What statistical analysis should I use for my data using SAS?

Authored by

stats writer

June 28, 2024

The appropriate statistical analysis for your data using SAS would depend on the characteristics and objectives of your data. SAS offers a wide range of statistical procedures and techniques to analyze and interpret data, including descriptive statistics, hypothesis testing, regression analysis, and data mining. It is recommended to carefully assess the type of data, research question, and desired outcome before selecting the appropriate statistical analysis in SAS. Additionally, consulting with a statistician or conducting a thorough literature review can also aid in determining the best approach for your data analysis.

What statistical analysis should I use? Statistical analyses using SAS

Introduction

One sample t-test

**A one sample t-test allows us to test whether a sample mean (from a normally
distributed interval variable) significantly differs from a hypothesized
value. For example, using the hsb2 data file, say we wish to test
whether the average writing score (write) differs significantly from 50. We
can do this as shown below.**

```
proc ttest data = "c:/mydata/hsb2" h0 = 50;
var write;
run;
```

The TTEST Procedure

Statistics

Lower CL Upper CL Lower CL Upper CL
Variable N Mean Mean Mean Std Dev Std Dev Std Dev Std Err
write 200 51.453 52.775 54.097 8.6318 9.4786 10.511 0.6702

T-Tests

Variable DF t Value Pr > |t|
write 199 4.14 <.0001

The mean of the variable write for this particular sample of students is 52.775,
which is statistically significantly different from the test value of 50. We would
conclude that this group of students has a significantly higher mean on the writing test
than 50.

One sample median test

A one sample median test allows us to test whether a sample median differs

**significantly from a hypothesized value. We will use the same variable, write,
as we did in the one sample t-test example above, but we do not need
to assume that it is interval and normally distributed (we only need to assume that write
is an ordinal variable). We will test whether the median writing score (write)
differs significantly from 50. The loccount option on the proc
univariate statement provides the location counts of the data shown at the
bottom of the output.**

```
proc univariate data = "c:/mydata/hsb2" loccount mu0 = 50;
var write;
run;
```

Basic Statistical Measures

Location Variability

Mean 52.77500 Std Deviation 9.47859
Median 54.00000 Variance 89.84359

Mode 59.00000 Range 36.00000

Interquartile Range 14.50000

Tests for Location: Mu0=50

Test -Statistic- -----p Value------

Student's t t 4.140325 Pr > |t| <.0001

Sign M 27 Pr >= |M| 0.0002

Signed Rank S 3326.5 Pr >= |S| <.0001

Location Counts: Mu0=50.00

Count Value

Num Obs > Mu0 12

Num Obs ^= Mu0 198

Num Obs < Mu0 72

You can use either the sign test or the signed rank test. The
difference between these two tests is that the signed rank requires that the
variable be from a symmetric distribution. The results indicate that the median of the variable write for this group is

statistically significantly different from 50.

Binomial test

A one sample binomial test allows us to test whether the proportion of successes on a two-level categorical dependent variable significantly differs from a hypothesized value. For example, using the hsb2 data file, say we wish to test whether the proportion of females (female) differs significantly from 50%, i.e., from .5. We will use the exact statement to produce the exact p-values.

```
proc freq data = "c:/mydata/hsb2";
tables female / binomial(p=.5);
exact binomial;
run;
```

The FREQ Procedure

female	Frequency	Percent	Cumulative Frequency	Cumulative Percent

----

0 91 45.50 91 45.50

1 109 54.50 200 100.00

Binomial Proportion for female = 0

Proportion (P) 0.4550

ASE 0.0352

95% Lower Conf Limit 0.3860

95% Upper Conf Limit 0.5240

Exact Conf Limits

95% Lower Conf Limit 0.3846

95% Upper Conf Limit 0.5267

Test of H0: Proportion = 0.5

ASE under H0 0.0354

Z -1.2728

One-sided Pr < Z 0.1015

Two-sided Pr > |Z| 0.2031

Exact Test

One-sided Pr <= P 0.1146

Two-sided = 2 * One-sided 0.2292

Sample Size = 200

The results indicate that there is no statistically significant difference (p = .2292). In other words, the proportion of females in this sample does not significantly differ from the hypothesized value of 50%.

Chi-square goodness of fit

A chi-square goodness of fit test allows us to test whether the observed proportions for a categorical variable differ from hypothesized proportions. For example, let's suppose that we believe that the general population consists of 10% Hispanic, 10% Asian, 10% African American and 70% White folks. We want to test whether the observed proportions from our sample differ significantly from these hypothesized proportions.

The hypothesized proportions are placed in parentheses after the testp= option on the tables statement.

```
proc freq data = "c:/mydata/hsb2";
tables race / chisq testp=(10 10 10 70);
run;
```

```
The FREQ Procedure

Test Cumulative Cumulative
race Frequency Percent Percent Frequency Percent
---------------------------------------------------------------
1 24 12.00 10.00 24 12.00
2 11 5.50 10.00 35 17.50
3 20 10.00 10.00 55 27.50
4 145 72.50 70.00 200 100.00

Chi-Square Test
for Specified Proportions
-----------------------
Chi-Square 5.0286
DF 3
Pr > ChiSq 0.1697

Sample Size = 200
```

These results show that racial composition in our sample does not differ significantly

from the hypothesized values that we supplied (chi-square with three degrees of freedom = 5.0286, p = .1697).

Two independent samples t-test

An independent samples t-test is used when you want to compare the means of a normally distributed interval dependent variable for two independent groups. For example, using the hsb2 data file, say we wish to test whether the mean for write is the same for males and females.

```
proc ttest data = "c:/mydata/hsb2";
class female;
var write;
run;
```

The TTEST Procedure
Statistics

Lower CL Upper CL Lower CL Upper CL
Variable female N Mean Mean Mean Std Dev Std Dev Std Dev Std Err

write 0 91 47.975 50.121 52.267 8.9947 10.305 12.066 1.0803
write 1 109 53.447 54.991 56.535 7.1786 8.1337 9.3843 0.7791
write Diff (1-2) -7.442 -4.87 -2.298 8.3622 9.1846 10.188 1.3042

T-Tests

Variable Method Variances DF t Value Pr > |t|
write Pooled Equal 198 -3.73 0.0002
write Satterthwaite Unequal 170 -3.66 0.0003

Equality of Variances
Variable Method Num DF Den DF F Value Pr > F
write Folded F 90 108 1.61 0.0187

The results indicate that there is a statistically significant difference between the
mean writing score for males and females (t = -3.73, p = .0002). In other words,
females have a statistically significantly higher mean score on writing
(54.991) than males (50.121).

Wilcoxon-Mann-Whitney test

The Wilcoxon-Mann-Whitney test is a non-parametric analog to the independent samples t-test and can be used when you do not assume that the dependent variable is a normally distributed interval variable (you need only assume that the variable is at least ordinal). We will use the same data file (the hsb2 data file) and the same variables in this example as we did in the independent t-test example above and will not assume that write, our dependent variable, is normally distributed.

```
proc npar1way data = "c:/mydata/hsb2" wilcoxon;
class female;
var write;
run;
```


The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable write Classified by Variable female

female	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score

---

0 91 7792.0 9145.50 406.559086 85.626374

1 109 12308.0 10954.50 406.559086 112.917431

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic 7792.0000

Normal Approximation

Z -3.3279

One-Sided Pr < Z 0.0004

Two-Sided Pr > |Z| 0.0009

t Approximation

One-Sided Pr < Z 0.0005

Two-Sided Pr > |Z| 0.0010

Z includes a continuity correction of 0.5.

The results suggest that there is a statistically significant difference

between the underlying distributions of the write scores of males and

the write scores of females (z = -3.329, p = 0.0009).

Chi-square test

A chi-square test is used when you want to see if there is a relationship between two categorical variables. In SAS, the chisq option is used on the tables statement to obtain the test statistic and its associated p-value. Using the hsb2 data file, let's see if there is a relationship between the type of school attended (schtyp) and students' gender (female). Remember that the chi-square test assumes that the expected value for each cell is five or higher. This assumption is easily met in the examples below. However, if this assumption is not met in your data, please see the section on Fisher's exact test below.

```
proc freq data = "c:/mydata/hsb2";
tables schtyp*female / chisq;
run;
```

The FREQ Procedure

Table of schtyp by female

schtyp(type of school)
female
Frequency|
Percent |
Row Pct |
Col Pct | 0| 1| Total
---------+--------+--------+
1 | 77 | 91 | 168
| 38.50 | 45.50 | 84.00
| 45.83 | 54.17 |
| 84.62 | 83.49 |
---------+--------+--------+
2 | 14 | 18 | 32
| 7.00 | 9.00 | 16.00
| 43.75 | 56.25 |
| 15.38 | 16.51 |
---------+--------+--------+
Total 91 109 200
45.50 54.50 100.00

Statistics for Table of schtyp by female

Statistic DF Value Prob

Chi-Square 1 0.0470 0.8283

Likelihood Ratio Chi-Square 1 0.0471 0.8281

Continuity Adj. Chi-Square 1 0.0005 0.9815

Mantel-Haenszel Chi-Square 1 0.0468 0.8287

Phi Coefficient 0.0153

Contingency Coefficient 0.0153

Cramer's V 0.0153

Sample Size = 200

These results indicate that there is no statistically significant relationship between the type of school attended and gender (chi-square with one degree of freedom = 0.0470, p = 0.8283).

Let's look at another example, this time looking at the relationship between gender (female) and socio-economic status (ses). The point of this example is that one (or both) variables may have more than two levels, and that the variables do not have to have

the same number of levels. In this example, female has two levels (male and female) and ses has three levels (low, medium and high).

```
proc freq data = "c:/mydata/hsb2";
tables female*ses / chisq;
run;
```

```
The FREQ Procedure

Table of female by ses

female ses

Frequency|
Percent |
Row Pct |
Col Pct | 1| 2| 3| Total
---------+--------+--------+--------+
0 | 15 | 47 | 29 | 91
| 7.50 | 23.50 | 14.50 | 45.50
| 16.48 | 51.65 | 31.87 |
| 31.91 | 49.47 | 50.00 |
---------+--------+--------+--------+
```

1 | 32 | 48 | 29 | 109
| 16.00 | 24.00 | 14.50 | 54.50
| 29.36 | 44.04 | 26.61 |
| 68.09 | 50.53 | 50.00 |
---------+--------+--------+--------+
Total 47 95 58 200
23.50 47.50 29.00 100.00

Statistics for Table of female by ses

Statistic DF Value Prob

Chi-Square 2 4.5765 0.1014
Likelihood Ratio Chi-Square 2 4.6789 0.0964
Mantel-Haenszel Chi-Square 1 3.1098 0.0778
Phi Coefficient 0.1513
Contingency Coefficient 0.1496
Cramer's V 0.1513

Sample Size = 200

Again we find that there is no statistically significant relationship between the
variables (chi-square with two degrees of freedom = 4.5765, p = 0.1014).

Fisher's exact test

The Fisher's exact test is used when you want to conduct a chi-square test, but one or more of your cells has an expected frequency of less than five. Remember that the chi-square test assumes that each cell has an expected frequency of five or more, but the Fisher's exact test has no such assumption and can be used regardless of how small the expected frequency is. In the example below, we have cells with observed frequencies of two and one, which may indicate expected frequencies that could be below five, so we will use Fisher's exact test with the fisher option on the tables statement.

```
proc freq data = "c:/mydata/hsb2";
tables schtyp*race / fisher;
run;
```

The FREQ Procedure

Table of schtyp by race

schtyp(type of school) race

Frequency|
Percent |
Row Pct |
Col Pct | 1| 2| 3| 4| Total
---------+--------+--------+--------+--------+
1 | 22 | 10 | 18 | 118 | 168
| 11.00 | 5.00 | 9.00 | 59.00 | 84.00
| 13.10 | 5.95 | 10.71 | 70.24 |
| 91.67 | 90.91 | 90.00 | 81.38 |
---------+--------+--------+--------+--------+
2 | 2 | 1 | 2 | 27 | 32
| 1.00 | 0.50 | 1.00 | 13.50 | 16.00
| 6.25 | 3.13 | 6.25 | 84.38 |
| 8.33 | 9.09 | 10.00 | 18.62 |
---------+--------+--------+--------+--------+
Total 24 11 20 145 200
12.00 5.50 10.00 72.50 100.00

Statistics for Table of schtyp by race

Statistic DF Value Prob
--
Chi-Square 3 2.7170 0.4373

Likelihood Ratio Chi-Square 3 2.9985 0.3919
Mantel-Haenszel Chi-Square 1 2.3378 0.1263
Phi Coefficient 0.1166
Contingency Coefficient 0.1158
Cramer's V 0.1166

WARNING: 38% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test

Table Probability (P) 0.0077
Pr <= P 0.5975

Sample Size = 200

These results suggest that there is not a statistically significant relationship
between race and type of school (p = 0.5975). Note that the Fisher's exact test does not
have a "test statistic", but computes the p-value directly.

One-way ANOVA

A one-way analysis of variance (ANOVA) is used when

**you have a categorical independent
variable (with two or more categories) and a normally distributed interval dependent
variable and you wish to test for differences in the means of the dependent variable
broken down by the levels of the independent variable. For example, using the hsb2 data file, say we wish to test whether the mean of write
differs between the three program types (prog). We will also use
the means statement to output the mean of write for each level of program
type. Note that this will not tell you if there is a statistically significant
difference between any two sets of means.**

```
proc glm data = "c:/mydata/hsb2";
class prog;
model write = prog;
means prog;
run;
quit;
```

The GLM Procedure

Class Level Information

Class Levels Values

prog 3 1 2 3

Number of observations 200

Dependent Variable: write writing score

Sum of

Source DF Squares Mean Square F Value Pr > F

Model 2 3175.69786 1587.84893 21.27 <.0001

Error 197 14703.17714 74.63542

Corrected Total 199 17878.87500

R-Square Coeff Var Root MSE write Mean

0.177623 16.36983 8.639179 52.77500

Source DF Type I SS Mean Square F Value Pr > F

prog 2 3175.697857 1587.848929 21.27 <.0001

Source DF Type III SS Mean Square F Value Pr > F

prog 2 3175.697857 1587.848929 21.27 <.0001

Level of ----------write----------

prog N Mean Std Dev

1 45 51.3333333 9.39777537

2 105 56.2571429 7.94334333

3 50 46.7600000 9.31875441

The mean of the dependent variable differs significantly among the levels of program type. However, we do not know if the difference is between only two of the levels or all three of the levels. (The F test for the model is the same as the F test for prog because prog was the only variable entered into the model. If other variables had also been entered, the F test for the Model would have been different from prog.) We can also see that the students in the academic program have the highest mean writing score, while students in the vocational program have the lowest.

Kruskal Wallis test

The Kruskal Wallis test is used when you have one independent variable with two or more levels and an ordinal dependent variable. In other

words, it is the non-parametric version
of ANOVA. It is also a generalized form of the Mann-Whitney test method,
as it permits
two or more
groups. We will use the same data file as the one way ANOVA
example above (the hsb2 data file) and the same variables as in the
example above, but we will not assume that write is a normally distributed interval
variable.

```
proc npar1way data = "c:/mydata/hsb2";
class prog;
var write;
run;
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable write
Classified by Variable prog

prog	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score

---

1 45 4079.0 4522.50 340.927342 90.644444

3 50 3257.0 5025.00 353.525185 65.140000

2 105 12764.0 10552.50 407.705133 121.561905

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square 34.0452

DF 2

Pr > Chi-Square <.0001

The results indicate
that there is a statistically significant difference among the three type of
programs (chi-square with two degrees of freedom = 34.0452, p = 0.0001).

Paired t-test

A paired (samples) t-test is used when you have two related observations
(i.e., two observations per subject) and you want to see if the means on these two normally
distributed interval variables differ from one another.

**For example, using the hsb2 data file we will test whether the mean of read is equal to
the mean of write.**

```
proc ttest data = "c:/mydata/hsb2";
paired write*read;
run;
```

The TTEST Procedure
Statistics

Lower CL Upper CL Lower CL Upper CL
Difference N Mean Mean Mean Std Dev Std Dev Std Dev Std Err
write - read 200 -0.694 0.545 1.7841 8.0928 8.8867 9.8546 0.6284

T-Tests

Difference DF t Value Pr > |t|

write - read 199 0.87 0.3868

These results indicate that the mean of read is not statistically significantly

different from the mean of write (t = 0.87, p = 0.3868).

Wilcoxon signed rank sum test

The Wilcoxon signed rank sum test is the non-parametric version of a paired samples t-test. You use the Wilcoxon signed rank sum test when you do not wish to assume that the difference between the two variables is interval and normally distributed (but you do assume the difference is ordinal). We will use the same example as above, but we will not assume that the difference between read and write is interval and normally distributed. We will first do a data step to create the difference of the two scores for each subject. This is necessary because SAS will not calculate the difference for you in proc univariate.

```
data hsb2a;
set 'c:/mydata/hsb2';
diff = read - write;
run;
```

proc univariate data = hsb2a;

var diff;

run;

The UNIVARIATE Procedure

Variable: diff

Basic Statistical Measures

Location Variability

Mean -0.54500 Std Deviation 8.88667

Median 0.00000 Variance 78.97284

Mode 6.00000 Range 45.00000

Interquartile Range 13.00000

Tests for Location: Mu0=0

Test -Statistic- -----p Value------

Student's t t -0.86731 Pr > |t| 0.3868

Sign M -4.5 Pr >= |M| 0.5565

Signed Rank S -658.5 Pr >= |S| 0.3677

The results suggest that there is not a statistically significant difference between read

and write.

If you believe the differences between read and write were not ordinal
but could merely be classified as positive and negative, then you may want to consider a
sign test in lieu of sign rank test. Note that the SAS output
gives you the results for both the Wilcoxon signed rank test and the sign test
without having to use any options. Using the sign test, we again
conclude that there is no statistically significant difference between read
and write (p=.5565).

McNemar test

You would perform McNemar's test
if you were interested in the marginal frequencies of two binary outcomes.
These binary outcomes may be the same outcome variable on matched pairs
(like a case-control study) or two outcome
variables from a single group. Let us consider two

**questions, Q1 and Q2, from a test taken by 200 students. Suppose 172
students answered both questions correctly, 15 students answered both
questions incorrectly, 7 answered Q1 correctly and Q2 incorrectly, and 6
answered Q2 correctly and Q1 incorrectly. These counts can be considered in
a two-way contingency table. The null hypothesis is that the two
questions are answered correctly or incorrectly at the same rate (or that
the contingency table is symmetric).**

```
data set1;
input Q1correct Q2correct students;
datalines;
1 1 172
0 1 6
1 0 7
0 0 15
run;

proc freq data=set1;
```

```
table Q1correct*Q2correct;
exact mcnem;
weight students;
run;

The FREQ Procedure

Table of Q1correct by Q2correct

Q1correct Q2correct

Frequency|
Percent |
Row Pct |
Col Pct | 0| 1| Total
---------+--------+--------+
0 | 15 | 6 | 21
| 7.50 | 3.00 | 10.50
| 71.43 | 28.57 |
| 68.18 | 3.37 |
---------+--------+--------+
1 | 7 | 172 | 179
| 3.50 | 86.00 | 89.50
| 3.91 | 96.09 |
| 31.82 | 96.63 |
```

---------+--------+--------+

Total 22 178 200

11.00 89.00 100.00

Statistics for Table of Q1correct by Q2correct

McNemar's Test

Statistic (S) 0.0769

DF 1

Asymptotic Pr > S 0.7815

Exact Pr >= S 1.0000

Simple Kappa Coefficient

Kappa 0.6613

ASE 0.0873

95% Lower Conf Limit 0.4901

95% Upper Conf Limit 0.8324

Sample Size = 200

McNemar's test statistic suggests that there is not a statistically significant difference in the proportions of correct/incorrect answers to these two questions.

One-way repeated measures ANOVA

You would perform a one-way repeated measures analysis of variance if you had one categorical independent variable and a normally distributed interval dependent variable that was repeated at least twice for each subject. This is the equivalent of the paired samples t-test, but allows for two or more levels of the categorical variable.

The one-way repeated measures ANOVA tests whether the mean of the dependent variable differs by the categorical variable. We have an example data set called rb4wide, which is used in Kirk's book Experimental Design. In this data set, y1 y2 y3 and y4 represent the dependent variable measured at the 4 levels of a, the repeated measures independent variable.

```
proc glm data = 'c:/mydata/rb4wide';
model y1 y2 y3 y4 = ;
repeated a ;
run;
```

quit;

The GLM Procedure
Repeated Measures Analysis of Variance

Repeated Measures Level Information

Dependent Variable Y1 Y2 Y3 Y4

Level of a 1 2 3 4

Manova Test Criteria and Exact F Statistics for the Hypothesis of no a Effect
H = Type III SSCP Matrix for a
E = Error SSCP Matrix

S=1 M=0.5 N=1.5

Statistic Value F Value Num DF Den DF Pr > F

Wilks' Lambda 0.24580793 5.11 3 5 0.0554
Pillai's Trace 0.75419207 5.11 3 5 0.0554
Hotelling-Lawley Trace 3.06821705 5.11 3 5 0.0554
Roy's Greatest Root 3.06821705 5.11 3 5 0.0554

Repeated Measures Analysis of Variance

Univariate Tests of Hypotheses for Within Subject Effects

Adj Pr > F

Source DF Type III SS Mean Square F Value Pr > F G - G H - F

a 3 49.00000000 16.33333333 11.63 0.0001 0.0015 0.0003

Error(a) 21 29.50000000 1.40476190

Greenhouse-Geisser Epsilon 0.6195

Huynh-Feldt Epsilon 0.8343

The results indicate that the model as well as both factors (a and s)

are statistically significant. The p-value given in this output for a

(0.0001) is the "regular" p-value and is the p-value that you would get if you assumed compound

symmetry in the variance-covariance matrix.

Repeated measures logistic regression

If you have a binary outcome

measured repeatedly for each subject and you wish to run a logistic

regression that accounts for the effect of multiple

**measures from single
subjects, you can perform a repeated measures logistic regression. In
SAS, this can be done by using the
genmod procedure and indicating binomial
as the probability distribution and logit as the link function to be used in
the model. The
exercise data file contains
three pulse measurements from each of 30 people assigned to two different diet regiments and
three different exercise regiments. If we define a "high" pulse as being over
100, we can then predict the probability of a high pulse using diet
regiment.**

```
proc genmod data='c:/mydata/exercise' descending;
class id diet / descending;
model highpulse = diet / dist = bin link = logit;
repeated subject = id / type = exch;
run;
```

Response Profile

Ordered Total
Value highpulse Frequency

1 1 27
2 0 63

PROC GENMOD is modeling the probability that highpulse='1'.

Parameter Information

Parameter Effect diet

Prm1 Intercept
Prm2 diet 2
Prm3 diet 1

Algorithm converged.

GEE Model Information

Correlation Structure Exchangeable
Subject Effect id (30 levels)
Number of Clusters 30

The GENMOD Procedure

GEE Model Information

Correlation Matrix Dimension 3
Maximum Cluster Size 3
Minimum Cluster Size 3

Algorithm converged.

Exchangeable Working
Correlation

Correlation 0.3306722695

GEE Fit Criteria

QIC 113.9859
QICu 111.3405

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Standard 95% Confidence
Parameter Estimate Error Limits Z Pr > |Z|

Intercept -1.2528 0.4328 -2.1011 -0.4044 -2.89 0.0038
diet 2 0.7538 0.6031 -0.4283 1.9358 1.25 0.2114

diet 1 0.0000 0.0000 0.0000 0.0000 . .

These results indicate that diet is not statistically significant (Z = -1.25, p = 0.2114).

Factorial ANOVA

A factorial ANOVA has two or more categorical independent variables (either with or without the interactions) and a single normally distributed interval dependent variable. For example, using the hsb2 data file we will look at writing scores (write) as the dependent variable and gender (female) and socio-economic status (ses) as independent variables, and we will include an interaction of female by ses. Note that in SAS, you do not need to have the interaction term(s) in your data set. Rather, you can have SAS create it/them temporarily by placing an asterisk between the variables that will make up the interaction term(s).

```
proc glm data = "c:/mydata/hsb2";
class female ses;
model write = female ses female*ses;
run;
quit;
```

The GLM Procedure

Dependent Variable: write writing score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2278.24419	455.64884	5.67	<.0001
Error	194	15600.63081	80.41562		
Corrected Total	199	17878.87500			

R-Square	Coeff Var	Root MSE	write Mean
0.127427	16.99190	8.967476	52.77500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
female	1	1176.213845	1176.213845	14.63	0.0002
ses	2	1080.599437	540.299718	6.72	0.0015
female*ses	2	21.430904	10.715452	0.13	0.8753

Source	DF	Type III SS	Mean Square	F Value	Pr > F

female 1 1334.493311 1334.493311 16.59 <.0001
ses 2 1063.252697 531.626349 6.61 0.0017
female*ses 2 21.430904 10.715452 0.13 0.8753

These results indicate that the overall model is statistically significant (F = 5.67, p = 0.001). The variables female and ses are also statistically significant (F = 16.59, p = 0.0001 and F = 6.61, p = 0.0017, respectively). However, that interaction between female and ses is not statistically significant (F = 0.13, p = 0.8753).

Friedman test

You perform a Friedman test when you have one within-subjects independent variable with two or more levels and a dependent variable that is not interval and normally distributed (but at least ordinal). We will use this test to determine if there is a difference in the reading, writing and math scores. The null hypothesis in this test is that the

distribution of the
ranks of each type of score (i.e., reading, writing and math) are the
same. To conduct a Friedman test, the data need
to be in a long format; we will use proc transpose to change our data
from the wide format that they are currently in to a long format. We create a variable to code for the type of score, which we will
call rwm (for read, write, math), and col1
that contains the score on the dependent variable, that is the reading,
writing or math score. To obtain
the Friedman test, you need to use the cmh2 option on the tables
statement in proc freq.

```
proc sort data = "c:/mydata/hsb2" out=hsbsort;
by id;
run;

proc transpose data=hsbsort out=hsblong name=rwm;
by id;
var read write math;
```

run;

proc freq data=hsblong;
tables id*rwm*col1 / cmh2 scores=rank noprint;
run;

The FREQ Procedure

Summary Statistics for rwm by COL1
Controlling for id

Cochran-Mantel-Haenszel Statistics (Based on Rank Scores)

Statistic Alternative Hypothesis DF Value Prob
---
1 Nonzero Correlation 1 0.0790 0.7787
2 Row Mean Scores Differ 2 0.6449 0.7244

Total Sample Size = 600

The Row Mean Scores Differ is the same as the Friedman's chi-square, and we
see that with a value of 0.6449 and a p-value of 0.7244, it is not statistically

significant. Hence, there is no evidence that the distributions of the
three types of scores are different.

Ordered logistic regression

Ordered logistic regression is used when the dependent variable is
ordered, but not continuous. For example, using the hsb2 data file we will create an ordered
variable called write3. This variable will have the values 1, 2
and 3, indicating a low, medium or high writing score. We do not
generally recommend categorizing a continuous variable in this way; we are
simply creating a variable to use for this example. We will use gender
(female), reading score (read) and social studies score (socst)
as predictor variables in this model. The desc option on the proc
logistic statement is used so that SAS models the odds of being in the
lower category. The Response Profile table in the

output shows the
value that SAS used when conducting the analysis (given in the Ordered Value
column), the value of the original variable, and the number of cases in each
level of the outcome variable. (If you want SAS to use the values that you
have assigned the outcome variable, then you would want to use the order
= data option on the proc logistic statement.) The note below
this table reminds us that the "Probabilities modeled are cumulated over the
lower Ordered Values." It is helpful to remember this when interpreting the
output. The expb option on the model statement tells SAS to show the exponentiated coefficients (i.e., the proportional odds ratios).

```
data hsb2_ordered;
set "c:/mydata/hsb2";
if 30 <= write <=48 then write3 = 1;
if 49 <= write <=57 then write3 = 2;
if 58 <= write <=70 then write3 = 3;
```

run;

proc logistic data = hsb2_ordered desc;
model write3 = female read socst / expb;
run;

The LOGISTIC Procedure

Model Information

Data Set WORK.HSB2_ORDERED
Response Variable write3
Number of Response Levels 3
Model cumulative logit
Optimization Technique Fisher's scoring

Number of Observations Read 200
Number of Observations Used 200

Response Profile

Ordered Total
Value write3 Frequency

1 3 78
2 2 61

3 1 61

Probabilities modeled are cumulated over the lower Ordered Values.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Score Test for the Proportional Odds Assumption

Chi-Square DF Pr > ChiSq

2.1211 3 0.5477

Model Fit Statistics

Intercept
Intercept and
Criterion Only Covariates

AIC 440.627 322.553
SC 447.224 339.044
-2 Log L 436.627 312.553

Testing Global Null Hypothesis: BETA=0

Test Chi-Square DF Pr > ChiSq

Likelihood Ratio 124.0745 3 <.0001
Score 93.1890 3 <.0001
Wald 76.6752 3 <.0001

Analysis of Maximum Likelihood Estimates

Standard Wald
Parameter DF Estimate Error Chi-Square Pr > ChiSq Exp(Est)

Intercept 3 1 -11.8007 1.3122 80.8702 <.0001 0.000
Intercept 2 1 -9.7042 1.2026 65.1114 <.0001 0.000
FEMALE 1 1.2856 0.3225 15.8901 <.0001 3.617
READ 1 0.1177 0.0215 29.8689 <.0001 1.125
SOCST 1 0.0802 0.0190 17.7817 <.0001 1.083

Odds Ratio Estimates

Point 95% Wald
Effect Estimate Confidence Limits

FEMALE 3.617 1.922 6.805
READ 1.125 1.078 1.173
SOCST 1.083 1.044 1.125

Association of Predicted Probabilities and Observed Responses

Percent Concordant 83.8 Somers' D 0.681
Percent Discordant 15.7 Gamma 0.685
Percent Tied 0.6 Tau-a 0.453
Pairs 13237 c 0.840

The results indicate that the overall model is statistically significant (p < .0001), as are each of the predictor variables (p < .0001). There are two intercepts for this model because there are three levels of the outcome variable. We also see that the test of the proportional odds assumption is non-significant (p = .5477). One of the assumptions underlying ordinal logistic (and ordinal probit) regression is that the relationship between each pair of outcome groups is the same. In other words, ordinal logistic regression assumes that the coefficients that

describe the relationship between, say, the lowest versus all higher
categories of the response variable are the same as those that describe the
relationship between the next lowest category and all higher categories,
etc. This is called the proportional odds assumption or the parallel
regression assumption. Because the relationship between all pairs of groups
is the same, there is only one set of coefficients (only one model). If
this was not the case, we would need different models (such as a generalized
ordered logit model) to describe the relationship between each pair of
outcome groups.

Factorial logistic regression

A factorial logistic regression is used when you have two or more categorical
independent variables but a dichotomous dependent variable. For example, using the hsb2 data file we will use female as our dependent variable,

because it is the only dichotomous variable in our data set; certainly not because it
common practice to use gender as an outcome variable. We will use type of program (prog)
and school type (schtyp) as our predictor variables. Because
neither
prog nor schtyp are continuous variables, we need to include them on the class statement. The desc option on the proc
logistic statement is necessary so that SAS models the odds of being female
(i.e., female = 1). The expb option on the model statement tells
SAS to show the exponentiated coefficients (i.e., the odds ratios).

```
proc logistic data = "c:/mydata/hsb2" desc;
class prog schtyp;
model female = prog schtyp prog*schtyp / expb;
run;
```

The LOGISTIC Procedure

Model Fit Statistics

Intercept
Intercept and
Criterion Only Covariates
AIC 277.637 284.490
SC 280.935 304.280
-2 Log L 275.637 272.490

Testing Global Null Hypothesis: BETA=0

Test Chi-Square DF Pr > ChiSq
Likelihood Ratio 3.1467 5 0.6774
Score 2.9231 5 0.7118
Wald 2.6036 5 0.7608

Type III Analysis of Effects

Wald
Effect DF Chi-Square Pr > ChiSq
prog 2 1.1232 0.5703
schtyp 1 0.4132 0.5203
prog*schtyp 2 2.4740 0.2903

Analysis of Maximum Likelihood Estimates

Standard Wald

Parameter DF Estimate Error Chi-Square Pr > ChiSq Exp(Est)

Intercept 1 0.3331 0.3164 1.1082 0.2925 1.395

prog 1 1 0.4459 0.4568 0.9532 0.3289 1.562

prog 2 1 -0.1964 0.3438 0.3264 0.5678 0.822

schtyp 1 1 -0.2034 0.3164 0.4132 0.5203 0.816

prog*schtyp 1 1 1 -0.6269 0.4568 1.8838 0.1699 0.534

prog*schtyp 2 1 1 0.3400 0.3438 0.9783 0.3226 1.405

The results indicate that the overall model is not statistically significant (LR chi2 = 3.1467, p = 0.6774). Furthermore, none of the coefficients are statistically significant either. In addition, there is no statistically significant effect of program (p = 0.5703), school type (p = 0.5203) or of the interaction (p = 0.2903).

Correlation

A correlation is useful when you want to see the linear relationship between two (or more) normally distributed interval variables. For example,

using the hsb2 data file we can run a correlation between two continuous variables, read and write.

```
proc corr data = "c:/mydata/hsb2";
var read write;
run;
```

The CORR Procedure

2 Variables: read write

Pearson Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

	read	write
read	1.00000	0.59678
reading score		<.0001
write	0.59678	1.00000
writing score	<.0001	

In the second example below, we will run a correlation between a dichotomous variable, female,

and a continuous variable, write. Although it is assumed that the variables are interval and normally distributed, we can include dummy variables when performing correlations.

```
proc corr data = "c:/mydata/hsb2";
var female write;
run;
```

The CORR Procedure

2 Variables: female write

Pearson Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

	female	write
female	1.00000	0.25649
		0.0002
write	0.25649	1.00000
writing score	0.0002	

**In the first example above, we see that the correlation between read and write
is 0.59678. By squaring the correlation and then multiplying by 100, you can
determine what percentage of the variability is shared. Let's round
0.59678 to be
0.6, which when squared would be .36, multiplied by 100 would be 36%. Hence read
shares about 36% of its variability with write. In the output for the second
example, we can see the correlation between write and female is
0.25649. Squaring this number yields .0657871201, meaning that female shares
approximately 6.5% of its variability with write.**

Simple linear regression

**Simple linear regression allows us to look at the linear relationship between one
normally distributed interval predictor and one normally distributed interval outcome
variable. For example, using the hsb2 data file, say we wish to**

look at the relationship between writing scores (write) and reading scores (read);
in other words, predicting write from read.

```
proc reg data = "c:/mydata/hsb2";
model write = read / stb;
run;
quit;
```

The REG Procedure
Model: MODEL1
Dependent Variable: write writing score

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6367.42127	6367.42127	109.52	<.0001
Error	198	11511	58.13866		
Corrected Total	199	17879			

Root MSE	7.62487	R-Square	0.3561
Dependent Mean	52.77500	Adj R-Sq	0.3529
Coeff Var	14.44788		

Parameter Estimates

Parameter Standard Standardized
Variable Label DF Estimate Error t Value Pr > |t| Estimate
Intercept Intercept 1 23.95944 2.80574 8.54 <.0001 0
read reading score 1 0.55171 0.05272 10.47 <.0001 0.59678

We see that the relationship between write and read is positive
(.55171)
and based on the t-value (10.47) and p-value (0.000), we conclude this
relationship is statistically significant. Hence, there is a statistically significant positive linear relationship between reading and writing.

Non-parametric correlation

A Spearman correlation is used when one or both of the variables are not assumed to be
normally distributed and interval (but are assumed to be ordinal). The values of the
variables are converted in ranks and then correlated. In

our example, we will look
for a relationship between read and write. We will not assume that
both of these variables are normal and interval. The spearman
option on the proc corr statement is used to tell SAS to perform a
Spearman rank correlation instead of a Pearson correlation.

```
proc corr data = "c:/mydata/hsb2" spearman;
var read write;
run;
```

The CORR Procedure

2 Variables: read write

Spearman Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

	read	write
read	1.00000	0.61675
reading score		<.0001

write 0.61675 1.00000
writing score <.0001

The results suggest that the relationship between read and write
(rho = 0.61675, p = 0.000) is statistically significant.

Simple logistic regression

Logistic regression assumes that the outcome variable is binary (i.e., coded as 0 and
1). We have only one variable in the hsb2 data file that is coded
0 and 1, and that is female. We understand that female is a silly
outcome variable (it would make more sense to use it as a predictor variable), but we can
use female as the outcome variable to illustrate how the code for this command is
structured and how to interpret the output. The first variable listed on
the model statement is the outcome (or dependent) variable, and all of the rest of
the variables are listed after the equals sign and are predictor (or independent) variables. You can use the

expb option on the model statement if you want to see the odds ratios. In our example, female will be the outcome
variable, and read will be the predictor variable. As with OLS regression,
the predictor variables must be either dichotomous or continuous; they cannot be
categorical.

```
proc logistic data = "c:/mydata/hsb2" desc;
model female = read / expb;
run;
```

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	0.7261	0.7420	0.9577	0.3278	2.067
read	1	-0.0104	0.0139	0.5623	0.4533	0.990

Odds Ratio Estimates

Point 95% Wald
Effect Estimate Confidence Limits
read 0.990 0.963 1.017

Association of Predicted Probabilities and Observed Responses

Percent Concordant 50.3 Somers' D 0.069
Percent Discordant 43.4 Gamma 0.073
Percent Tied 6.3 Tau-a 0.034
Pairs 9919 c 0.534

The results indicate that reading score (read) is not a statistically
significant predictor of gender (i.e., being female), Wald chi-square =
0.5623, p = 0.4533.

Multiple regression

Multiple regression is very similar to simple regression, except that in multiple
regression you have more than one predictor variable in the equation. For example,
using the hsb2 data file we will predict writing score

from gender (female),
reading, math, science and social studies (socst) scores. The stb
option on the model statement tells SAS to display the standardized
regression coefficients (seen on the far right of the output).

```
proc reg data = "c:/mydata/hsb2";
model write = female read math science socst / stb;
run;
quit;
```

The REG Procedure
Model: MODEL1
Dependent Variable: write writing score

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	10757	2151.38488	58.60	<.0001
Error	194	7121.95060	36.71109		
Corrected Total	199	17879			

Root MSE 6.05897 R-Square 0.6017
Dependent Mean 52.77500 Adj R-Sq 0.5914
Coeff Var 11.48075

Parameter Estimates

Parameter Standard Standardized
Variable Label DF Estimate Error t Value Pr > |t| Estimate
Intercept Intercept 1 6.13876 2.80842 2.19 0.0300 0
female 1 5.49250 0.87542 6.27 <.0001 0.28928
read reading score 1 0.12541 0.06496 1.93 0.0550 0.13566
math math score 1 0.23807 0.06713 3.55 0.0005 0.23531
science science score 1 0.24194 0.06070 3.99 <.0001 0.25272
socst social studies score 1 0.22926 0.05284 4.34 <.0001 0.25967

The results indicate that the overall model is statistically significant (F = 58.60, p = 0.0001). Furthermore, all of the predictor variables are statistically significant except for read.

Analysis of covariance

**Analysis of covariance is like ANOVA, except in addition to the categorical predictors
you have continuous predictors as well. For example, the one
way ANOVA example used write as the dependent variable and prog as the
independent variable. Let's add read as a continuous variable to this
model.**

```
proc glm data = "c:/mydata/hsb2";
class prog;
model write = prog read;
run;
quit;
```

The GLM Procedure

Dependent Variable: write writing score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7017.68123	2339.22708	42.21	<.0001

Error 196 10861.19377 55.41425
Corrected Total 199 17878.87500

R-Square Coeff Var Root MSE write Mean
0.392512 14.10531 7.444075 52.77500

Source DF Type I SS Mean Square F Value Pr > F
prog 2 3175.697857 1587.848929 28.65 <.0001
read 1 3841.983376 3841.983376 69.33 <.0001

Source DF Type III SS Mean Square F Value Pr > F
prog 2 650.259965 325.129983 5.87 0.0034
read 1 3841.983376 3841.983376 69.33 <.0001

The results indicate that even after adjusting for reading score (read), writing
scores still significantly differ by program type (prog) F = 5.87, p = 0.0034.

Multiple logistic regression

Multiple logistic regression is like simple logistic regression, except that there are
two or more predictors. The predictors can be interval variables or dummy variables,
but cannot be categorical variables. If you have

categorical predictors, they should be coded into one or more dummy variables. We have only one variable in our data set that is coded 0 and 1, and that is female. We understand that female is a silly outcome variable (it would make more sense to use it as a predictor variable), but we can use female as the outcome variable to illustrate how the code for this command is structured and how to interpret the output. In our example, female will be the outcome variable, and read and write will be the predictor variables. The desc option on the proc logistic statement is necessary so that SAS models the probability of being female (i.e., female = 1). The expb option on the model statement tells SAS to display the exponentiated coefficients (i.e., the odds ratios).

```
proc logistic data = "c:/mydata/hsb2" desc;
model female = read write / expb;
```

run;

The LOGISTIC Procedure

Model Information

Data Set WORK.HSB2
Response Variable female
Number of Response Levels 2
Number of Observations 200
Model binary logit
Optimization Technique Fisher's scoring

Response Profile

Ordered Total
Value female Frequency
1 1 109
2 0 91

Probability modeled is female=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Intercept
Intercept and
Criterion Only Covariates
AIC 277.637 253.818
SC 280.935 263.713
-2 Log L 275.637 247.818

Testing Global Null Hypothesis: BETA=0

Test Chi-Square DF Pr > ChiSq
Likelihood Ratio 27.8186 2 <.0001
Score 26.3588 2 <.0001
Wald 23.4135 2 <.0001

Analysis of Maximum Likelihood Estimates

Standard Wald
Parameter DF Estimate Error Chi-Square Pr > ChiSq Exp(Est)
Intercept 1 -1.7061 0.9234 3.4137 0.0647 0.182
read 1 -0.0710 0.0196 13.1251 0.0003 0.931
write 1 0.1064 0.0221 23.0748 <.0001 1.112

Odds Ratio Estimates

Point 95% Wald
Effect Estimate Confidence Limits
read 0.931 0.896 0.968
write 1.112 1.065 1.162

Association of Predicted Probabilities and Observed Responses

Percent Concordant 69.3 Somers' D 0.396
Percent Discordant 29.7 Gamma 0.400
Percent Tied 1.0 Tau-a 0.198
Pairs 9919 c 0.698

These results show that both read (Wald chi-square = 13.1251, p = 0.0003) and write (Wald chi-square = 23.0748, p
= 0.0001) are
significant predictors of female.

Discriminant analysis

Discriminant analysis is used when you have one or more normally
distributed interval independent

**variables and a categorical dependent variable. It is a multivariate technique that
considers the latent dimensions in the independent variables for predicting group
membership in the categorical dependent variable. For example, using the hsb2 data file, say we wish to use read, write and math
scores to predict the type of program (prog) to which a student belongs.**

```
proc discrim data = "c:/mydata/hsb2" can;
class prog;
var read write math;
run;
```

```
The SAS System
The DISCRIM Procedure

Observations 200 DF Total 199
Variables 3 DF Within Classes 197
Classes 3 DF Between Classes 2

Class Level Information

Variable Prior
```

prog Name Frequency Weight Proportion Probability

1 _1 45 45.0000 0.225000 0.333333

2 _2 105 105.0000 0.525000 0.333333

3 _3 50 50.0000 0.250000 0.333333

Pooled Covariance Matrix Information

Natural Log of the
Covariance Determinant of the
Matrix Rank Covariance Matrix

3 12.18440

Pairwise Generalized Squared Distances Between Groups

2 _ _ -1 _ _
D (i|j) = (X - X)' COV (X - X)
i j i j

Generalized Squared Distance to prog

From
prog 1 2 3

1 0 0.73810 0.31771

2 0.73810 0 1.90746

3 0.31771 1.90746 0

Canonical Discriminant Analysis

Adjusted Approximate Squared
Canonical Canonical Standard Canonical
Correlation Correlation Error Correlation
1 0.512534 0.502546 0.052266 0.262691
2 0.067247 0.031181 0.070568 0.004522

Test of H0: The canonical correlations in
the current row and all
Eigenvalues of Inv(E)*H that follow are zero
= CanRsq/(1-CanRsq)
Likelihood Approximate
Eigenvalue Difference Proportion Cumulative Ratio F Value Num DF Den DF Pr > F
1 0.3563 0.3517 0.9874 0.9874 0.73397507 10.87 6 390 <.0001
2 0.0045 0.0126 1.0000 0.99547788 0.45 2 196 0.6414

Total Canonical Structure

Variable Label Can1 Can2
read reading score 0.822122 -0.167318

write writing score 0.818851 0.572893
math math score 0.933429 -0.239761

Between Canonical Structure

Variable Label Can1 Can2
read reading score 0.999644 -0.026693
write writing score 0.995813 0.091410
math math score 0.999433 -0.033682

Pooled Within Canonical Structure

Variable Label Can1 Can2
read reading score 0.778465 -0.184093
write writing score 0.775344 0.630310
math math score 0.912889 -0.272463

Total-Sample Standardized Canonical Coefficients

Variable Label Can1 Can2
read reading score 0.299373057 -0.449624188
write writing score 0.363246854 1.298397979
math math score 0.659035164 -0.743012325

Pooled Within-Class Standardized Canonical Coefficients

Variable Label Can1 Can2

read reading score 0.272852441 -0.409793246

write writing score 0.331078354 1.183414147

math math score 0.581553807 -0.655657953

Raw Canonical Coefficients

Variable Label Can1 Can2

read reading score 0.0291987615 -.0438532096

write writing score 0.0383228947 0.1369822435

math math score 0.0703462492 -.0793100780

Class Means on Canonical Variables

prog Can1 Can2

1 -.3120021323 0.1190423066

2 0.5358514591 -.0196809384

3 -.8444861449 -.0658081053

Linear Discriminant Function

_ -1 _ -1 _

Constant = -.5 X' COV X Coefficient Vector = COV X

j j j

Linear Discriminant Function for prog

Variable Label 1 2 3

Constant -24.47383 -30.60364 -20.77468

read reading score 0.18195 0.21279 0.17451

write writing score 0.38572 0.39921 0.33999

math math score 0.40171 0.47236 0.37891

Generalized Squared Distance Function

2 _ -1 _

D (X) = (X-X)' COV (X-X)

j j j

Posterior Probability of Membership in Each prog

2 2

Pr(j|X) = exp(-.5 D (X)) / SUM exp(-.5 D (X))

j k k

Number of Observations and Percent Classified into prog

From

prog 1 2 3 Total

1 11 17 17 45

24.44 37.78 37.78 100.00

2 18 68 19 105
17.14 64.76 18.10 100.00

3 14 7 29 50
28.00 14.00 58.00 100.00

Total 43 92 65 200
21.50 46.00 32.50 100.00

Priors 0.33333 0.33333 0.33333

Error Count Estimates for prog

1 2 3 Total
Rate 0.7556 0.3524 0.4200 0.5093
Priors 0.3333 0.3333 0.3333

Clearly, the SAS output for this procedure is quite lengthy, and it is
beyond the scope of this page to explain all of it. However, the main
point is that two canonical variables are identified by the analysis, the
first of which seems to be more related to program type than the second.

One-way MANOVA

MANOVA (multivariate analysis of variance) is like ANOVA, except that there are two or more dependent variables. In a one-way MANOVA, there is one categorical independent variable and two or more dependent variables. For example, using the hsb2 data file, say we wish to examine the differences in read, write and math broken down by program type (prog). The manova statement is necessary in the proc glm to tell SAS to conduct a MANOVA. The h= on the manova statement is used to specify the hypothesized effect.

```
proc glm data = "c:/mydata/hsb2";
class prog;
model read write math = prog;
manova h=prog;
run;
quit;
```

The GLM Procedure

Dependent Variable: read reading score

Sum of
Source DF Squares Mean Square F Value Pr > F
Model 2 3716.86127 1858.43063 21.28 <.0001
Error 197 17202.55873 87.32263
Corrected Total 199 20919.42000

R-Square Coeff Var Root MSE read Mean
0.177675 17.89136 9.344658 52.23000

Source DF Type I SS Mean Square F Value Pr > F
prog 2 3716.861270 1858.430635 21.28 <.0001

Source DF Type III SS Mean Square F Value Pr > F
prog 2 3716.861270 1858.430635 21.28 <.0001

Dependent Variable: write writing score

Sum of
Source DF Squares Mean Square F Value Pr > F
Model 2 3175.69786 1587.84893 21.27 <.0001
Error 197 14703.17714 74.63542

Corrected Total 199 17878.87500

R-Square Coeff Var Root MSE write Mean
0.177623 16.36983 8.639179 52.77500

Source DF Type I SS Mean Square F Value Pr > F
prog 2 3175.697857 1587.848929 21.27 <.0001

Source DF Type III SS Mean Square F Value Pr > F
prog 2 3175.697857 1587.848929 21.27 <.0001

Dependent Variable: math math score
Sum of
Source DF Squares Mean Square F Value Pr > F
Model 2 4002.10389 2001.05194 29.28 <.0001
Error 197 13463.69111 68.34361
Corrected Total 199 17465.79500

R-Square Coeff Var Root MSE math Mean
0.229140 15.70333 8.267019 52.64500

Source DF Type I SS Mean Square F Value Pr > F
prog 2 4002.103889 2001.051944 29.28 <.0001

Source DF Type III SS Mean Square F Value Pr > F

prog 2 4002.103889 2001.051944 29.28 <.0001

Multivariate Analysis of Variance

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for prog
E = Error SSCP Matrix

Characteristic Characteristic Vector V'EV=1
Root Percent read write math
0.35628297 98.74 0.00208033 0.00273039 0.00501196
0.00454266 1.26 -0.00312441 0.00975958 -0.00565061
0.00000000 0.00 -0.00904826 0.00054800 0.00823531

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall prog Effect
H = Type III SSCP Matrix for prog
E = Error SSCP Matrix

S=2 M=0 N=96.5

Statistic Value F Value Num DF Den DF Pr > F
Wilks' Lambda 0.73397507 10.87 6 390 <.0001
Pillai's Trace 0.26721285 10.08 6 392 <.0001

Hotelling-Lawley Trace 0.36082563 11.70 6 258.23 <.0001
Roy's Greatest Root 0.35628297 23.28 3 196 <.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

This command produces four different test statistics that are used to evaluate the
statistical significance of the relationship between the independent variable and the
outcome variables. According to all four criteria, the students in the different
programs differ in their joint distribution of read, write and math.

Multivariate multiple regression

Multivariate multiple regression is used when you have two or more
dependent variables that are
to be predicted from two or more predictor variables. In our example, we will
predict write and read from female, math, science and social studies (socst) scores. The mtest statement in

the proc reg is used to test hypotheses in multivariate regression
models where there are several independent variables fit to the same dependent variables. If no equations or options are specified, the mtest
statement tests the hypothesis that all estimated parameters except the
intercept are zero. In other words, the multivariate tests test whether
the independent variable specified predicts the dependent
variables together, holding all of the other independent variables
constant. You can put a label in front of the mtest statement to
aid in the interpretation of the output (this is particularly useful when you have multiple
mtest statements).

```
proc reg data = "c:/mydata/hsb2";
model write read = female math science socst;
female: mtest female;
math: mtest math;
science: mtest science;
```

socst: mtest socst;

run;

quit;

The REG Procedure

Model: MODEL1

Dependent Variable: write writing score

Analysis of Variance

Sum of Mean

Source DF Squares Square F Value Pr > F

Model 4 10620 2655.02312 71.32 <.0001

Error 195 7258.78251 37.22453

Corrected Total 199 17879

Root MSE 6.10119 R-Square 0.5940

Dependent Mean 52.77500 Adj R-Sq 0.5857

Coeff Var 11.56076

Parameter Estimates

Parameter Standard

Variable Label DF Estimate Error t Value Pr > |t|

Intercept Intercept 1 6.56892 2.81908 2.33 0.0208

female 1 5.42822 0.88089 6.16 <.0001
math math score 1 0.28016 0.06393 4.38 <.0001
science science score 1 0.27865 0.05805 4.80 <.0001
socst social studies score 1 0.26811 0.04919 5.45 <.0001

Model: MODEL1
Dependent Variable: read reading score

Analysis of Variance

Sum of Mean
Source DF Squares Square F Value Pr > F
Model 4 12220 3054.91459 68.47 <.0001
Error 195 8699.76166 44.61416
Corrected Total 199 20919

Root MSE 6.67938 R-Square 0.5841
Dependent Mean 52.23000 Adj R-Sq 0.5756
Coeff Var 12.78840

Parameter Estimates

Parameter Standard
Variable Label DF Estimate Error t Value Pr > |t|
Intercept Intercept 1 3.43000 3.08624 1.11 0.2678
female 1 -0.51261 0.96436 -0.53 0.5956

math math score 1 0.33558 0.06999 4.79 <.0001
science science score 1 0.29276 0.06355 4.61 <.0001
socst social studies score 1 0.30976 0.05386 5.75 <.0001

Model: MODEL1
Multivariate Test: female

Multivariate Statistics and Exact F Statistics

S=1 M=0 N=96

Statistic Value F Value Num DF Den DF Pr > F
Wilks' Lambda 0.83011470 19.85 2 194 <.0001
Pillai's Trace 0.16988530 19.85 2 194 <.0001
Hotelling-Lawley Trace 0.20465280 19.85 2 194 <.0001
Roy's Greatest Root 0.20465280 19.85 2 194 <.0001

Model: MODEL1
Multivariate Test: math

Multivariate Statistics and Exact F Statistics

S=1 M=0 N=96

Statistic Value F Value Num DF Den DF Pr > F
Wilks' Lambda 0.84006791 18.47 2 194 <.0001

Pillai's Trace 0.15993209 18.47 2 194 <.0001
Hotelling-Lawley Trace 0.19037995 18.47 2 194 <.0001
Roy's Greatest Root 0.19037995 18.47 2 194 <.0001

Model: MODEL1
Multivariate Test: science

Multivariate Statistics and Exact F Statistics

S=1 M=0 N=96

Statistic Value F Value Num DF Den DF Pr > F
Wilks' Lambda 0.83357462 19.37 2 194 <.0001
Pillai's Trace 0.16642538 19.37 2 194 <.0001
Hotelling-Lawley Trace 0.19965265 19.37 2 194 <.0001
Roy's Greatest Root 0.19965265 19.37 2 194 <.0001

Model: MODEL1
Multivariate Test: socst

Multivariate Statistics and Exact F Statistics

S=1 M=0 N=96

Statistic Value F Value Num DF Den DF Pr > F
Wilks' Lambda 0.77932902 27.47 2 194 <.0001

Pillai's Trace 0.22067098 27.47 2 194 <.0001
Hotelling-Lawley Trace 0.28315509 27.47 2 194 <.0001
Roy's Greatest Root 0.28315509 27.47 2 194 <.0001

With regard to the univariate tests, each of the independent variables is statistically significant predictor for writing. All of the independent variables are also statistically significant predictors for reading except female (t = -0.53, p = 0.5956). All of the multivariate tests are also statistically significant.

Canonical correlation

Canonical correlation is a multivariate technique used to examine the relationship between two groups of variables. For each set of variables, it creates latent variables and looks at the relationships among the latent variables. It assumes that all variables in the model are interval and normally distributed. In SAS, one group of variables is placed on the var statement and the other

group on the with statement. There need not be an equal number of variables in the two groups. The all option on
the proc cancorr statement provides additional output that many researchers
might find useful.

```
proc cancorr data = "c:/mydata/hsb2" all;
var read write;
with math science;
run;
```

The CANCORR Procedure

VAR Variables 2
WITH Variables 2
Observations 200

Means and Standard Deviations

Variable	Mean	Standard Deviation	Label
read	52.230000	10.252937	reading score
write	52.775000	9.478586	writing score
math	52.645000	9.368448	math score

science 51.850000 9.900891 science score

Correlations Among the Original Variables

Correlations Among the VAR Variables

read write
read 1.0000 0.5968
write 0.5968 1.0000

Correlations Among the WITH Variables

math science
math 1.0000 0.6307
science 0.6307 1.0000

Correlations Between the VAR Variables and the WITH Variables

math science
read 0.6623 0.6302
write 0.6174 0.5704

Canonical Correlation Analysis

Adjusted Approximate Squared

Canonical Canonical Standard Canonical Correlation Correlation Error Correlation
1 0.772841 0.771003 0.028548 0.597283
2 0.023478 . 0.070849 0.000551

Test of H0: The canonical correlations in the current row and all Eigenvalues of Inv(E)*H that follow are zero = CanRsq/(1-CanRsq)

Likelihood Approximate Eigenvalue Difference Proportion Cumulative Ratio F Value Num DF Den DF Pr > F
1 1.4831 1.4826 0.9996 0.9996 0.40249498 56.47 4 392 <.0001
2 0.0006 0.0004 1.0000 0.99944876 0.11 1 197 0.7420

Multivariate Statistics and F Approximations

S=2 M=-0.5 N=97

Statistic Value F Value Num DF Den DF Pr > F
Wilks' Lambda 0.40249498 56.47 4 392 <.0001
Pillai's Trace 0.59783426 42.00 4 394 <.0001
Hotelling-Lawley Trace 1.48368501 72.58 4 234.16 <.0001
Roy's Greatest Root 1.48313347 146.09 2 197 <.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Raw Canonical Coefficients for the VAR Variables

V1 V2

read reading score 0.063261313 0.1037907932

write writing score 0.0492491834 -0.12190836

Raw Canonical Coefficients for the WITH Variables

W1 W2

math math score 0.0669826768 -0.120142451

science science score 0.0482406314 0.1208859811

Standardized Canonical Coefficients for the VAR Variables

V1 V2

read reading score 0.6486 1.0642

write writing score 0.4668 -1.1555

Standardized Canonical Coefficients for the WITH Variables

W1 W2

math math score 0.6275 -1.1255

science science score 0.4776 1.1969

Canonical Structure

Correlations Between the VAR Variables and Their Canonical Variables

V1 V2

read reading score 0.9272 0.3746

write writing score 0.8539 -0.5205

Correlations Between the WITH Variables and Their Canonical Variables

W1 W2

math math score 0.9288 -0.3706

science science score 0.8734 0.4870

Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables

W1 W2

read reading score 0.7166 0.0088

write writing score 0.6599 -0.0122

Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables

V1 V2
math math score 0.7178 -0.0087
science science score 0.6750 0.0114

Canonical Redundancy Analysis

Raw Variance of the VAR Variables Explained by Their Own The Opposite
Canonical Variables Canonical Variables
Canonical
Variable Cumulative Canonical Cumulative
Number Proportion Proportion R-Square Proportion Proportion
1 0.7995 0.7995 0.5973 0.4775 0.4775
2 0.2005 1.0000 0.0006 0.0001 0.4777

Raw Variance of the WITH Variables Explained by Their Own The Opposite
Canonical Variables Canonical Variables
Canonical

Variable Cumulative Canonical Cumulative
Number Proportion Proportion R-Square Proportion Proportion
1 0.8100 0.8100 0.5973 0.4838 0.4838
2 0.1900 1.0000 0.0006 0.0001 0.4839

Standardized Variance of the VAR Variables Explained by
Their Own The Opposite
Canonical Variables Canonical Variables
Canonical
Variable Cumulative Canonical Cumulative
Number Proportion Proportion R-Square Proportion Proportion
1 0.7944 0.7944 0.5973 0.4745 0.4745
2 0.2056 1.0000 0.0006 0.0001 0.4746

Standardized Variance of the WITH Variables Explained by
Their Own The Opposite
Canonical Variables Canonical Variables
Canonical
Variable Cumulative Canonical Cumulative
Number Proportion Proportion R-Square Proportion

Proportion

1 0.8127 0.8127 0.5973 0.4854 0.4854

2 0.1873 1.0000 0.0006 0.0001 0.4855

Squared Multiple Correlations Between the VAR Variables and
the First M Canonical Variables of the WITH Variables

M 1 2

read reading score 0.5135 0.5136

write writing score 0.4355 0.4356

Squared Multiple Correlations Between the WITH Variables
and the First M Canonical Variables of the VAR Variables

M 1 2

math math score 0.5152 0.5153

science science score 0.4557 0.4558

The output above shows the linear combinations corresponding to the first canonical
correlation. At the bottom of the output are the two canonical correlations.

These results indicate that the first canonical correlation is .772841. The F-test in this output tests the hypothesis that the first canonical correlation is equal to zero. Clearly, F = 56.47 is statistically significant. However, the second canonical correlation of .0235 is not statistically significantly different from zero (F = 0.11, p = 0.7420).

Factor analysis

Factor analysis is a form of exploratory multivariate analysis that is used to either reduce the number of variables in a model or to detect relationships among variables. All variables involved in the factor analysis need to be continuous and are assumed to be normally distributed. The goal of the analysis is to try to identify factors which underlie the variables. There may be fewer factors than variables, but there may not be more factors than variables. For our example, let's suppose that we think that there are some common factors underlying the various test

scores. We will use the principal components method of extraction,
use a varimax rotation, extract two factors and obtain a scree plot of the
eigenvalues. All of these options are listed on the proc factor
statement.

```
proc factor data = "c:/mydata/hsb2" method=principal
rotate=varimax nfactors=2 scree;
var read write math science socst;
run;
```

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1

Eigenvalue	Difference	Proportion	Cumulative
1	3.38081982	2.82344156	0.6762	0.6762
2	0.55737826	0.15058550	0.1115	0.7876
3	0.40679276	0.05062495	0.0814	0.8690

4 0.35616781 0.05732645 0.0712 0.9402

5 0.29884136 0.0598 1.0000

2 factors will be retained by the NFACTOR criterion.

The FACTOR Procedure

Initial Factor Method: Principal Components

Factor Pattern

Factor1 Factor2

READ reading score 0.85760 -0.02037

WRITE writing score 0.82445 0.15495

MATH math score 0.84355 -0.19478

SCIENCE science score 0.80091 -0.45608

SOCST social studies score 0.78268 0.53573

Variance Explained by Each Factor

Factor1 Factor2

3.3808198 0.5573783

Final Communality Estimates: Total = 3.938198

READ WRITE MATH SCIENCE SOCST

0.73589906 0.70373337 0.74951854 0.84945810 0.89958900

The FACTOR Procedure

Rotation Method: Varimax

Orthogonal Transformation Matrix

1 2

1 0.74236 0.67000

2 -0.67000 0.74236

Rotated Factor Pattern

Factor1 Factor2

READ reading score 0.65029 0.55948

WRITE writing score 0.50822 0.66742

MATH math score 0.75672 0.42058

SCIENCE science score 0.90013 0.19804

SOCST social studies score 0.22209 0.92210

Variance Explained by Each Factor

Factor1 Factor2

2.1133589 1.8248392

Final Communality Estimates: Total = 3.938198

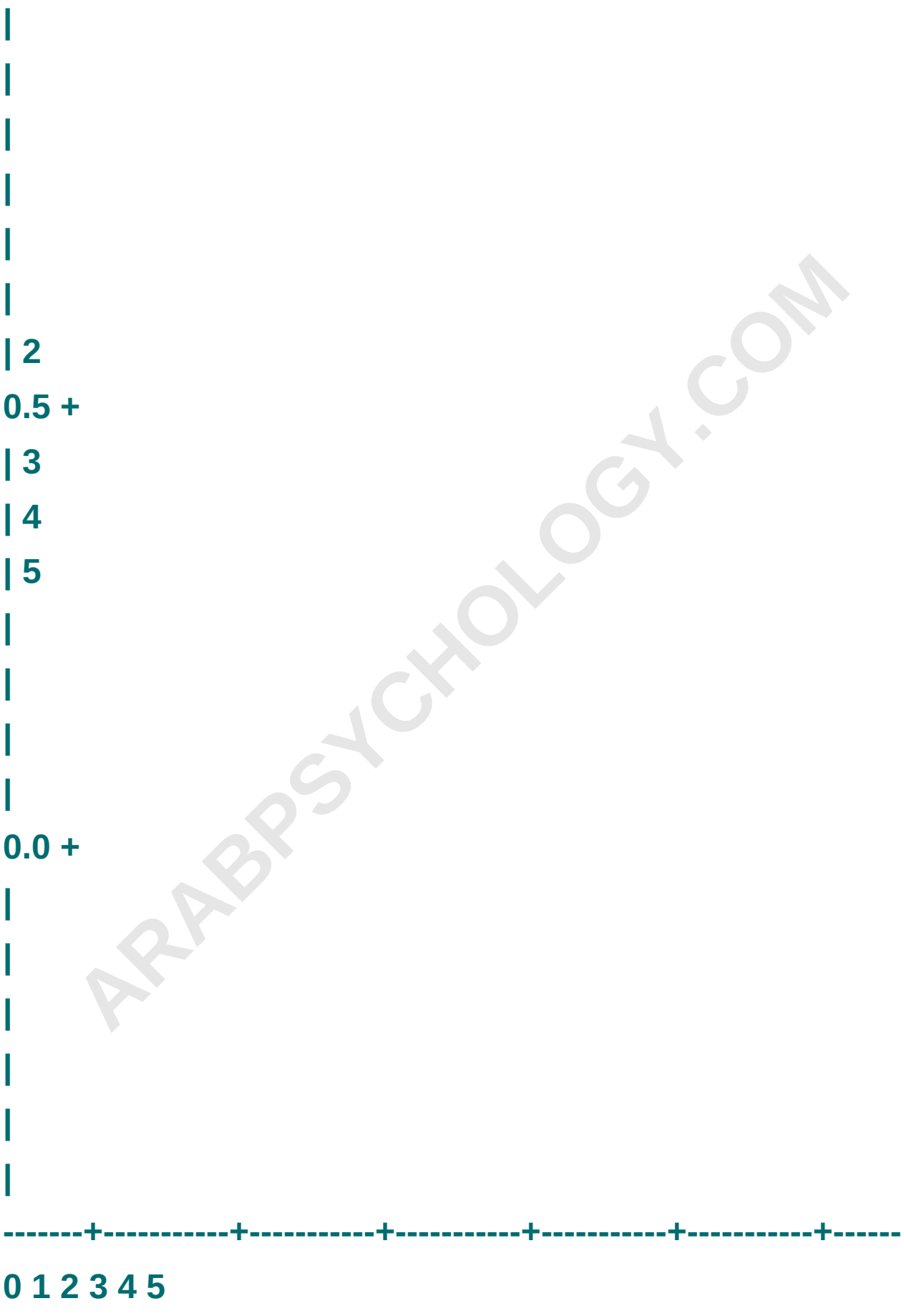
READ WRITE MATH SCIENCE SOCST

0.73589906 0.70373337 0.74951854 0.84945810

0.89958900

Communality (which is the opposite of uniqueness) is the proportion of variance of the variable (i.e., read) that is accounted for by all of the factors taken together, and a very low communality can indicate that a variable may not belong with any of the factors. From the factor pattern table, we can see that all five of the test scores load onto the first factor, while all five tend to load not so heavily on the second factor. The purpose of rotating the factors is to get the variables to load either very high or very low on each factor. In this example, because all of the variables loaded onto factor 1 and not on factor 2, the rotation did not aid in the interpretation. Instead, it made the results even more difficult to interpret. The scree plot may be useful in determining how many factors to retain.

Screen Plot of Eigenvalues



Number