards therefore exhibit glass hydration and sodium depletion in some sectors. We would like to study the distribution of sodium (in wt%) in the 51 analyses. The data are stored in the file *sodiumcontent.txt*.

1. Open a new script called “*sodium\_LASTNAME.m*” and store the data from *sodiumcontent.txt* in a variable called *Na* using MATLAB’s *load* function.
2. We will first calculate and display some properties of the distribution. Display this information as follows:

CENTRAL TENDENCY

Sample size: XX

Sample mean: XX.XX (UNIT)

Sample median: XX.XX (UNIT)

DISPERSION

Range: X.XX (UNIT)

Standard deviation: X.XX (UNIT)

DISTRIBUTION'S SHAPE

Skewness: X.XX

Kurtosis: X.XX

1. Now we would like to see a histogram , making sure to satisfy all the rules below:
   * Use an appropriate number of histogram bins for the number of data values
   * Allow MATLAB to auto-distribute the bin centers according to the range of the different datasets.
   * Add a label to the x axis “Na (UNIT)” with the appropriate unit.
2. Discuss how the distribution’s shape is consistent with the properties calculated in question 11 (3 sentences max)

The distribution’s shape is consistent with the properties calculated in question 11 because the mean is between the dispersion displayed, the median is similar to the mean, the range is the highest bin minus the lowest bin, the standard deviation is reflected in the spread, and the skewness has a negative tail, and the kurtosis is peaked and therefore greater than 3.

1. The dataset seems to have one outlier. Modify your script so that the outlier is removed from the dataset right after the load command. What properties are most affected by this change (3 sentences max)

The properties most effected by the removal of the outlier are the range and mean. This is because the range is a measure of the highest value minus the lowest value and the mean is the average of all the values.

Part 3: What to Hand in?

Like previous labs, upload all of your files to Canvas. For your convenience, I have provided a list of the required files below.

**Script Files**: EarthHist\_LASTNAME.m, Earthinfo\_LASTNAME.m, sodium\_LASTNAME.m

**Lab report**: lab5\_LASTNAME.pdf