**HH/HLST 2300:**  **Statistical Methods in Health Studies**

**Fall Assignment**

Assigned: Friday November 5, 2021; Due 5PM Tuesday December 7, 2021

Submit to Crowdmark via eClass (Instructions will be posted by Monday November 8, 2021)

A researcher has collected data for 7681 **patients** who had a procedure performed at Clinic ABC in 2009 that serves the Greater Toronto Area regions (Excel file: Fall Assignment Data.xls). The data includes the unique **patient identifier**, **sex** (male = 1; female = 2), **age** (years) and **patient region** (Durham = 1; Halton = 2; Peel = 3; York = 4). The researchers also collected some lifestyle data including smoking frequency (every day = 1; some days = 2; not at all = 3) and whether the patient eats herbal supplements (yes = 1; no = 2). The researcher also collected data on the number of times in a week the patient performs vigorous activity (where 99 is the user-defined missing value with the label ‘unable to do vigorous activity), the number of times in a week they perform moderate activity (where 99 is the user-defined missing value with the label ‘unable to do moderate activity), and the number of times in a week they perform strength activity (where 99 is the user-defined missing value with the label ‘unable to do strength activity). Note that for these three activity variables, there are also system-missing values in the dataset. Finally, we have the procedure order date/time, procedure completion date/time and the procedure score.

Before proceeding with any analysis, be sure to:

* Ensure that variables are of correct measure (nominal, ordinal, scale). Remember to also do this for any newly created variables
* Add labels to all categorical variables (remember to also do this for any newly created categorical variables)
* Add missing value labels for the three activity variables (this will be covered in Lab 11)
* For the procedure order and completion date/time variables, in the variable view, under column ‘type’, ensure that the Date format is set to dd-mmm-yyyy hh:mm.

1. Generate Variable Information and Variable Values tables for the dataset. Copy and paste these generated SPSS tables. Note, while you already know how to generate these tables, you may want to hold off on this question until you learn more about missing values so that you don’t have to generate these tables twice.
2. Create a new variable, ‘Age\_cat1’ that groups the variable ‘Age’ into two categories: <65yrs and ≥65 yrs. Copy and paste the screenshots of the SPSS window(s) used to show how you created the new variable. Remember to add labels to all values. Generate a pie chart of this new variable where the slices represent proportions belonging to each category. Show data labels to display the percentages in each slice (round to 2 decimal places and show the % symbol). Copy and paste the generated SPSS pie chart.
3. Create a new variable, ‘Smoking\_binary’ that groups the variable ‘Smoking Frequency’ in two categories: yes (combines categories every day and some days) and no (category not at all). Copy and paste the screenshots of SPSS window(s) used to show how you created the new variable. Remember to add labels to all values. Generate a pivot table that includes the observed counts of ’Smoking\_binary’ (placed as rows) by ‘Region’ (placed as columns), layered by ‘Sex’. Copy and paste the generated SPSS pivot table. Answer the following question: How many female Halton region residents smoke?
4. Create a new variable called ‘Age\_cat2’ that groups the variable ‘Age’ into the following: 18-24 yrs, 25-44 yrs, 45-64 yrs, 65+ yrs. Generate a stacked bar graph such that each bar represents a different ‘Age\_cat2’ category and the stacks represent the percentage of the various categories within the ‘Smoking Frequency’. Order the stacks such that the ‘every day’ category is the bottom or base stack and the ‘not at all’ category is the top stack. Add data labels (1 decimal place with % trailing character). Copy and paste the generated SPSS stacked bar graph.
5. Create a new variable ‘zProcedurescore’ that is the standardized variable of ‘Procedure Score’. But, before you generate the z-scores for ‘Procedure Score’, ensure that the variable is not markedly non-normal by copying and pasting the appropriate generated SPSS output table and showing your calculations to determine skewness. After you’ve created the new variable ‘zProcedurescore’, generate case summaries of the variables ‘Procedure Score’ and ‘zProcedurescore’ for the first 10 cases in your dataset. Copy and paste the generated SPSS table and answer the following question: Is case 1’s procedure score (.1578) above or below the mean procedure score? Justify your answer with the Case Summary table.
6. Create a new variable ‘rndScoretwodecimal’ that rounds the ‘Procedure Score’ to its nearest second decimal place. Copy and paste the SPSS compute variable window used to create this new variable. Generate a frequency histogram for this new variable paneled by ‘Herbal Supplements’ (rows) and overlaying the normal curve. Set the bin width = 0.1. Display data labels in each bar to add clarity. Copy and paste the generated SPSS frequency histograms.
7. Create a new variable “Yrs\_till\_retirement\_peel” that is the number of years until a Peel region patient retires with the condition that only Peel patients that are younger than retirement age (retirement age begins at age 65 yrs) are selected. Copy and paste the SPSS compute variable window used to create this new variable and highlight/place a box around the (1) numeric expression and (2) conditional expression used to create this variable. Generate a boxplot graph of “Yrs\_till\_retirement\_peel” which contains the boxplots for males and females side-by-side. Copy and paste the generated SPSS boxplot graph.
8. Create a new variable called ‘Procedure\_duration’ that is the length of time between the procedure order date/time and procedure completion date/time, measured in hours with the fractional part retained. Copy and paste the screenshots of the SPSS window(s) used to show how you created the new variable. Generate a descriptive statistics table for ‘Procedure\_duration’ that includes the mean, standard deviation, median, quartiles, 80th percentile and skewness. Copy and paste the generated SPSS output table.
9. Create a new variable called ‘Procedure\_order\_hour’ that extracts the hour that the procedure was ordered. Copy and paste the screenshots of the SPSS window(s) used to show how you created the new variable. Generate a frequency table for ‘Procedure\_order\_hour’. Copy and paste the generated SPSS output table. Answer the following question: What is the mode of ‘Procedure\_order\_hour’ (state using AM or PM).
10. Generate a frequency table for ‘Weekly Vigorous Activity’. Copy and paste the generated SPSS output table. Conduct a Missing Values Analysis with EM estimation for the variables ‘Weekly Vigorous Activity’, ‘Weekly Moderate Activity’ and ‘Weekly Strength Activity’. Are values missing completely at random? Justify your answer by copying and pasting the appropriate SPSS output table to support your answer.