**EPI 624 TAKE-HOME FINAL**

**SPRING 2022**

This take-home final exam accounts for 30 points total (plus 5 points for Extra Credit).

This assignment is due at the end of Module 14, on Tuesday, April 26 at 11:59pm. Please copy and paste all of your SAS code, clean log, and output, fill out the below table, complete final report (described below) and submit the document through Canvas.

The objective is to determine whether there are differences in selected demographic or clinical characteristics among 2019 National Health Interview Survey respondents who are current cigarette smokers.

Attachments: NHIS19\_everydaysmk.sas7bdat, NHIS19\_sometimessmk.sas7bdat

Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Participant Characteristics |  | Smoking Frequency? | | P value |
| All | Every Day | Sometimes |
| (N =3890 ) | (N = 3044 ) | (N = 846 ) |
| Age\* | 49.18(15.44) | 50.25(15.07) | 45.33(16.14) | <0.0001 |
| Sex, n (%)\*\* |  |  |  |  |
| Male |  |  |  |  |
| Female |  |  |  |  |
| Education Level, n (%)\*\* |  |  |  |  |
| <HS Graduate |  |  |  |  |
| HS Graduate/GED |  |  |  |  |
| Some College |  |  |  |  |
| Associate Degree |  |  |  |  |
| Bachelor’s Degree or Higher |  |  |  |  |
| BMI\* |  |  |  |  |
| Any Cardiovascular Disease, n (%)\*\* |  |  |  |  |
| Yes | 432(11.11) |  |  |  |
| No |  |  |  |  |
| Please fill in total N for each column heading where it says (N = )  \*For continuous variables, report Mean (SD) or Median (IQR) as appropriate for each column (All, Yes, No) at top of table. Specify which values you are reporting on the left hand side of the table.  \*\*For categorical variables, report n and % of total N for each column (All, Every Day, Sometimes) at top of table.  ***Do not fill numbers into the shaded cells.*** | | | | |

**Tasks**

* Bring data into SAS (2 points)
  + NOTE: If you're having issues seeing the dataset after bringing it into SAS, please do the following:
    - Open the word document containing proc format code (named "Selected NHIS Formats", attached to the final exam)
    - copy/paste code into your program.
    - add proc format statement before pasted code (and run statement after)
    - select and run
* Combine datasets for every day and sometimes cigarette smokers (2 points)
  + Make sure that in the combined dataset, there is a variable that will allow you to determine if an observation an every day smoker or not.
* Prepare and clean combined data set (15 points)
  + Please check variable names and values (also listed in Codebook). All included variables are numeric.
  + Use the provided stroke, heart attack, angina and coronary heart disease variables to create a new composite variable for any cardiovascular disease (named *any\_cvd*)
    - Participants who reported that they had at least one of these conditions would be coded as “Yes”
    - Participants who answered “No” to all of the questions asking if they had any of these conditions would be coded as “No”
    - All other participants should be coded as missing
  + Exclude anyone from the dataset if the person has a missing value for any\_cvd or invalid (Refused, Not Available, Don’t Know, Not Ascertained) value for one or more of the other variables.
    - "Other variables" means ***other variables not involved in the creation of the any\_cvd variable***(i.e. education, age, sex, height, weight).
  + Recode education variable into a new categorical variable (named *educ\_cat*), new categories are listed below, with values of original variable in parentheses
    - <HS Graduate (00, 01, 02)
    - HS Graduate/GED (03, 04)
    - Some College (05)
    - Associate Degree (06, 07)
    - Bachelors Degree or Higher (08, 09, 10, 11)
  + Use height and weight variables and the following equation to calculate body mass index (BMI) (named *bmi\_calc*), rounded to nearest tenth.
    - Note: SAS expression for exponentiation to the nth power: variable\*\*n
* Save final cleaned dataset as a permanent dataset (1 point)
* Produce descriptive statistics (2.5 points, 0.5 point for each characteristic)
* Produce p-values (2.5 points, 0.5 point for each characteristic)
* Complete final report (5 points). The final report should address the following:
  + Describe the process you went through to clean/recode the data. Essentially, your report should allow someone else to replicate your work, based on information provided in the report, and come to the same findings/conclusions.
    - For each original variable:
      * How many observations had missing or invalid values (e.g., Don’t Know, Refused)? How did you handle this?
      * Did you have to recode this variable? If yes, how did you do it?
      * Did you use this variable to create another variable?
      * What is the impact of the approaches you used to clean the data?
    - What are the attributes of the cleaned data set that you saved permanently?
      * How many total included observations?
      * Number of observations having missing/invalid/implausible values for one or more variables
      * Distribution of each variable
    - What procedures did you use to obtain the descriptive statistics and the p-values?
      * Why?

**Extra Credit (5 points):**

The data for the extra credit question come from the *2020 Pew Research Center’s American Trends Panel: Wave 67 Climate and Coronavirus*, conducted April 29 – May 5, 2020. Please complete the following:

1. Import the data set (2 points)
2. Please use ARRAY and DO loops to recode variables corresponding to the following questions in the questionnaire (3 points):
   1. COVID\_SCI6\_a\_W67, COVID\_SCI6\_b\_W67, COVID\_SCI6\_c\_W67, COVID\_SCI6\_d\_W67
   2. Recode “Definitely will happen” and “Probably will happen” to “Likely to Happen” (create new variables)
   3. Recode “Probably will NOT happen” and “Definitely will NOT happen” to “Unlikely to happen” (create new variables).
   4. Please create 4 frequency tables for each original variable against its newer version.