# ECE 3710-002 – Applied Probability and Statistics for Engineers and Scientists

## Spring 2024

## Project 1

## Due February 23, 2024, 11:59 pm

You should work on this project in a team of 2 to 3 people. Please send me your team member's names by February 5th, 2024.

## Getting started

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States. As its name implies, the BRFSS is designed to identify risk factors in the adult population and report emerging health trends. For example, respondents are asked about their diet and weekly physical activity, their HIV/AIDS status, possible tobacco use, and even their level of healthcare coverage. The BRFSS Web site (http://www.cdc.gov/brfss) contains a complete description of the survey, including the research questions that motivate the study and many interesting results derived from the data.

We will focus on a random sample of 20,000 people from the BRFSS survey conducted in 2000. While there are over 200 variables in this data set, we will work with a small subset.

First, load the cdc.csv file into MATLAB:

```
brfssData = readtable('cdc.csv');
```

This line of code will read the data into a table.

Note, there are other ways to read such data into MATLAB. You can use the "Import Data" tool on the Home ribbon. Or you can use the command `uiimport('-file')`.

You have several options on how to structure the imported data. The instructions below assume that you import the data as a table. You could read the data in as column vectors, for example, but you will need to alter the syntax accordingly.

The data now shows up in your workspace is a table with each row representing a case and each column representing one variable (i.e., `smoke100`).

Look at the names of the variables in your Workspace. They should be `genhlth`, `exerany`, `hlthplan`, `smoke100`, `height`, `weight`, `wtdesire`, `age`, and `gender`. Each one of these variables corresponds to a question that was asked in the survey.

For example, for `genhlth`, respondents were asked to evaluate their general health, responding either excellent, very good, good, fair, or poor. The `exerany` variable indicates whether the respondent exercised in the past month (‘`1`’) or did not (‘`0`’). Likewise, `hlthplan` indicates whether the respondent had some form of health coverage (‘`1`’) or did not (‘`0`’). The `smoke100` variable indicates whether the respondent had smoked at least 100 cigarettes in her lifetime. The other variables record the respondent’s height in inches, weight in pounds as well as their desired weight (`wtdesire`), age in years, and gender.

## **Summaries and tables**

The BRFSS questionnaire is a massive trove of information. A good first step in any analysis is to distill all of that information into a few summary statistics and graphics. As a simple example, the functions `mean()`, `median()`, `iqr()`, `min()`, and `max()` return useful summary statistics.

To determine the median weight of people in this dataset use:

```
weightMedian = median(brfssData.weight)
```

While it makes sense to describe a quantitative variable like `weight` in terms of these statistics, what about categorical data? We would instead consider the sample frequency or relative frequency distribution. The function `tablulate()` does this for you by counting the number of times each kind of response was given. For example, to see the number of people who have smoked 100 cigarettes in their lifetime, use:

```
smokeTbl = tabulate(brfssData.smoke100)
```

This function creates an array with the counts and percentages for each category.

Next, we make a bar plot of the entries in the table using the `bar()` command and select out the second column (the counts) from the `smokeTbl` variable:

```
bar (smokeTbl(:,2))
```

Note that if the categories are non-numeric, as in gender, then you may need to convert cell values to matrix values. For example:

```
genderTbl = tabulate(brfssData.gender)

bar(cell2mat(genderTbl(:,2)))
```

**Question 1:**

a. Create a numerical summary for `height` and `age`:
   i. Plot boxplots for each and compute the interquartile range (the `iqr()` function will help).

b. Compute the relative frequency distribution for `gender` and `exerany`:
   i. Create a barplot for both the `gender` and `exerany` variables.
   ii. What proportion of the sample is female?
   iii. What proportion has exercised in the past month?

**Question 2:**

The `crosstab()` command can be used to tabulate any number of variables that you provide. For example, to examine which participants have smoked across each gender, we could use the following.

```
crosstab(brfssData.gender,brfssData.smoke100)
```

This command returns a table in which the rows represent `gender` (‘`m`’, ‘`f`’) and the columns represent `smoke100` (‘`0`’, ‘`1`’). Recall that ‘`1`’ indicates a respondent has smoked at least 100 cigarettes. So, the number in row #1 and column #1 is the males (‘`m`’) that has not smoked 100 cigarettes (‘`0`’).

What does the output of the `crosstab()` function reveal about smoking habits and gender?

**Question 3:**

a. Find the mean and standard deviation of weight.
b. What proportion of the weights is within one standard deviation of the mean? Does this match up with what you would expect?

## What to turn in

One of the team members should upload a zip file containing the project report (in PDF) and MATLAB code (in .m) to CANVAS by the deadline, and all your team members' names must be on the first page of the report.

Note: you can also complete this project in Python, however, assistance may not be available. You will submit Python code (in .py) instead of MATLAB code in this case.

You will also turn in a report summarizing your findings. For each question,

- Describe your approach to answering these questions.
- Describe any assumptions you are making.
- Provide the results of your analysis using tables, plots, equations, etc. if applicable.
- Briefly explain the meaning of each result.

Submission without MATLAB code (or Python code) will result in zero scores for this project assignment.

You can discuss this project assignment with another team, but the work submitted must be your team's own effort.