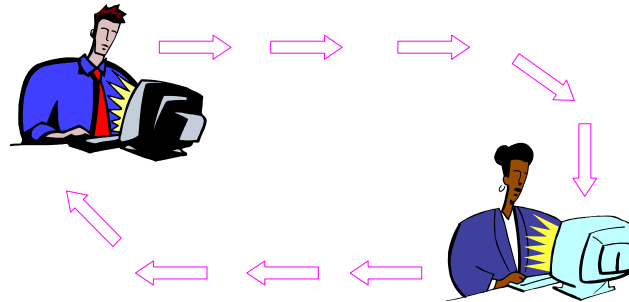


U S I N G O D S W I T H P R O C U N I V A R I A T E

Rebecca G . Frederick
Louisiana State University
Department of Experimental Statistics



August03

South Central SAS Users Group
2003

1

ABSTRACT

- Proc Univariate is used by many statisticians to get a handle on the type of data that is to be analyzed. Regardless of whether I am conducting a Regression, General Linear Model, or a Survival Analysis, I have to investigate the data. Outliers are a natural occurrence in many datasets. A analysis with Proc Univariate should occur first on the data before any other procedure. Hypertext markup language or HTML using the ODS statement permits the results to easily shared with collaborators.

August03

South Central SAS Users Group
2003

2

INTRODUCTION

- Typically, a statistician will be handed data of a large number of observations that has been given by a client. The client has assured the statistician that all of the observations have been checked and re-checked for outliers and all problems have been eliminated. But, no data set has been cleaned-up until the data set has been analyzed through SAS[®] Proc Univariate. Further, suppose that the client is located at another location and so, all information will be emailed between the client and the statistician. This is where the SAS[®] ODS statements will come in handy for use.



August 03

South Central SAS Users Group
2003

3

INTRODUCTION (CONTINUE)

- A quantile is defined as the amount of area under a density curve or the area to the left of specified fraction of the total unit area. For instance, a given value of p is the p th percentile such that the area to the left of it is p .

<u>Q uantile</u>	<u>E stim ate</u>	} <i>The E stim ates are from the data set, Students, and are the Test Scores from a particular section of class.</i>
100% (M ax Q uantile)	94	
95%	94	
75% (T hird Q uantile)	79	
50% (M edian Q uantile)	77	
25% (F irst Q uantile)	58	
0% (M in Q uantile)	52	

August 03

South Central SAS Users Group
2003

4

ADDITIONS TO THE PROC UNIVARIATE : PROBPLOT, & QQPLOTS

- If the data distribution matches the theoretical distribution, the points on the plot form a linear pattern, $y=x$. Thus, you can use a Q-Q plot or a probability plot to determine how well a theoretical distribution models a set of measurements.
- For example,

Graph A : No Linear Relationship Graph B : Linear Relationship



August 03

South Central SAS Users Group
2003

5

ADDITIONS TO THE PROC UNIVARIATE : PROBPLOT, & QQPLOTS (CONTINUE)

- The slope and intercept are visual estimates of the scale and location parameters of the theoretical distribution.
- Q-Q plots are more convenient than probability plots for graphical estimation of the location and scale parameters because the axis of a Q-Q plot is scaled linearly.
- On the other hand, probability plots are more convenient for estimating percentiles or probabilities.

August 03

South Central SAS Users Group
2003

6

PROBPLOT

- Probplot - C reates a probability plot by using high-resolution graphs, which compare ordered variable values with percentiles of a specified theoretical distribution.
- Syntax of the statement:
`PROBPLOT variable(s) / option(s);`
- Probplot statement options can request a distribution (Beta, Exponential, Gamma, LogNormal, Normal, and Weibull) and each of the distribution parameters (Alpha, Beta, C, MU, Sigma, Slope, Theta, and Zeta)

August03

South Central SAS Users Group
2003

7

PROBPLOT (continue)

- The distribution parameters can compute a maximum likelihood estimate by specifying:
`"distribution _parameter=est".`
- Probplot can control the appearance of distribution reference line, general plot layout, enhance the probability plot or comparative plot.
- INSET statement - Places a box or table of summary statistics directly in the high-resolution graphics.

August03

South Central SAS Users Group
2003

8

PROBPLOT (continue)

- PROGRAM:

```
LIBNAME STUDENT 'C:\MYSASDIR\SESSION7';

PROC SORT DATA = STUDENT.STUDENTS OUT=SORTED;
  BY SECTION;
RUN;

GOPTIONS HTITLE=2 HTEXT=1 FTEXT=SWISSB FTITLE=SWISSB;
SYMBOL VALUE=STAR;

PROC UNIVARIATE DATA=SORTED NOPRINT;
  BY SECTION;
  VAR EXAM;
  PROBPLOT EXAM /NORMAL(MU=EST SIGMA=EST);
  INSET MEAN STD / HEADER='Normal Parameters'
        POSITION=(95,5) REFPOINT=BR;
  TITLE1 '100 Obs Sampled from a Normal Distribution';
  TITLE2 'Normal Probability Plot';
RUN;
```

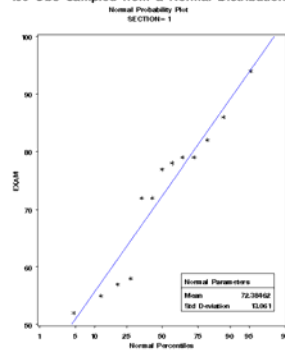
August 03

South Central SAS Users Group
2003

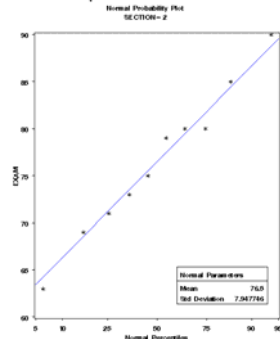
9

PROBPLOT OUTPUT: (continue)

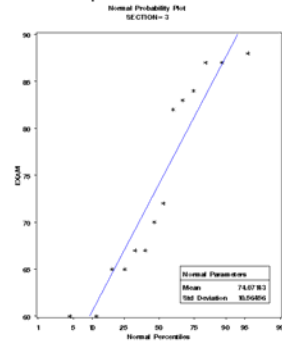
100 Obs Sampled from a Normal Distribution



100 Obs Sampled from a Normal Distribution



100 Obs Sampled from a Normal Distribution



August 03

South Central SAS Users Group
2003

10

QQPLOT

- Q Q plot - Creates a quantile-quantile plot by using high-resolution graphs, which compare ordered variable values with quantiles of a specified theoretical distribution.
- Q Q plot statement options can request a distribution (Beta, Exponential, Gamma, LogNormal, Normal, and Weibull) and each of the distribution's parameters (Alpha, Beta, C, Mu, Sigma, Slope, Theta, and Zeta).
- The distribution parameters can compute a maximum likelihood estimate by specifying:
"distribution _parameter=est".
- Q Q plot can control the appearance of distribution reference line, general plot layout, enhance the probability plot or comparative plot.

August 03

South Central SAS Users Group
2003

11

QQPLOT (continue)

- PROGRAM :

```
LIBNAME STUDENT 'C:\MYSASDIR\SESSION7';

PROC SORT DATA = STUDENT.STUDENTS OUT=SORTED;
  BY SECTION;
RUN;

GOPTIONS HTITLE=2 HTEXT=1 FTEXT=SWISSB FTITLE=SWISSB;
SYMBOL VALUE=STAR;

PROC UNIVARIATE DATA=SORTED NOPRINT;
  BY SECTION;
  VAR EXAM;
  QQPLOT EXAM /NORMAL(MU=EST SIGMA=EST);
  INSET MEAN STD / HEADER='Normal Parameters'
          POSITION=(95,5) REFPOINT=BR;
  TITLE1 '100 Obs Sampled from a Normal Distribution';
  TITLE2 'Normal Probability Plot';
RUN;
```

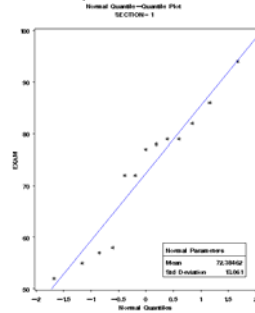
August 03

South Central SAS Users Group
2003

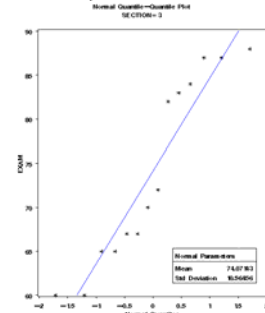
12

QQ PLOT OUTPUT (continue)

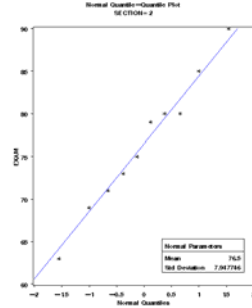
100 Obs Sampled from a Normal Distribution



100 Obs Sampled from a Normal Distribution



100 Obs Sampled from a Normal Distribution



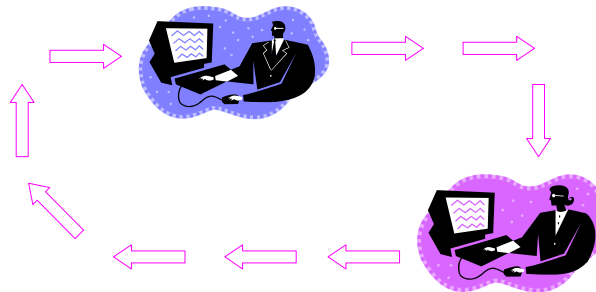
August 03

South Central SAS Users Group
2003

13

OUTPUT DELIVERY SYSTEM (ODS)

- ODS - User can combine raw data with one or more table definitions to produce output to a printer or formatted in Hypertext Markup Language (HTML).



August 03

South Central SAS Users Group
2003

14

ODS (continue)

- ODS breaks down the procedures into separate pieces so that the user can print out only sections of the report. ODS, in version 9, can currently support many destinations but here are at least four destinations:
 - The Output destination produces SAS dataset.
 - The Listing destination produces monospace output, which is formatted like traditional SAS procedure output.
 - The HTML destination produces output that is formatted in Hypertext Markup Language.
 - The Printer destination produces output that is formatted for high-resolution printers.

August 03

South Central SAS Users Group
2003

15

ODS (continue)

•PROGRAM :

```
LIBNAME STUDENT 'C:\MYSASDIR\SESSION7';

PROC SORT DATA = STUDENT.STUDENTS OUT=SORTED;
  BY SECTION;
RUN;

OPTIONS HTITLE=2 HTEXT=1 FTEXT=swissb FTITLE=SWISSB;
SYMBOL VALUE=STAR;

ODS HTML FILE='ODSHTML_BODY.HTM'
          CONTENTS='ODSHTML_CONTENTS.HTM'
          PAGE='ODSHTML_PAGE.HTM'
          FRAME='ODSHTML_FRAME.HTM';
```

August 03

South Central SAS Users Group
2003

16

ODS (continue)

•PROGRAM :

```
PROC UNIVARIATE DATA=SORTED NOPRINT;
  BY SECTION;
  VAR EXAM;
  QQPLOT EXAM /NORMAL(MU=EST SIGMA=EST);
  INSET MEAN STD / HEADER='Normal Parameters'
        POSITION=(95,5) REFPOINT=BR;
  TITLE1 '100 Obs Sampled from a Normal Distribution';
  TITLE2 'Normal Quantile-Quantile Plot';
RUN;

ODS HTML CLOSE;
```

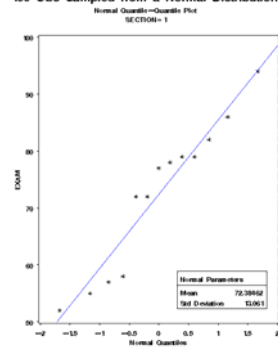
August03

South Central SAS Users Group
2003

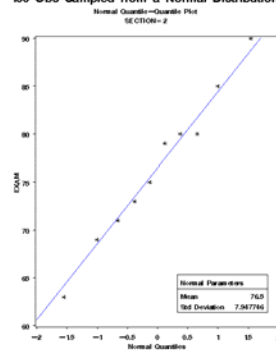
17

ODS OUTPUT (continue)

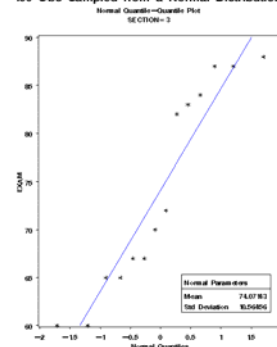
100 Obs Sampled from a Normal Distribution



100 Obs Sampled from a Normal Distribution



100 Obs Sampled from a Normal Distribution



August03

South Central SAS Users Group
2003

18

CONCLUSION :

- The Output Delivery System (ODS) is an extremely powerful tool that statisticians can use to communicate results of analyses with clients. High-resolution graphics of the probability plots and Q-Q plots display theoretical distributions relative to the actual data. Performance of these is displayed through a very simple procedure named, SAS[®] Proc Univariate.



August03

South Central SAS Users Group
2003

19

REFERENCES:

- Meeker, William Q. and Luis A. Escobar. (1998), Statistical Methods For Reliability Data, New York: John Wiley & Sons.
- SAS Institute, Inc (2000). SAS OnlineDoc, Version 8. Cary, NC : SAS Institute, Inc.
- SAS Institute, Inc (2003). SAS Help and Documentation, Version 9. Cary, NC : SAS Institute, Inc.

August03

South Central SAS Users Group
2003

20

TRADE MARKS:

- SAS is a registered trademark or a trademark of SAS Institute Inc. in the USA and other countries. ® Indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.

August03

South Central SAS Users Group
2003

21

CONTACT INFORMATION:

Your comments and questions are valued and encouraged.
Contact the author at:

Rebecca G. Frederick
Louisiana State University Agricultural Center
Department of Experimental Statistics
Baton Rouge, LA 70803-5606
Work Phone: 225-578-8303
Fax: 225-578-8344
Email: rfrederi@lsu.edu
<http://www.stat.lsu.edu/faculty/frederick>

August03

South Central SAS Users Group
2003

22