**PubH 6325**

**Data Processing with PC-SAS**

**University of Minnesota - School of Public Health**

**Fall 2021**

**Assignment #1: Checking and Describing Data**

**Submit your answers as a word document including your X500 ID and the assignment number in the document title (ex: demm0009\_Assigment1.doc). The word document should contain all SAS code necessary to generate the answer and when requested any additional text used to report results or summarize findings. Please also submit the complete SAS editor file used to complete the assignment; include your X500 ID and the assignment number in the title of the SAS editor file (ex: demm0009\_Assigment1.SAS).** In the editor file, make sure to label questions and include commentary and notes as appropriate, using the following syntax, as discussed in class: */\*COMMENTS HERE\*/.*

**Introduction**

In this assignment, you will be asked to utilize the SAS procedures you have learned during your first week of the course. This assignment is worth 100 points (20% of the final grade) and is to be completed independently in compliance with the Student Conduct Code. **There should be no active collaboration or discussion between students to complete the assignments**.

The following are examples of unacceptable collaboration but in general if there is any question about appropriate discussion between students, ask the instructor, TA or simply refrain from discussion: i) Multiple students turning in duplicates of one editor file; ii) copy/pasting answers from another student’s assignment; iii) working jointly on the assignment in one editor file and then duplicating the file prior to submitting the assignment; iv) setting up a central document (e.g. Google docs) for multiple individuals to edit/comment; v) group work in which there is explicit discussion of the assignment but each student completes and submits their assignment independently.

Throughout the course graded assignments we will be progressing through an analysis similar to a previously published paper by Edwards et al., “Combined associations of sedentary behavior and cardiorespiratory fitness on cognitive function among older adults”. PDFs of this paper and other papers on this topic have been posted on Canvas. Please read them to gain an understanding of the underlying hypotheses and goals of the analysis.

For this assignment, you are interested in looking into smoking and its association with blood biomarkers and demographic characteristics in adults, using NHANES 2001-2002 data.

**This exercise uses adapted versions of the data sets DEMO\_B.sas7bdat, SMQ\_B.sas7bdat and L06\_B.sas7bdat, available on Canvas.**

**In the code provided below, replace “**FILEPATH**” with the directory in your computer in which you saved the files** (ex: /Username/Documents/PUBH6325/).

**Assignment 1**

**Part I. Checking the Data.**

1. At the top of your SAS Editor file, please type your name and X.500 in the following format:

/\*First Last X.500 \*/  
**(1 points)**

1. Check the contents (PROC CONTENTS) of the three SAS data sets provided.  
   **(21 points)**
   1. How many variables are there in each data set? How many labels are there in each data set? What is the function of labels?  
      **(8 points)**
   2. How many observations are there in each of the data sets?  
      **(3 points)**
   3. What information is contained in the column labeled “#”? What option would you use to order this column in ascending order? Write the necessary code (use the data set DEMO\_B).  
      **(4 points)**
   4. You are interested in knowing more about current cigarette smokers and levels of cotinine, a biomarker associated to smoking.  
      How can you use PROC CONTENTS to find the two variables in the datasets that contain the information you are interested in? Report the variables names and which data set(s) contain these variables.

**(6 points)**

1. Use the appropriate procedure to print the data to the SAS output window (\**NOTE: don’t confuse “print” with traditional hard copy printing - you should not print paper copies of the data for this question*). For the following questions, use data set **SMQ\_B.sas7bdat**.

Respondent Sequence Number (SEQN) is the variable that identifies participants by number. Run the following command first, to order the dataset by descending SEQN.   
**(25 points)**

**PROC SORT** data = "FILEPATH/SMQ\_B.sas7bdat";

by DESCENDING SEQN;

**run**;

* 1. Based on the large number of observations in this dataset, it can be advantageous to only print a small number of observations. What code would you use to look at only the first (i.e. highest SEQN) 50 observations?  
     **(4 points)**
  2. The variable SMD030 is labeled *Age started smoking cigarets (sic) regularly*. Among the first 50 observations, which two participants reported starting to smoke at the earliest ages? Report their SEQN and ages.

What statement can you use to assure only the relevant variables (SEQN and SMD030) will be printed? Include the code used in your answer.  
**(5 points)**

* 1. What is the total number of daily cigarettes smoked by the participants in the first 50 observations?   
     Use the PRINT procedure, the SUM statement and the variable SMD070 (# cigarettes smoked per day now) to answer this question. Include the code used in your answer.  
     **(5 points)**
  2. What Statement option would you use to print only observations corresponding to those who smoke cigarettes “some days”? Write down the code used to include only the first 50 “**some days” cigarette smokers (SMQ040 = 2)** in your output.  
     **(3 points)**
  3. The variable SMD090 contains information on the average number of cigarettes smoked per day over the past 30 days.   
     Using this variable and the correct statements within the PRINT procedure, report the total number of cigarettes smoked (over the past 30 days) among the first 50 observations of those who reported smoking “**some days” (SMQ040 = 2)**.  
     Calculate (by hand) the mean number of cigarettes smoked daily among these 50 participants.  
     **(4 points)**
  4. Repeat the previous step among those who reported smoking **daily (SMQ040 = 1).** Using the variable SMD090 and the correct statements within the PRINT procedure, report the total number of cigarettes smoked daily among first 50 observations.   
     Calculate (by hand) the mean number of cigarettes smoked daily among these 50 participants.  
     **(4 points)**

**PART II. Describing the data**.

1. Although PROC PRINT is a great command to visualize the dataset, SAS has commands that are more suited to observe the distribution of variables in the entire dataset. Use PROC FREQ to report the frequency (absolute number) and the percentage of individuals corresponding to each value of the below listed variables. Include the code in your answers.  
   Ignore missing data when answering the following questions, unless otherwise requested.

For the following questions, use data set **SMQ\_B.sas7bdat**.  
**(26 points)**

* 1. Filters (SMD100FL)  
     **(6 points)**

|  |  |  |
| --- | --- | --- |
| **SMD100FL** | **Frequency** | **Percentage** |
| 0 =Non-filtered |  |  |
| 1 =Filtered |  |  |

* 1. Filter smokers, among those who smoke daily (SMQ040=1)  
     **(8 points)**

|  |  |  |
| --- | --- | --- |
| **SMD100FL** | **Frequency** | **Percentage** |
| 0 =Non-filtered |  |  |
| 1 =Filtered |  |  |

* 1. Among those who reported smoking daily, how many participants **did not** answer whether they smoke filtered or unfiltered cigarettes? How many provided an answer?  
     **(4 points)**
  2. One of the most common uses for PROC FREQ is to cross-tabulate two variables (ie; to create 2x2 tables). This is done by adding an asterisk ( \*) between the two variables of interest (ex: A\*B) listed in the TABLE statement.

Use this syntax to cross-tabulate the variables SMQ077 (time between waking up and smoking) and SMD100FL (Filter type indicator). Include all participants. Fill the table below with the frequencies of each combination of response to these two variables.  
**(8 points)**

|  |  |  |
| --- | --- | --- |
| **SMQ077** | **SMD100FL** | |
| **0 = NON-FILTERED** | **1 = FILTERED** |
| 1 =within 5 minutes |  |  |
| 2 =from 6 to 30 minutes |  |  |
| 3 =from 30 to 60 minutes |  |  |
| 4 =more than 60 minutes |  |  |

1. PROC FREQ is particularly useful for categorical variables. For continuous numeric variables PROC MEANS is commonly used. Answer the following questions regarding the age at which participants reported last smoking cigarettes regularly (SMD055). Note that this does not include second-hand smoking, or consumption of other tobacco/nicotine products.  
   **(27 points)**
   1. Use the appropriate procedure to report the mean, standard deviation, minimum and maximum values for this variable.  
      Include the code used.  
      **(7 points)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Mean** | **St Deviation** | **Minimum** | **Maximum** |
|  |  |  |  |

* 1. Often researchers will select number(s) outside of the expected range to represent non-numerical responses, such as refusal to answer, and/or not knowing the answer.  
     Why can this be a problem? Do you think this could be a problem with the variable SMD055?  
     **(5 points)**
  2. Independent of your answer in b), describe what you could do to deal with having these out-of-range values in a continuous variable.  
     **(2 points)**
  3. Use the FREQ procedure to check if values of SMD055 might be out range. Remember that unit of the variable is years, and that participants were asked at what age they last smoked cigarettes regularly.  
     Include the code used.  
     **(6 points)**
  4. If appropriate, use the WHERE command to remove value(s) outside of the expected range of the variable SMD055.

Include the code used.  
(HINT: if you need to remove more than one value choose an appropriate syntax, e.g. where VARIABLE ^= **X** and VARIABLE ^= **Y**;)  
**(7 points)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Mean** | **St Deviation** | **Minimum** | **Maximum** |
|  |  |  |  |