Elementary Statistics Using Base SAS®

Deborah Babcock Buck, D. B. & P. Associates, Houston, TX

ABSTRACT

Base SAS includes a number of procedures that will allow you to perform elementary statistical analyses. Most of these PROCs produce descriptive statistics, but there are some capabilities for inferential statistics (hypothesis testing) as well. This paper will discuss which Base SAS PROCs will provide you with the desired statistics and show examples of code and output for a research study. The output shown was generated using SAS Version 8, however, the code is also valid for SAS Version 6.12.

INTRODUCTION

Almost all projects utilizing SAS require some form of summary information or statistical analyses. These may include complicated analyses of variance or regression techniques, or simple frequency counts to verify information in the data set. This paper will focus on the elementary statistics available in Base SAS. Which PROC is most appropriate for a given situation depends upon the type of variables to be analyzed and the desired statistics. Some of the most commonly used PROCs in Base SAS for generating statistics are the following.

- UNIVARIATE – produces descriptive and/or inferential statistics on measures of location, variability, and distribution, as well as percentiles, frequency counts and plots
- MEANS – produces descriptive and some inferential statistics across all observations and within subgroups
- CORR – produces descriptive statistics and correlation coefficients
- FREQ – produces one to n-way frequency tables including counts and percentages, as well as tests and measures of association and agreement such as Chi-square, Fisher's Exact and Cochran-Mantel-Haenszel tests

Additional procedures frequently used in Base SAS include PROC SUMMARY, which has the same functionality as PROC MEANS except that the default is no printed output. PROC TABULATE and REPORT can also provide descriptive statistics arranged in a customized form for report writing. PROC CHART and PLOT are available to view the data graphically. This paper will focus on the procedures listed in the bullet points above.

RESEARCH STUDY

In order to demonstrate potential uses for Base SAS PROCs in analyzing the data from a study, we will examine the objectives and analyses of a small research study.

A group of researchers in education have developed a new teaching method that they believe will improve learning skills in math. Therefore, they decide to conduct a study to measure and compare change in math scores using their new method and the standard method. The developers would like to show that their new method results in a significant improvement from baseline in the math test score. Also, they would like to demonstrate a greater improvement in the new method than in the standard method. However, since they are unsure of the potential results, all tests will be two-sided.

These researchers only have access to Base SAS, therefore some frequently used inferential statistics, such as analysis of variance, are not available to them.

Previous studies have shown that girls and boys respond differently to various teaching methods, so gender is also recorded in the data. Data for 30 students, 15 on each teaching method, are collected and entered into a SAS data set called SCHOOL. This original data set includes the variables ID, STYLE (teaching method), GENDER, PRE and POST, where PRE and POST are the pre-teaching method math scores and post-teaching method scores, respectively.

The researchers would also like to see if pre-score and change in score are related. In addition, they would like to show that a larger proportion of students on the new method have a greater than 5% improvement from baseline. Therefore, three new variables are added to the original data set using the following code.

```
DATA METHOD;
 SET SCHOOL;
  CHANGE=POST-PRE;
  PCHANGE=ROUND(((CHANGE/PRE)*100),1);
  IF PCHANGE GT 5 THEN GT5='YES';
   ELSE IF PCHANGE NE . AND PCHANGE LE 5
        THEN GT5='NO';
  LABEL CHANGE='Change in Math Score'
        PCHANGE='Percent Change in Math Score'
        GT5='Change Greater than 5%';
RUN;
```

To summarize, the goals for this project are to demonstrate or examine whether:

- the new teaching method results in a statistically significant change from baseline in math test score,
- the change from baseline using the new method is significantly different from that using the standard method,
- male and female students respond in the same way to the two teaching methods,
- the change in score is related to the math pre-score,
- a larger proportion of students on the new method experience an improvement of greater than 5%.

USING PROC UNIVARIATE

The first thing the researchers want to do is to examine the distribution of the variables PRE and CHANGE to see whether these variables have a normal distribution so that parametric test assumptions are met. They decide to use PROC UNIVARIATE because UNIVARIATE will provide tests of normality when the

NORMAL option is included on the PROC statement. The following code is used to produce the output shown in Appendix Table 1.

```
TITLE1 'APPENDIX TABLE 1';
TITLE2 'PROC UNIVARIATE OUTPUT';
PROC UNIVARIATE DATA=METHOD NORMAL;
 VAR PRE CHANGE;
RUN;
```

The output for the CHANGE variable is shown in Appendix Table 1. This output contains a number of useful pieces of information. UNIVARIATE produces descriptive and inferential statistics on measures of central tendency, including the mean, median and mode. It also generates measures of variability, percentiles and extreme values. The tests of normality, requested in the NORMAL option, indicate that the CHANGE variable can be examined using parametric methods.

The overall mean change from baseline in math score for all students is 1.5, which is statistically significantly different from 0, as is indicated in the probability level shown with the results of the Student's t-test (testing H_O.: μ=0). UNIVARIATE also produces the 5 lowest and highest values by default, which is helpful in checking for possible outliers in the data. Additional useful options available in UNIVARIATE are FREQ which produces a frequency listing of the variable's values, and PLOT which produces a frequency plot (histogram).

USING PROC MEANS

Although UNIVARIATE provides many of the same statistics as MEANS, PROC MEANS will easily allow a limited number of statistics to be printed in the report. To determine whether the change in test score differs from zero in each of the teaching methods, the researchers then decide to use PROC MEANS with a CLASS statement. The data does not need to be sorted before the CLASS statement is included. In this study, the researchers are interested in seeing the mean change in each teaching method group, the standard deviation, the 95% confidence intervals around the mean change, and the t-statistic and probability associated with the t test of whether the mean is significantly different from zero.

By default, PROC MEANS generates the number of observations, mean, standard deviation, minimum and maximum for each numeric variable. Specific statistics can be requested as options on the PROC MEANS statement. Because the researchers only want the mean, standard deviation, 95% confidence limits, t value and probability associated with the t, they specify those statistics in the MEANS statement. The CLASS statement specifies whether and how PROC MEANS should group the summary statistics into subgroups. Since this analysis should only include the desired statistics for each teaching method, the CLASS statement specifies this. A VAR statement will control for which variables statistics are produced – in this case, CHANGE. The MAXDEC= option limits the number of decimals to 2 for each of the statistics except the probability.

The following code will produce the desired summary statistics for each teaching method, as well as generate two new data sets.

```
TITLE1 'APPENDIX TABLE 2';
TITLE2 'PROC MEANS OUTPUT';
TITLE3 'FOR EACH TEACHING METHOD';
PROC MEANS DATA=METHOD
     MEAN STD CLM T PRT MAXDEC=2;
 CLASS STYLE;
 VAR CHANGE;
 OUTPUT OUT=NEW (WHERE=(STYLE='NEW'))
        N=NNUM
        MEAN=NMEAN
        VAR=NVAR;
 OUTPUT OUT=OLD (WHERE=(STYLE='OLD'))
        N=ONUM
        MEAN=OMEAN
        VAR=OVAR;
RUN;
```

Output from Appendix Table 2 shows that the mean change from pre-test in math test score for the new teaching method is significantly different from zero (p<.0001) with a mean change of 2.67. The mean change for the standard method, on the other hand, does not different statistically significantly from zero (p=.6092) with a mean change of only .33. Therefore, the researchers would like to be able to claim that their new method performs significantly better than the standard method.

They would like to show that there is a significant difference between teaching methods. PROC TTEST (in SAS/STAT) would easily provide them with the t-test for difference in means for two independent samples (H_O: $\mu_1 = \mu_2$). However, since the researchers do not have access to SAS/STAT, they need to find an alternate method to conduct a two-sided t-test. One of the researchers suggests that they write a SAS DATA step program, using the output from the PROC MEANS and a probability function for the t test (PROBT) to conduct the t-test for difference between teaching method means.

The code for PROC MEANS shown above produces two new data sets which include the number of observations, mean, and variance for the two teaching methods separately. Multiple OUTPUT statements in the MEANS produce the desired statistics by utilizing the WHERE = option to specify the appropriate teaching method statistics for each new data set. That information is then merged into the data set TTEST which contains the code to calculate the t statistic and associated probability.

The formulae for calculating the t value comparing the two means is as follows:

$$d = Mean_1 - Mean_2;$$
$$s_d = \sqrt{((Variance_1/n_1) + (Variance_2/n_2))}$$
$$t = d/s_d$$

The following code with PROC PRINT produces the t-test information which compares the change in test score for the two teaching methods.

```
DATA TTEST (DROP=STYLE _TYPE_ _FREQ_);
 MERGE NEW OLD;
  DF=(NNUM + ONUM)-2;
  DMEAN=NMEAN-OMEAN;
  DSE=SQRT((NVAR/NNUM)+(OVAR/ONUM));
  T=DMEAN/DSE;
  P=(1-PROBT(ABS(T),DF))*2;
  P=ROUND(P,.0001);
RUN;

TITLE1 'APPENDIX TABLE 3';
TITLE2 'PROC PRINT OUTPUT';
TITLE3 'T-TEST COMPARING TEACHING METHODS';

 LABEL NNUM='No. Students in New Method'
       NMEAN='Mean for New Method'
       NVAR='Variance for New Method'
       ONUM='No. Students in Old Method'
       OMEAN='Mean for Old Method'
       OVAR='Variance for Old Method'
       DF='Degrees of Freedom'
       DMEAN='Difference in Means'
```

```
    DSE='Standard Deviation for Difference'
    T='T-value'
    P='Probability > T';
RUN;
```

The output in Appendix Table 3 shows that the t-test indicates that the new teaching method is significantly better than the standard method in improving the math test score, based on the p-value of .0081.

The researchers are also interested in testing whether the mean change from baseline is significantly different from zero within each gender in each teaching method group. An additional PROC MEANS with the CLASS statement including both STYLE and GENDER will supply this information. Again, the researchers want the mean, standard deviation, 95% confidence limits, t and probability of t. However, they find that the additional class variable causes the output to span more than one line. Therefore, they add the FW= option to limit the width of each field to 8 instead of the default of 12.

```
TITLE1 'APPENDIX TABLE 4';
TITLE2 'PROC MEANS OUTPUT';
TITLE3 'FOR EACH GENDER-TEACHING METHOD';
PROC MEANS DATA=METHOD
    MEAN STD CLM T PROBT MAXDEC=2 FW=8;
 CLASS STYLE GENDER;
 VAR CHANGE;
RUN;
```

Based on the results shown in Appendix Table 4, it does appear that males and females may react differently to the two teaching methods.

For the new teaching method, both genders show an improvement in math score. At first look, it appears that males have a greater improvement in math score. However, curiously, that mean does not show a significant difference from zero at the alpha=.05 level (p=.0643), while the female mean of 2.5 does show a significant difference from zero with a p-value of .0013. Closer examination of the data shows that the male data has slightly greater variability among the data points, but a major factor influencing the significance level is the much smaller sample size in the male group. These results point up the importance of having a sufficient sample size (and the associated power with a larger sample size) to detect a difference where a true difference exists.

For the old teaching method, males show an improvement in math scores of 1.5 (which is approaching significance with p=.0636), while females show a decrease of –1.0 in score. This would tend to support the theory that males and females perform differently on the two teaching methods. However, the appropriate way to analyze this data would be to conduct an analysis of variance on the change in math scores with the main effects of STYLE and GENDER and the interaction term of STYLE by GENDER. SAS/STAT will provide those capabilities.

USING PROC CORR

Another objective of this study was to examine whether there was a relationship between pre-teaching method math score and change in math score. To conduct this analysis the researchers decide to use PROC CORR to generate a correlation analysis.

PROC CORR provides descriptive statistics for all variables specified in VAR or WITH statements in addition to some measures of association, including correlation coefficients. Because the distribution of the pre-teaching method test score may not be normally distributed (based on the output from PROC UNIVARIATE), it was decided to include both the Pearson correlation coefficient and the Spearman rank correlation to test H_0: r=0. The following code generates the correlation coefficients for pre-score with change in score over both teaching methods.

```
TITLE1 'APPENDIX TABLE 5';
TITLE2 'PROC CORR OUTPUT';
TITLE3 'ALL STUDENTS';
PROC CORR DATA=METHOD PEARSON SPEARMAN;
 VAR PRE CHANGE;
RUN;
```

The output from Appendix Table 5 shows a very weak negative correlation between pre-teaching method math score and change in score, whether using parametric or nonparametric correlation techniques.

The researchers then suggest that perhaps the correlation differs between the two teaching methods, so the correlation analysis is conducted separately for each teaching method. The following code generates the correlation analysis for each teaching method separately. (Note that the data must be sorted by STYLE in order to include the BY statement with the PROC CORR.)

```
PROC SORT DATA=METHOD;
 BY STYLE;


TITLE1 'APPENDIX TABLE 6';
TITLE2 'PROC CORR OUTPUT';
TITLE3 'FOR EACH TEACHING METHOD';
PROC CORR DATA=METHOD;
 BY STYLE;
 VAR PRE CHANGE;
RUN;
```

Output from Appendix Table 6 shows an almost zero correlation between the pre-score and change on the new teaching method. A weak negative correlation exists in the standard teaching method.

USING PROC FREQ

The researchers would like to market their new teaching method by claiming that the new method has a significantly higher proportion of students who experience a greater than 5% improvement in math scores using their technique. PROC FREQ is used to conduct the analysis since FREQ will provide inferential statistics on the association between variables in two-way tables.

The CHISQ option on the TABLES statement is included to examine the null hypothesis of no association between the STYLE variable and the GT5 variable (which dichotomizes the percent change in math score into more than 5% improvement versus 5% or less improvement). The following code produces the output for Appendix Table 7.

```
TITLE1 'APPENDIX TABLE 7';
TITLE2 'PROC FREQ OUTPUT';
TITLE3 'METHOD BY GREATER THAN 5% IMPROVEMENT';

TABLES STYLE*GT5/CHISQ;
RUN;
```

Output from Appendix Table 7 shows the number and percent of students in each STYLE by GT5 variable combination. It produces the Chi-Square statistics, as well as a number of other measures of association. In examining the output, the researchers note that FREQ gives a warning message about the small cell sizes. Therefore, they decide that the two-sided Fisher's Exact Test is a more appropriate test to evaluate the relationship.

Examination of the results from the two-sided Fisher's test show that the probability level does not indicate that the new training method has a statistically significantly greater number of students with a greater than 5% improvement in math test score (p=.169). However, there is a trend for the new training method to show a greater proportion of students with a greater than 5% improvement (33% on the new method versus 7% on the standard.) Once again, a larger sample size might have resulted in finding a statistically significant difference.

CONCLUSIONS FROM THE RESEARCH STUDY

Using Base SAS PROCs, the researchers are able to come to the following conclusions from their study.

- The new teaching method results in a statistically significant increase from baseline in math test score, while the standard method does not.
- The change from baseline using the new method is significantly better than that using the standard method.
- Male and female students do not appear to respond in the same way to the two teaching methods. However, larger sample sizes and /or more advanced statistical analyses are necessary to verify this hypothesis.
- The change from baseline math score does not appear to be related to the math test pre-score, or if it is, they are only weakly negatively related.
- A larger proportion of students on the new method did experience an improvement of greater than 5%. However, this was not statistically significant. A study with a larger sample size might provide this result.

OTHER BASE PROCS FOR ANALYSES

PROC TABULATE and PROC REPORT are both frequently used for reporting descriptive statistics. They both incorporate some of the functionality of PROC MEANS, FREQ and UNIVARIATE. PROC REPORT also includes some DATA step capabilities. Both of the procedures allow more flexibility in customizing the "look" of a report with descriptive statistics, and thus are powerful report-writing tools.

For a pictorial look at your data, PROC CHART AND PROC PLOT will produce basic graphics. PROC CHART provides bar charts, block charts, and pie charts. PROC PLOT produces scatter plots. These procedures can also assist in data checking. The SAS/GRAPH product provides high resolution graphics for presentations.

An additional procedure in Base SAS that can be used to generate statistics is PROC SQL. This procedure is not covered in this paper due to time limitations, but can perform many of the DATA step and PROC step functions and should also be considered for producing descriptive statistics.

CONCLUSION

This paper has attempted to provide some useful information on generating elementary descriptive and inferential statistics using Base SAS procedures. This paper utilized SAS programming under Version 8. However, the approaches covered in this paper are all valid for Version 6.12. Version 8 offers a number of additional capabilities, including long variables names, the Enhanced Editor and the ODS (Output Delivery System) features which were not addressed in this paper . Hopefully, this paper will help guide you in how to analyze your data in a logical and organized manner using Base SAS.

REFERENCES

SAS Institute Inc. (1990) *SAS Procedures Guide, Version 6, Third Edition*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1990) *SAS Language and Procedures Guide, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1999) *"SAS Procedures", SAS Version 8 Online Documentation*, Cary, NC: SAS Institute Inc.

AUTHOR CONTACT INFORMATION

Debbie Buck
D. B. & P. Associates
10418 Indian Paintbrush Lane
Houston, TX 77095
Voice: 281-256-1619
Fax: 281-256-1634
Email: debbiebuck@houston.rr.com

APPENDIX TABLE 1
PROC UNIVARIATE OUTPUT

The UNIVARIATE Procedure
Variable: CHANGE (Change in Math Score)

Moments

N	30	Sum Weights	30
Mean	1.5	Sum Observations	45
Std Deviation	2.50172354	Variance	6.25862069
Skewness	-0.6583419	Kurtosis	0.22869676
Uncorrected SS	249	Corrected SS	181.5
Coeff Variation	166.78157	Std Error Mean	0.45675014

Basic Statistical Measures

Location		Variability	
Mean	1.500000	Std Deviation	2.50172
Median	2.000000	Variance	6.25862
Mode	4.000000	Range	11.00000
		Interquartile Range	4.00000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	3.284071	Pr > |t|	0.0027
Sign	M	7.5	Pr >= |M|	0.0059
Signed Rank	S	118	Pr >= |S|	0.0025

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.947526	Pr < W	0.1451
Kolmogorov-Smirnov	D	0.125609	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.092671	Pr > W-Sq	0.1376
Anderson-Darling	A-Sq	0.597485	Pr > A-Sq	0.1122

Quantiles (Definition 5)

Quantile	Estimate
100% Max	6
99%	6
95%	4
90%	4
75% Q3	4
50% Median	2
25% Q1	0
10%	-2
5%	-3
1%	-5
0% Min	-5

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
-5	12	4	22
-3	14	4	23
-2	10	4	24
-2	8	4	28
-1	21	6	17

APPENDIX TABLE 2
PROC MEANS OUTPUT
FOR EACH TEACHING METHOD

The MEANS Procedure

Analysis Variable : CHANGE Change in Math Score

Teaching Method	N Obs	Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean	t Value	Pr > |t|
NEW	15	2.67	1.99	1.57	3.77	5.19	0.0001
OLD	15	0.33	2.47	-1.03	1.70	0.52	0.6092

APPENDIX TABLE 3
PROC PRINT OUTPUT
T-TEST COMPARING TEACHING METHODS

No. Students in New Method	Mean for New Method	Variance for New Method	No. Students in Old Method	Mean for Old Method	Variance for Old Method
15	2.66667	3.95238	15	0.33333	6.09524

Degrees of Freedom	Difference in Means	Standard Deviation for Difference	T-value	Probability > T
28	2.33333	0.81844	2.85096	.0081

APPENDIX TABLE 4
PROC MEANS OUTPUT
FOR EACH GENDER-TEACHING METHOD

The MEANS Procedure

Analysis Variable : CHANGE Change in Math Score

Teaching Method	Gender	N Obs	Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean	t Value	Pr > |t|
NEW	F	10	2.50	1.72	1.27	3.73	4.61	0.0013
	M	5	3.00	2.65	-0.29	6.29	2.54	0.0643
OLD	F	7	-1.00	2.45	-3.27	1.27	-1.08	0.3216
	M	8	1.50	1.93	-0.11	3.11	2.20	0.0636

APPENDIX TABLE 5
PROC CORR OUTPUT
ALL STUDENTS

The CORR Procedure

2 Variables: PRE CHANGE

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum	Label
PRE	30	83.50000	8.37793	84.50000	64.00000	95.00000	Pre-test Math Score
CHANGE	30	1.50000	2.50172	2.00000	-5.00000	6.00000	Change in Math Score

Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	PRE	CHANGE
PRE Pre-test Math Score	1.00000	-0.20812 0.2698
CHANGE Change in Math Score	-0.20812 0.2698	1.00000

Spearman Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	PRE	CHANGE
PRE Pre-test Math Score	1.00000	-0.13892 0.4641
CHANGE Change in Math Score	-0.13892 0.4641	1.00000

APPENDIX TABLE 6
PROC CORR OUTPUT
FOR EACH TEACHING METHOD

---------- Teaching Method=NEW ----------

The CORR Procedure

2 Variables: PRE CHANGE

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
PRE	15	83.00000	8.72599	1245	64.00000	94.00000	Pre-test Math Score
CHANGE	15	2.66667	1.98806	40.00000	-1.00000	6.00000	Change in Math Score

Pearson Correlation Coefficients, N = 15
Prob > |r| under HO: Rho=0

	PRE	CHANGE
PRE Pre-test Math Score	1.00000	-0.04117 0.8842
CHANGE Change in Math Score	-0.04117 0.8842	1.00000

---------- Teaching Method=OLD ----------

The CORR Procedure

2 Variables: PRE CHANGE

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
PRE	15	84.00000	8.28941	1260	68.00000	95.00000	Pre-test Math Score
CHANGE	15	0.33333	2.46885	5.00000	-5.00000	4.00000	Change in Math Score

Pearson Correlation Coefficients, N = 15
Prob > |r| under HO: Rho=0

	PRE	CHANGE
PRE Pre-test Math Score	1.00000	-0.34553 0.2072
CHANGE Change in Math Score	-0.34553 0.2072	1.00000

APPENDIX TABLE 7
PROC FREQ OUTPUT
METHOD BY GREATER THAN 5% IMPROVEMENT

The FREQ Procedure

Table of STYLE by GT5

STYLE(Teaching Method)
GT5(Change Greater than 5%)

Frequency Percent Row Pct Col Pct	NO	YES	Total
NEW	10	5	15
	33.33	16.67	50.00
	66.67	33.33	
	41.67	83.33	
OLD	14	1	15
	46.67	3.33	50.00
	93.33	6.67	
	58.33	16.67	
Total	24	6	30
	80.00	20.00	100.00

Statistics for Table of STYLE by GT5

Statistic	DF	Value	Prob
Chi-Square	1	3.3333	0.0679
Likelihood Ratio Chi-Square	1	3.5808	0.0585
Continuity Adj. Chi-Square	1	1.8750	0.1709
Mantel-Haenszel Chi-Square	1	3.2222	0.0726
Phi Coefficient		-0.3333	
Contingency Coefficient		0.3162	
Cramer's V		-0.3333	

WARNING: 50% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

Fisher's Exact Test

Cell (1,1) Frequency (F)	10
Left-sided Pr <= F	0.0843
Right-sided Pr >= F	0.9916
Table Probability (P)	0.0759
Two-sided Pr <= P	0.1686