

USING SAS STATISTICAL PROCEDURES IN MEDICAL RESEARCH

Joe Cholewczynski Murray Electronics

The SASTM system provides statistical analysis procedures that are very useful for decision making in medical research.

This paper is a guide on when to use a statistical procedure with examples from medical research. Emphasis has been placed on the appropriate SAS procedure and options rather than statistical inference of the examples. If the reader is interested in learning more about statistical inference with regard to a specific procedure, I recommend that the reader look at the references listed in the SAS/STATTM Guide for Personal Computers or the SAS Procedures Guide for the procedure in question. It is also recommended to have a good intermediate statistics text book handy as part of your reference library.

Statistical procedures are used to analyze data sets from the following areas of medical research: Clinical trials, or the analysis of patients being treated for a specific disease; vital statistics, or the analysis of birth, death, disease, and mortality rates; pharmacokinetics, or the analysis of the absorption, metabolism, and excretion of medicines on healthy normal patients; and biological experiments, or the analysis of experiments performed with animals or humans.

The SAS statistical procedures can perform the following analyses: descriptive statistics, analysis of variance, categorical analysis, regression analysis, nonparametric statistics, and survival analysis.

Descriptive statistics are methods used to describe and summarize data. The best SAS procedure to use for descriptive statistics is PROC UNIVARIATE. This procedure and its options can give measures of central tendency (mean, median, mode), measures of dispersion (range, standard deviation, coefficient of variation), frequency tables, stem-and-leaf plots, box plots, a test to check if data are from a normal distribution, and the five highest and five lowest values for the data. Other SAS procedures that can give descriptive statistics are PROC MEANS and PROC SUMMARY. These two procedures are not as comprehensive as PROC UNIVARIATE, but are useful in situations where you only require a limited amount of descriptive measures for your report.

The following is an example of the use of PROC UNIVARIATE for descriptive statistics. Serum calcium levels were measured in 47 osteoarthritis patients. The biostatistician wants to summarize and describe the data using descriptive statistics. The program below creates a SAS data set called CALCIUM and gives different descriptive analyses in the output report:

```
DATA CALCIUM;
INPUT PATIENT SCL @@;
CARDS;
```

1	7.4	13	7.7	25	7.3	37	7.9
2	8.1	14	7.5	26	7.4	38	8.2
3	7.0	15	8.0	27	8.3	39	7.7
4	6.9	16	6.8	28	7.6	40	7.3
5	7.3	17	6.9	29	7.7	41	7.5
6	7.7	18	7.3	30	7.4	42	7.7
7	6.5	19	7.7	31	7.0	43	7.3
8	7.2	20	7.0	32	7.1	44	7.6

```

9 7.0 21 7.3 33 7.1 45 7.8
10 8.3 22 7.6 34 7.5 46 7.1
11 7.4 23 7.5 35 7.7 47 7.2
12 7.4 24 6.9 36 8.0
;
PROC UNIVARIATE DATA=CALCIUM
FREQ PLOT NORMAL;
VAR SCL;
ID PATIENT;
TITLE 'DESCRIPTIVE STATISTICS';
RUN;

```

The FREQ, PLOT, and NORMAL options in PROC UNIVARIATE creates a frequency table, a stem-and-leaf plot, a box plot, a normal probability plot, and the Shapiro-Wilk statistic and its respective significance probability (w:Normal and PROB < W) for testing whether the data are from a normal distribution. The VAR statement contains the variable whose data will be summarized to create descriptive statistics. The ID statement contains the variable identifying the five lowest and five highest values in the summarized variable data.

Besides being used for descriptive statistics, PROC UNIVARIATE can generate other important information that is used in medical research. PROC UNIVARIATE will analyze paired observations (usually before and after measurements) if you create a new variable in the data set that is the difference of the two measurements (diff=after-before). The analyses that PROC UNIVARIATE can perform on the new variable are the paired sample t-test (Prob >T) and the Wilcoxon signed rank test (Prob >S).

The following is an example of the use of PROC UNIVARIATE for analyzing paired observations. The weights of fifteen male patients with ages ranging from 30 to 40 years old were taken at

a clinic. The physician puts the patients on an exercise and diet program. Six weeks later, the weights of the 15 male patients are again taken and recorded. Did the exercise and diet program have an effect on the weight of the patients? The program below can help us answer this question:

```

DATA WEIGHT;
INPUT PATIENT BEFORE AFTER;
DIFF=AFTER-BEFORE;
CARDS;
1 227.3 226.4
2 219.6 213.3
3 208.9 209.2
4 196.7 195.3
5 201.3 201.9
6 207.5 207.9
7 209.3 209.1
8 216.6 215.6
9 222.7 221.9
10 198.5 199.9
11 184.3 184.1
12 192.6 192.9
13 227.7 226.3
14 205.8 204.9
15 211.3 211.0
;

```

```

PROC UNIVARIATE DATA=WEIGHT;
VAR DIFF;
TITLE 'PAIRED COMPARISON TEST';
RUN;

```

The data set WEIGHT is created with values for the patient number, the patient's weight at the first visit, and the patient's weight at the second visit. A new variable was created by subtracting the BEFORE measurement from the AFTER measurement. The PROC UNIVARIATE statement requires only the VAR statement with the new variable to be analyzed to do a paired comparison analysis.

Analysis of variance is a method used to compare two or more means from independent groups. If you want to compare means

from two independent groups, use PROC TTEST. If you want to compare means from two or more independent groups with an equal number of nonmissing observations in each group, use PROC ANOVA. If you want to compare means from two or more independent groups with an unequal number of nonmissing observations in each group, use PROC GLM. PROC ANOVA and PROC GLM are also used to perform repeated measures analysis on medical research data.

The following is an example of the use of PROC TTEST. A parallel study was performed by a pharmaceutical company. They tested the maximum concentration of drug A in 29 patients against the maximum concentration of drug B in 23 patients. The program below is used to analyze the data:

```
DATA STUDY;
  INPUT PATIENT CMAX DRUG $ @@;
  CARDS;
1 27.4 A 19 19.7 A 37 18.0 B
2 31.6 B 20 21.6 B 38 17.5 A
3 20.3 A 21 20.0 A 39 18.7 A
4 17.6 A 22 19.1 B 40 22.2 B
5 22.9 B 23 21.1 A 41 31.3 A
6 33.8 A 24 22.5 A 42 27.5 B
7 19.7 B 25 23.7 B 43 26.6 A
8 30.5 A 26 28.2 A 44 28.2 B
9 21.3 B 27 25.6 B 45 25.7 A
10 18.4 A 28 27.8 A 46 22.3 B
11 17.3 B 29 23.9 B 47 24.1 A
12 26.0 A 30 21.6 A 48 26.7 B
13 24.7 A 31 15.3 B 49 21.2 A
14 18.9 B 32 23.0 A 50 19.9 B
15 29.1 A 33 26.7 B 51 30.4 A
16 28.3 B 34 27.7 A 52 17.6 A
17 26.5 A 35 29.5 B
18 25.9 B 36 30.3 A
;
PROC TTEST DATA=STUDY;
CLASS DRUG;
VAR CMAX;
TITLE 'PROC TTEST';
RUN;
```

The data set STUDY is created

with variables for patient number, maximum concentration and drug type. The CLASS statement in PROC TTEST identifies the variable that specifies the two groups. The VAR statement identifies the variable you wish to analyze.

The following is an example of the use of PROC GLM. Three groups of patients were given different treatments and blood samples were taken from the patients 24 hours after administration. Twenty-one patients took drug P, nineteen patients took drug Q, and fifteen patients took drug R. Is there a difference in concentrations of drug at 24 hours among the three medications? The program below is used to analyze the data:

```
DATA BLOOD;
  INPUT PATIENT DRUG $ CONC @@;
  CARDS;
1 P 5.2 20 Q 7.8 39 R 5.1
2 Q 3.7 21 R 3.9 40 P 4.8
3 R 4.8 22 P 4.1 41 Q 5.5
4 P 5.6 23 Q 4.5 42 R 4.7
5 Q 4.9 24 R 4.9 43 P 5.1
6 R 5.0 25 P 6.3 44 Q 5.0
7 P 3.9 26 Q 5.8 45 R 4.9
8 Q 4.4 27 R 5.3 46 P 6.1
9 R 4.9 28 P 4.1 47 Q 3.9
10 P 5.1 29 Q 5.5 48 P 5.2
11 Q 5.7 30 R 6.6 49 Q 4.4
12 R 7.9 31 P 3.7 50 P 4.7
13 P 4.0 32 Q 5.8 51 Q 4.3
14 Q 2.5 33 R 6.2 52 P 4.9
15 R 3.3 34 P 4.1 53 Q 4.6
16 P 4.7 35 Q 5.6 54 P 5.2
17 Q 5.6 36 R 6.3 55 P 5.7
18 R 7.7 37 P 4.7
19 P 4.9 38 Q 5.6
;
PROC GLM DATA = BLOOD;
CLASS DRUG;
MODEL CONC = DRUG;
LSMEANS DRUG / STDERR PDIF;
MEANS DRUG / T BON;
TITLE 'PROC GLM';
```

RUN;

The data set BLOOD is created with variables for patient number, drug type, and concentration of drug at 24 hours. The CLASS statement in PROC GLM identifies the variable that specifies the multiple groups. The MODEL statement in PROC GLM for a one-way analysis of variance has the variable you wish to analyze on the left side of the equal sign, and the variable that specifies the multiple groups, on the right side of the equal sign. The LSMEANS statement in PROC GLM calculates adjusted means for each drug group, and the STDERR and PDIF options after the slash print the standard errors of each estimated mean, the t statistic to test if each least squares mean is equal to zero, and the probability of significance for all pairs of drug group least squares means. The MEANS statement in PROC GLM calculates unadjusted means for each drug group, and the T and BON options after the slash perform multiple t-tests and Bonferroni t-tests for all pairs of drug group unadjusted means.

Categorical analysis refers to analysis of data that can be put into frequency tables. PROC FREQ is used to analyze a frequency table and will give chi-square statistics and Fisher's exact test. PROC CATMOD is used to fit a linear model to ordered categorical data.

The following is an example of the use of PROC FREQ. A clinical trial was performed at a hospital to study the effect of dyazide on reducing blood pressure. Hypertensive patients were given either dyazide or placebo tablets for five weeks.

At the end of the study, 25 of 55 patients on dyazide experienced clinically meaningful changes in blood pressure, while 11 of 58 patients on placebo experienced clinically meaningful changes in blood pressure. The program below is used to analyze the data:

```
DATA TRIAL;
  INPUT DRUG $ CHANGES $ COUNT;
  CARDS;
DYAZIDE YES 25
DYAZIDE NO 30
PLACEBO YES 11
PLACEBO NO 47
;
PROC FREQ DATA = TRIAL;
  TABLES CHANGES*DRUG /
  EXPECTED CHISQ;
  WEIGHT COUNT;
  TITLE 'CHI-SQUARE TEST';
RUN;
```

The data set TRIAL is created with values for drug type, whether the patient experienced clinically meaningful changes in blood pressure, and the number of patients in each category. The TABLES statement in PROC FREQ creates a contingency table, and the EXPECTED and CHISQ options after the slash give expected cell frequencies and the chi-square test for independence. The WEIGHT statement in PROC FREQ allows each observation of the variable COUNT to correspond to a cell frequency in the contingency table.

Regression analysis refers to analysis of the relationships between variables. PROC CORR is used to find the correlation between pairs of variables. PROC REG is used to fit a linear regression model by the method of least squares.

The following is an example of

the use of PROC REG. Several clinicians wanted to know if the age (in years) and weight (in pounds) of male patients were important in determining their systolic blood pressure. These three variables were measured in 48 males. The program below is used to analyze the data:

```

DATA VALUES;
INPUT PATIENT AGE
WEIGHT SYSTOLIC @@;
CARDS;
1 34 179 127 25 36 181 99
2 56 221 143 26 31 209 114
3 44 208 153 27 33 197 126
4 26 169 118 28 40 199 118
5 49 215 123 29 38 217 139
6 29 203 110 30 44 222 166
7 31 237 156 31 47 241 161
8 35 210 143 32 45 198 144
9 44 226 167 33 28 163 97
10 32 200 119 34 33 198 144
11 24 188 102 35 34 191 120
12 29 223 116 36 37 188 110
13 47 186 113 37 27 195 108
14 39 193 126 38 31 202 123
15 40 185 130 39 53 176 117
16 38 163 114 40 51 181 123
17 44 207 183 41 38 201 153
18 49 156 117 42 33 173 142
19 37 164 122 43 26 148 119
20 30 133 118 44 28 144 131
21 32 146 140 45 35 158 109
22 52 191 150 46 34 177 123
23 41 186 145 47 36 182 127
24 43 185 147 48 39 192 153
;
PROC REG DATA=VALUES;
MODEL SYSTOLIC=AGE WEIGHT;
TITLE 'PROC REG';
RUN;

```

The data set VALUES is created with variables for patient number, age, weight, and systolic blood pressure. The MODEL statement is created in PROC REG by placing the dependent variable on the left side of the equal sign and the independent variables on the right side of the equal sign.

In this example, the variables AGE and WEIGHT are independent variables, and the variable SYSTOLIC is the dependent variable.

Nonparametric statistics are methods used to perform distribution-free analyses. PROC NPAR1WAY is used to do nonparametric analysis of two or more independent groups. PROC RANK will create a data set with a new variable that is the rank (smallest to largest) of each observation in the data set. Nonparametric tests should only be used when you suspect that the data are not normally distributed.

The following is an example of the use of PROC NPAR1WAY. The data set from the example program for PROC TTEST is used, only this time we will run a Wilcoxon rank sum test on the data. The program below is used to analyze the data:

```

DATA STUDY;
INPUT PATIENT CMAX DRUG $ @@;
CARDS;
1 27.4 A 19 19.7 A 37 18.0 B
2 31.6 B 20 21.6 B 38 17.5 A
3 20.3 A 21 20.0 A 39 18.7 A
4 17.6 A 22 19.1 B 40 22.2 B
5 22.9 B 23 21.1 A 41 31.3 A
6 33.8 A 24 22.5 A 42 27.5 B
7 19.7 B 25 23.7 B 43 26.6 A
8 30.5 A 26 28.2 A 44 28.2 B
9 21.3 B 27 25.6 B 45 25.7 A
10 18.4 A 28 27.8 A 46 22.3 B
11 17.3 B 29 23.9 B 47 24.1 A
12 26.0 A 30 21.6 A 48 26.7 B
13 24.7 A 31 15.3 B 49 21.2 A
14 18.9 B 32 23.0 A 50 19.9 B
15 29.1 A 33 26.7 B 51 30.4 A
16 28.3 B 34 27.7 A 52 17.6 A
17 26.5 A 35 29.5 B
18 25.9 B 36 30.3 A
;
PROC NPAR1WAY DATA=STUDY
WILCOXON;

```

```
CLASS DRUG;  
VAR CMAX;  
TITLE 'PROC NPAR1WAY';  
RUN;
```

In the PROC NPAR1WAY statement, the WILCOXON option will do a Wilcoxon rank sum test on the data. The CLASS and VAR statements are the same as in PROC TTEST.

Survival analysis refers to analysis of the length of time to specific event. Survival analysis is also called life table analysis. PROC LIFETEST is a procedure that can perform survival analysis.

The methods described in this paper cover many problems that are a part of medical research. These procedures can be used to analyze complicated problems in data analysis. Whenever there is doubt on the procedure to run on a data set, it is best to contact a professional statistician on the methods that would be appropriate to a specific situation.

SASTM is a registered trademark of SAS Institute, Inc., Cary, N.C.

REFERENCES:

Agresti, Alan, Categorical Data Analysis, John Wiley & Sons, Inc., New York, NY, 1990.

Freund, Rudolf, J., Littell, Ramon C., and Spector, Philip C., SAS System for Linear Models, 1986 Edition, SAS Institute, Inc., Cary, NC.

Kirkwood, Betty R., Essentials of Medical Statistics, Blackwell Scientific Publications, Oxford, England, 1988.

Snedecor, George, W., and Cochran, William G., Statistical Methods, Eighth Edition, Iowa State University Press, Ames, IA, 1989.

Schlotzhauer, Sandra D., and Littell, Ramon C., SAS System for Elementary Statistical Analysis, SAS Institute, Inc., Cary, NC, 1989.

SAS Procedures Guide for Personal Computers, Version 6 Edition, SAS Institute, Inc. Cary, NC, 1987.

SAS/STAT Guide for Personal Computers, Version 6 Edition, SAS Institute, Inc., Cary, NC, 1987.

SAS Technical Report: P-179 Additional SAS/STAT Procedures, Release 6.03, SAS Institute, Inc., Cary, NC, 1988.