**Assignment 3**  ***Biostatistics 203A***

***Due date: Week 4***

In the previous homework, we used the dataset MDFACW02\_d1*.csv* and made a SAS file. The source of the original data is from <https://oriseapps.orau.gov/cedr/working_dfs.aspx>.

Data File Set: MDFACW02\_d1

File Name: Working personnel file for Mound Plant.

It is claimed that the data set is ‘working dataset’, which means that the data is not completely clean.

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.

In this homework, we will work on the quality check (QC) of data using summary statistics and graph.

**Exercise 1. missing numeric value & flag lists [10 points]**

Using the data set with containing age, run proc mean command of age. This data set contains one row per subject and there are no duplication. Generate a new variable name of age of death, age\_dth, using age\_dth being equal to (death date – birth date)/365.

1. Graph histogram of age using proc sgplot command; also try to graph a box-Whisker plot for the variable of age. Which plot do you prefer?

Proc sgplot data= your\_data\_file;

histogram your\_variable\_name\_of\_age\_at\_hire;

histogram your\_variable\_name\_of\_age\_at\_death;

run;

Proc sgplot data= your\_data\_file;

vbox your\_variable\_name\_of\_age;

run;

1. Find out the number of subjects where *the age of death* is less than 15.

Proc means data =your\_data\_file n mean std noobs maxdec=2;

where *the age of death<15*;

var age at hiring *the age of death*;

run;

Fill out the table where ***age at ~~hiring~~ death is less than 15***

|  |  |  |  |
| --- | --- | --- | --- |
|  | N | Mean | Standard deviation |
| Age at Hiring |  |  |  |
| Age\_at\_death |  |  |  |

Comment(s) regarding output: You will found an odd result, where the number of subjects who died in younger than 15 is large. This is due to SAS defined the missing value as negative infinite. What will be an appropriate way to generate the age at death?

Either you can define in the data procedure or proc command.

Let’s add another condition of no missing values in the next question.

1. **With consideration of missing values of the date of death, we can add condition in the “Where” step.**

Find out the number of subjects where age of death is less than 15 with **no missing information on birth dates and death dates.**

Proc means data =your\_data\_file n mean std noobs maxdec=2;

where …… ;

var age;

run;

Fill out the table where ***age at hiring is less than 15 and no missing in the age of death***

|  |  |  |  |
| --- | --- | --- | --- |
|  | N | Mean | Standard deviation |
| Age at Hiring |  |  |  |
| Age\_at\_death |  |  |  |

Comment(s) regarding output: We expect that number of observation is 3 (N=3) for both variables.

1. Print the subject ID (orauid), birth date (bdate), sex, education level, hiring and terminating dates, and age where ***age at hiring is less than 15* with no missing information regarding birth dates and hiring dates.**

Here are the pseudo code:

Proc print data =your\_data\_file;

where …… ;

var orauid bdate sex educ hiredate termdate ddate age;

run;

Copy the output of proc print command

Copy the output of print including the variables of orauid bdate sex educ hiredate termdate ddate age.

|  |
| --- |
| orauid bdate sex educ hiredate termdate ddate age@Hiring difference of hiring date |
|  |

1. print the subject ID (orauid), birth date (bdate), sex, education level, hiring and terminating dates, and age where **age at hiring is less than 15 or *greater than 80* years old** with no missing information regarding the birth dates and hiring dates.

Proc print data =your\_data\_file;

where …… ;

var orauid, bdate, ….., age;

run;

Copy the output of print including the variables of orauid bdate sex educ hiredate termdate ddate age.

|  |
| --- |
| orauid bdate sex educ hiredate termdate ddate age@Hiring difference of hiring date |
|  |

Comment(s) regarding output: The total number of the expected flag list is 24.

Big question: What patterns do you see in the date variables? With that thought, let’s move on the next question.

**Exercise 2 Data QC [10pts]**

Let’s try another logic check.

|  |
| --- |
| Logic 1: the date hiring a person occurred earlier than the termination date. The date of hiring can be theoretically same as the date of termination. |
| Flag List: flag the cases if the hiring date is later than the date of termination. |

1. Use the condition in “where” of the proc means found out the number of the flagged cases using age variable and print out the list.

Proc means data =your\_data\_file n mean std noobs maxdec=2;

where …… ;

var age;

run;

Proc print data =your\_data\_file;

where …… ;

var orauid, bdate, ….., age;

run;

Fill out table where the hiring date occurred later than the date of termination.

|  |  |
| --- | --- |
|  | N |
| Age at Hiring |  |
| Age\_at\_death |  |

Copy the output of print including the variables of orauid bdate hiredate termdate ddate age.

|  |
| --- |
| orauid bdate hiredate termdate ddate age@Hiring difference of hiring date |
|  |

1. We found that the main possible cause of flagged cases are wrong censoring date of termination on 9/15/1999. Generate the variable of date of 09/15/1999, which seems to the source of the problem. Using this date, generate a difference of dates where the hiring date are stated as 09/15/1999.

|  |
| --- |
| Pattern 1: It is plausible that the cause of the flagged list can be associated with censoring dates of 09/15/1999 |
| Flag List: flag the cases if the date of hiring was on 09/15/1999. |

data your\_new\_file;

set your\_last\_file;

/\*defining the date below: input is the function for converting character into numeric\*/

date09151999=”09/15/1999”;

date09151999n=input(date09151999 ,type\_the\_date\_of\_format);

diff\_date= hire\_date- date09151999n;

format diff\_date type\_the\_date\_of\_format.;

drop date09151999;

rename date0915199n= date0915199;

run;

Find the number of observation where the date of hire is on 09/15/1999 and print the list including subject ID, date of hire, age, and the difference of hiring date from the date of 09/15/1999

Fill out table where the hiring dates were on 09/15/1999

|  |  |
| --- | --- |
|  | N |
| Age at Hiring |  |
| Age\_at\_death |  |

Copy the output of print including the variables of orauid bdate hiredate termdate ddate age,

difference of hiring date from the date of 09/15/1999.

|  |
| --- |
| orauid bdate hiredate termdate ddate age@Hiring difference of hiring date |
|  |

1. Find the list of subjects who had the age of death being negative values.

By identifying the cases from (b) of the date of 9/15/1999, is the list of the subjects with the negative age included in the list from (b)?

If no, what type of check can be performed?

|  |
| --- |
| Logic 2: Age cannot be negative. |
| Flag List 2: flag the cases if the dates of death are negative. |

Fill out table of where the ages on the date of death are negative

|  |  |
| --- | --- |
|  | N |
| Age at Hiring |  |
| Age\_at\_death |  |

Copy the output of print including the variables of orauid bdate hiredate termdate ddate age,

difference of hiring date where the ages on the date of death are negative.

|  |
| --- |
| orauid bdate hiredate termdate ddate age@Hiring difference of hiring date |
|  |

1. According to (c) above, what criteria can be added? You can use the similar procedure of (b) using the odd date of birth.

Using the criteria of the date of birth, how many list of subjects were flagged?

Using this flagged date of birth date and hiring date, how many cases are in total? You can use summary statistics (e.g. proc means);

|  |
| --- |
| Pattern 2: It is plausible that the cause of the flagged list can be associated with censoring dates of hiring on 09/15/1999 and the date of birth on 07/01/1999 |
| Flag List: flag the cases if the date of hiring was on 09/15/1999 & the date of birth was on 07/01/1999. |

List the subject ID, date of birth, date of hiring, and age that had the common flagged cases from 2.(b) and 2.(d) above.

Fill out table where the date of birth was on 07/01/1999 and hiring date was on 09/15/1999.

|  |  |
| --- | --- |
|  | N |
| Age at Hiring |  |
| Age\_at\_death |  |

Copy the output of print of this flag list including the variables of orauid bdate hiredate termdate ddate age, difference of hiring date from the date of 09/15/1999.

|  |
| --- |
| orauid bdate hiredate termdate ddate age@Hiring difference of hiring date |
|  |

In this homework, you found plausible causes of error in the age when a subject was hired at Mound Plant. This can be a systematic error from an organization who collected the data set.