Introductory Computing for Statistics

Lecture 3:Descriptive Statistics,Charts and Plots

Xiao Li

October 23, 2017

Review of Lecture 2

- INFILE, INPUT statement
- FILENAME, PUT, FILE statement
- PROC SORT, PROC MERGE, and BY VARIABLE
- conditional statement (if...else...), and KEEP, DROP statement

Today's topics

- Creating titles and variable labels
  - LABEL and TITLE statements
  - Generate descriptive statistics
  - PROC UNIVARIATE
  - PROC MEANS
  - PROC FREQ
- Create charts and plots
  - PROC CHART
  - PROC PLOT

Corresponding reading (Elliott and Morrell): Ch 5 and Ch 7

LABEL statement

LABEL statement is to attach an extended label to a **variable**.

Syntax

LABEL variable1 = 'label1' variable2 = 'label2' .... ;

```
LABEL time = time needed to complete exam
      score = 'exam score'
      student = "student's name"
      teacher = 'teacher"s name'
;
```

LABEL statement

Notes about LABEL statement

- If you want the LABELS to show in the output, you must add the LABEL option in the PROC PRINT statement.
- Labels can contain up to 255 characters. Any characters are valid for the labels.
- If a label contains single quotes or apostrophes, you have two choices:
    - Use single quotes around the entire text and **two** single quotes in place of the apostrophe
    - Use double quotes around the entire text and **one** single quote inside test.

LABEL statement

LABEL statements are valid in both DATA steps and PROC steps.

- If in a DATA step, the label is **permanent** and can only be changed by a subsequent LABEL statement.
- If in a PROC step, the label is only **temporary** and is valid only for that procedure (ie. that PROC statement).

TITLE statement

TITLE statement is the title of the **output**. By default this title is "The SAS System". To change this title use the following code:

Syntax

TITLE 'title content';

```
      TITLE 'Survey Report';
      TITLE2 'Linear Regression';
      ;
```

Note: The use of single quotations is the same here as in LABEL statements.

TITLE statement

Notes about TITLE statement

- TITLE statements can appear in DATA steps or PROC steps.
- You can create more than one title per page by numbering the TITLE statements TITLE1, TITLE2, etc.
- TITLE statements are permanent unless they are changed by adding subsequent TITLE statements.
- Changing TITLE*n* will delete all titles with a number greater than *n*.

LABEL and TITLE statements example

Example 3.1

```
DATA one;
INPUT name $ weight height @@;
LABEL name=‘Name’ weight=‘Weight(lb)’
      height=‘Height(cm)’ ;
TITLE ‘Survey Data’;
TITLE2 ‘FROM NJ’;
DATALINES;
John 200 175 Jeffrey 160 180
Tom 140 162 Chris 155 170
;
RUN;
PROC PRINT data=one LABEL NOOBS; /* TITLE ‘new’ ; */
RUN;
```

Notes: The option "NOOBS" tells SAS to not print out the observation number in the output.

Three PROCs for descriptive statistics

- PROC UNIVARIATE
- PROC MEANS
- PROC FREQ

PROC UNIVARIATE

Syntax

PROC UNIVARIATE data=dataset options;
BY vaiables;
VAR variables;
ID variables;

By default, SAS prints the following descriptive statistics for every numerical variable:

Moments mean, variance, skewness, kurtosis, etc.

Basic statistical measures median, mode, range, IQR, etc.

Tests for location t-test, sign test, signed-rank test

Quantiles 0%,1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, 99%,100%

Extreme values the five largest and five smallest observations

PROC UNIVARIATE options

Syntax

PROC UNIVARIATE data=dataset options;
BY vaiables;
VAR variables;
ID variables;

Options:

1. NORMAL—— Generates several statistics to test for normality and their corresponding p-values.
   - The null hypothesis is that the data are normally distributed vs. the alternative that the data are non-normal.
   - As a general rule of thumb, if the p-value from this test is less than 5%, then there is enough statistical evidence to conclude that the data are **not** normally distributed.

PROC UNIVARIATE options

Syntax

PROC UNIVARIATE data=dataset options;
BY vaiables;
VAR variables;
ID variables;

2. PLOT —— Add a stem-and-leaf plot, a box plot and a normal probability plot.
   - Stem-and-leaf plot: for large datasets this may be replaced by a horizontal bar chart by default.
   - Box plot: the top and bottom of the box are the 75th and 25th percentiles, respectively. The median is denoted by a bar, and the mean is denoted by a plus sign.
   - Normal probability plot: data points are represented by asterisks, and plus signs denote a straight line for reference. If the data are normally distributed, the data points should fall along the straight line.

PROC UNIVARIATE statements

Syntax

PROC UNIVARIATE data=dataset options;
BY vaiables;
VAR variables;
ID variables;

- VAR statement tells SAS to provide results for a specific list of variables rather than for all of the variables.
- ID (means "identifier") specifies a variable whose value is printed next to the smallest and largest observations. This statement makes it easy to identify extreme observations.

PROC UNIVARIATE statements

Syntax

PROC UNIVARIATE data=dataset options;
BY vaiables;
VAR variables;
ID variables;

- VAR statement tells SAS to provide results for a specific list of variables rather than for all of the variables.
- ID (means "identifier") specifies a variable whose value is printed next to the smallest and largest observations. This statement makes it easy to identify extreme observations.
- By statement specifies a variable for which every different value defines a subgroup of observations. The data **must** first be sorted by the same BY variable before using the BY statement in PROC UNIVARIATE.

PROC UNIVARIATE statements

Syntax

PROC UNIVARIATE data=dataset options;
BY vaiables;
VAR variables;
ID variables;

- VAR statement tells SAS to provide results for a specific list of variables rather than for all of the variables.
- ID (means "identifier") specifies a variable whose value is printed next to the smallest and largest observations. This statement makes it easy to identify extreme observations.
- By statement specifies a variable for which every different value defines a subgroup of observations. The data **must** first be sorted by the same BY variable before using the BY statement in PROC UNIVARIATE.

PROC UNIVARIATE example

Example 3.2 ▸ Example 3.2 code

PROC MEANS

PROC MEANS is good procedure to use when you are only interested in the basic descriptive statistics. It is more concise than PROC UNIVARIATE.

Syntax

PROC UNIVARIATE data=dataset options;
VAR variables;
BY vaiables;

Default Setting: without any specified options, SAS provides a 5-value summary - *sample size (n), minimum value (min), maximum value (max), mean*, and *standard deviation* (std).

PROC MEANS options

nmiss – the number of missing observations.

range – the range (MAX - MIN).

sum – the sum.

var – the variance.

stderr – the standard error of the mean.

t – t-statistic for testing whether the mean is significantly different from 0.

probt – provides the p-value for t-test (above).

PROC MEANS example

Example 3.3 ▸ Example 3.3 code

PROC FREQ

PROC FREQ generates tables for categorical data.

Syntax

```
PROC FREQ data=dataset options;
BY variables;
TABLES var1 var1*var2 / options;
```

PROC FREQ statements

Syntax

```
PROC FREQ data=dataset options;
BY variables;
TABLES var1 var1*var2 / options;
```

TABLES – This specifies which tables to create either by a single variable name or by two variable names separated by the star sign.

Tables Options:

chisq – Compute the chi-square statistic for testing independence/homogeneity in a two-way table.

exact – Perform Fisher's exact test for tables larger than 2 x 2.

nocol – Omit the column percents from the table.

nocum – Omit the cumulative frequencies from the table.

nofreq – Omit the cell frequencies from the table.

nopercent – Omit any percents from the table.

norow – Omit the row percents from the table.

PROC FREQ statements

Syntax

PROC FREQ data=dataset options;
BY variables;
TABLES var1 var1*var2 / options;

TABLES – This specifies which tables to create either by a single variable name or by two variable names separated by the star sign.

Tables Options:

chisq – Compute the chi-square statistic for testing independence/homogeneity in a two-way table.

exact – Perform Fisher's exact test for tables larger than 2 x 2.

nocol – Omit the column percents from the table.

nocum – Omit the cumulative frequencies from the table.

nofreq – Omit the cell frequencies from the table.

nopercent – Omit any percents from the table.

norow – Omit the row percents from the table.

PROC FREQ example

Example 3.4 (Example 3.3 continued) Two-way tables.

```
PROC freq data=grade;
TITLE 'Grades from Statistics Class';
run;

PROC freq data=grade;
TABLES grade grade*year gender*year /chisq;
run;
```

PROC CHART

Reference: http://galsterhome.com/stats/Tutorial/SAS12.htm

Syntax

```
PROC CHART data=dataset;
BY variables;
VBAR variables / options ;
HBAR variables / options ;
```

PROC CHART statements

- BY statement produces a separate chart for each BY group.
- VBAR statement creates a vertical bar chart (Histogram), while HBAR creates a horizontal one (Rotated histogram).

PROC CHART statements

- BY statement produces a separate chart for each BY group.
- VBAR statement creates a vertical bar chart (Histogram), while HBAR creates a horizontal one (Rotated histogram).

PROC CHART statements

- BY statement produces a separate chart for each BY group.
- VBAR statement creates a vertical bar chart (Histogram), while HBAR creates a horizontal one (Rotated histogram).
  - Options:

    SUBGROUP = variable – The bar can be divided into parts representing the values of the specified variable. The first character of variable is used.

    TYPE = freq (or pct) – The chart contains frequencies (default) or percents.

PROC CHART statements

- BY statement produces a separate chart for each BY group.
- VBAR statement creates a vertical bar chart (Histogram), while HBAR creates a horizontal one (Rotated histogram).
  - Options:

    SUBGROUP = variable – The bar can be divided into parts representing the values of the specified variable. The first character of variable is used.

    TYPE = freq (or pct) – The chart contains frequencies (default) or percents.

PROC CHART example

Example 3.5 (Example 3.3 continued) Histograms

```
PROC CHART data=grade;
BY year;
VBAR grade;
TITLE1 ‘Histogram of grades from statistics class’;
RUN;

PROC CHART data=grade;
VBAR grade/subgroup=year;
TITLE1 ‘Histogram of grades from statistics class’;
RUN;
```

*Attention to the difference between these two steps. Try to understand how BY statement and SUBGROUP option work.

PROC PLOT

Syntax

PROC PLOT data=dataset;
BY variables;
PLOT plot-requests;

- In the simplest type of plot we have: yvar * xvar. Automatically the y-variable goes to the vertical axis and x-variable goes to the horizontal axis.
- The default symbols for the points:
  - Character A represents the value of one observation in the data set.
  - When a point represents the values of two observations, the character B appears, and so on through the alphabet.
  - The character Z is used for the occurrence of 26 or more observations at the same printing position.

PROC PLOT plot-requests

- yvar * xvar = 'char' - Observations are plotted using the character specified, such as '+', '*', or '.'.
- yvar * xvar = variable – Observations are plotted using the first character of the value of variable.
- yvar * (xvar1 xvar2) yvar * xvar1 yvar * xvar2 – Two plots appear on separate pages.
- (yvar1 yvar2) * xvar yvar1 * xvar yvar2 * xvar – Two plots appear on separate pages.
- (yvar xvar1 xvar2) yvar*xvar1 yvar*xvar2 xvar1*xvar2 – Not all combinations; Order counts.
- yvar1 * xvar1 = 'char1' yvar2 * xvar2 = 'char2' /overlay – Two plots yvar1*xvar1 and yvar2*xvar2 appear on the same plot. They are overlaid

PROC PLOT example

Example 3.6 (Example 3.3 continued) Plots

```
PROC PLOT data=grade;
plot course*finexam;
run;
PROC PLOT data=grade;
plot course*finexam=grade;
run;
PROC PLOT data=grade;
plot course*finexam='f' course*quiz='q' /overlay;
run;
PROC PLOT data=grade;
plot (course finexam quiz);
run;
```