Examining Data for Analysis

Fikret Isik,
North Carolina State University,
Cooperative Tree Improvement Program, Raleigh, NC, U.S.A.

1.1 Introduction

This is **the most critical part** of data handling. Even one outlier can ruin all the results and misled you. Take your time and check the data before proceeding to more complicated data analysis steps. **Summary statistics** gives you an overall picture about the data. Create summary statistics files and save them. You will always need to go back to the descriptive statistics during report writing. In this chapter, we will present a few simple but useful SAS procedures to examine the data.

1.2 Getting to know data

If you are not familiar with the data set, the CONTENT procedure of SAS is handy tool to get to know data. The procedure tells you;

1. The name and the location of data (What is the full name of the data set? which computer do data are stored, the name of folders data are located etc.)
2. When the data were created, when data were last time modified,
3. Number of variables, whether they are numeric or character
4. Which environment the data were produced (i.e., Windows 1998, XP, Unix, etc..)

Code 1: Getting to know data

```
Proc contents data=mydata;
run;
```

Output 1:

The CONTENTS Procedure

Data Set Name	HBOOK.MYDATA	Observations	4887
Member Type	DATA	Variables	19
Engine	V9	Indexes	0
Created	Sunday, February 12, 2006	Observation Length	160
Last Modified	Sunday, February 12, 2006	Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

This is a SAS data set as it is clear from the data set name (HBOOK.MYDATA). The first part in the data set name 'HBOOK' is the SAS library name. The second part is the actual name. Total number of observations (number of rows 4887) and total number of variables (number of columns 19) are given along with some other details about the data.

Engine/Host Dependent Information

Data Set Page Size	16384
Number of Data Set Pages	49
First Data Page	1
Max Obs per Page	102
Obs in First Data Page	83
Number of Data Set Repairs	0
File Name	c:\handbook\mydata.sas7bdat
Release Created	9.0101M3
Host Created	XP_PRO

The physical location of the data set is given as '**c:\handbook**'. If you look at the folder 'HANDBOOK', the data set will be seen as **mydata.sas7bdat** because it is a SAS data set.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
5	Col	Num	8	BEST12.	BEST32.
7	Family	Char	16	$16.	$16.
11	Fork	Num	8	BEST12.	BEST32.
10	Ht	Num	8	BEST12.	BEST32.
14	Qual	Num	8	BEST12.	BEST32.

Alphabetic List of Variables and Attributes table shows the list of variables, whether they are numeric (NUM) or character (CHAR), their length (the column width) and format. During the data analysis it is important to know whether the variable is character or numeric.

1.3 Examination of numeric variables

The UNIVARIATE Procedure gives the most complete information for a **numeric** variable. The procedure; (1) examines the distribution of numeric variables, (2) calculates descriptive statistics of numeric variables, (3) tabulates extreme observations of numeric variables, (4) plots the distribution of numeric variables etc…

Code 2: Examining data for outliers, errors

```
Proc univariate data=mydata plot;
   var height;
   ID family;
run;
```

(1) The PLOT statement creates low-resolution stem-and-leaf, box, and normal probability plots. They are useful to inspect data visually.

(2) Use VAR statement to specify which numeric variables to examine. If you do not specify which variables to examine, the code produces results for all the numeric variables in the data.

(3) ID specifies one or more variables whose values identify the extreme observations. If you do not use ID, the code gives only the ROW number of the extreme observation in data.

Output 2:

By default, the procedure produces many tables. Here, we only gave three important tables.

The UNIVARIATE Procedure
Variable: Ht

Moments

N	3611	Sum Weights	3611
Mean	193.652728	Sum Observations	699280
Std Deviation	33.2983763	Variance	1108.78186
Skewness	-0.0044524	Kurtosis	0.93414794
Uncorrected SS	139420182	Corrected SS	4002702.52
Coeff Variation	17.1948914	Std Error Mean	0.554127

Basic Statistical Measures

Location		Variability	
Mean	193.6527	Std Deviation	33.29838
Median	194.0000	Variance	1109
Mode	200.0000	Range	387.00000
		Interquartile Range	43.00000

The MOMENTS shows the overall descriptive statistics for the numeric variable (height). Large standard deviation could suggest existence of outliers. Check the Skewness also. It is a departure from normality. A perfect normal distribution has zero skewness. The reason for large negative or positive skewness could be existence of outliers. If the absolute value of Skewness is greater than 3.0, then, a linear model may not be the best way to analyze data.

Basic Statistical Measures

Location		Variability	
Mean	193.6527	Std Deviation	33.29838
Median	194.0000	Variance	1109
Mode	200.0000	Range	387.00000
		Interquartile Range	43.00000

Tests for Location: Mu0=0

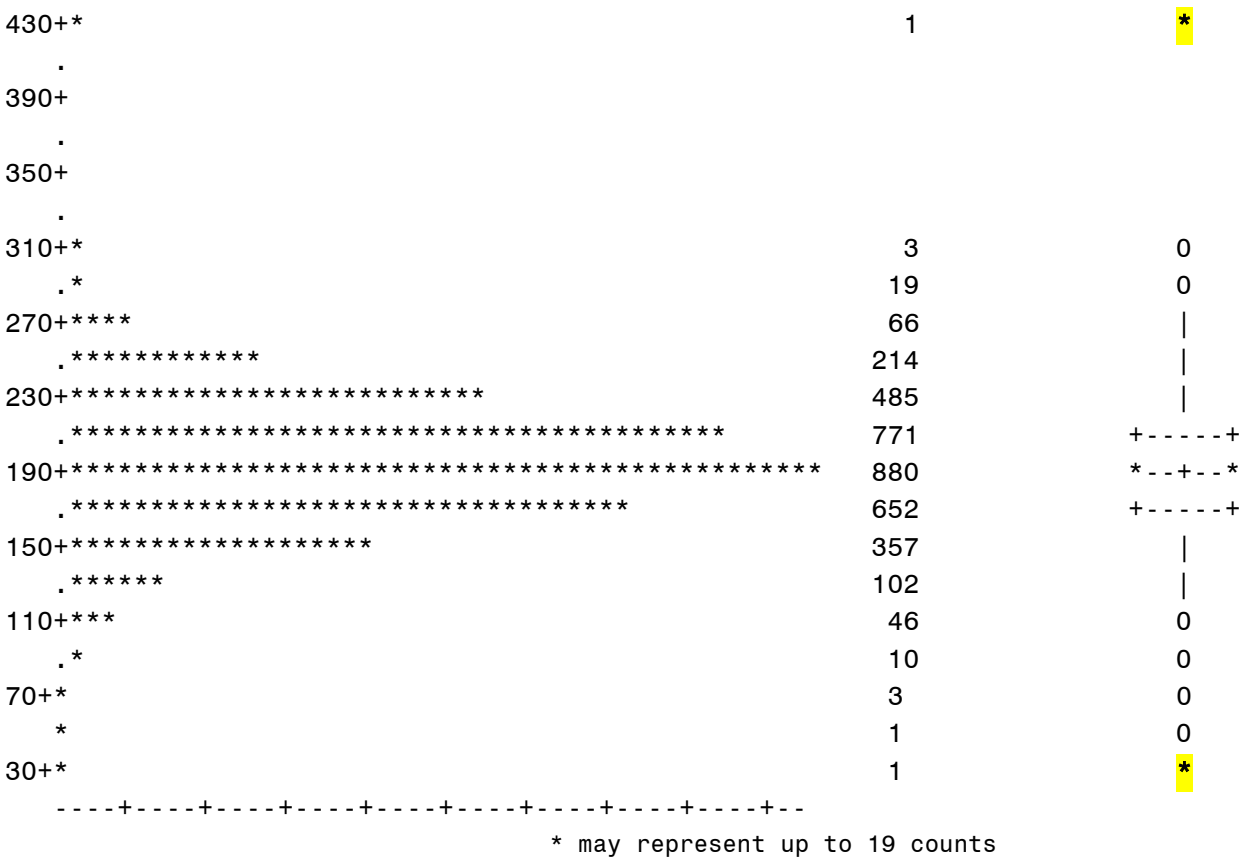
Test	-Statistic-		-----p Value------	
Student's t	t	349.4735	Pr > |t|	<.0001
Sign	M	1805.5	Pr >= |M|	<.0001
Signed Rank	S	3260733	Pr >= |S|	<.0001

Extreme Observations

----------Lowest----------			----------Highest----------		
Value	Family	Obs	Value	Family	Obs
33	1-1716	4278	292	1-549	2726
48	1-1545	3309	300	11-1542	171
70	1-1545	887	300	11-1540	191
76	1-1545	2912	318	1-1548	251
78	1-1712	3643	420	VA-NC-0-85-14	170

The lowest five and highest five extreme observations (VALUE) were given for the numeric variable. The smallest value 33 is if for Family 1-1716. It is located in the row number 4278. The highest value is 420. Are they outliers?

Histogram # Boxplot



STEM-LEAF PLOT shows the distribution like a bar chart.

You see frequencies after the stem-leaf plot.

BOX PLOT: The median in the middle of the box (*----*) and the mean (+) are overlapped, suggesting a symmetric distribution (or very small skewness).

End of the box are (+----+) the first (Q1) and third (Q3) quantiles. The difference between Q3-Q1 is the Inter Quartile Range (IQR). IQR is important, because, observations that lie beyond 3*IQR distance from the box are extreme outliers and are denoted with '*' symbol.

Mild outliers lie 1.5*IQR distance from the ends of the box and are denoted with the '0'symbol. There are 2 extreme outliers in the data.

The normal probability plot is another plot to examine the distribution for normality. For a normally distributed data without any outlier, the trend would be straight between y and x axis. Here there is one too large and a few too small observations distorting this.

1.4 Examination of character/discrete variables

The FREQ procedure is a useful tool for detecting errors in discrete or character variables.

Code 3: Obtaining number of observations

```
Proc freq data=mydata ;
table block family/out=FreqTable nocol norow nocum nopercent;
run;
```

(1) TABLE statement creates one-way tables for block and family
(2) An output data set called FreqTables (OUT=FreqTables) is created. The new output file contains specified statistics
(3) Column and row percentages were excluded from the output with NOCOL, NOROW NOCUL and NOPERCENT statements

Output 3:

block	Frequency
1	120
2	115
3	116
.	.
34	118
35	116
36	120
118	1

Frequency Missing = 765

⇑ Numbers of observations per block is given. The number of trees per block is expected to be 120 but because of low survival, the actual number of trees (frequency) is usually less. The block number 118 must be a typo error because blocks are numbered from 1 to 36 in the field experiment. In addition, there could not be one tree in a block. This needs to be checked out and corrected in the raw data.

Family	Frequency

```
11-1531                  31
11-1532                   7
11-1533                  35
...                     ...
TX-HAB-62                35
filler                    1
```

Numbers of observations per family across all replications is expected to be 36 because there are 36 blocks and each of 120 families had one tree per block.

Here, for family 11-1532 there are only 7 trees. This family needs to be checked. The 'filler' in the family list is not an experimental family but a random tree to fill the gap of a dead tree after one year of planting. Filler trees should be removed from the data in the final analysis.

The data are checked now. The outliers are mostly typo errors. In this case it is easy to correct them. You should have a good reason to change a value in the data. We are now ready to summarize the data.

1.5 Data Summary

The MEANS procedure of SAS is a powerful tool to summarize data. The statistics can be saved as an output file for later reference. See the help system for more information about the MEANS procedure. The following code produces overall descriptive statistics for height and diameter of trees.

Code 4: Producing overall basic statistics

```
Proc means data=hbook.mydata mean std cv n maxdec=2;
   var Height diameter;
run;
```

(1) Here, the means (MEAN), standard deviations (STD), coefficients of variation (CV) and the numbers of observations per variable (N) are requested.

(2) With the VAR option two variables are selected, the HEIGHT and the DIAMETER.

Output 4:

The MEANS Procedure

Variable	Mean	Std Dev	Coeff of Variation	N
Height	193.65	33.30	17.19	3611
Diameter	12.08	8.20	67.85	3615

Height has the mean of 193.65 with standard deviation of 33.30. Diameter has the mean of 12.08 with the standard deviation of 8.20.

Code 5: Basic statistics for each level of a factor

```
Proc means data=hbook.mydata mean std cv n maxdec=2;
class block ;
var Height ;
run;
```

(2) Here, the means, standard deviation (std), coefficient of variation (cv) and the number of observations (n) are requested for each block.

Output 5:

Analysis Variable : Height

block	N Obs	Mean	Std Dev	Coeff of Variation	N
1	125	194.40	32.81	16.88	94
2	115	179.58	38.08	21.21	85
3	116	192.44	30.42	15.81	85
		... more lines			
34	118	189.40	26.15	13.81	90
35	120	212.79	30.41	14.29	99

Coefficients of variation for blocks range from 21.21 (block 2) to 14.29 (block 35). Block 2 is the most variable. Survival is the highest in block 35 (99 trees out of 120) and the lowest in block 2.