Epidemiology 640 – Fall 2017
SAS for Epidemiological Research

NAME: Stephanie Mecham

SECTION: 2

Homework 4: Merging and Descriptive Statistics

This homework assignment continues our replication of **Bansil et al., 2011** using the 2013-2014 data from NHANES. You may discuss the assignment with others but you MUST run your own code and write up the answers in your own words. Please turn in a printed copy of your write-up that includes all relevant SAS code (attached as a printout to the back of your write-up) in lecture at **9:30AM on Friday, Friday, October 27th**, and **submit an electronic copy** of your write-up and code via Canvas Assignment tab. ONLY the relevant output that is **requested by the questions** should be included in the paper copy of your homework. This homework is worth a total of 8 points.

Be sure that:
1) **Your SAS code runs from start to finish.**
2) Your results make sense (check your sample size and look for unreasonable, unlikely, or impossible answers).
3) Your homework has your name on EACH page. Submit your write-up on Canvas as an attachment to Assignment 4. Your file should be turned in as either a .doc, .docx or .pdf in the following format:
   LASTNAME_FIRSTNAME_HW4.docx
4) Your code is **well commented** (the top of your file should include the homework number and your name, each question should be identified in the code, and each new task should be described by comments) and formatted (use indentation and carriage returns to improve readability). Turn in your SAS code as an attachment to Assignment 4 on Canvas. Your file should be turned in as a .sas in the following format:
   LASTNAME_FIRSTNAME_HW4.sas
5) 5% will be deducted if either of tasks 3 or 4 above is not completed.

NOTE: The codebook created in your Homework 2 will help you with this assignment.

Create a new folder on your M: drive or flash drive specifically for Homework 4. Create a permanent library called **nhanes** that points to the folder you just created. Download the included datasets from Canvas and place them in the folder on your computer that corresponds to your **nhanes** library.

Make sure that your **nhanes** library contains all of the following downloaded datasets and corresponding variables below. (Notice that the **samplenew** dataset contains only those who are **aged 18 years or more** and has been **restricted on pregnancy and hypertension status**, as we did in **Homework 3**, but without created variables).

samplenew.sas7bdat **diq_h.sas7bdat**
slq_h.sas7bdat **smq_h.sas7bdat**

NAME: Stephanie Mecham

SECTION: 2

bmx_h.sas7bdat **hiq_h.sas7bdat**

1. Describe why the following would or would not be a good choice as the unique identifier in a study with 5,000 participants.

   a. **Social Security Number (0.2 pts):**
   It would be guaranteed that each SSN would represent a unique individual. However there may be security concerns because SSN's are usually sensitive information.

   b. **Home Address (0.2 pts):**

   It may double-count distinct individuals that live at the same address. It also excludes people who do not have access to permanent housing.

   c. **A number between 10000 and 50000 (0.2 pts):**
   This would be fine as long as each individual was assigned a distinct number and there were no repeats.

2. To recreate the analysis used in the Bansil et al. paper, we will want to combine all of our datasets. Before we merge these various datasets together, which procedure should we run to ensure proper merging? (0.2 pts)

One to one merging by ID via proc sort statements and then a data step to conduct the merge.

3. Create a new temporary dataset called **merge1** consisting of the following datasets: **slq_h, bmx_h, diq_h, smq_h, and hiq_h.** You can do this in one step so long as the identifier is the same across all datasets. Keep only those variables that are in the paper, i.e. those in the Homework 2 codebook corresponding to these datasets. Paste the log below. How many individuals are in the merge1 dataset? How many variables? (1 pt)

```
14   data merge1;
15       merge  nhanes.slq_h nhanes.bmx_h nhanes.diq_h nhanes.smq_h nhanes.hiq_h;
16       by SEQN;
17       keep SEQN SLQ060 SLD010H BMXBMI DIQ010 SMQ020 SMQ040 HIQ011;
18       run;

NOTE: There were 6464 observations read from the data set NHANES.SLQ_H.
NOTE: There were 9813 observations read from the data set NHANES.BMX_H.
NOTE: There were 9770 observations read from the data set NHANES.DIQ_H.
NOTE: There were 7168 observations read from the data set NHANES.SMQ_H.
NOTE: There were 10175 observations read from the data set NHANES.HIQ_H.
NOTE: The data set WORK.MERGE1 has 10175 observations and 8 variables.
NOTE: DATA statement used (Total process time):
      real time           0.03 seconds
      cpu time            0.01 seconds
```

10175 individuals and 8 variables.

4. Now combine the **merge1** file with the **samplenew** dataset, matching by **SEQN**, into a new temporary file called **merge2**. Making sure the two datasets are successfully merged and show the log below. How many observations and variables are there in the merge2 dataset? (1 pt)

```
24   data merge2;
25       merge  merge1 nhanes.samplenew;
26       by SEQN;
27       run;

NOTE: There were 10175 observations read from the data set WORK.MERGE1.
NOTE: There were 5708 observations read from the data set NHANES.SAMPLENEW.
NOTE: The data set WORK.MERGE2 has 10175 observations and 25 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.00 seconds
```

10175 observations and 25 variables.

5. Look at the **merge2** dataset. Do you see complete records for all observations with respect to demographic variables that we know are collected on everyone? Why is that? (1 pt)

No, some are missing their demographic variables (gender, race, age, education level). This is because some of the individuals were only included in the merge1 dataset and not the samplenew dataset, so when the sets were merged together, their demographic information showed up as missing.

6. Redo the merge from question 4 again, this time into a new permanent dataset called **sleepbp.** Use a DATA Step with an **IN** command to merge and restrict to those participants

who were included in the **samplenew** and the **merge1** datasets. Paste the relevant portion of your log showing that this was successfully done. (Please make sure your log is readable!) (1 pt)

```
38   data nhanes.sleepbp;
39       merge  merge1 (in=a) nhanes.samplenew (in=b);
40       by SEQN;
41       if a and b;
42       run;

NOTE: There were 10175 observations read from the data set WORK.MERGE1.
NOTE: There were 5708 observations read from the data set NHANES.SAMPLENEW.
NOTE: The data set NHANES.SLEEPBP has 5708 observations and 25 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
```

5708 observations and 25 variables.

7. Based on the counts in question 5 and 6, were all of the people in the **samplenew** dataset also in the **merge1** dataset. How do you know? (0.6 pts)

Yes, all of the people in samplenew were in merge1 (but not all of the people in merge1 were in samplenew). We know this because our log shows that there were 10175 individuals in merge1 and 5708 in samplenew. When we did the merge step that restricted the number of people to those present in both datasets (represented in the sleepbp file), we got 5708 observations which was the same number as in the samplenew set. This tells us that all of the samplenew individuals were also present in merge1, otherwise we would have had an even smaller number for the sleepbp dataset.

8. In the **samplenew** dataset there are over 4500 individuals with missing records for **RIDEXPRG.** Did these individuals make it into your final **sleepbp** merged file? What does that tell you about how an IN statement differs from deleting persons with missing information? (1 pt)

Yes, the individuals with missing data for RIDEXPRG were still included in the sleepbp file. This tells us that an IN statement doesn't look for individual missing variables and delete those individuals from the set, it simply looks for the presence of an individual within specified datasets and excludes those who were not present in the ones specified by your data step.

9. Check the variables **SLQ060, HIQ011, RIAGENDR, and DMDEDUC2** for any "don't know" or "refused" responses and recode them to missing within the **sleepbp** dataset. After

doing this, check them again to make sure your recoding worked properly. Be sure to include your code that checks/rechecks these variables. (1 pt)

```
*CHECK;

proc means data=nhanes.sleepbp n nmiss min max;
var SLQ060 HIQ011 RIAGENDR DMDEDUC2;
run;

*RE-CODE;

data nhanes.sleepbp;
        set nhanes.sleepbp;

        if SLQ060 = 7 then newSLQ060=.;
        else newSLQ060=SLQ060;

        if SLQ060 = 9 then newSLQ060=.;
        else newSLQ060=SLQ060;

        if HIQ011 = 7 then newHIQ011=.;
        else newHIQ011=HIQ011;

        if HIQ011= 9 then newHIQ011=.;
        else newHIQ011=HIQ011;

        if RIAGENDR = 7 then newRIAGENDR=.;
        else newRIAGENDR=RIAGENDR;

        if RIAGENDR = 9 then newRIAGENDR=.;
        else newRIAGENDR=RIAGENDR;

        if DMDEDUC2 = 7 then newDMDEDUC2=.;
        else newDMDEDUC2=DMDEDUC2;

        if DMDEDUC2 = 9 then newDMDEDUC2=.;
        else newDMDEDUC2=DMDEDUC2;

        run;

*RE-CHECK;
```

```
proc means data=nhanes.sleepbp n nmiss min max;
var newSLQ060 newHIQ011 newRIAGENDR newDMDEDUC2;
run;
```

The SAS System

The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Maximum
SLQ060	Ever told by doctor have sleep disorder?	5708	0	1.0000000	9.0000000
HIQ011	Covered by health insurance	5708	0	1.0000000	9.0000000
RIAGENDR	Gender	5708	0	1.0000000	2.0000000
DMDEDUC2	Education level - Adults 20+	5385	323	1.0000000	9.0000000

The SAS System

The MEANS Procedure

Variable	N	N Miss	Minimum	Maximum
newSLQ060	5696	12	1.0000000	2.0000000
newHIQ011	5701	7	1.0000000	2.0000000
newRIAGENDR	5708	0	1.0000000	2.0000000
newDMDEDUC2	5380	328	1.0000000	5.0000000

10. In the same DATA step as above, recode SLD010H into an indicator variable (dummy code) for short sleep duration called **shortsleep** defined as in the Bansil et al. paper. Be sure to account for coding for missing values of SLD010H so that **shortsleep** is missing for those records as well. Then combine **shortsleep** with **SLQ060** into a combination of sleep problems variable, called **sleepcombo**, with four mutually exclusive categories. Paste a PROC FREQ table output of this new variable below and be sure to save **sleepbp** with these newly created variables to your **nhanes** library. (0.6 pts)

The SAS System

The FREQ Procedure

sleepcombo	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	3288	57.76	3288	57.76
2	262	4.60	3550	62.36
3	1862	32.71	5412	95.06
4	281	4.94	5693	100.00
Frequency Missing = 15				

CODE:

```
*Homework 4: Merging and Descriptive Statistics | Stephanie Mecham | EPID 640 Section 2 ;

*Creating a permanent library;

    libname nhanes "C:\Users\smecham\Desktop\nhanes";
    run;

*Question 3: Merging datasets in merge1;

    proc sort data= nhanes.bmx_h;
    by seqn;
    run;

    data merge1;
        merge  nhanes.slq_h nhanes.bmx_h nhanes.diq_h nhanes.smq_h nhanes.hiq_h;
        by SEQN;
        keep SEQN SLQ060 SLD010H BMXBMI DIQ010 SMQ020 SMQ040 HIQ011;
        run;

*Question 4: Merging merge1 with samplenew;

    data merge2;
        merge  merge1 nhanes.samplenew;
        by SEQN;
        run;

*Question 6: Merging into sleepbp with individuals only in both datasets;

data nhanes.sleepbp;
        merge  merge1 (in=a) nhanes.samplenew (in=b);
        by SEQN;
        if a and b;
        run;

*Question 9: Recoding refused/missing data points to missing in sleepbp set;

*Check;

proc means data=nhanes.sleepbp n nmiss min max;
```

```
var SLQ060 HIQ011 RIAGENDR DMDEDUC2;
run;

*Re-coding;

data nhanes.sleepbp;
	set nhanes.sleepbp;

	if SLQ060 = 7 then newSLQ060=.;
	else if SLQ060 = 9 then newSLQ060=.;
	else newSLQ060=SLQ060;

	if HIQ011 = 7 then newHIQ011=.;
	else if HIQ011= 9 then newHIQ011=.;
	else newHIQ011=HIQ011;

	if RIAGENDR = 7 then newRIAGENDR=.;
	else if RIAGENDR = 9 then newRIAGENDR=.;
	else newRIAGENDR=RIAGENDR;

	if DMDEDUC2 = 7 then newDMDEDUC2=.;
	else if DMDEDUC2 = 9 then newDMDEDUC2=.;
	else newDMDEDUC2=DMDEDUC2;

	if SLD010H = . then shortsleep =.;
	else if SLD010H < 7 then shortsleep=1;
	else shortsleep=0;

run;


*Re-checking;

proc means data=nhanes.sleepbp n nmiss min max;
var newSLQ060 newHIQ011 newRIAGENDR newDMDEDUC2;
run;



*Question 10: Creating sleepcombo from shortsleep and SLQ060 and running a PROC FREQ;

data nhanes.sleepbp;
```

```
set nhanes.sleepbp;
if shortsleep=1 and SLQ060=1 then sleepcombo=4;
else if shortsleep=1 and SLQ060=2 then sleepcombo=3;
else if shortsleep=0 and SLQ060=1 then sleepcombo=2;
else if shortsleep=0 and SLQ060=2 then sleepcombo=1;
run;

proc freq data= nhanes.sleepbp;
        tables sleepcombo;
        run;
```