Introduction to SAS
Basic programming and procedures

John P. Connolly

Office of Information Technology
The University of Texas at Arlington
connolly@uta.edu

Oct 11, 2018
Table of contents

1 Introduction

2 Basic Data Analysis
   - Using PROC PRINT, PROC SORT, and PROC MERGE
   - Using PROC FREQ and PROC MEANS

3 Basic Statistical Procedures
   - Using PROC UNIVARIATE
   - Using PROC CORR/ PLOT/ REG/ GLM/ TTEST
   - Performing a Chi Square test
The SAS slides can be downloaded at
https://mavspace.uta.edu:
443/people/c/co/connolly/SASfiles.
Developed throughout the 1960's by graduate students at North Carolina State University.

SAS Institute, Inc., incorporated in 1976 by Barr, Goodnight, Sall, and Helwig.
Product information

  - SAS/STAT, SAS/ETS, SAS/IML, SAS/QC.
Availability

- Available in the some OIT labs.
- Available for purchase by UT students, faculty, and staff.
  - Each license costs $85.00 and may only be installed on (1) machine.
  - Licenses are valid for the fiscal year beginning September 1 and ending August 31.
  - A UT Electronic ID (UT-EID) is required to use this on-line service. To obtain a UT-EID go to http://www.utexas.edu/eid.
Writing a SAS program

- **Two key steps:**
  - **Data step**
    - Instructs SAS how to process your data
      - where to find it . . .
      - what variables it contains . . .
      - how it’s formatted, etc.
  - **Proc step**
    - Tells SAS to perform a specific type of analysis on the data
      - e.g., `proc means` will compute means.
The Data Step

- **Data** name <Options> ;
- Optional Data Step commands, e.g., *input*, *infile*, *keep*, *drop*, . . .
- Optional Data programming statements, e.g., Age_in_months = age*12;
- Optional *instream* data:
  - **Datalines**;
  - . . . rows of data . . .
  - ;
- **Run**;
Simple SAS program

```
proc means;
var essay midterm final;
run;
```

```
title "My First SAS Program Ever";

data student_scores;
  input name $ essay midterm final;
  datalines;
  james 50 90 83
  sarah 55 79 80
  rebecca 59 98 97
  robert 48 66 76
  charles 54 78 90
;
run;
```

```
title "Means of Students' Scores";
```

---

**Introduction**

Basic Data Analysis
Basic Statistical Procedures

John P. Connolly

Introduction to SAS
### Means of Students' Scores

- **essay**: N = 5, Mean = 53.200000, Std Dev = 4.3243497, Minimum = 48.0000000, Maximum = 59.0000000
- **midterm**: N = 5, Mean = 82.2000000, Std Dev = 12.2556110, Minimum = 66.0000000, Maximum = 98.0000000
- **final**: N = 5, Mean = 85.2000000, Std Dev = 8.3486526, Minimum = 76.0000000, Maximum = 97.0000000
Enter data directly into the SAS program file.
- This is known as instream data.

Read data from an external text file.
- Use infile statement to specify file path.
  - e.g. infile 'C:\healthdata\March2005.dat'.

SAS Dataset.
- A SAS dataset is a file previously created by SAS.
- Use libname statement to specify path to library that contains the SAS dataset.
- Use libname.filename to reference SAS dataset.
Infile Example

```sas
data crime_data;
  infile 'crimeData';
  input state $1-15 murder rape robbery assault burglary larceny auto;
run;

proc standard data=crime_data out=stand mean=0 std=1;
  var murder rape robbery assault burglary larceny auto;
run;

proc fastclus data=stand maxc=5 least=2 out=clus;
  id state;
  var murder rape robbery assault burglary larceny auto;
run;

proc candisc data=clus out=Can noprprint;
  class Cluster;
  var murder rape robbery assault burglary larceny auto;
run;

proc sort data=Can;
  by Cluster;
run;

proc print;
  var state Cluster Can1 Can2;
run;
```
Introduction to SAS

SAS dataset Example

```sas
data cd;
    infile 'CrimeData';
    input state $1-15 murder rape robbery assault burglary larceny auto;
run;
libname mylib './datalib';
data mylib.cd_sds;
    set cd;
run;

libname mylib './datalib';
proc standard data=mylib.cd_sds out=stand mean=0 std=1;
    var murder rape robbery assault burglary larceny auto;
run;
proc fastclus data=stand maxc=5 least=2 out=clus;
    id state;
    var murder rape robbery assault burglary larceny auto;
run;
proc candisc data=clus out=Can noprime:
    class Cluster;
    var murder rape robbery assault burglary larceny auto;
run;
ods graphics on;
proc sgplot data=Can:
    scatter y=Can2 x=Can1 / group=Cluster ;
run;
proc sort data=Can;
    by Cluster;
run;
proc print:
    var state Cluster Can1 Can2;
run;
```

John P. Connolly
Data Step Statements

- **Data** statement
  - Name the dataset
  - Specify any options.

- **Data step statements, e.g.,**
  - **Input** statement
    - Names each variable in your dataset
    - Identifies the format of the data
    - Not needed if the data is being read from a SAS data set.

- Optional data programming statements
  - Optional actions to carry out on your data, e.g.,
    - create new variables out of existing variables
    - $\text{BMI} = 700\times\text{weight}/(\text{height}\times\text{height})$;
Basic Input Styles

- **List**
  - Column numbers are not specified
  - Data values are separated by blanks
  - E.g., Input name $ exam_score;

- **Column**
  - Specify column positions of data
  - E.g., Input name $ 1-8 exam_score 11-12;

- **Formatted**
  - Specify the data type and field width of each input value
  - Input name $char8. +2 exam_score 6;
Common Optional Data Step Statements

- **Set** statement
  - Specifies a dataset that the current dataset will be based on.

- **Keep** statement
  - List variables to keep.

- **Drop** statement
  - List variables to drop.

- **Datalines** (or **Cards**)
  - Used only when your data is *instream*
  - Tells SAS that the next line of the program contains data.
Common Optional Data Step Programming Statements

- The statements are executed
  - Each time a record of data has been read in or processed by SAS.
- Create new variable, e.g:
  - ht_in_inches = ht_in_feet/12;
- Conditional execution of statements, e.g:
  - if gender='M' then . . .
- Loops
The SAS Internal Dataset

- The dataset that is created by the data step
- Provides the initial input for the proc statements
- Is temporary by default
  - But can be permanently saved as a SAS dataset.
- The default SAS dataset for procs
  - Is the one that was created by the most recently executed data step.
The SAS Data Step - Example

- Data cars;
  - infile 'C:\car_data.dat';
  - input make $ model $ num_cylinders mpg price;
  - run;

- Data cars2;
  - set cars;
  - keep make mpg price;
  - run;

- Proc means
  - var mpg price;
  - run;
The Proc Step

- Mainly used to carry out statistical analyses on your data
- Includes a **proc** statement
- Example procs:
  - **proc princomp**: performs principal components analysis
  - **proc factor**: performs factor analysis
  - **proc cluster**: performs cluster analysis.

Syntax:

- `proc proc-name <proc-options>;`
- `<proc step optional statements>`
- `run;`
Proc Statement Options

- Used to tailor the statistical analysis
- Many statement options are specific to a particular proc, e.g.,
  `proc factor priors=max;`
- A few options are common across a no. of different procs, e.g.,
  - The `data` option
  - The `plots` option
  - The `out` option
  - The `outstat` option.
Common Proc Step Optional Statements

- **var** var-list;
  - Specify the numeric variables to analyze
- **by** group-variable(s);
  - Perform separate analyses for each level of a group variable
- **class**
  - Identify variables that are to be treated as categorical in an analysis
- **model**
  - Specify a model to test.
You can access SAS manuals and user guides on the Internet.

Example search prompt:
- SAS <version no.> keyword(s)
- e.g., SAS 9.3 proc princomp cov
The SAS Windowing Environment

This is the environment in which you:

- Create and execute a SAS program
- Locate and fix errors
- View the results
- Save the SAS program and any desired results.
The SAS Windowing Environment

Provides the following windows:

- **Editor Window**
  - Where you enter the program.

- **Log Window**
  - Displays status and error messages regarding program execution.

- **Output Window**
  - Displays the results of the statistical analysis.

- **Results Window**
  - Organizes the results obtained from one or more analyses.

- **Explorer Window**
  - Allow the user to locate/access other files.
The SAS Windowing Environment - The Editor

- Libraries
- File Shortcuts
- Favorite Folders
- My Computer

Explorer

Contents of SAS Environment

Log - (Untitled)

OTE: Copyright (c) 2002-2010 by SAS Institute Inc., Cary, NC, USA.
OTE: SAS (r) Proprietary Software 9.3 (TS1M1)
LICENSED TO UNIVERSITY OF TEXAS AT AUSTIN-WORKSHOP TRAINING, SITE 70121965.
OTE: This session is executing on the XP_PRO platform.

OTE: Updated analytical products:
- INTROSTAT 9.3_M1, SAS/ETS 9.3 M1

OTE: SAS initialization used:
- real time: 0.01 seconds
- cpu time: 0.04 seconds

Editor - Untitled1 *

data try_it;
  input x y;
  dataraw:
    20 30
    40 50
    55 65
    76 32
  ;
  run:

proc print;
  run:

Results

Explorer

Output - (Untitled)

Log - (Untitled)

Editor - Untitled1 *

C:\Documents and Settings\dbs59
Ln 3, Col 10
The SAS Windowing Environment - Submitting programs
The SAS Windowing Environment - Displaying results
The SAS Windowing Environment - Saving programs

- Click on File > New Program
- Click on File > Save As
- Enter the name of the program
- Click on File > Save
OPTIONS LS=72;
TITLE 'CLASS EXAMPLE 1';
DATA ONE;
   INPUT NAME $ 1-4 AGE 6-7 HEIGHT 9-10 WEIGHT 12-14 ID 16-18 SEX $ 20;
CARDS;
    JOHN 20 72 180 001 M
    MARY 17 65 103 003 F
    BOB  25 76 200 004 M
    JO   7 36  50 002 F
    ERIC 15 67 135 006 M
    PHIL 18 70 195 005 M
;
PROC PRINT DATA=ONE;
run;
Example 2

OPTIONS LS=72;
TITLE 'CLASS EXAMPLE 2';
DATA ONE;
    INPUT NAME $ AGE HEIGHT WEIGHT ID SEX $ ;
CARDS;
JOHN 20 72 180 001 M
MARY 17 65 103 003 F
BOB  25 76 200 004 M
JO   7  36  50 002 F
ERIC 15 67 135 006 M
PHIL 18 70 195 005 M
;
PROC SORT;
    BY AGE;
*;
* PRINT THE DATA SET IN ORDER BY AGE;
*;
PROC PRINT DATA=ONE;
run;
OPTIONS LS=72;
TITLE 'CLASS EXAMPLE 3';
DATA ONE;
   INPUT NAME $ 1-4 AGE 6-7 HEIGHT 9-10 WEIGHT 12-14 ID 16-18 SEX $ 20;
CARDS;
JOHN 20 72 180 001 M 
MARY 17 65 003 F 
BOB  25 76 200 004 M 
JO  36 50 002 F 
ERIC 15 67 135 006 
PHIL 18 70 195 005 M 
;
PROC SORT;
   BY NAME;
PROC PRINT DATA=ONE;
DATA TWO;
   SET ONE;
IF(AGE GE 20);
PROC PRINT DATA=TWO;
run;
Example 4

OPTIONS LS=72;
TITLE 'CLASS EXAMPLE 4';
DATA ONE;	
   INPUT NAME $ AGE HEIGHT WEIGHT ID SEX $ ;
CARDS;
   JOHN 20 72 180 001 M
   MARY 17 65 103 003 F
   BOB 25 76 . 004 M
   JO . 36 50 002 F
   ERIC 15 67 135 006 .
   PHIL 18 70 195 005 M
;
PROC SORT DATA=ONE;	
   BY ID;
PROC PRINT DATA=ONE;

DATA TWO;	
   INPUT ID CLASS $;
CARDS;
   001 A
   002 B
   003 A
   004 C
   006 B
   005 A
;
contd. from previous slide...

PROC SORT DATA=TWO;
   BY ID;

DATA THREE;
   MERGE ONE TWO;
   BY ID;

PROC PRINT DATA=THREE;
run;
Example 5

OPTIONS LS=72;
TITLE 'CLASS EXAMPLE 5';

PROC FORMAT;
  VALUE OPINION 1='VERY BAD'
                2='BAD'
                3='NEUTRAL'
                4='GOOD'
                5='VERY GOOD'
                OTHER='DATA ERROR';
  VALUE $SX 'M'='MALE'
           'F'='FEMALE';

DATA ONE;
  INPUT NAME $ 1-4 AGE 6-7 HEIGHT 9-10 WEIGHT 12-14 ID 16-18 SEX $ 20
              @22 (Q1-Q10) (1.);
  CARDS;
  JOHN 20 72 180 001 M 1123215422
  MARY 17 65 003 F 1243533422
  BOB  25 76 200 004 M 2455324431
  JO   36 50 002 F 3211454432
  ERIC 15 67 135 006 2.24536533
  PHIL 18 70 195 005 M 13224355..
  ;
PROC PRINT DATA=ONE;
PROC FREQ DATA=ONE;
  TABLE Q1-Q10;
Example 5 - contd.

contd. from previous slide ...

PROC MEANS DATA=ONE;
   VAR AGE HEIGHT WEIGHT Q1-Q10;

DATA ONE;
   SET ONE;

LABEL
   Q1='TEACHING METHODS'
   Q2='CLASS INFORMATION'
   Q3='CLOTHING STYLE'
   Q4='ROOM ARRANGEMENT'
   Q5='NUMBER OF CHAIRS'
   Q6='VISUAL AIDS'
   Q7='BREAKS'
   Q8='REFRESHMENTS'
   Q9='HANDOUTS'
   Q10='ROOM TEMPERATURE';

FORMAT Q1-Q10 OPINION. SEX SX.;

PROC FREQ DATA=ONE;
   TABLES Q1-Q10 SEX;
run;
**Example 6**

```sas
OPTIONS LS=72;
TITLE 'CLASS EXAMPLE 6';
DATA ONE;
   INPUT NAME $ 1-4 AGE 6-7 HEIGHT 9-10 WEIGHT 12-14 ID 16-18 SEX $ 20
       @22 (Q1-Q10) (1.);
CARDS;
   JOHN 20 72 180 001 M 1123215422
   MARY 17 65 003 F 1243533422
   BOB 25 76 200 004 M 2455324431
   J0 36 50 002 F 3211454432
   ERIC 15 67 135 006 2.24536533
   PHIL 18 70 195 005 M 13224355..
   EDNA 21 67 155 007 F 4444555111
;
PROC SORT;
   BY SEX;
PROC UNIVARIATE DATA=ONE;
   VAR AGE HEIGHT;
   BY SEX;
run;
```
Example 7

OPTIONS LS=72;
TITLE 'CLASS EXAMPLE 7';
DATA ONE;
  INPUT NAME $ 1-4 AGE 6-7 HEIGHT 9-10 WEIGHT 12-14 ID 16-18 SEX $ 20 @22 (Q1-Q10) (1.);
  RATIO=HEIGHT/WEIGHT;
  IF(AGE EQ .) THEN AGEGRP=.;
  ELSE IF(AGE LT 20) THEN AGEGRP=1;
  ELSE AGEGRP=2;
  IF(SEX EQ ' ') THEN NSEX=.;
  ELSE IF(SEX EQ 'F') THEN NSEX=1;
  ELSE IF(SEX EQ 'M') THEN NSEX=2;
CARDS;
JOHN 20 72 180 001 M 1123215422
MARY 17 65 003 F 1243533422
BOB 25 76 200 004 M 2455324431
JO 36 50 002 F 3211454432
ERIC 15 67 135 006 2.24536533
PHIL 18 70 195 005 M 13224355..
PAUL 25 66 175 007 M 1243554221
JEAN 26 64 135 008 F 1133554211
CARL 10 55 78 009 M 5554334221
LYNN 15 60 130 010 F 3342554311
TED 30 69 155 011 M 1113445334
contd. from previous slide ...

MARY  26  66  125  012  F  1145333221
SID   19  77  205  013  M  3443232211
PETE  27  75  195  014  M  1111444455
LIZ   21  66  125  015  F  4432311231
ROB   18  69  155  016  M  1123223433
;

PROC PRINT DATA=ONE;

PROC CORR DATA=ONE;
   VAR RATIO HEIGHT WEIGHT;

PROC PLOT DATA=ONE;
   PLOT RATIO*WEIGHT RATIO*HEIGHT;
   TITLE2 'SCATTER PLOT';

PROC REG DATA=ONE;
   MODEL RATIO=WEIGHT HEIGHT;
   OUTPUT OUT=M_ONE RESIDUAL=R PREDICTED=P;
   TITLE2 'REGRESSION OF RATIO AGAINST WEIGHT AND HEIGHT';

PROC PLOT DATA=M_ONE;
   PLOT R*WEIGHT R*HEIGHT /VREF=0;
   TITLE2 'RESIDUAL PLOTS FROM REGRESSION MODEL M_ONE';

John P. Connolly  Introduction to SAS
contd. from previous slide...

PROC PLOT DATA=M_ONE;
   PLOT RATIO*WEIGHT='A' P*WEIGHT='P'/OVERLAY;
   TITLE2 'ACTUAL VS. PREDICTED';

PROC GLM DATA=ONE;
   CLASS AGEGRP NSEX;
   MODEL HEIGHT=AGEGRP NSEX AGEGRP*NSEX;
   TITLE2 'AN ANOVA USING GLM - PROC ANOVA SHOULD NOT BE USED IF';
   TITLE3 'THE DATA IS UNBALANCED';

PROC TTEST DATA=ONE;
   CLASS SEX;
   VAR HEIGHT;
   TITLE2 'TTEST TO SEE IF MEAN HEIGHT OF FEMALES IS DIFFERENT';
   TITLE3 'FROM MEAN HEIGHT OF MALES';

run;
Example 8

OPTIONS LS=72;
TITLE 'CLASS EXAMPLE 8';
DATA ONE;
   INPUT A B;
CARDS;
1 2
1 2
1 3
1 2
1 3
1 3
1 2
1 3
1 2
1 2
2 2
2 3
2 4
2 3
2 2
2 4
2 4
2 4
2 3
2 2
2 4
Example 8 - contd.

contd. from previous slide . . .

1 2
1 2
1 2
1 2
1 2
1 2
1 3
1 3
1 3
1 3
1 3
1 3
1 3
1 3
1 4
1 4
1 4
1 4
1 4
1 4
Example 8 - contd.

contd. from previous slide ...

```
2 3
2 3
2 3
2 3
2 4
2 4
2 4
2 4
2 2
2 2
2 2
2 2
2 2
;
PROC FREQ DATA=ONE;
   TABLE A*B/CHISQ;
TITLE2 'CHI SQUARE TEST';
run;
```