STAT 270 Assignment 1

Upload to Crowdmark by 5pm on Monday, May 25, 2020

Instructions: Be sure to label the axes appropriately on all graphs. Also, you must provide any code you use (final computed values and plots are not sufficient). **I strongly recommend that you create a Jupyter Notebook for this assignment**, embedding your written comments in Markdown cells. However, if you wish to complete all or part of the assignment by hand, you may scan or take a picture of your pages and then upload these (in .jpg format) to Crowdmark.

1. (7 marks) A study is conducted to investigate the prevalence of Piscine Orthoreovirus (PRV) in wild BC chinook salmon. The researchers catch 50 salmon in the Georgia Strait, recording on each the presence/absence of the virus, the distance from the nearest commercial fish farm when caught (in km), and a rating of the degree of inflammation in the heart muscle tissue (a number from 0 to 3, where 0 represents no inflammation and 3 represents severe inflammation).

   a. What is the likely target population in this study? (1 mark)
   b. What is the sample? (1 mark)
   c. Is the sample likely to be representative of the target population? Explain. (2 marks)
   d. List each variable in this study and identify its type (according to the flow chart given in Lecture 3). (3 marks)

2. (6 marks) A group decides to investigate residual pesticides found on fives types of fruits and vegetables (strawberry, kale, avocado, orange, and mushroom) sold in grocery stores in Vancouver. The file `pesticide.csv` contains the total number of samples of each fruit/vegetable collected (`Total`) and the number of those samples that tested positive for residual pesticides (`Pesticide`). NOTE: In R, if the data are stored in a data frame called `data`, you can add a column representing the *proportion* of samples that tested positive for pesticides using the command

   ```
   data$Proportion=data$Pesticide/data$Total
   ```

   a. Create an appropriate plot to display the *frequency* associated with each fruit/vegetable that tested positive for pesticides. (3 marks)
   b. Create an appropriate plot to display the *relative frequency* associated with each fruit/vegetable that tested positive for pesticides. (3 marks)

3. (15 marks) Consider again the scenario described in Question 2. The file `mushroom.txt` lists the number of pesticides (`Number`) found on each of the 82 mushroom samples.

   a. Create an appropriate plot to display the *proportion* (or *percentage* – your choice) of samples that tested positive for 0, 1, 2, …, 15 pesticides. (3 marks)
   b. Using your plot, describe the apparent shape, spread, and "centre" of the number of pesticides. Comment on unusual points, if any. (4 marks)
   c. What is the sample mode of the number of pesticides? (1 mark)
   d. Create a boxplot of the number of pesticides. How would you interpret the 3 horizontal lines in the box for someone who has never taken a statistics course? (3 marks)
   e. Compute the interquartile range of the number of pesticides. Do you expect this value to be sensitive to outliers? Explain. (2 marks).
   f. Compute the sample standard deviation of the number of pesticides. Do you expect this value to be sensitive to outliers? Explain. (2 marks).

4. (6 marks) Say we are interested in the maximum daily temperature on August 1 in Vancouver. We look up the maximum temperature observed on August 1 (in ºC) over the previous *n* years. We record these values as $x_1, x_2, \dots, x_n$. We compute the sample mean as $\bar{x}$ and the sample standard deviation as s_x. We then convert the temperatures to ºF (recording them as $y_1, \dots, y_n$) via the following formula:

$$y_i = \frac{9}{5}x_i + 32$$

   a. Give the sample mean of $y_1, \dots, y_n$ in terms of $\bar{x}$ and/or s_x. (3 marks)
   b. Give the sample standard deviation of $y_1, \dots, y_n$ in terms of $\bar{x}$ and/or s_x. (3 marks)

5. (6 marks) We can *centre* observations of a numeric variable, $x_1, x_2, \dots, x_n$, by computing $x_i^* = x_i - \bar{x}, i = 1, \dots, n$. Let the sample mean and standard deviation of $x_1, x_2, \dots, x_n$ be $\bar{x}$ and s_x, respectively.
   a. What is the sample mean of $x_1^*, x_2^*, \dots, x_n^*$? (3 marks)
   b. Give the sample standard deviation of $x_1^*, x_2^*, \dots, x_n^*$ in terms of $\bar{x}$ and/or s_x. (3 marks)