Linear Statistical Modeling Methods with SAS

Descriptive Statistics Using SAS

Xuemao Zhang
East Stroudsburg University

January 22, 2024

Introduction

- Before doing any statistical inferences, we must be able to describe the data in a straight-forward, easy-to-comprehend fashion. One way is to display the data graphically and the second way is to show the descriptive summary statistics: sample size, mean, median, minimum, maximum, variance and standard deviation etc.
- These can be done by using the MEANS, UNIVARIATE and PLOT procedures in SAS.

Descriptive Statistics

Example. Consider the data set foot.txt with 6 variables and 40 observations each.

```
data foot;
infile "foot.txt" firstobs=2; /*please change directory*/
input Sex$ Age FootLength ShoePrint ShoeSize Height;
/*Sex is categorical or nominal*/
run;

proc print data= foot;
run;

proc means data=foot;
TITLE "Simple Descriptive Statistics";
run;
```

Descriptive Statistics

- We can specify which statistics we want to compute by specifying options for PROC MEANS.

```
proc means data=foot N MEAN MEDIAN VAR;
run;
```

- We may also wish to specify for which numerical variables in our data set we want to compute descriptive statistics.

```
proc means data=foot N MEAN MEDIAN VAR;
VAR Age FootLength Height;
run;
```

- For options of PROC MEANS, please go to the following link for more information. https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/proc/p0f0fjpjeuco4gn1ri963f683mi4.htm

Descriptive Statistics and plots

- If we would like a more extensive list of statistics, including tests of normality, stem-and-leaf plots, and boxplots, PROC UNIVARIATE is the way to go.

```
proc univariate data=foot plot;
TITLE "More Descriptive Statistics";
VAR Age FootLength Height;
run;
```

Descriptive Statistics and plots

- BOXPLOT can be used to generate a box-plot as well. However, BOXPLOT is used to create sideby- side box-and-whiskers plots of measurements organized in groups. To get a single box plot for a variable you need to create a constant **grouping variable**. For example, to create a box-plot for the variable Height, we

```
data foot;
set foot;
group=1;
run;
```

```
proc boxplot data=foot;
plot Height*group/ BOXSTYLE=SCHEMATIC;
run;
```

- The `BOXSTYLE=SCHEMATIC` option is used to draw a modified box plot that identifies outliers.

Descriptive Statistics and plots

- To request, additionally, a test of normality (Shapiro-Wilk test, Kolmogorov-Smirnov test, Cramer-von Mises test and Anderson-Darling test), a stem-and-leaf plot, and a boxplot, we would add the options NORMAL and PLOT as follows.

```
proc univariate data=foot NORMAL PLOT;
TITLE "More Descriptive Statistics";
VAR Height;
run;
```

Descriptive Statistics and plots

- We can ask PROC UNIVARIATE to produce histograms, QQ (Quantile-Quantile) plots, and probability plots by adding appropriate statements.

```
proc univariate data=foot;
TITLE "Histogram for Height";
histogram Height;
run;
```

```
proc univariate data=foot;
TITLE "QQ-plot for Height";
qqplot Height;
run
```

Descriptive Statistics and plots

- The PROC PLOT (or PROC GPLOT) can be used to generate a scatter plot for us to investigate the relationship between two variables.

```
proc plot data=foot;
TITLE "Scatter-plot of Show Size by Height";
plot ShoeSize*Height;  /*y verus x*/
run;
```

Descriptive Statistics and plots

- PROC SGPLOT is a more powerful procedure for plots
  - https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/grstatproc/p1t32i8511t1gfn17sw07yxtazad.htm
- Box plot of `Height`

```
proc sgplot data=foot;
vbox Height;
run;
```

Descriptive Statistics and plots

- Box plot of `Height` by `Sex`

```
proc sgplot data=foot;
vbox Height /category=Sex;
run;
```

- Horizontal Box plot of `Height` by `Sex`

```
proc sgplot data=foot;
hbox Height /category=Sex;
run;
```

Descriptive Statistics and plots

- Scatter plot of ShoeSize and Height

```
proc sgplot data=foot;
 scatter x=Height  y = ShoeSize;
run;
```

Descriptive Statistics and plots

- Scatter plot of ShoeSize and Height by “Sex

```
proc sgplot data=foot;
 scatter x=Height  y = ShoeSize / group=Sex;
run;
```

License