CSC423 Data Analysis and Regression
SAS Procedures for Exploratory Data Analysis (EDA)

PROC MEANS

The **PROC Means** provides data summarization tools to compute descriptive statistics for variables across all observations and within groups of observations. For example, it calculates the descriptive statistics based on moments, estimates the percentiles and the median value, calculates the confidence limits for the mean and performs a **t** test on the population average.

> **PROC MEANS DATA**=*dataset-name* <*statistic-keyword(s)*>;
> **VAR** *variables*; ← one or more measurement variables in the dataset
> **RUN;**
>
> The **statistics keywords** specify which statistics to compute and the order to display them in the output.
> The available keywords in the PROC statement include: MEAN (average), STD (standard deviation), STDERR (standard error), P50 (median), Q1 (first quartile), Q3 (third quartile).
> Keywords for confidence intervals: CLM, ALPHA=*value* specifies the confidence level (1-*value*)%
> Keywords for t-tests: T (t-statistic), PRT (p-value). Notice that SAS provides a test only for a zero population average, i.e. H0: μ=0.

PROC UNIVARIATE

You can use the PROC UNIVARIATE to compute the following:

- Descriptive statistics based on moments (including skewness and kurtosis), quantiles or percentiles (such as the median), frequency tables, and extreme values
- Histograms. Optionally, these can be fitted with probability density curves for various distributions and with kernel density estimates. (HISTOGRAM statement)
- Probability plots. These plots facilitate the comparison of a data distribution with various theoretical distributions. (QQPLOT statement or PROBPLOT)

Syntax:

```
PROC UNIVARIATE < options >;
     BY variables ;
     VAR variables ;
     CLASS variable-1;
     HISTOGRAM < variables > /normal;
     OUTPUT OUT=new-SAS-data-set;
     PROBPLOT < variables > / normal (mu=est sigma=est);
     RUN;
```

<OPTIONS> Options for the PROC UNIVARIATE are *normal* and *plot*. The *normal* option is used to compute tests on normality to determine if variables defined in VAR come from a normal distribution (Shapiro-Wilk test, Kolmogorov – Smirnov test…). The *plot* option creates low-level boxplot and normal probability plot.

BY variables ; Calculate separate statistics for each BY group

VAR variables **;** Select the analysis variables and determine their order in the report. Variables are one or more measurement variables in the dataset. If you do not use the VAR statement, all numeric variables in the data set are analyzed.

CLASS variable-1**;** specify one or two variables that group the data into classification levels. The analysis is carried out for each combination of level.

HISTOGRAM < variables > / **NORMAL CFILL** = white **PFILL** = solid **MIDPOINTS** = <list >;

Creates a histogram, the NORMAL option displays a fitted normal curve on the histogram, the MIDPOINTS= option specifies midpoints for the histogram, cfill & pfill control the appearance of the histogram. If no variables are specified, histograms are created for each variable defined in the VAR statement.

OUTPUT OUT=new-SAS-data-set**;**

The OUTPUT statement saves statistics and BY variables in an output data set.

PROBPLOT < variables > **/ NORMAL (MU=EST SIGMA=EST);**

Creates a normal probability plot

RUN;

EXAMPLE

Consider the following data set on the time between machine failures. Data were collected during a study on machine performance that involved 39 similar machines. The producing company states that on average the time between failures is 20 hours. The researchers believe that on average the time between failures is longer than 20 hours, so they want to estimate the average time between failures and test the claim of the producing company.

DATA: 21.6 21.7 22.7 21.2 21.9 21.6 24.8 22.5 21.9 23.6 23.0 22.3 23.3 24.2 25.5 22.5 23.1 24.7 26.2 24.7 23.6 21.5 23.7 24.3 26.2 22.5 22.7 21.5 24.3 24.7 25.7 27.3 22.4 20.1 26.3 23.9 21.7 23.3 22.2

STEP 1 – Read the data into SAS and create the SAS data set “failure”

```
Title 'Time between failures';

data failure;
infile "c:/…/faildata.dat";
input time;
timecent=time-20;
label time = 'time between failures' timecent = time-20 hours;
```

STEP 2 – Compute some descriptive statistics about the data and a 95% confidence interval for the average time between failures.

```
proc means mean std stderr clm p25 p50 p75;
var time;
run;
```

```
                         Time between failures
                          The MEANS Procedure

         Analysis Variable : time time between failures
                                              Lower 95%     Upper 95%
      Mean        Std Dev     Std Error    CL for Mean   CL for Mean
-------------------------------------------------------------------
23.3564103     1.6676165     0.2670323    22.8158315    23.8969890
-------------------------------------------------------------------
         Analysis Variable : time time between failures

           25th Pctl      50th Pctl       75th Pctl
      -----------------------------------------------
          21.9000000     23.1000000      24.7000000
      -----------------------------------------------
```

The estimated average time between failures is 23.356 hours, with standard error equal to 0.267 hours. The average time is between 22.81 hours and 23.9 hours.

STEP 3 – Test the company's claim that the average time between failures is 20 hours.
Null hypothesis: **Ho: μ=20** hours against the alternative hypothesis that **Ha: μ > 20 hours.**
To use SAS, we need to compute the variable timecent=time-20 and express the test as:
Ho: μ=0 vs Ha: μ>0 where μ is now the population average for the new variable timecent.
Note: Examine the data histogram and the normal probability plot to check the normality assumptions, before carrying out the statistical test.

```
proc univariate normal;
var timecent;
histogram /cfill=WHITE pfill=SOLID name='HIST' normal;
probplot/normal(mu=est sigma=est color=BLUE l=1 w=1);
run;
```

```
                     The UNIVARIATE Procedure
                 Variable:  timecent  (time-20 hours)

                             Moments
   N                           39    Sum Weights               39
   Mean                3.35641026    Sum Observations       130.9
   Std Deviation       1.66761646    Variance          2.78094467
   Skewness            0.47112614    Kurtosis          -0.3745406
   Uncorrected SS          545.03    Corrected SS      105.675897
   Coeff Variation     49.6845241    Std Error Mean    0.26703235

                   Basic Statistical Measures
         Location                     Variability
     Mean      3.356410    Std Deviation            1.66762
     Median    3.100000    Variance                 2.78094
     Mode      2.500000    Range                    7.20000
                           Interquartile Range      2.80000
NOTE: The mode displayed is the smallest of 2 modes
    with a count of 3.

                   Tests for Location: Mu0=0

           Test           -Statistic-    -----p Value------

           Student's t    t    12.5693    Pr > |t|    <.0001
           Sign           M       19.5    Pr >= |M|   <.0001
           Signed Rank    S        390    Pr >= |S|   <.0001
```

RESULT: The t test is highly significant, since the p-value is very small (<.0001/2=.00005). Thus the data do not support the company's claim and are consistent with the researchers' hypothesis. Note that the t-statistic is positive and very large, indicating that the actual time between failures is sensibly larger than 20 hours.

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	**W**	**0.965106**	**Pr < W**	**0.2628**
Kolmogorov-Smirnov	**D**	**0.114608**	**Pr > D**	**>0.1500**
Cramer-von Mises	W-Sq	0.078669	Pr > W-Sq	0.2166
Anderson-Darling	A-Sq	0.512045	Pr > A-Sq	0.1921

Quantiles (Definition 5)

Quantile	Estimate
100% Max	7.3
99%	7.3
95%	6.3
90%	6.2
75% Q3	4.7
50% Median	3.1
25% Q1	1.9
10%	1.5
5%	1.2
1%	0.1
0% Min	0.1

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
0.1	34	5.7	31
1.2	4	6.2	19
1.5	28	6.2	25
1.5	22	6.3	35
1.6	6	7.3	32

Parameters for Normal Distribution

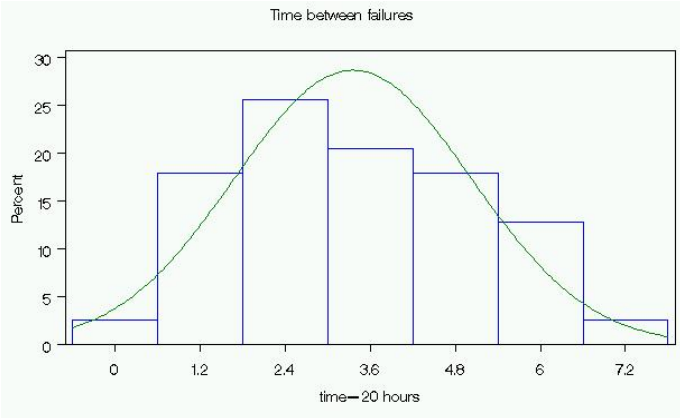
Parameter	Symbol	Estimate
Mean	Mu	3.35641
Std Dev	Sigma	1.667616

Goodness-of-Fit Tests for Normal Distribution

Test	---Statistic----		-----p Value-----	
Kolmogorov-Smirnov	D	0.11460833	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.07866879	Pr > W-Sq	0.217
Anderson-Darling	A-Sq	0.51204470	Pr > A-Sq	0.192

The Shapiro-Wilk test supports the assumption that the data arise from a normal population. The normal probability plot confirms this result, because the points lie close to a line. The histogram, however, is skewed. We assume that data come from a normally distributed population and we use the t-test. Notice that both the sign test and the t test produce the same result.

Histogram of the data



Normal probability plot