14 Assignment, Summary Reports 1

**Exercise 1**. Counting Levels of a Variable with PROC FREQ

The data set orders is in the STA5066 sub directory. Use this data set to produce two reports.

1. Use a PROC FREQ step to display the number of distinct levels of Customer\_ID and Employee\_ID for retail orders.

(a) Limit the report to retail sales (Order\_Type=1).

(b) Display the title: Unique Customers and Salespersons for Retail Sales.

(c) Do not display the counts for individual levels of Customer\_ID and Employee\_ID.

2. Use a PROC FREQ step to display the number of distinct levels for Customer\_ID for catalog and Internet orders.

(a) Limit the report to catalog and Internet orders (Order\_Type values other than 1).

(b) Display the title: Unique Customers for Catalog and Internet.

(c) Do not display the counts for individual levels of Customer\_ID and Employee\_ID.

**Exercise 2**. Producing Frequency Reports with PROC FREQ

The data set orders in in the STA5066 subdirectory.

1. Use the following PROC FORMAT step to create a user-defined format.

proc format;

value ordertypes 1=’Retail’ 2=’Catalog’ 3=’Internet’;

run;

2. Use a PROC FREQ step to produce three frequency reports.

(a) Number of orders in each year. Apply the YEAR4. format to the Order\_Date variable to combine all orders within the same year.

(b) Number of orders of each order type. Apply the ordertypes. format defined in the PROC FORMAT step to the Order\_Type variable. Do not display the cumulative frequency and percentages.

(c) Number of orders for each combination of year and order type. Suppress all percentages that normally appear in each cell of an n-way table.

**Exercise 3**. Creating an Output Data Set with PROC FREQ

The data sets order\_fact and product\_list are in the STA sub directory.

1. Use a PROC FREQ step with the data set order\_fact to create a table of frequencies based on Product\_ID. Create an output data set called freqcount containing the frequency counts based on Product\_ID. This is done with an OUT= option in the TABLES statement.

2. Merge the data set created above with the data set product\_list to obtain the product name (the variable Product\_name) for each product id code.

3. Sort the merged data by the variable Count so that the most frequently ordered products appear at the top of the resulting data set.

4. Print the first 10 observations, those that represent the ten products ordered most often.

**Exercise 4**. Creating a Summary Report with PROC MEANS

The data set order\_fact is in the STA5066 sub directory.

Use the following PROC FORMAT step to create a user-defined format:

proc format;

value ordertypes 1=’Retail’ 2=’Catalog’ 3=’Internet’;

run;

Use a PROC MEANS step with the order\_fact data set to prepare a report listing the total sales by Order\_Date and Order\_Type.

1. Title the report: Revenue (in U.S. Dollars) Earned from All Orders

2. Display only the SUM statistic for the Total\_Retail\_Price variable.

3. Display separate statistics for each combination of Order\_Date and Order\_Type.

4. Apply the ordertypes. format so that the order types are displayed as text descriptions, not numbers.

5. Apply the YEAR4. format so that order dates are displayed as years, not individual dates.

**Exercise 5**. Analyzing Missing Numeric Values with PROC MEANS

The data set staff is in the STA5066 sub directory.

Use a PROC MEANS step to display (only) the number of missing values and the number of nonmissing values for the variables Birth\_Date, Emp\_Hire\_Date, and Emp\_Term\_Date variables on this data set.

1. Display separate statistics for each value of the variable Gender.

2. Suppress the output column that displays the total number of observations in each classification group.

3. Title the report: Number of Missing and Non-Missing Date Values

**Exercise 6**. Creating an Output Data Set with PROC MEANS

The data sets order\_fact and product\_list are in the STA5066 sub directory.

1. Use a PROC MEANS step to produce a data set using an output statement. The data set should contain the total (sum) of the variable Total\_Retail\_Price for each unique value of Product\_ID

2. Merge the data set created above with the data set product\_list to obtain the Product\_Name value for each Product\_ID code.

3. Sort the merged data so that the products with higher revenues appear at the top of the resulting data set.

4. Print the first 10 observations, those that represent the ten products with the most revenue.

**Exercise 7**. Proc Freq Nhanes 3

1. Use a data step to read the permanent SAS dataset Analysis (in the STA5066 sub directory) and create a temporary SAS dataset called work.AnalysisTmp, that contains only the following variables:

variable name description

seqn sequence number

dmaracer Race of Participant

dmarethn Race/Ethnicity

dmaethnr Ethnicity (coding below)

hssex Sex of Participant (1=male, 2=female)

hsageir Age of participant (yrs, 90=90+)

1. Use a PROC FREQ step with work.AnalysisTmp to obtain one-way tables for each of the variables: dmaracer, dmarethn, and hssex. The report should include only the cell frequencies.
2. Repeat the PROC FREQ step, to obtain one-way tables for each of the variables: dmaracer and dmarethn, and hssex, restricting the analysis to females under 50 years of age (hssex is code 2 for females).
3. Use the following code to create a user-defined format for age to define the age groups <45,45- 59, and 60+:

proc format;

value agef low-<45=”<45”

45-59=”45-59”

60-high=”60+”;

Run;

1. Use a PROC FREQ step with work.AnalysisTmp to obtain two-way tables that cross-tabulates each of the variables dmaracer, dmarethn, and hssex with hsageir. Use the format agef. to format age into the specified age groups.

**Exercise 8**. PROC UNIVARIATE, heart data set

The data set heart is in the sashelp library (pre-defined when you start SAS).

Write a program using PROC UNIVARIATE to analyze the variable cholesterol on the data set.

1. Include a histogram with a normal overlay that includes an inset providing the mean (with one decimal place), the standard deviation (with one decimal place), and the number of observations analyzed. Position the inset in the upper right hand corner of the graph.

2. Include a qqplot of the variable cholesterol that includes an inset providing the mean (with one decimal place), the standard deviation (with one decimal place), and the number of observations analyzed. Position the inset in the middle of the top of the graph. See the following link to see how to position the inset:

https://v8doc.sas.com/sashtml/proc/z1216811.htm#z1226578