```
/*********************************************************
    Week 5: PROC MEANS,PROC FREQ, and PROC UNIVARIATE
    Author: Cheng Peng
      Date: 2/15/2020
 Topics  1. Loading tab and csv file to SAS
         2. PROC MEANS Basics
         3. PROC MEANS - beyond averages
         4. PROC FREQ Basics
         5. PROC UNIVARIATE basics
*********************************************************/
DM "CLEAR LOG";
DM "CLEAR OUT";

LIBNAME w05 "C:\STA311\w05";
OPTIONS PS = 94 LS =74 NODATE NONUMBER NOCENTER;



/*******************************************************
      Topic 1: Loading Data with Data and PROC step

      Read text and csv data into SAS with DATA STEP
      and PROC IMPORT - Pay attention to the DELIMITER
      when using INFILE-INPUT statement in a data step
*******************************************************/

/** txt - tab delimitered data **/
DATA TXTIRIS;
LENGTH variety $ 10;
/* txt files are tab delimitered data. The delimiter of this data is '09'x
   that has to be specifed in the following INFILE statement            */
INFILE "C:\STA311\w05\w05-iris-text.txt"
        delimiter = '09'x  FIRSTOBS = 2;
INPUT SepalLengt SepalWidth  PetalLength  PedalWidth  variety $;
RUN;

PROC CONTENTS DATA = TXTIRIS;
RUN;
```

```
/** CSV - comma separated values, dlm =","  should be specified in the
    INFILE statement.                                                  **/
DATA CSVIRIS;
LENGTH variety $ 10;
INFILE "C:\STA311\w05\w05-iris-csv.csv"
          dlm="," FIRSTOBS = 2;
INPUT SepalLength SepalWidth PetalLength PedalWidth variety $;
RUN;

PROC CONTENTS DATA = CSVIRIS;
RUN;

/** PROC IMPORT Methods **/
/** PROC IMPORT TXT **/
PROC IMPORT OUT= WORK.txt_iris
            DATAFILE= "C:\STA311\w05\w05-iris-text.txt"
            DBMS=TAB REPLACE;    /* .txt is a tab delimitered data  */
     GETNAMES=YES;
     DATAROW=2;          /** data record starts from row 2, first row lists variables names  **/
     GUESSINGROWS =150;  /** This option is required since the
                              length of variety is more than 8 bytes **/
RUN;

/* PROC IMPORT - CSV */
PROC IMPORT OUT= WORK.iris_csv
            DATAFILE= "C:\STA311\w05\w05-iris-csv.csv"
            DBMS=CSV REPLACE;  /* database management system identifier: CSV  */
     GETNAMES=YES;
     DATAROW=2;  /*  */
       GUESSINGROWS =MAX; /* Need to specified whenever the data set has character
                             variables. The maximum number of row SAS can guess is
                             32767. */
RUN;


/*************************************************************
                  Topic 2. PROC MEANS
```

```
 Certain SAS procedures can only be performed on numeric data.
 Two such procedures - PROC MEANS and PROC UNIVARIATE - are
 illustrated here using the height/weight SAS data set.
 (Note that PROC SUMMARY generates output similar to PROC MEANS.)

  Several Statements to know:
  1. CLASS
  2. BY
  3. OUTPUT
*************************************************************/
DM "CLEAR LOG";
DM "CLEAR OUT";

/** Example 2.1: default descriotive statistics -
               N Mean Std Dev Minimum Maximum     **/
PROC MEANS DATA = iris_csv;    /* Begin the PROC step */
  /* Add 2 titles */
  TITLE1 'PROC MEANS:  Example 2.1';
  TITLE2 'No keywords specified';
RUN;             /* End the PROC step */

/** Explore the information in the data set **/
PROC CONTENTS DATA = IRIS_CSV;
RUN;


/*************************************************************
   Example 2.2: available key words - all descriptive statistics

   When using PROC MEANS, the CLASS statement avoids having
   to sort the data first, but the CLASS statement is more
   suited to smaller data sets or when just a few CLASS
   variables are to be used.

 Some of the keywords available with PROC MEANS:
        N - number of observations
        MEAN - mean value
        MIN - minimum value
```

```
        Q1 - first quartile
        Median - middle value of the sorted data set
        Q3 - third quartile
        MAX - maximum value
        SUM - the total of values
        NMISS - number of missing values
        MAXDEC = n - setting the maximum number of decimal places

*****************************************************************/

PROC MEANS DATA = IRIS_CSV N MIN Q1 MEDIAN MEAN Q3 MAX SUM NMISS ;
     VAR sepal_length;
     /*Separate the analysis by values of variety */
     CLASS variety; /* CLASS statement produces a single table */

     /* Add 3 titles */
     TITLE1 'PROC MEANS:  Example 2.2';
     TITLE2 'Use of VAR, CLASS, and TITLE statements';
     TITLE3 'CLASSED by variety';
RUN;

/* Example 2.3: BY- statement to replace CLASS statement!
                BY statement will be used frequently in the future.
                In order to use BY statement, the data set has to
                be sorted by the CLASS variable - variety!      */
PROC SORT DATA = IRIS_CSV;
BY variety;
RUN;

OPTION LS=150;

/* pay attention to the  */
PROC MEANS N MIN Q1 MEAN MEDIAN Q3 MAX STD SUM NMISS MAXDEC = 1 DATA = iris_csv MAXDEC = 3;
     /*Separate the analysis by values of variety */
     BY variety;   /* BY statement produces three separate tables! */
     /*Apply analysis only to "sepal_length" variable*/
     VAR sepal_length;
     /* Add 3 titles */
```

```
     TITLE1 'PROC MEANS:  Example 2.3';
     TITLE2 'Use of VAR, BY, and TITLE statements';
     TITLE3 'CLASSED by variety';
       TITLE4 "BY Statement - Produces 3 Tables";
RUN;

/** Other descriptive statistics with PROC MEANS:

    SUM     Sum of observations
    MEDIAN  50th percentile
    P1      -> 1st percentile
    P5      -> 5th percentile
    P10     -> 10th percentile
    P90     -> 90th percentile
    P95     -> 95th percentile
    P99     -> 99th percentile
    Q1      -> First Quartile
    Q3      -> Third Quartile
    *****
    VAR     -> Variance
    RANGE   -> Range
    USS     -> Uncorr. sum of squares
    CSS     -> Corr. sum of squares
    STDERR  -> Standard Error
    T -> Student's t value for testing Ho: md = 0
    PRT     -> P-value associated with t-test above
    SUMWGT  -> Sum of the WEIGHT variable values
    QRANGE  -> Quartile range
    *****
    CLM -> confidence limits on the mean
    LCLM -> lower confidence limit on the mean (one-sided)
    UCLM -> upper confidence limit on the mean (one-sided)
**/

/** Example 2.4: confidence interval for mean sepal_length and
                 sepal_width by variety                    **/

PROC MEANS DATA = IRIS_CSV   MAXDEC = 3
```

```
                  N             /* sample size */
                  MEAN          /* sample mean */
                  MAXDEC = 1    /* decimal places to keep */
                  ALPHA = 0.01  /* significant level  */
                  CLM           /* two sided confidence limits: 99%  */
                  LCLM          /* lower confidence limit: 99% */
                  UCLM          /* upper confidence limit: 99% */
                  ;
     /*Separate the analysis by values of variety */
     CLASS variety;   /* want to create a single table */
     /*Apply analysis only to "sepal_length" and "sepal_width" variables*/
     VAR sepal_length sepal_length;
     /* Add 3 titles */
     TITLE1 'PROC MEANS:  Example 2.4';
     TITLE2 'Confidence Intervals: 99%';
     TITLE3 'CLASSED by variety';
RUN;

PROC PRINT; RUN;


OPTIONS PS = 90 LS =70 NONUMBER NODATE;   /* global options */

/* Example 2.5: OUTPUT- statement creates a SAS data set that stores
                the output information from PROC MEANS! - This is another
                method for creating SAS data sets.                  */

PROC MEANS DATA = IRIS_CSV ALPHA=0.01 NOPRINT; /* Setting up confidence level.
                              if not specified, default level 0.05  */
                  *CLM           /* suppress the output statistics */
                  N              /* sample size */
                  MEAN           /* sample mean */
                  MAXDEC = 1     /* decimal places to keep */
                  ALPHA = 0.01   /* significant level  */
                  CLM            /* two sided confidence limits: 99%  */
                  LCLM           /* lower confidence limit: 99% */
                  UCLM           /* upper confidence limit: 99% */
                  ;
```

```
     /*Separate the analysis by values of variety */
     CLASS variety;   /* want to create a single table */
     /*Apply analysis only to "sepal_length" and "sepal_width" variables*/
     VAR sepal_length sepal_length;
     /* output an SAS data set */
       OUTPUT OUT = My_CLM            /* output SAS data set --> temp lib*/
                  N = size            /* sample size */
                MEAN = avg
                STDERR = sd_err      /* standard error */
                LCLM = LCI           /* lower confidence limit  */
                UCLM = UCI;          /* upper confidence limit  */
     /* Add 3 titles */
     TITLE1 'PROC MEANS:  Example 2.5';
     TITLE2 'Confidence Intervals: 99%';
     TITLE3 'CLASSED by variety';
RUN;

TITLE "";  /* this clears all previous titles */

PROC PRINT; RUN;

/************************************************************
            Topic 3:  PROC FREQ

Frequency tables and crosstabulation tables provide a way to
summarize data for ordinal and categorical variables.
Frequency tables show the distribution of a variable's values.
************************************************************/

/**  Example 3.1. Simple frequency table **/
PROC FREQ DATA = IRIS_CSV;
TABLE VARIETY;
TITLE "Frequency Table of Variety";
RUN;

/** Example 3.2. Output frequency table to a SAS data set
    Frequencies and percentages calculated using Proc Freq
    can also be saved to an output dataset using the OUT option
```

```
    combined with the TABLES statement.

    The OUTCUM option can also be added to include the
    cumulative frequencies in the output dataset          **/

PROC FREQ DATA = IRIS_CSV;
TABLE VARIETY / OUT = Variety_FREQ OUTCUM;
TITLE1 "Frequency Table of Variety";
TITLE2 "Store the output to a SAS data set";
RUN;

PROC PRINT;
TITLE "Print out the output Frequency table";
RUN;

TITLE "";

/** We will revisit PROC FREQ after we do some data manipulation **/

/*************************************************************
            Topic 4:  PROC UNIVARIATE

PROC UNIVARIATE produces descriptive statistics on continuous
variables just like proc means, but many more of them, and also
can produce some univariate plots.

*************************************************************/

OPTION NOCENTER LS = 90;

/* Example 4.1:  Univariate analysis of a numerical variable
                 -- a simple example                        */
PROC UNIVARIATE DATA = IRIS_CSV
    NORMAL  /* normality test */
      FREQ    /* frequency table of sepal_length: not useful*/
    PLOT;   /* plot of the histogram of sepal_length */
    VAR Sepal_length;
      /* specification of details in the histogram
```

```
          1. place a normal curve on the top of the histogram
    */
    HISTOGRAM Sepal_length/NORMAL;
    TITLE 'PROC UNIVARIATE EXAMPLE';
    FOOTNOTE 'Evaluate the distribution of Sepal_length';
RUN;

TITLE "";
FOOTNOTE "";


/* Example 4.2:  PROC UNIVARIATE -- OUTPUT SAS Data set
   This is a very simple example with a few descriptive statistics  */
PROC UNIVARIATE  DATA = IRIS_CSV NOPRINT;
    VAR Sepal_length;
      OUTPUT OUT = UNIVAR_OUT_AVG          /* output data set #1 */
             MEAN = Sepal_length_avg;
      OUTPUT    OUT = UNIV_OUT_MORE        /* output data set #2 */
                /* three descriptive statistics were written to the SAS data set */
             MEAN =Sepal_L_AVG
                   STD = Sepal_L_STD
                   MIN = Sepal_L_MIN;
RUN;


/*****************************************************
  Available Statistics for OUT in PROC UNIVARIATE

*** Descriptive Statistics ***
CSS   ==> Sum of squares corrected for the mean
CV ==> Percent coefficient of variation
KURTOSIS | KURT   ==> Measurement of the heaviness of tails
MAX   ==> Largest (maximum) value
MEAN ==> Arithmetic mean
MIN   ==> Smallest (minimum) value
MODE ==> Most frequent value (if not unique, the smallest mode)
N==> Number of observations on which calculations are based
NMISS==> Number of missing observations
```

```
NOBS==> Total number of observations
RANGE==> Difference between the maximum and minimum values
SKEWNESS | SKEW==> Measurement of the tendency of the deviations to be larger in one direction than in the
other
STD | STDDEV==> Standard deviation
STDMEAN | STDERR==> Standard error of the mean
SUM==> Sum
SUMWGT==> Sum of the weights
USS==> Uncorrected sum of squares
VAR==> Variance

**** Quantile Statistics  ****
MEDIAN | Q2 | P50==> middle value (50th percentile)
P1==> 1st percentile
P5==> 5th percentile
P10==> 10th percentile
P90==> 90th percentile
P95==> 95th percentile
P99==> 99th percentile
Q1 | P25==> Lower quartile (25th percentile)
Q3 | P75==> Upper quartile (75th percentile)
QRANGE==> Difference between the upper and lower quartiles
           (also known as the inner quartile range)

*** Robust Statistics ***
GINI==> Gini's mean difference
MAD==> Median absolute difference
QN==> 2nd variation of median absolute difference
SN==> 1st variation of median absolute difference
STD_GINI==> Standard deviation for Gini's mean difference
STD_MAD==> Standard deviation for median absolute difference
STD_QN==> Standard deviation for the second variation of
          the median absolute difference
STD_QRANGE==> Estimate of the standard deviation, based on
              interquartile range
STD_SN==> Standard deviation for the first variation of the
          median absolute difference
```

```
*** Hypothesis Test Statistics ***
MSIGN==> Sign statistic
NORMAL==> Test statistic for normality. If the sample size is less than or
          equal to 2000, this is the Shapiro-Wilk W statistic. Otherwise,
          it is the Kolmogorov D statistic.
PROBM==> Probability of a greater absolute value for the sign statistic
PROBN==> Probability that the data came from a normal distribution
PROBS==> Probability of a greater absolute value for the signed-rank statistic
PROBT==> Two-tailed p-value for Student's t statistic with  degrees of freedom
SIGNRANK==> Signed rank statistic
T==> Student's t statistic to test the null hypothesis that the
     the population mean is equal to mu_0

**************************************************/
```