**Assignment2**  ***Biostatistics 203A***

In homework 2, we will use the dataset MDFACW02\_d1*.csv*. The source of data is https://oriseapps.orau.gov/cedr/working\_dfs.aspx

Data File Set: MDFACW02\_d1

File Name: Working personnel file for Mound Plant.

Date fields with only 2 digits for the year must be handled with care and caution.

When importing the data, users must examine the data element as well as the site associated with the data to determine the correct century.

For example, some birth dates go back to the 1800s, whereas employment dates, exposure dates, and vital status dates will be in the 1900s or even the 2000s.

You also have an access to the dictionary of dataset for labels and the corresponding code book for formats. The file name of dictionary is MDFACW02\_d1\_dictionary.csv. The code book is a zipped folder.

**Exercise 1 Import and read data [5 points]**

Write the following codes into SAS.

Step 1: Use library name and prod import commands and import this MDFACW02\_d1.csv file.

libname CEDR "\folderes\Your folder\Biostat203A\LabHomework\Lab2";

**run**;

**PROC** **IMPORT** OUT= CEDR.xxx

DATAFILE= " \folderes\Your folder\Biostat203A\LabHomework\Lab2\MDFACW02\_d1.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=**2**;

**RUN**;

Step2: Read the log window output, and utilize the output of log to create data command.

**data** MDFACW (label="Working personnel file for Mound Plant");

infile " \folderes\Your folder\Biostat203A\LabHomework\Lab2\MDFACW02\_d1.csv" DSD firstobs=**2**

informat orauid $8. ;

…

informat seq\_no **4** ;

format orauid $8. ;

…

format seq\_no **4** ;

input orauid $

…

seq\_no ;

label autopsy = "Autopsy"

…

seq\_no = "Sequence Number of Row"

**run**;

**proc** **means** data=CEDR.MDFACW ;

**run**;

**proc** **print** data=CEDR.MDFACW (obs=**5**) label;

**run**;

We specify the location of the data using the "infile" statement. The correct path to the folder with your data is "/folders/myfolders/datafilename". The "dsd" statement tells SAS that the delimiter for this data set is a comma. The "firstobs" statement tells SAS that the data begins on the second line of the csv file.

The default length of a character variable in SAS is 8 characters. The character strings in our data set are longer than 8 characters. We used the "length statement" to set the length of the variables name and location to be 50.

**Exercise 2 Proc means [3pts]**

Use proc means to fill in values for the following table.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | Standard deviation | Median | Minimum | Maximum |
| Date of death |  |  |  |  |  |  |
| Date last alive |  |  |  |  |  |  |
| Sequence Number of Row |  |  |  |  |  |  |

**Exercise 3 Proc format and freq [3pts]**

Use proc freq to fill in values for the following table.

|  |  |  |
| --- | --- | --- |
| Undergraduate Enrollment | Frequency | Percent |
| Grade school |  |  |
| Some high school |  |  |
| High school graduate |  |  |
| Associates Degree |  |  |
| Unknown |  |  |

**Exercise 4 Proc format and freq [3 pts]**

Write a SAS program to fill in the values in the following table, where N refers to frequencies and % refers to percentages.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Alive | | Dead | | Unknown | | Total | |
|  | N | % | N | % | N | % | N | % |
| Grade school |  |  |  |  |  |  |  |  |
| Some high school |  |  |  |  |  |  |  |  |
| High school graduate |  |  |  |  |  |  |  |  |
| Associates Degree |  |  |  |  |  |  |  |  |
| Unknown |  |  |  |  |  |  |  |  |
| Total |  |  |  |  |  |  |  |  |

**Exercise 5 Proc means**

An alternative approach to exploring the potential association between education and age is to consider the mean and median rank within each category of enrollment. Using a proc mean program with a class statement, fill in the values in the following table.

A variable of age can be generated using the date of birth and the date of first hired at Mound

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Statistics for Variable: Age at first hired at Mound | | | | | |
|  | N | Mean | Std Dev | Median | Min | Max |
| Education |  |  |  |  |  |  |
| Grade school |  |  |  |  |  |  |
| Some high school |  |  |  |  |  |  |
| High school graduate |  |  |  |  |  |  |
| Associates Degree |  |  |  |  |  |  |
| Unknown |  |  |  |  |  |  |

**Exercise 6 Proc contents [3 pts]**

Read the University Rankings Data Set in using SAS one more time. Make sure that all of the following attributes are applied:

-ID number, Sex, ICDa8, autopsy, and up to the results of a 1986 SSA submission (SSA861) should be the only character variables

- dates, education level, and sequence number of row should numerical variables.

-None of the values should be truncated when reading the data in

-Every variable should be labeled

- Formats should be applied to variables in the code list (e.g. cvs, dmvflag, drace, …, ssa861) except for the state of death.

Run the following code and examine the table titled 'List of Variables and Attributes’.

**proc** **contents** data=xxx order=varnum ;**run**;