Histogram of Numeric Data Distribution from the UNIVARIATE Procedure

Chauthi Nguyen, GlaxoSmithKline, King of Prussia, PA

ABSTRACT

The UNIVARIATE procedure from the Base SAS® Software has been widely used for the descriptive statistic tables and reports. With the use of HISTOGRAM statement in PROC UNIVARIATE, we can have a fast and simple way to review the overall distribution of a quantitative variable in a graphical display. This will be helpful for the interpretation of the statistic reports.

The paper will demonstrate how to create a simple frequency histogram using PROC UNIVARIATE and the HISTOGRAM statement. Using the CLASS statement together with the HISTOGRAM statement will create comparative histograms. In clinical trials, the comparative histograms could be used to compare the treatment effects of a test drug against Placebo.

INTRODUCTION

The word *histogram* is derived from Greek: *histos* 'anything set upright'; *gramma* 'drawing, record, writing'. In statistics, a histogram is a graphical display of tabulated frequency. The histogram differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height. The bars must be adjacent [1].

Histograms allow you to explore your data by displaying the distribution of a continuous variable (percentage of sample) against categories of the value. You can obtain the shape of the distribution and whether the data are distributed symmetrically.

In SAS, the histograms can be produced using PROC UNIVARIATE, PROC CHART, or PROC GCHART. The paper will demonstrate the use of PROC UNIVARIATE with the HISTOGRAM statement and its options, and other related statements that affect the histograms.

THE UNIVARIATE PROCEDURE WITH THE HISTOGRAM STATEMENT

Let's start by creating a simple histogram of the WEIGHT variable. The data is taken from SAS Online Help [2]. Sample of the first 10 observations in the dataset **sashelp.class**:

Obs	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69.0	112.5
2	Alice	F	13	56.5	84.0
3	Barbara	F	13	65.3	98.0
4	Carol	F	14	62.8	102.5
5	Henry	M	14	63.5	102.5
6	James	M	12	57.3	83.0
7	Jane	F	12	59.8	84.5
8	Janet	F	15	62.5	112.5
9	Jeffrey	M	13	62.5	84.0
10	John	M	12	59.0	99.5

Here is a basic code to create the histogram (**Figure 1**):

```
TITLE 'Summary of Weight Variable (in pounds)';
PROC UNIVARIATE DATA = sashelp.class NOPRINT;
   HISTOGRAM weight / NORMAL;
RUN;
```

We can have more than one analysis variable in the HISTOGRAM statement. Each variable will have a separate histogram. NOPRINT option suppress the summary statistics, NORMAL option overlays the normal curve.

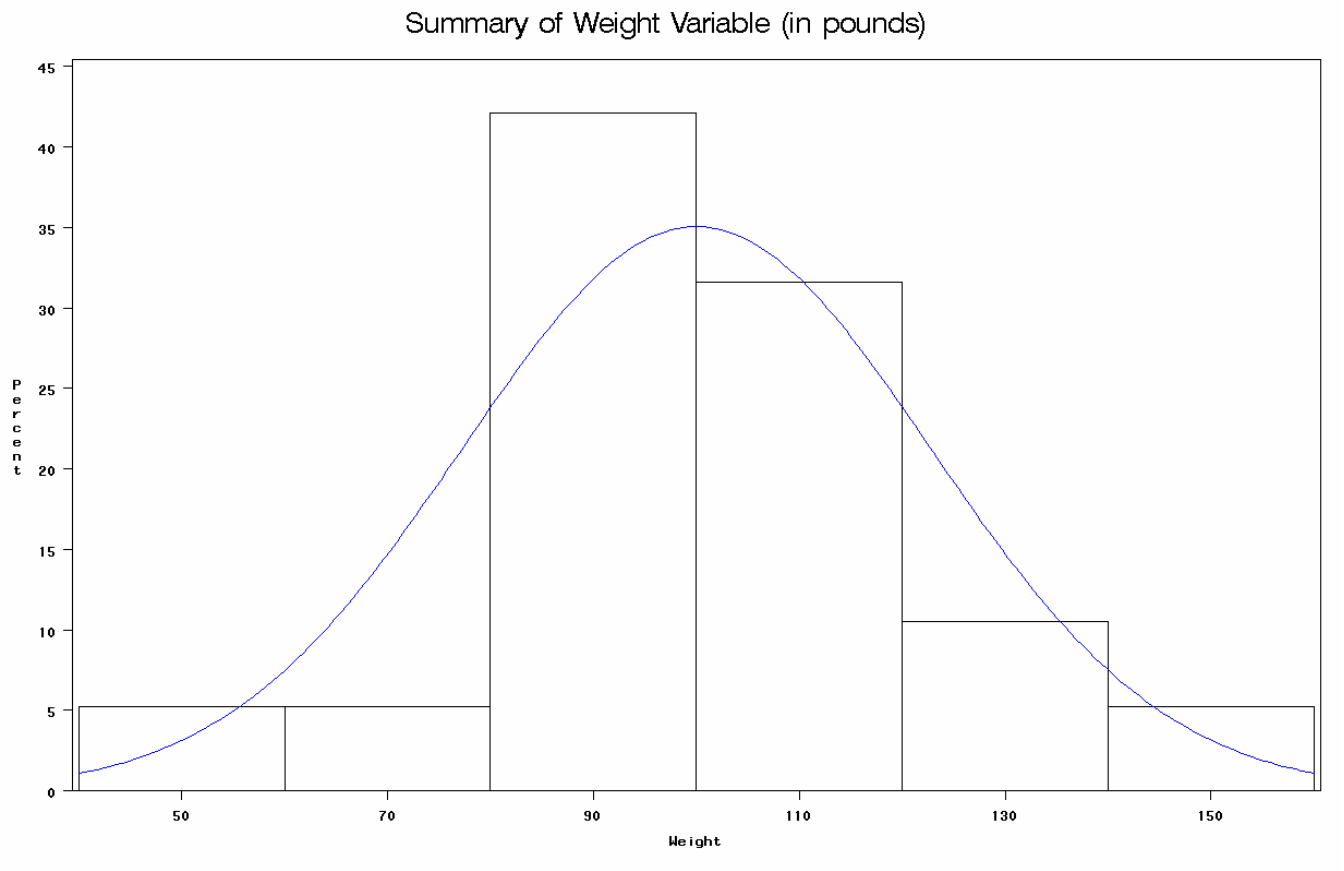

Figure 1

By default the histogram is created in the SAS graphic window. Without NOPRINT option, the summary statistics are created in the SAS output window (or .lst file if run in batch mode). We can use FILENAME statement to direct the output to an external file, and use GOPTIONS statements to enhance text presentation.

```
FILENAME outfig "C:\NESUG2007\histogram.html";

GOPTIONS RESET=all CTITLE=cx000080 FTEXT=swiss ROTATE CTEXT=black
      HTEXT=0.85 HTITLE=2.5 DEVICE=html  GSFNAME=outfig SFMODE=replace;
```

We can add the CFILL option to fill color for the histogram and INSET statement to insert a box of the summary statistics directly in the graph (**Figure 2)**. By default the font of the text in the inset box(es) inside the graph is FONT=SIMPLEX.

```
PROC UNIVARIATE DATA = sashelp.class;
   HISTOGRAM weight / NORMAL CFILL = ltgray;
   INSET N = 'Number of students' MEDIAN (8.2) MEAN (8.2) STD='Standard Deviation' (8.3)
         / POSITION = ne;
RUN;

/*Figure 2 is created in the file 'C:\NESUG2007\histogram.html'. It is saved as a .GIF file, and
is inserted into this paper as a picture.*/
```

Figure 2

To create a comparative histogram, use a **CLASS** statement to specify one or two classification variables.

```
TITLE 'Summary of Weight Variable (in pounds) - added a CLASS statement';
PROC UNIVARIATE DATA = sashelp.class NOPRINT;
   CLASS sex;
   HISTOGRAM weight / NORMAL (COLOR = red)
                      CFILL = ltgray
                      CTEXT = blue;
    INSET N = 'Number of students'  MEDIAN (8.2) MEAN (8.2) STD = 'Standard Deviation' (8.3)
          / POSITION = ne;
    LABEL sex = 'Gender';
RUN;
```

In **Figure 3** below, the normal curve is now changed to red with COLOR = red option. The text were changed to blue with CTEXT = blue option.

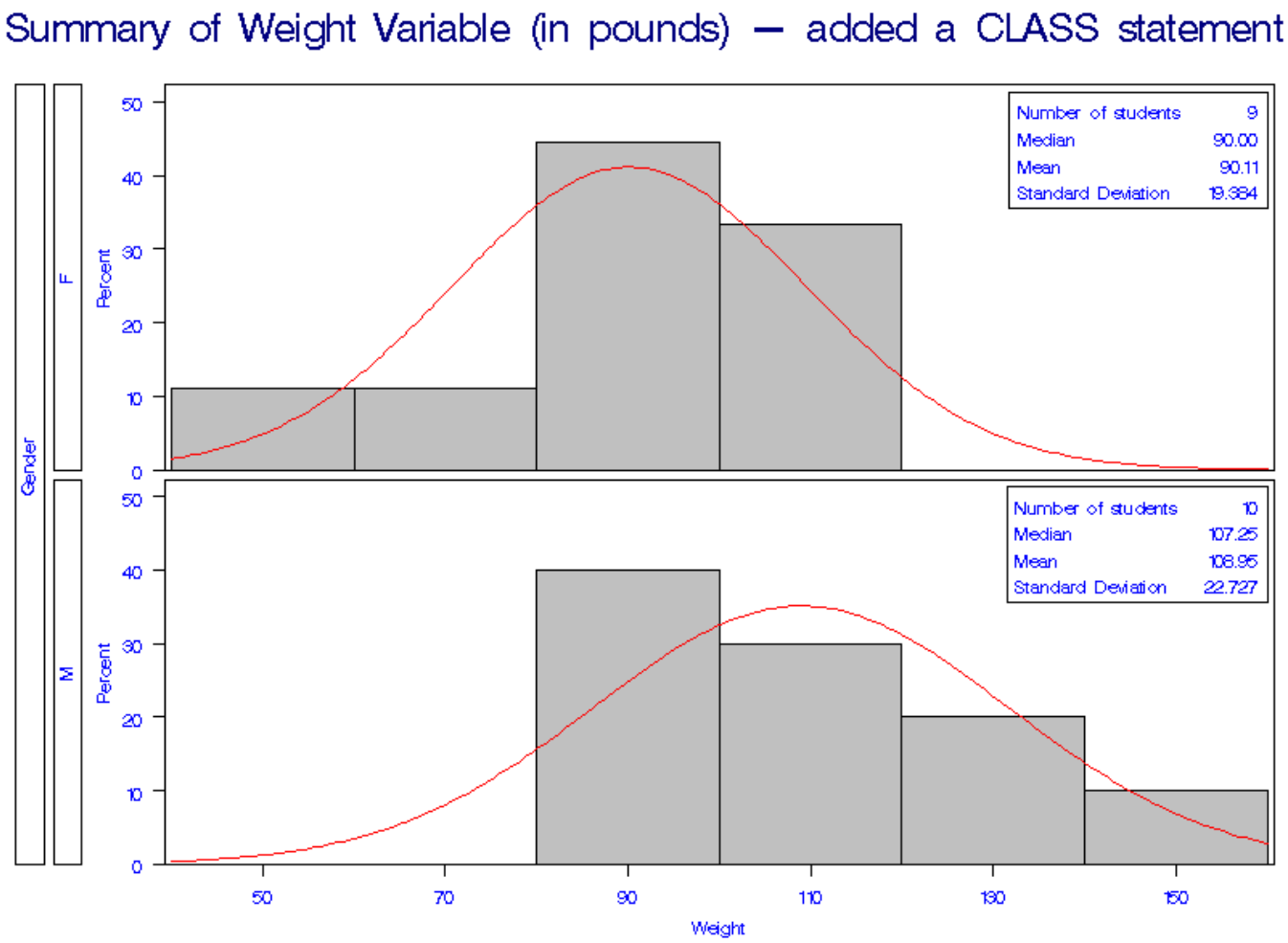

Figure 3

To control the display of the fitted curve, we use the secondary options that are specified in the parentheses after the primary distribution option, e.g.
Codes for **Figure 4:**

```
PROC UNIVARIATE DATA = sashelp.class NOPRINT;
   CLASS sex;
   HISTOGRAM weight / NORMAL (COLOR = red NOPRINT FILL W=3)
                      CFILL = ltgray
                      CTEXT = blue;
    INSET N = 'Number of students'  MEDIAN (8.2) MEAN (8.2) STD = 'Standard Deviation' (8.3)
          / POSITION = ne;
    LABEL sex = 'Gender';
RUN;
```

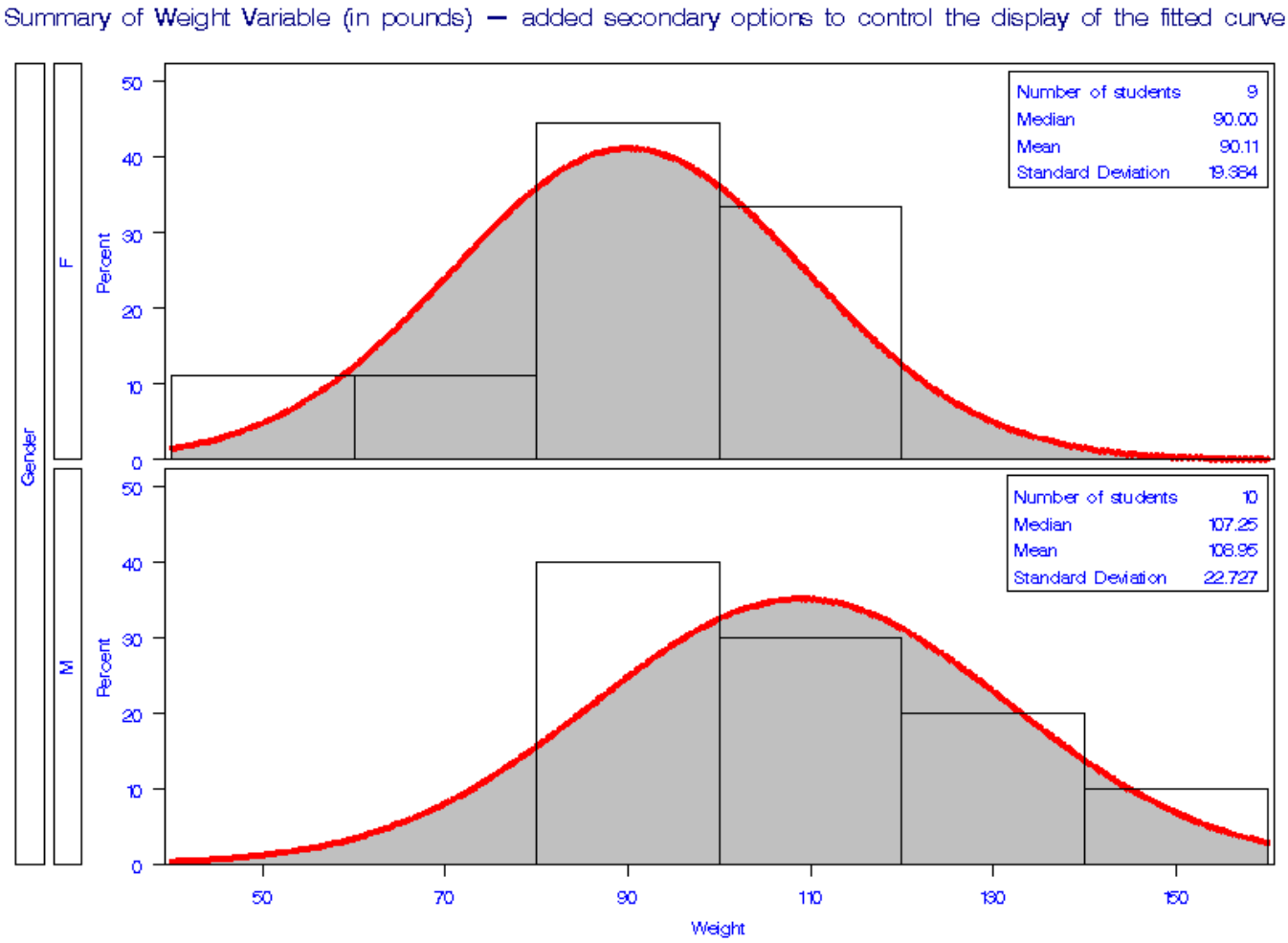

Figure 4

There are many options of the HISTOGRAM statement (primary options) for enhancing histograms. We also have the option for requesting an output data set, e.g. / OUTHISTOGRAM=

Example codes for **Figure 5**: the histogram bars were suppressed with NOBARS option, output dataset **_outhist** contains information on histogram intervals.

```
TITLE 'Summary of Weight - More options for enhancing the histograms';
PROC UNIVARIATE DATA = sashelp.class NOPRINT;
  CLASS sex;
  HISTOGRAM weight / NORMAL (COLOR=green FILL W=5)
                     CFILL = ltgray
                     CTEXT = red
                     NOBARS
                     VAXISLABEL = "Percentage of patients"
                     CGRID = black
                     OUTHISTOGRAM =  _outhist
                     ;
  INSET N = 'Number of students'  MEDIAN (8.2) MEAN (8.2) STD= 'Standard Deviation' (8.3)
        / POSITION = ne;
  LABEL sex = 'Gender';
RUN;
```

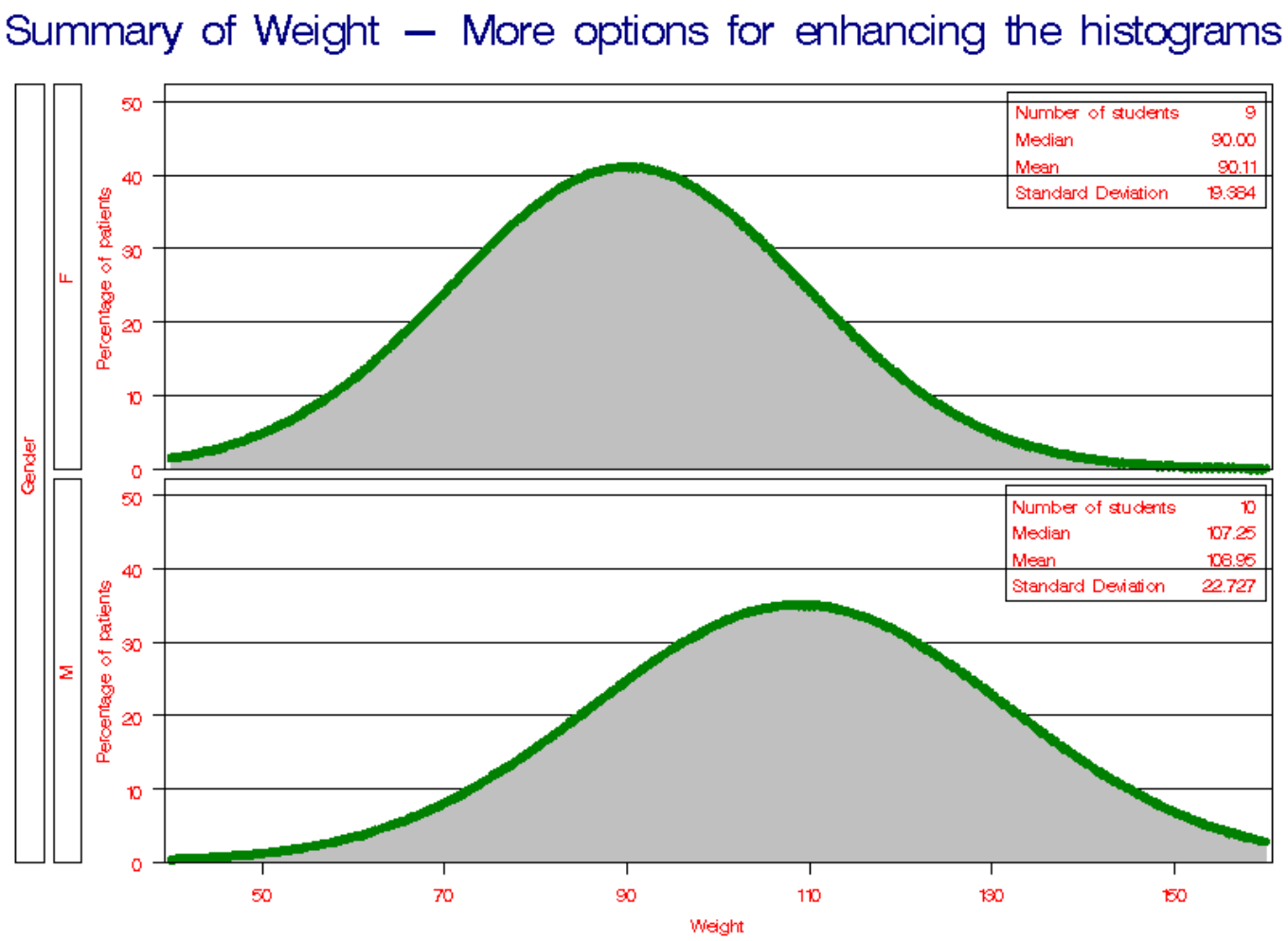

Figure 5

Data in the output dataset _outhist:

_outhist 10

Obs	Sex	_VAR_	_MIDPT_	_OBSPCT_	_CURVE_	_EXPPCT_
1	F	Weight	50	11.1111	NORMAL	5.5297
2	F	Weight	70	11.1111	NORMAL	24.0803
3	F	Weight	90	44.4444	NORMAL	39.4064
4	F	Weight	110	33.3333	NORMAL	24.3426
5	F	Weight	130	0.0000	NORMAL	5.6513
6	F	Weight	150	0.0000	NORMAL	0.4875
7	M	Weight	50	0.0000	NORMAL	1.4420
8	M	Weight	70	0.0000	NORMAL	8.5739
9	M	Weight	90	40.0000	NORMAL	24.5498
10	M	Weight	110	30.0000	NORMAL	33.9724
11	M	Weight	130	20.0000	NORMAL	22.7475
12	M	Weight	150	10.0000	NORMAL	7.3592

CONCLUSIONS

With the HISTOGRAM statement and its options, PROC UNIVARIATE can produce high-resolution figures in an easy way to summarize and visualize the distributions of the numeric variables. We would not need to use additional graphical statements, i.e. PATTERNS, AXIS1, AXIS2, etc. The plots option of PROC UNIVARIATE could also create the basic histograms, box plots and normal plots, however the output is not so good and it can be difficult to read normal plots.

In practice, the box plots are more popular for the reports than the histograms. We hope that PROC UNIVARIATE will have a BOXPLOTS statement in the near future.

REFERENCES:

[1] Wikipedia, the free encyclopedia (http://en.wikipedia.org/wiki/Histogram)

[2] SAS 9.1 Online Documentation (Chapter 3. The UNIVARIATE Procedure)

ACKNOWLEDGMENTS

The author has special thanks to Dr. Shi-Tao Yeh for his encouragement on the paper's topic.

CONTACT INFORMATION

Chauthi Nguyen
GlaxoSmithKline Pharmaceutical, Inc.
2301 Renaissance Boulevard, King of Prussia, PA 19406
Work phone: (610) 787 - 3853
E-mail: Chauthi.Nguyen@gsk.com