BIOS 511 Lab 2
SAS Procedures: PRINT, FREQ, UNIVARIATE, MEANS

Please read the following instructions carefully before beginning this lab. You will need to start by downloading the DM datasets from the LAB-02 assignment on the Sakai site. This dataset will be the basis of all activites for this lab.

Instructions:

- All tasks should be completed in a single SAS program named lab-02-PID.sas where PID is your student PID number. Please make sure to include an appropriate header in your SAS program.
- All the output that your SAS program produces in this lab should be should be delivered to a *single PDF file* named lab-02-PID-output.pdf.
- In your code, before starting a new task, include a block comment with the task number

```
/*************************************************************
                        SAS Code for Task # X
 *************************************************************/
```

- The output from the 8 tasks (7 producing output) should be put on separate pages.
- All the output from a single task should be put on the same page.
    - Recommendation: While initially writing the SAS program, do not concern yourself with creating the PDF output. Simply view results in the results window to verity that your program has created the desired output. Once your program is essentially complete, add the appropriate ODS statements to create the PDF file based on the requirements above.
- You will upload the SAS program, SAS log, and PDF output file to document completion of the lab.

Background Information on Data

The ECHO clinical trial was a (hypothetical) randomized, placebo-controlled trial designed to demonstrate efficacy of a novel therapy (4mg "Echo Max" tablets) for the treatment of hypertension compared to a placebo. A formal definition of a randomized, controlled trial (commonly termed an RCT) is given here https://www.cancer.gov/publications/dictionaries/cancer-terms/def/randomized-clinical-trial.

In this lab you will work with the demographics dataset (DM). Over the semester we will work with several other datasets from this hypothetical trial. The structure of the ECHO Trial datasets (e.g., variable names, types, and labels) is defined according a set of standard guidelines for organizing clinical trials data: *The Study Data Tabulation Model (SDTM)*. The SDTM defines a standard structure for study data that are to be submitted as part of a product application to a regulatory authority such as the United States Food and Drug Administration (FDA). This course is not designed to teach you all about the SDTM but it is helpful to be exposed to the idea of data standards and to this model in particular due to its popularity.

To read more about the current version of the SDTM, see the PDFs on the Sakai site (CDISC folder) named "SDTM v1.5.pdf" (defines the SDTM but is somewhat abstract) and "SDTMIG v3.2.pdf" (the implementation guide designed to give users practical advice implementing the SDTM). For students interested in working on pharmaceutical drug/device trials, I encourage you to read through these documents. Experience with the SDTM is very marketable for that industry.

DM – Demographic and baseline data for the ECHO clinical trial (# obs = 602)

The DM dataset includes the set of essential standard variables that describe each subject in a clinical study (e.g., race, age, treatment group, date treatment first taken, etc). It is the parent dataset for all other observations (i.e., data) on human subjects in the trial. Every subject in the trial must have an observation in the DM dataset. The structure of the DM dataset is one observation per subject. For more details see page 62 of 398 (and following) of the SDTM implementation guide.

The below table is an example of a dataset specification. Such tables are typically created to document the contents of a dataset for downstream users and the "Description of Values" column in the specification is designed to give users insight into the values the respective variables can take in the data.

#	Variable	Type	Len	Label	Description of Values
1	STUDYID	Char	10	Study Identifier	Values are always equal to "ECHO"
2	USUBJID	Char	30	Unique Subject Identifier	Values are of the form "ECHO-XXX-YYY" where XXX and YYY are integers
3	RFXSTDTC	Char	20	Date/Time of First Study Treatment	The date of first dose of study treatment in YYYY-MM-DD format
4	RFXENDTC	Char	20	Date/Time of Last Study Treatment	The date of last dose of study treatment in YYYY-MM-DD format
5	RFICDTC	Char	20	Date/Time of Informed Consent	The date that informed consent was obtained for study participation in YYYY-MM-DD format
6	AGE	Num	8	Age	Age at enrollment in years, if provided
7	AGEU	Char	6	Age Units	Values are always equal to "YEARS"
8	SEX	Char	10	Sex	Values are "M", "F", or missing
9	RACE	Char	50	Race	Race is collected as one of six possible standardized values. Print the data to see the possible values.
10	ARMCD	Char	20	Planned Arm Code	Code for planned treatment group – values are "ECHOMAX" or "PLACEBO"
11	ARM	Char	50	Description of Planned Arm	Planned treatment group – values are "4mg Echo Max Tablets" or "Placebo"
12	COUNTRY	Char	20	Country	Values are "USA", "CAN", or "MEX"
13	VISITNUM	Num	8	Visit Number	Always equal to -1
14	VISIT	Char	50	Visit Name	Always equal to "Screening"
18	DMDTC	Char	20	Date/Time of Collection	Equal to the date of the screening visit in YYYY-MM-DD format

BIOS 511 Lab 2
SAS Procedures: PRINT, FREQ, UNIVARIATE, MEANS

Task 1: To better understand the structure of the DM dataset, please write a PROC PRINT step to print the data for the first 10 subjects (i.e., first 10 observations).

- Initially, print all variables in the dataset (achieved by omitting a VAR statement). Next, add a VAR statement to ensure that only the following variables are printed: Unique Subject ID, Date/Time of First Treatment, Age, Sex, Race, Planned Arm Code, Description of Planned Arm, and Country. Your "final" program should use the VAR statement.

- Use *two* options on the PROC PRINT statement to suppress the default observation numbering and request that variable labels be printed instead of variable names.

- **EVERYONE COMPLETE THIS EXERCISE:** If you are not aware of the two options that are needed, you can search the online SAS documentation.
    - To do this, first google "SAS 9.4 Product Documentation" and follow the SAS 9.4 Product Documentation web link.
    - Click on the link "SAS Procedures by Name and Product"
    - Click on the link "SAS Procedures by Name"
    - Use the Alphabetical navigation links to find The "PRINT" Procedure
    - Click on the "Syntax" tab and then click "PROC PRINT Statement"
    - From here you should be able to read about *all* options for the PROC PRINT statement.

  You would be wise to bookmark the SAS 9.4 Product Documentation page as a resource for now and in the future.

- Make sure the output is titled
  Task1: Demographics Data for Select ECHO Trial Subjects

- The following SAS code is a template for the required PROC PRINT step:

```
proc print data = libref.filename(obs=10) <options>;
      var <list of variables to print>;
run;
```

  Note: The option "obs=10" is a dataset option and such options are included in parentheses adjacent to the dataset to which they apply. We will learn more about using dataset options.

BIOS 511 Lab 2
SAS Procedures: PRINT, FREQ, UNIVARIATE, MEANS

Task 2: Create a *one-way frequency table* that computes the *number of ECHO trial subjects* in each treatment group. Make sure your output includes the percentage of subjects in each treatment group but no other percentages.

- The treatment group to which a subject belongs is identified by the *ARM* and *ARMCD* variables.

- Use a PROC FREQ step to count the number of observations having each value of *ARMCD* (or ARM) to determine the number of subjects in each treatment group.

- Use a **TABLE** statement option to suppress undersired percentages (if any are present in the output). To identify relevant table statement options, you can use the same strategy described in task # 1. Under the "Syntax" tab, you should be able to find a section on the TABLE statement where information on all possible options can be found. Read through the documentation to find the correction option (if one is needed).

- The following SAS code is a template for the required PROC FREQ step:

```
proc freq data = libref2.filename2 ;
      table <table request> / <options>;
run;
```

- Make sure the output is titled
  Task2: Number and Percent of ECHO Trial Subjects by Treatment Group

BIOS 511 Lab 2
SAS Procedures: PRINT, FREQ, UNIVARIATE, MEANS

Task 3: Produce a *two-way frequency table* to determine the number and percentage of ECHO Trial subjects in each country *within* each treatment group (*ARMCD or ARM*).

- Make the treatment group display as the column in the output.
- By using the appropriate **TABLE** statement option, suppress all percentages other than the percentages of subjects in each country *within* each treatment group. This is to say that the percentage of subjects in each treatment group should sum up to 100% when you add up the value for the individual countries. These percentages are designed to allow comparison of the country distribution across treatment groups.
- Make sure the output is titled
  Task3: Number and Percent of ECHO Trial Subjects by Treatment Group and Country
- The same SAS code template from the previous task is relevant here.

Task 4: Produce a *two-way frequency table* to determine the number and percentage of ECHO Trial subjects for age categories <65 versus >=65 *within* each treatment group (*ARMCD or ARM*).

- First, copy this DATA step into your program for this task.

```
data DM;
 set echo.DM;

  length ageCat $10;
  if not missing(age) and age <65 then ageCat = '1: <65';
  else if age >= 65               then ageCat = '2: >= 65';

run;
```

  You are not required to fully understand the DATA step at this time but the gist of it is that the DATA step copies the permanent DM (my libref is named ECHO) into the work library to create a temporary dataset named DM (or work.DM). The DATA step also creates a new character variable named "ageCat" that has a length of 10 characters.

- Make the treatment group display as the column in the output.
- By using the appropriate **TABLE** statement option, suppress all percentages other than the percentages of subjects in each age category *within* each treatment group.
- Make sure that you table includes counts for any subjects who have missing values of age category (i.e., missing values of age) but DO NOT include those subjects in the calculation of percentages. This is achieved by adding another option to the **TABLE** statement.
- Make sure the output is titled
  Task3: Number and Percent of ECHO Trial Subjects by Treatment Group and Age Category
- Using a LABEL statement in the PROC FREQ step, add a temporary label to the variable "ageCat" that is equal to "Age Category".

Task 5: Using a PROC MEANS step, compute a five number summary for age. A five number summary includes the sample size (N), the mean, the standard deviation, the minimum, and the maximum. Also request PROC MEANS to compute the number of subjects with missing age. The order of the key words listed on the PROC MEANS statement determines their presentation order in output. The common ordering is N, # missing, mean, standard deviation, minimum, and maximum.

- To identify the relevant statistics keywords needed for the PROC MEANS statement to get the five number summary (and number of missing values), use the SAS 9.4 Product Documentation as described in the first task.

- Be sure to include a VAR statement that only lists the variable age otherwise all numeric variables in the dataset will be summarized! Temporarily omit the VAR statement so that you can observe this behavior.

- Make sure the output is titled
  Task5: Summary of Age for ECHO Trial Subjects

- The following SAS code is a template for the required PROC MEANS step:

```
proc means data = libref.filename <list of statistics to compute>;
      var <variables to analyze>;
run;
```

Task 6: This task mirrors the previous task with the exception that for this task we want to produce a five number summary *for each treatment group*.

- To create the desired summary, add a CLASS statement to the previous task's SAS code and list either the ARMCD or ARM variable in the CLASS statement (in fact you can list both and it won't hurt). Doing so instructs SAS to perform the analysis for each distinct value of the variable(s) listed in the class statement.

- No SAS code template should be needed for this task as only a single additional statement is required compared to the previous task.

- Use the FW option on the PROC MEANS statement to specify a field width of 5 (and compare to the previous tasks output).

- Make sure the output is titled
  Task6: Summary of Age for ECHO Trial Subjects by Treatment Group

BIOS 511 Lab 2
SAS Procedures: PRINT, FREQ, UNIVARIATE, MEANS

Task 7: The MEANS Procedure is optimal for computing basic summaries of data (i.e., five number summaries). The UNIVARIATE Procedure can do the same things but the tabular ODS output is not as concise as that from PROC MEANS. However, there are many things that PROC UNIVARIATE can do that PROC MEANS cannot do and so it is important to know both procedures.

For this task you will produce a histogram of the age variable for each treatment group as well as an estimate of the normal distribution that best matches the age distribution for each treatment group. Pretty fancy!

This task is an example of creating what are often called "default" ODS graphics. The term default means that for certain analyses, by default the PROC will produce graphs. Later in the semester we will learn how to make graphs all by ourselves using SAS procedures designed to create graphs (i.e., that do nothing by make graphs).

- The following SAS code can be used as a template for this task.

```
proc univariate data = libref.filename ;
 class <classification variable>;
 var <analysis variable>;
 histogram <analysis variable> / <options>;
 inset <list of statistics to display on histograms> / format=5.2;
run;
```

  In this template, since all we want is a histogram, inclusion of the VAR statement is not really necessary. In general, if we do not include the VAR statement, the UNIVARIATE Procedure will analyze every numeric variable in the dataset! This could result in a lot of unwanted output. However, we are going to use an ODS SELECT statement to select only the graph and not any tabular output (which is why the VAR statement is irrelevant).

- Look up the HISTOGRAM statement in the SAS 9.4 Product Documentation and identify the keyword necessary for SAS to draw the "best-fitting" normal distribution curve over the top of the histogram. Note this option is only a single word (although further customization is possible).

- Look up the INSET statement in the SAS 9.4 Product Documentation and identify the statistics keywords that should be listed on that statement so that the histogram displays the mean and standard deviation. Note that I have already included on option on the inset statement, the FORMAT= option that controls the number of decimal places displayed for the summary statistics.

- When you are happy with your output, use an ODS SELECT statement so that ONLY the histogram is selected for inclusion in your output file. You can use ODS TRACE ON/OFF to identify the name of the ODS object corresponding to the graph or you can simply look up the name of the ODS object in the SAS documentation. Just as there is a "Syntax" tab for each procedure in the documentation, there is a "Details" tab. The "Details" tab will have a section of ODS Tables and ODS Graphics giving the names required for the ODS SELECT/EXCLUDE statements.

- Make sure the output is titled: *Task7: Distribution of Age for ECHO Trial Subjects by Treatment Group*

  If you dislike the fact that the image has a title embedded in it (likely "Distribution of Age"), you can remove that title by adding the appropriate option on the histogram statement. In particular, you need to add the following: odstitle=" "
  That option basically sets the title to missing.

Task 8: Once your program is complete, add a footnote to *all* output that documents the data stamp of the data you are analyzing. Such is common practice when writing programs that will be used frequently to summary updated datasets. For example, you may get new data transferred to you once a month and it is helpful to know what version of the data is being summarized. The footnote should read: "ECHO Data Extract Date: 2017-10-10".