MACQUARIE University
SYDNEY·AUSTRALIA

STAT2170 AND STAT6180
APPLIED STATISTICS

SGTA Week 7 : Data Management and Presentation 2023

Part 1

Question 1

Imagine that you are about to go on an overseas trip to a destination where it is easy to buy fashionable clothing at great prices. Your extended family have all requested that you bring some home and you need to keep a detailed list of their measurements and favourite colours so that you can bring home just the right garment for them!

We are going to set up an Excel spreadsheet to record all of this information. It will contain the following EIGHT fields (columns):

- name (of family member)
- gender (male/female)
- age (years)
- favourite colour (black/blue/red/green/white/other)
- height (cm)
- weight (kg)
- waist (cm)
- arm length (cm)

Note: It's usually recommended to add a numeric ID field to uniquely identify each record (i.e., each family member) in the data. Here, for simplicity, their name may be used as identifier as long as each name in your extended family is unique, which is only acceptable if there are ONLY a small number of subjects.

You may follow the steps suggested below.

a. Before even opening Excel program, you need think very carefully how you are going to enter each variable, particularly for categorical variables. E.g.: For the variable Gender, you may either enter as a text (string) variable as, Male or Female, M or F, or 1 (for male) or 0 (for female).
b. Open Excel (Look under Microsoft Office) and name each of the eight columns on the top of each column (i.e. using first row), one for each variable above.
c. Once you have set those up for the spreadsheet, start typing in each piece of the data collected on each member, one record per subject (family member). Please do it for a few members (or may just make up a few). Always double check each data entry.

Question 2

Use the file created in Excel created above to do the following:

a. Import the data from Excel to R. If you saved the file as an xlsx file it can be converted to csv with the following:
    - Save the Excel file as a .csv file
        - Goto File → Save As

  - Choose location
  - Click "Save as Type" and select "CSV (Comma delimited) (*.csv)"
- Once the csv is created, load the newly saved data file into RStudio by either
  - Using the Tools → Import Dataset menus
  - Command Line: `read.csv("myfilename.csv", header = TRUE)`

b. Practice some simple skills in `RStudio`.

- Compute BMI (= kg/m^2)
  - Place BMI as a new variable in the dataframe
  - Optional: truncate the BMI to 2 decimal places using the `round` function (see `? round`)
- Re-code a variable
  - E.g. recode M/F to Male/Female
- Calculating mean height in cm using an appropriate formula
- Use the `which` command to find a particular family member in your dataset.
  - Display their record using a index reference (e.g. `mydata[myfamilymember, ]` where `myfamilymember` is the index computed with the `which` command)
- Get the subset of the data with only males (or females) using
  - cell referencing
  - subset function

c. Author a simple Rmarkdown document that creates a description of your dataset with a boxplot and histogram of some of the variables. (*Note*: you can create a visual summary of the categorical variables such as gender or favourite colour variables with the `barplot` command)

Part 2

`R` and `RStudio` provide us with options to change various attributes of a plot such as caption, label, font size, and colour. See `help(par)` for more information on graphical parameters. To explore these options further, let's first import into `R` or `RStudio` the dataset `axial.txt`, which could be found on iLearn, and then name the imported dataset `axial`. Recall that the first column of `axial` gives out the `Axial Stiffness Index`.

Question 1

Using the option `main`, which controls the caption/title of a plot, make a histogram of the first column of `axial` and change this histogram's caption into `Axial Stiffness Index`.

Question 2

The option `col` controls the colour of the elements of a plot. Please correct all the errors in the codes below, which try to

- produce a comparative boxplot of `Axial Stiffness Index`, the first column of `axial`, for each group of treatment
- colour the boxes with red, blue, green, yellow, and pink, respectively.

```
boxplot(asi ~ treatment, col = c(red, blue, green, yellow, pink)
```

Question 3

The option `cex.lab` controls the font size of the axis labels.

a) Use the function `qqnorm()` to generate a normal QQ-plot of `Axial Stiffness Index`.

b) Use `cex.lab`, change the font size of the axis labels to 150% of the original size.

Question 4

Now we try to generate the normal QQ plot in the previous question in a somewhat manual way.

a) Use the function `ppoints()` to generate a probability sequence of length 35, the number of observations we have in `axial`, that is symmetric and all values are within $(0, 1)$. You may want to save these probabilities in a vector called `s`.

   `ppoints()` is the function called by `qqnorm()` to generate the set of probabilities at which to evaluate the inverse distribution. `ppoints()` is preferred over `seq()` in this context as it excludes the boundaries.

b) Use the function `quantile()` to find the quantiles of `Axial Stiffness Index` with the probabilities you generated in part a), and put these quantiles into a vector called `y`.

c) Use the function `qnorm()` to find the quantiles of the standard normal distribution with the probabilities you generated in part a), and put these quantiles into a vector called `x`.

d) Generate a scatter plot with the horizontal axis being `x` and the vertical axis being `y`.

e) Use the option `xlab` and `ylab` to change the label of the horizontal axis and vertical axis of the scatter plot in the previous sub-question into `Theoretical Quantiles` and `Sample Quantiles`, respectively.

Question 5

Now please try to produce a graph where we put the 2 plots generated in the last sub-questions of Question 3 and Question 4 side by side.

- Hint: To draw 3 plots side by side, one could enter `par(mfrow=c(1,3))` in the `R` console before drawing the 3 plots one by one.