Data Analysis Workshop - Activity 2

Introduction

This activity is designed to be completed using MATLAB. You will be provided with a code snippet which will contain a working example. Your job is to modify this example code in order to complete the same task for a different data set.

To get started:

1. You will need to log on to a virtual machine which has MATLAB on it - click on the green VM icon on the desktop of your computer
2. Double click on the virtual machine called *desktop*
3. Once the virtual machine has loaded, open MATLAB from the windows *Start* menu

The following is a table which lists the dataset file names for each activity and part.

Part	**Type**	**Filename**
Part 1	Example	`energy_efficiency_data.csv`
Part 1	Task to Complete	`car_data.csv`
Part 2	Example	`car_data.csv`
Part 2	Task to Complete	`energy_efficiency_data_east.csv`
		`energy_efficiency_data_west.csv`

Part 1

The Task

This part should take you about 10 minutes or so, and is designed to get you familiar with estimating population parameters using confidence intervals in MATLAB.

An example code snippet has been provided for a different data set. The code snippet can be found in the learning materials on Learnline, it is called `activity2_part1_example.m`. This example code relies on a data set called `energy_efficiency_data.csv`. The first thing

you should do is to load `activity2_part1_example.m` into MATLAB and run the script. Your job will be to create a similar script which modifies the example code to accept and run with a different data set.

Overview of the Data Set

The dataset that you will use for this task is called `car_data.csv`. It is a very simple data set based around fuel efficiency of cars from the US and cars from Japan. The first column in the data set is kilometres per litre for cars designed and manufactured in the U.S., and the second column is kilometres per litre for cars designed and manufactured in Japan.
If you would like to know more about this data set, please visit the source:

`http://www.itl.nist.gov/div898/handbook/eda/section3/eda3531.htm`

The data is provided in .csv format - note that sometimes you will have to format your own data prior to beginning your analysis. **There are code snippets for a similar exercise that you can use to help you with the MATLAB syntax that you need to use and to help you understand what to do.**

Instructions

1. Load the `car_data.csv` data set into MATLAB. To do this you will need to modify the following example code snippet:

   ```
   % Load the .csv data into MATLAB and store it in a variable called data
   data = csvread('energy_efficiency_data.csv',1,0);
   ```

   Note that the `car_data.csv` dataset does not have names in the first row, so doesn't require the `,1,0)` part. See the first activity for help if you get stuck.

2. Create a single vector of fuel efficiency for the US data only. Call your vector `u_fuel_eff`. You will need to modify the following code snippet:

   ```
   % Create a vector of data for the heating load
   heating_load = data(:,9);
   ```

3. Plot a histogram of the dataset to get an idea of the distribution. You need to modify the following code snippet:

   ```
   % Plot histogram for the heating load data
   histogram(heating_load)
   xlabel('Heating Load (Kilowatts)')
   ylabel('Frequency')
   ```

   How many observations are there in your data? Is the variable we are looking at roughly normally distributed? Why are these things important to know prior to calculating the confidence interval?

4. Calculate the mean of the vector and store this in a variable called `xbar`. You will need to modify the following code snippet:

```
% Find x-bar (the sample statistic for the mean)
xbar = mean(heating_load)
```

5. Calculate the standard error - check the example code to help you with this. Call your standard error `SEM`. You will need to modify the following code snippet:

```
SEM = std(heating_load)/sqrt(length(heating_load));     % Standard Error
```

6. Calculate your t-scores (in a vector) for 2.5% and 97.5% - this will give you a 95% confidence interval. Call your t score vector `ts`. You will need to modify the following code snippet:

```
ts = tinv([0.025  0.975],length(heating_load)-1);     % t-Score
```

7. Calculate the 95% confidence interval for the true population mean. You will need to modify the following code snippet:

```
CI = xbar + ts*SEM      % Confidence Intervals
```

8. How would you structure a sentence to report this confidence interval?

Part 2

The Task

This part should take you about 10 minutes or so, and is designed to get you familiar with hypothesis testing of population parameters using two sample t-tests in MATLAB.

An example code snippet has been provided for a different data set. The code snippet can be found in the learning materials on Learnline, it is called `activity2_part2_example.m`. This example code relies on a data set called `car_data.csv`. The first thing you should do is to load `activity2_part2_example.m` into MATLAB and run the script. Your job will be to create a similar script which modifies the example code to accept and run with a different data set.

Overview of the Data Set

The datasets that you will use for this task represents two samples:

- `energy_efficiency_data_east.csv`
- `energy_efficiency_data_west.csv`

The samples are from a larger data set, and the first sample is of easterly facing dwellings, and the second sample represents westerly facing dwellings. The larger data set comes from an energy analysis which was performed using 12 different building shapes simulated in Ecotect. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters. We simulate various settings as functions of the afore-mentioned characteristics to obtain 768 building shapes. The dataset comprises 768 samples and 8 features, aiming to predict two real valued responses.

The datasets contains eight attributes and two responses (or outcomes). This variables listed across the columns are:

Column	Variable
1	Relative Compactness
2	Surface Area
3	Wall Area
4	Roof Area
5	Overall Height
6	Orientation
7	Glazing Area
8	Glazing Area Distribution
9	Heating Load (Measured Response)
10	Cooling Load (Measured Response)

If you would like to know more about this data set, please visit the source:

`http://archive.ics.uci.edu/ml/datasets/Energy+efficiency`

The data is provided in .csv format - note that sometimes you will have to format your own data prior to beginning your analysis. **There are code snippets for a similar exercise that you can use to help you with the MATLAB syntax that you need to use and to help you understand what to do.**

Instructions

To complete this part of the activity, complete the following steps:

1. Load the 2 samples of data into MATLAB. We interested in the heating load for easterly facing buildings and westerly facing buildings. You will need to modify the following code snippet:

```
% Load the .csv data into MATLAB and store it in a variable called data
data = csvread('car_data.csv');
```

> **ANSWER**
>
> Note that this is a bit tricky because we are using two datasets (one for each sample). Your code should be:
> `data1 = csvread('energy_efficiency_data_east.csv');`
> `data2 = csvread('energy_efficiency_data_west.csv');`

2. We need to create vectors of data which we would like to compare statistically. Create a vector for heating load for east facing buildings, and one for western facing buildings. You need to modify the following code snippet:

```
% Seperate the samples into two variables us_cars and japanese_cars
us_cars = data(:,1);
japanese_cars = data(:,2);
```

   You will need to determine which column the heating load is in - check the data overview.

3. How many observations are in each data set? Why is this important? Create a side by side histogram plot of the two data sets to see their distributions. You will need to modify the following code snippet:

```
% Plot histogram to get an understanding of the distribution
subplot(1,2,1)
histogram(us_cars,10) % histogram plot for easterly heat loading
xlabel('Efficiency (km/L)')
ylabel('Frequency')
title('US')
subplot(1,2,2)
histogram(japanese_cars,10) % histogram plot for westerly heat loading
xlabel('Efficiency (km/L)')
ylabel('Frequency')
title('Japan')
```

Why is it important to know what the distribution looks like? Is it appropriate to perform a two sample t-test?

4. We would like to see if there is a statistically significant difference between the heating load for eastern facing buildings and western facing buildings. What would your null hypothesis be?

   H_0 :

5. We are running a two tailed test at the 5% significance level. What is your alternative hypothesis?

   H_a :

6. Run the two sample t-test. You need to modify the following code snippet:

```
% Run the two sample t-test which provides a statistically indication
% if the two samples are from the same or different populations
% (That is, do the cars from japan perform the same as the cars
% from the US in terms of fuel efficiency, statistically speaking)

% The important value here is the p-value
% low p-value means statistically different
% high p-value means not statistically different

[h,p,ci,stats] = ttest2(us_cars,japanese_cars)
```

7. What is your p-value?

   `p-value:`

8. What is the conclusion of your statistical test? Write a sentence which accurately conveys your findings using appropriate language.