MATH 4044 – Statistics for Data Science

Practical Week 2

Exercise 1

Statistics can be used to filter spam from incoming email messages. By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy. One of those characteristics is whether the email contains no numbers, small numbers, or big numbers. Another characteristic is whether or not an email has any HTML content.

Data file for this exercise is based on a sample of 50 emails stored in a SAS data file called `email50.sas7bdat` located in `mydata` library on the SAS OnDemand server. The data statement to access this file is `data=mydata.email50`

Some of the variables in that file are as follows:

Variable	Description
spam	Specifies whether the message was spam; 0 = no, 1 = yes
num_char	The number of characters in the email
line_breaks	The number of line breaks in the email (not including text wrapping)
format	Indicates if the email contained special formatting, such as bolding, tables or links, which would indicate the message is in html format; 1 = html, 0 = text
number	Indicates whether the email contained no number, a small number (under one million) or a large number; none = no number, small = number under one million, big = large number

(a) Obtain a frequency distribution table of variables *spam* and *format*, with *spam* as the row variable. Which would be more helpful to someone hoping to classify email as spam or regular email: row or column percentages?

(b) Obtain the clustered bar chart and 100% stacked bar chart of the same variables. Which would be more helpful?

(c) Repeat parts (a) and (b) with variable *number* instead of *format*.

(d) Would either characteristic, *format* or *number*, alone be effective in identifying spam email? Explain briefly.

Exercise 2

Data file for this exercise is called `marathon.sas7bdat` and stored in `mydata` library. The data statement to access this file is `data=mydata.marathon`

It contains finishing times, in hours, for male and female winners of the New York marathon between 1980 and 1999.

(a) Obtain a histogram and boxplot of finishing times. What features of the distribution are apparent in the histogram and not in the boxplot? What features of the distribution are apparent in the boxplot and not in the histogram?

(b) The distribution of finishing times is bimodal – it has two distinct peaks. What may be the reason for the bimodal distribution? Explain.

(c) Obtain a boxplot of finishing times by gender and compare the distribution of marathon times for men and women. Comment briefly.

Exercise 3

Data file for this exercise is called `cars.sas7bdat` and comes from the `sashelp library`.

The data statement to access this file is `data=sashelp.cars`

Suppose we wish to investigate fuel economy of cars in city vs highway driving conditions based on their origin (Asia, Europe and US). Variables of interest are therefore *Origin*, *MPG_City* and *MPG_Highway*.

(a) Obtain Descriptive Statistics, histograms and boxplots of *MPG_City* by *Origin*. Use a variable of your choice to identify outliers.

(b) Obtain Descriptive Statistics, histograms and boxplots of *MPG_Highway* by *Origin*. Use a variable of your choice to identify outliers.

(c) Discuss your results from parts (a) and (b). What are some of your key observations?

For histograms in separate panels of the same graph you can use the following code:

```
proc univariate data=sashelp.cars noprint;
     histogram / nrows=3;
     var MPG_City;
     class Origin;
run;
```