2.2 Introduction to SAS

SAS stands for "Statistical Analysis System." Even though it runs on PCs and Macs as well as on bigger computers, it is truly the last of the great old mainframe statistical packages. The first beta release was in 1971, and the SAS Institute, Inc. was spun off from North Carolina State University in 1976, the year after Bill Gates dropped out of Harvard. This is a serious pedigree, and it has both advantages and disadvantages.

The advantages are that the number of statistical procedures SAS can do is truly staggering, and the most commonly used ones have been tested so many times by so many people that their correctness and numerical efficiency is beyond any question. For the purposes of this class, there are no bugs. The disadvantages of SAS are all related to the fact that it was *designed* to run in a batch-oriented mainframe environment. So, for example, the SAS Institute has tried hard to make SAS an "interactive" program, but has not really worked. It's as if someone painted an eighteen-wheel transport truck yellow, and called it a school bus. Yes, you can take the children to school in that thing, but would you want to?

2.2.1 The Four Main File Types

A typical SAS job will involve four main types of file.

- **The Raw Data File**: A file consisting of rows and columns of numbers; or maybe some of the columns have letters (character data) instead of numbers. The rows represent observations and the columns represent variables, as described at the beginning of Section 1.1. In the first example we will consider below, the raw data file is called `drp.dat`.

- **The Program File**: This is also sometimes called a "command file," because it's usually not much of a program. It consists of commands that the SAS software tries to follow. You create this file with a text editor like `pico` or `emacs`. The command file contains a reference to the raw data file (in the `infile` statement), so SAS knows where to find the data. In the first example we will consider below, the command file is called `reading.sas`. SAS expects program files to have the extension `.sas`, and you should always follow this convention.

- **The Log File**: This file is produced by every SAS run, whether it is successful of unsuccessful. It contains a listing of the command file, as well any error messages or warnings. The name of the log file is automatically generated by SAS; it combines the first part of the command file's name with the extension `.sas`. So for example, when SAS executes the commands in `reading.sas`, it writes a log file named `reading.log`.

- **The List File**: The list file contains the output of the statistical procedures requested by the command file. The list file has the extension `.lst` — so, for example, running SAS on the command file `reading.sas` will produce `reading.lst` as well as `reading.log`. A successful SAS run will almost always produce a list file. The absence of a list file indicates that there was at least one fatal error. The presence of a list file does not mean there were no errors; it just means that SAS was able to

do *some* of what you asked it to do. Even if there are errors, the list file will usually not contain any error messages; they will be in the log file.

2.2.2 Running SAS from the Command Line

There are several ways to run SAS. We will run SAS from the unix command line. In my view, this way is simplest and best.

If, by accident or on purpose, you type SAS without a filename, then SAS assumes you want to initiate an interactive session, and it tries to start the SAS Display Manager. If you are logged in through an ordinary telnet or ssh session, SAS terminates with an error: `ERROR: Cannot open X display. Check display name/server access authorization.` SAS assumes you are using the unix X-window graphical interface, so it will not work if your computer is emulating a (semi) dumb terminal. If you are in an X-window session, after a while several windows will open up. The only suggestion I have is this: Make sure the SAS Program Editor window is selected. From the File menu, choose Exit. Whew.

If you choose to ignore this advice and actually try to use the Display Manager, you are on your own. You will have my sympathy, but not my help. The joke about painting the transport truck yellow applies, and the joke is on you.

The following illustrates a simple SAS run from the command line. Initially, there are only two files in the (sub)directory — `reading.sas` (the program file) and `drp.dat` (the raw data file). The command `sas reading` produces two additional files — `reading.log` and `reading.lst`. In this and other examples, the unix prompt is `tuzo.erin` (the name of the unix machine used to produce the examples), followed by a `>` sign.

```
tuzo.erin > ls
drp.dat         reading.sas
tuzo.erin > sas reading
tuzo.erin > ls
drp.dat         reading.log     reading.lst     reading.sas
```

2.2.3 Structure of the Program File

A SAS program file is composed of units called *data steps* and *proc steps*. The typical SAS program has one data step and at least one proc step, though other structures are possible.

- Most SAS commands belong either in data step or in a proc step;they will generate errors if they are used in the wrong kind of step.
- Some statements, like the `title` and `options` commands, exist outside of the data and proc steps, but there are relatively few of these.

The Data Step The data step takes care of data acquisition and modification. It almost always includes a reference to the raw data file, telling SAS where to look for the data. It specifies variable names and labels, and provides instructions about how to read the data; for example, the data might be read from fixed column locations. Variables from the raw data file can be modified, and new variables can be created.

Each data step creates a **SAS data set**, a file consisting of the data (after modifications and additions), labels, and so on. Statistical procedures operate on SAS data sets, so you must create a SAS data set before you can start computing any statistics.

A SAS data set is written in a binary format that is very convenient for SAS to process, but is not readable by humans. In the old days, SAS data sets were always written to temporary scratch files on the computer's hard drive; these days, they may be maintained in RAM if they are small enough. In any case, the default is that a SAS data set disappears after the job has run. If the data step is executed again in a later run, the SAS data set is re-created.

Actually, it is possible to save a SAS data set on disk for later use. We won't do this much (there will be just one example), but it makes sense when the amount of processing in a data step is large relative to the speed of the computer. As an extreme example, one of my colleagues uses SAS to analyze data from Ontario hospital admissions; the data files have millions of cases. Typically, it takes around 20 hours of CPU time on a very strong unix machine just to read the data and create a SAS data set. The resulting file, hundreds of gigabytes in size, is saved to disk, and then it takes just a few minutes to carry out each analysis. You wouldn't want to try this on a PC.

To repeat, SAS data *steps* and SAS data *sets* sound similar, but they are distinct concepts. A SAS data *step* is part of a SAS program; it generates a SAS data *set*, which is a file – usually a temporary file.

SAS data sets are not always created by SAS data steps. Some statistical procedures can create SAS data sets, too. For example, `proc univariate` can take an ordinary SAS data set as input, and produce an output data set that has all the original variables, and also some of the variables converted to z-scores (by subtracting off the mean and dividing by the standard deviation). `Proc reg` (the main multiple regression procedure) can produce a SAS data set containing residuals for plotting and use in further analysis; there are many other examples.

The Proc Step "Proc" is short for procedure. Most procedures are statistical procedures; the main exception is `proc format`, which is used to provide labels for the values of categorical independent variables. The proc step is where you specify a statistical procedure that you want to carry out. A statistical procedures in the proc step will take a SAS data sets as input, and write the results (summary statistics, values of test statistics, p-values, and so on) to the list file. The typical SAS program includes one data step and several proc steps, because it is common to produce a variety of data displays, descriptive statistics and significance tests in a single run.

2.2.4 A First Example: `reading.sas`

Earlier, we ran SAS on the file `reading.sas`, producing `reading.log` and `reading.lst`. Now we will look at `reading.sas` in some detail. This program is very simple; it has just one data step and one proc step. More details will be given later, but it's based on a study in which one group of grade school students received a special reading programme, and a control group did not. After a couple of months, all students were given a reading test. We're just going to do an independent groups t-test, but first take a look at the raw data file. You'd do this with the unix `more` command.

Actually, it's so obvious that you should look at your data that nobody ever says it. But experienced data analysts always do it — or else they assume everything is okay and get a bitter lesson in something they already knew. It's so important that it gets the formal status of a **data analysis hint**.

Data Analysis Hint 1 *Always look at your raw data file. It the data file is big, do it anyway. At least page through it a screen at a time, looking for anything strange. Check the values of all the variables for a few cases. Do they make sense? If you have obtained the data file from somewhere, along with a description of what's in it, never believe that the description you have been given is completely accurate.*

Anyway, here is the file `drp.dat`, with the middle cut out to save space.

Treatment	24
Treatment	43
Treatment	58
⋮	⋮
Control	55
Control	28
Control	48
⋮	⋮

Now we can look at `reading.sas`.

```
/****************** reading.sas *********************
*  Simple SAS job to illustrate a two-sample t-test    *
*******************************************************/

options linesize=79 noovp formdlim='_';
title 'More & McCabe (1993) textbook t-test Example 7.8';

data reading;
     infile 'drp.dat';
     input group $ score;
     label group = 'Get Directed Reading Programme?'
           score = 'Degree of Reading Power Test Score';
proc ttest;
     class group;
     var score;
```

Here are some detailed comments about `reading.sas`.

- The first three lines are a comment. Anything between a `/*` and `*/` is a comment, and will be listed on the log file but otherwise ignored by SAS. Comments can appear anywhere in a program. You are not required to use comments, but it's a good idea.

The most common error associated with comments is to forget to end them with `*/`. In the case of `reading.sas`, leaving off the `*/` (or typing
by mistake) would cause the whole program to be treated as a comment. It would generate no errors, and no output — because as far as SAS would be concerned, you never requested any. A longer program would eventually exceed the default length of a comment (it's some large number of characters) and SAS would end the "comment" for you. At exactly that point (probably in the middle of a command) SAS would begin parsing the program. Almost certainly, the first thing it examined would be a fragment of a legal command, and this would cause an error. The log file would say that the command caused an error, and not much else. It would be *very* confusing, because probably the command would be okay, and there would be no indication that SAS was only looking at part of it.

- The next two lines (the `options` statement and the `title` statement) exist outside the proc step and the data step. This is fairly rare.

- All SAS statements end with a semi-colon (`;`). SAS statements can extend for several physical lines in the program file (for example, see the `label` statement). Spacing, indentation, breaking up s statement into several lines of text – these are all for the convenience of the human reader, and are not part of the SAS syntax.

- The most common error in SAS programming is to forget the semi-colon. When this happens, SAS tries to interpret the following statement as part of the one you tried to end. This often causes not one error, but a cascading sequence of errors. The rule is, *if you have an error and you do not immediately understand what it is, look for a missing semi-colon.* It will probably be *before* the portion of the program that (according to SAS) caused the first error.

- Cascading errors are not caused just by the dreaded missing semi-colon. They are common in SAS; for example, a runaway comment statement can easily cause a chain reaction of errors (if the program is long enough for it to cause any error messages at all). *If you have a lot of errors in your log file, fix the first one and don't waste time trying to figure out the others.* Some or all of them may well disappear.

- `options linesize=79 noovp formdlim='_';`

  These options are highly recommended. The `linesize=79` option is so highly recommended it's almost obligatory. It causes SAS to write the output 79 columns across, so it can be read on an ordinary terminal screen that's 80 characters across. You specify an output width of 79 characters rather than 80, because SAS uses one column for printer control characters, like page ejects (form feeds).

  If you do not specify `options linesize=79;`, SAS will use its default of 132 characters across, the width of sheet of paper from an obsolete line printer you probably have never seen. Why would the SAS Institute hang on to this default, when changing it to match ordinary letter paper would be so easy? It probably tells you something about the computing environments of some of SAS's large corporate clients.

- The `noovp` option makes the log files more readable if you have errors. When SAS finds an error in your program, it tries to *underline* the word that caused the error. It does this by going back and *overprinting* the offending word with a series of "underscores" (_ characters). On many printers this works, but when you try to look at the log file on a terminal screen (one that is *not* controlled by the SAS Display Manager), what often appears is a mess. The `noovp` option specifies `no overprinting`. It causes the "underlining" to appear on a separate line under the program line with the error. If you're running SAS from the unix command line and looking at your log files with the `more` command (or the `less` command or the `cat` commmand), you will probably find the `noovp` option to be helpful.

- The `formdlim='_'` option specifies a "form delimiter" to replace most form feeds (new physical pages) in the list file. This can save a lot of paper (and page printing charges). You can use any string you want for a form delimiter. The underscore (the one specified here) causes a solid line to be printed instead of going to a new sheet of paper.

- `title` This is optional, but recommended. The material between the single quotes will appear at the top of each page. This can be a lifesaver when you are searching through a stack of old printouts for something you did a year or two ago.

- `data reading;` This begins the data step, specifying the name of the SAS data set that is being created.

- `infile` Specifies the name of the raw data file. The file name, enclosed in single quotes, can be the full unix path to the file, like `/dos/brunner/public/senic.raw`. If you just give the name of the raw data file, as in this example, SAS assumes that the file is in the same directory as the command file.

- `input` Gives the names of the variables.

  - A character variable (the values of `group` are "Treatment' and "Control") must be followed by a dollar sign.
  - Variable names must be eight characters or less, and should begin with a letter. They will be used to request statistical procedures in the `proc` step. They should be meaningful (related to what the variable *is*), and easy to remember.
  - This is almost the simplest form of the `input` statement. It can be very powerful; for example, you can read data from different locations and in different orders, depending on the value of a variable you've just read, and so on. It can get complicated, but if the data file has a simple structure, the input statement can be simple too.

- `label` Provide descriptive labels for the variables; these will be used to label the output, usually in very nice way. Labels can be quite useful, especially when you're trying to recover what you did a while ago. Notice how this statement extends over two physical lines.

- `proc ttest;` Now the proc step begins. This program has only one data step and one proc step. We are requesting a two-sample t-test.

- `class` Specifies the independent variable.
- `var` Specifies the dependent variable(s). You can give a list of dependent variables. A separate univariate test (actually, as you will see, *collection* of tests is performed for each dependent variable.

reading.log Log files are not very interesting when everything is okay, but here is an example anyway. Notice that in addition to a variety of technical information (where the files are, how long each step took, and so on), it contains a listing of the SAS program — in this case, `reading.sas`. If there were syntax errors in the program, this is where the error messages would appear.

```
tuzo.erin > cat reading.log
1                                                       The SAS System                              11:08 Friday, January 2,

NOTE: Copyright (c) 1989-1996 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 6.12  TS020
      Licensed to UNIVERSITY OF TORONTO/COMPUTING & COMMUNICATIONS, Site 0008987001.



This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: AUTOEXEC processing beginning; file is /local/sas612/autoexec.sas.

NOTE: SAS initialization used:
      real time           0.780 seconds
      cpu time            0.152 seconds


NOTE: AUTOEXEC processing completed.

1          /****************** reading.sas *********************
2          *  Simple SAS job to illustrate a two-sample t-test    *
3          ******************************************************/
4
5          options linesize=79 noovp formdlim='_';
6          title 'More & McCabe (1993) textbook t-test Example 7.8';
7          data reading;
8               infile 'drp.dat';
9               input group $ score;
10              label group = 'Get Directed Reading Programme?'
11                    score = 'Degree of Reading Power Test Score';

NOTE: The infile 'drp.dat' is:
      File Name=/res/jbrunner/442s04/notesSAS/drp.dat,
      Owner Name=jbrunner,Group Name=research,
      Access Permission=rw-------,
      File Size (bytes)=660

NOTE: 44 records were read from the infile 'drp.dat'.
      The minimum record length was 14.
      The maximum record length was 14.
NOTE: The data set WORK.READING has 44 observations and 2 variables.
NOTE: DATA statement used:
      real time           0.190 seconds
      cpu time            0.051 seconds


12         proc ttest;
```

```
13              class group;
14              var score;
NOTE: The PROCEDURE TTEST printed page 1.
NOTE: PROCEDURE TTEST used:
      real time           0.030 seconds
      cpu time            0.009 seconds


2                               The SAS System    11:08 Friday, January 2, 2004


NOTE: The SAS System used:
      real time           1.120 seconds
      cpu time            0.233 seconds

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
```

reading.lst Here is the list file. Notice that the title specified in the `title` statement appears at the top, along with the time and date the program was executed. Then we get means and standard deviations, and several statistical tests — including the one we wanted. We get other stuff too, whether we want it or not. This is typical of SAS, and most other mainstream statistical packages as well. The default output from any given statistical procedures will contain more information than you wanted, and probably some stuff you don't understand at all. There are usually numerous options that can add *more* information, but almost never options to reduce the default output. So, you just learn what to ignore. It is helpful, but not essential,to have at least a superficial understanding of everything in the default output from procedures you use a lot.

```
---------------------------------------------------------------------------

               More & McCabe (1993) textbook t-test Example 7.8              1
                                          11:08 Friday, January 2, 2004

                                TTEST PROCEDURE

Variable: SCORE        Degree of Reading Power Test Score

GROUP          N                 Mean              Std Dev            Std Error
---------------------------------------------------------------------------
Control       23          41.52173913          17.14873323          3.57575806
Treatmen      21          51.47619048          11.00735685          2.40200219

Variances        T       DF    Prob>|T|
---------------------------------------
Unequal    -2.3109     37.9      0.0264
Equal      -2.2666     42.0      0.0286

For H0: Variances are equal, F' = 2.43    DF = (22,20)    Prob>F' = 0.0507
```

Now here are some comments about `reading.lst`.

- `Variable: SCORE` This tells you what the dependent variable is – particularly useful if you have more than one. Notice the nice use of the variable label that was supplied in the `label` statement.
- `GROUP` The independent variable. Underneath are the values of the independent variable. We also have the sample size n for each group, and the group mean, standard deviation, and also the standard error or the mean ($\frac{s}{\sqrt{n}}$, the estimated standard deviation of the sampling distribution of the sample mean).

- Well actually, if you look carefully, you see that we do *not* quite get the values of the independent variable under `GROUP`. The values of the (alphanumeric, or character-valued) variable `group` are `Control` and `Treatment`, but the printout says "`Treatmen`." This is not a printing error; it is a subtle error in the reading of the data. The default length of an alphanumeric data value is 8 characters, but "Treatment" has 9 characters. So SAS just read the first eight. No error message was generated and no harm was done in this case, but in other circumstances this error can turn a data file into a giant pile of trash, without warning. Later we will see how to override the default and read longer strings if necessary.

- Next we get a table whose first column is entitled "Variances." This gives t statistics for testing equality of means, which was what we are interested in. The traditional t-test assumes equal variances, and it is given in the column entitled "Equal."

  - The value of the test statistic is `-2.2666`.
  - The degrees of freedom $n_1 + n_2 - 2$ is given in the `DF` column.
  - The column `Prob>|T|` gives the two-tailed (two-sided) p-value. It is less than the traditional value of 0.05, so the results are statistically significant.

Sample Question 2.2.1 *What do we conclude from this study? Say something about reading, using non-technical language.*

Answer to Sample Question 2.2.1 *Students who received the Directed Reading Program got higher average reading scores than students in the control condition.*

It's worth emphasizing here that the main objective of doing a statistical analysis is to draw conclusions about the data — or to refrain from drawing such conclusions, for good reasons. The question "What do we conclude from this study?" will always be asked. The right answer will always be either "Nothing; the results were not statistically significant," or else it will be something about reading, or fish, or potatoes, or AIDS, or whatever is being studied. Many students, even when they have been warned, respond with a barrage of statistical terminology. They go on and on about the null hypothesis and Type I error, and usually say nothing that would tell a reasonable person what actually happened in the study. In the working world, a memo filled with such garbage could get you fired. Here, it will get you a zero for the question, even if the technical details you give are correct.

Remember, the purpose of writing up a statistical analysis is not to sound impressive and technical, but to impart information. To say things in a simple way is a virtue. It shows you understand what is going on. Now back to the printout.

- The row entitled "Unequal" gives a sort of t-test that does not assume equal variances. Well, it's not really a t-test, because the test statistic does not really have a t distribution, even when the data are exactly normal. But, the (very unpleasant) distribution of the test statistic is well approximated by a t distribution with the right degrees of freedom — not $n_1 + n_2 - 2$, but something messy that depends on the data. See the odd fractional degrees of freedom? See [2] for details. In any case, it does not matter much in this case, because the p-value is almost the same as the

p-value from the traditional test. They lead to the same conclusions, and there is no problem. What should you do when they disagree? I'd go with the test that makes fewer assumptions.

- Next we see `For H0: Variances are equal` and an F-test. This is the traditional test for whether the variances of two groups are equal, and it's *almost* significant. This test is provided so people can test for differences between variances; if it is significantly different they can use the unequal variance t-test, and otherwise they can use the traditional test. This seems reasonable, except for the following.

  Both the two-sample t-test and the F-test for equality of variances assume that the data are normally distributed. However, the normality assumption does not matter much for the t-test when the sample sizes are large, while for the variance test it matters a *lot*, regardless of how much data you have. When the data are non-normal, the test for variances will be significant more than 5% of the time even when the population variances are equal. If you have equal population variances and a large sample of non-normal data, the F-test for variances could easily be significant, leading you to worry unnecessarily about the validity of the t-test.

2.2.5 Background of the First Example

We don't do statistical analysis in a vacuum. Before proceeding with more computing details, let's find out more about the reading data. This first example is from an introductory text. It's Example 7.8 (p. 534) in More and McCabe's excellent *Introduction to the practice of statistics* [2]. We are interested in analyzing *real* data, not in doing textbook exercises. But we will not turn up our noses just yet, because

Data Analysis Hint 2 *When learning how to carry out a procedure using unfamiliar statistical software, always do a textbook example first, and compare the output to the material in the text. Regardless of what the manual might say, never assume you know what the software is doing until you see an example.*

More and McCabe do a great job of explaining the t-test with unequal variances, something SAS produces (along with usual t-test that assumes equal variances) without being asked when you request a t-test. Besides, the data actually come from someone's Ph.D. thesis, so there is an element of realism. Here is Moore and McCabe's description of the study.

> An educator believes that new directed reading activities in the classroom will help elementary school pupils improve some aspects of their reading ability. She arranges for a third grade class of 21 students to take part in these activities. A control classroom of 23 third graders follows the same curriculum without the activities. At the end of 8 weeks, all students are given a Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the program is designed to improve.

Sample Question 2.2.2 *What's wrong with this study?*

Answer to Sample Question 2.2.2 *The independent variable was manipulated by the experimenter, but it is not an experimental study. Even if classrooms were assigned randomly to conditions (it is impossible to tell whether they were, from this brief description), a large number of unobserved variables are potentially confounded with treatment. The teacher in the classroom that received the treatment might be better than the teacher in the control classroom, or possibly there was a particularly aggressive bully in the control classroom, or maybe a mini-epidemic of some childhood disease hit the control classroom —vdots. The list goes on. The point here is that there are many ways in which the classroom experiences of children in the treatment group differ systematically from the experiences of children in the control group.*

Sample Question 2.2.3 *How could the problem be fixed?*

Answer to Sample Question 2.2.3 *Assign classrooms at random to treatments. The unit of analysis should be the classroom, not the individual student.*

2.2.6 SAS Example Two: The statclass data

These data come from a statistics class taught many years ago. Students took eight quizzes, turned in nine computer assignments, and also took a midterm and final exam. The data file also includes gender and ethnic background; these last two variables are just guesses by the professor, and there is no way to tell how accurate they were. The data file looks like this. There are 21 columns and 62 rows of data; columns not aligned.

```
tuzo.erin >  more statclass.dat
1  2  9  1  7  8  4  3  5  2  6  10  10  10  5  0  0  0  0  55  43
0  2  10  10  5  9  10  8  6  8  10  10  8  9  9  9  9  10  10  66  79
1  2  10  10  5  10  10  10  9  8  10  10  10  10  10  10  9  10  10  94  67
1  2  10  10  8  9  10  7  10  9  10  10  10  9  10  10  9  10  10  81  65
0  1  10  1  0  0  8  6  5  2  10  9  0  0  10  6  0  5  0  54  29
```

⋮

Here is the SAS program.

```
tuzo.erin >  cat statmarks.sas
options linesize=79 pagesize=35;
title 'Grades from STA3000 at Roosevelt University:  Fall, 1957';
title2 'Illustrate Elementary Tests';

proc format; /* Used to label values of the categorical variables */
     value sexfmt    0 = 'Male'   1 = 'Female';
     value ethfmt    1 = 'Chinese'
                     2 = 'European'
                     3 = 'Other' ;
data grades;
     infile 'statclass.dat';
     input sex ethnic quiz1-quiz8 comp1-comp9 midterm final;
     /* Drop lowest score for quiz & computer  */
     quizave = ( sum(of quiz1-quiz8) - min(of quiz1-quiz8) ) / 7;
     compave = ( sum(of comp1-comp9) - min(of comp1-comp9) ) / 8;
     label ethnic = 'Apparent ethnic background (ancestry)'
           quizave = 'Quiz Average (drop lowest)'
           compave = 'Computer Average (drop lowest)';
     mark = .3*quizave*10 + .1*compave*10 + .3*midterm + .3*final;
     label mark = 'Final Mark';
     diff = quiz8-quiz1; /* To illustrate matched t-test */
     label diff = 'Quiz 8 minus Quiz 1';
     format sex sexfmt.;       /* Associates sex & ethnic    */
     format ethnic ethfmt.;    /* with formats defined above */

proc freq;
     tables sex ethnic;
proc means n mean std;
     var quiz1 -- mark;        /*  single dash only works with numbered
                                  lists, like quiz1-quiz8    */
proc ttest;
     title 'Independent t-test';
     class sex;
     var mark;
proc means n mean std t;
     title 'Matched t-test: Quiz 1 versus 8';
     var quiz1 quiz8 diff;
proc glm;
     title 'One-way anova';
     class ethnic;
     model mark = ethnic;
     means ethnic / Tukey Bon Scheffe;
proc freq;
     title 'Chi-squared Test of Independence';
     tables sex*ethnic / chisq;
proc freq; /*  Added after seeing warning from chisq test above */
```

```
     title 'Chi-squared Test of Independence: Version 2';
     tables sex*ethnic / norow nopercent chisq expected;
proc corr;
     title 'Correlation Matrix';
     var final midterm quizave compave;
proc plot;
     title 'Scatterplot';
     plot final*midterm; /* Really should do all combinations */
proc reg;
     title 'Simple regression';
     model final=midterm;

/* Predict final exam score from midterm, quiz & computer */
proc reg simple;
     title 'Multiple Regression';
     model final = midterm quizave compave / ss1;
     smalstuf: test quizave = 0, compave = 0;
```

Noteworthy features of this program include

- `options linesize=79 pagesize=35;` Good for $8\frac{1}{2}$ by 11 paper.
- `title2` Subtitle
- `proc format`
- `quiz1-quiz8`
- Creating new variables with assignment statements
- `sum(of quiz1-quiz8)`
- `diff = quiz8-quiz1`
- `format sex sexfmt.;`
- `quiz1 -- mark`
- Title inside a procedure labels just that procedure
- `proc freq` For frequency distributions
- `proc means` To get means and standard deviations
- `proc ttest` We've seen
- `proc means n mean std t` A matched t-test is just a single-variable t-test carried out on differences, testing whether the mean difference is equal to zero.
- `proc glm`
  - `class` Tells SAS that ethnic is categorical.

  - `model` Dependent variable(s) = independent variable(s)
  - `means ethnic / Tukey Bon Scheffe`

- `chisq` option on `proc freq`
- `chisq` option on `proc freq`
- `tables sex*ethnic / norow nopercent chisq expected;` In version 2 of `proc freq`
- `proc corr`
- `proc plot; plot final*midterm;` Scatterplot: First variable named goes on the y axis.
- `proc reg`: `model` Dependent variable(s) = independent variable(s) again
- `simple` option on `proc reg` gives simple descriptive statistics. This last procedure is an example of multiple regression, and we will return to it later once we have more background.

statmarks.lst

```
          Grades from STA3000 at Roosevelt University:  Fall, 1957          1
                        Illustrate Elementary Tests
                                              10:20 Friday, January 4, 2002

                                          Cumulative  Cumulative
             SEX   Frequency   Percent   Frequency    Percent
            ---------------------------------------------------
            Male           39      62.9          39       62.9
            Female         23      37.1          62      100.0

                    Apparent ethnic background (ancestry)

                                          Cumulative  Cumulative
              ETHNIC   Frequency   Percent   Frequency    Percent
            ---------------------------------------------------
            Chinese          41      66.1          41       66.1
            European         15      24.2          56       90.3
            Other             6       9.7          62      100.0

^L          Grades from STA3000 at Roosevelt University:  Fall, 1957
2
                        Illustrate Elementary Tests
                                              10:20 Friday, January 4, 2002

   Variable  Label                              N          Mean       Std Dev
   ------------------------------------------------------------------------
   QUIZ1                                       62     9.0967742     2.2739413
   QUIZ2                                       62     5.8870968     3.2294995
   QUIZ3                                       62     6.0483871     2.3707744
   QUIZ4                                       62     7.7258065     2.1590022
   QUIZ5                                       62     9.0645161     1.4471109
   QUIZ6                                       62     7.1612903     1.9264641
   QUIZ7                                       62     5.7903226     2.1204477
   QUIZ8                                       62     6.3064516     2.3787909
   COMP1                                       62     9.1451613     1.1430011
   COMP2                                       62     8.8225806     1.7604414
   COMP3                                       62     8.3387097     2.5020880
   COMP4                                       62     7.8548387     3.2180168
   COMP5                                       62     9.4354839     1.7237109
   COMP6                                       62     7.8548387     2.4350364
```

```
   COMP7                                    62     6.6451613     2.7526248
   COMP8                                    62     8.8225806     1.6745363
   COMP9                                    62     8.2419355     3.7050497
   MIDTERM                                  62    70.1935484    13.6235557
   FINAL                                    62    49.4677419    17.5141327
   QUIZAVE   Quiz Average (drop lowest)     62     7.6751152     1.1266917
   COMPAVE   Computer Average (drop lowest) 62     8.8346774     1.1204997
   MARK      Final Mark                     62    67.7584101    11.0235746
   ------------------------------------------------------------------------
^L                            Independent t-test
3
                                                 10:20 Friday, January 4, 2002

                                TTEST PROCEDURE

Variable: MARK         Final Mark

   SEX       N         Mean      Std Dev    Std Error      Minimum      Maximum
-------------------------------------------------------------------------------
  Male      39  67.62097070  10.11112521   1.61907581  43.61428571  89.93214286
Female      23  67.99145963  12.65945704   2.63967927  48.48214286  95.45714286

Variances        T       DF    Prob>|T|
----------------------------------------
Unequal    -0.1196     38.5      0.9054
Equal      -0.1268     60.0      0.8995

For H0: Variances are equal, F' = 1.57    DF = (22,38)    Prob>F' = 0.2190
^L                       Matched t-test: Quiz 1 versus 8
4
                                                 10:20 Friday, January 4, 2002

  Variable  Label                 N          Mean       Std Dev             T
  ------------------------------------------------------------------------------
  QUIZ1                           62     9.0967742     2.2739413    31.4995252
  QUIZ8                           62     6.3064516     2.3787909    20.8749114
  DIFF      Quiz 8 minus Quiz 1   62    -2.7903226     3.1578011    -6.9576965
  ------------------------------------------------------------------------------
^L                                One-way anova
5
                                                 10:20 Friday, January 4, 2002

                        General Linear Models Procedure
                            Class Level Information

                  Class    Levels    Values

                  ETHNIC        3    Chinese European Other

                   Number of observations in data set = 62

^L                                One-way anova
6
                                                 10:20 Friday, January 4, 2002

                        General Linear Models Procedure

Dependent Variable: MARK   Final Mark
                                     Sum of            Mean
Source                   DF          Squares          Square  F Value    Pr > F

Model                     2     1478.9595320     739.4797660     7.35    0.0014

Error                    59     5933.7115164     100.5713816

Corrected Total          61     7412.6710484

                  R-Square             C.V.        Root MSE          MARK Mean

                  0.199518         14.80042       10.028528          67.758410
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ETHNIC	2	1478.9595320	739.4797660	7.35	0.0014

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ETHNIC	2	1478.9595320	739.4797660	7.35	0.0014

General Linear Models Procedure

Tukey's Studentized Range (HSD) Test for variable: MARK

NOTE: This test controls the type I experimentwise error rate.

Alpha= 0.05 Confidence= 0.95 df= 59 MSE= 100.5714
Critical Value of Studentized Range= 3.400

Comparisons significant at the 0.05 level are indicated by '***'.

ETHNIC Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
European - Other	-5.108	6.539	18.185	
European - Chinese	4.252	11.528	18.803	***
Other - European	-18.185	-6.539	5.108	
Other - Chinese	-5.550	4.989	15.528	
Chinese - European	-18.803	-11.528	-4.252	***
Chinese - Other	-15.528	-4.989	5.550	

General Linear Models Procedure

Bonferroni (Dunn) T tests for variable: MARK

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than Tukey's for all pairwise comparisons.

Alpha= 0.05 Confidence= 0.95 df= 59 MSE= 100.5714
Critical Value of T= 2.46415

Comparisons significant at the 0.05 level are indicated by '***'.

ETHNIC Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
European - Other	-5.398	6.539	18.476	
European - Chinese	4.071	11.528	18.985	***
Other - European	-18.476	-6.539	5.398	
Other - Chinese	-5.813	4.989	15.790	
Chinese - European	-18.985	-11.528	-4.071	***
Chinese - Other	-15.790	-4.989	5.813	

General Linear Models Procedure

Scheffe's test for variable: MARK

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than Tukey's for all pairwise comparisons.

Alpha= 0.05 Confidence= 0.95 df= 59 MSE= 100.5714
Critical Value of F= 3.15312

Comparisons significant at the 0.05 level are indicated by '***'.

ETHNIC Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
European - Other	-5.626	6.539	18.704	
European - Chinese	3.928	11.528	19.127	***
Other - European	-18.704	-6.539	5.626	
Other - Chinese	-6.019	4.989	15.997	
Chinese - European	-19.127	-11.528	-3.928	***
Chinese - Other	-15.997	-4.989	6.019	

TABLE OF SEX BY ETHNIC

SEX ETHNIC(Apparent ethnic background (ancestry))

Frequency| Expected | Col Pct	Chinese	European	Other	Total
Male	27	7	5	39
	25.79	9.4355	3.7742	
	65.85	46.67	83.33	
Female	14	8	1	23
	15.21	5.5645	2.2258	
	34.15	53.33	16.67	
Total	41	15	6	62

STATISTICS FOR TABLE OF SEX BY ETHNIC

Statistic	DF	Value	Prob
Chi-Square	2	2.921	0.232
Likelihood Ratio Chi-Square	2	2.996	0.224
Mantel-Haenszel Chi-Square	1	0.000	0.995
Phi Coefficient		0.217	
Contingency Coefficient		0.212	
Cramer's V		0.217	

Sample Size = 62
WARNING: 33% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

Correlation Analysis

4 'VAR' Variables: FINAL MIDTERM QUIZAVE COMPAVE

Simple Statistics

Variable	N	Mean	Std Dev	Sum
FINAL	62	49.467742	17.514133	3067.000000
MIDTERM	62	70.193548	13.623556	4352.000000
QUIZAVE	62	7.675115	1.126692	475.857143
COMPAVE	62	8.834677	1.120500	547.750000

Simple Statistics

Variable	Minimum	Maximum	Label
FINAL	15.000000	89.000000	
MIDTERM	44.000000	103.000000	
QUIZAVE	4.571429	9.714286	Quiz Average (drop lowest)

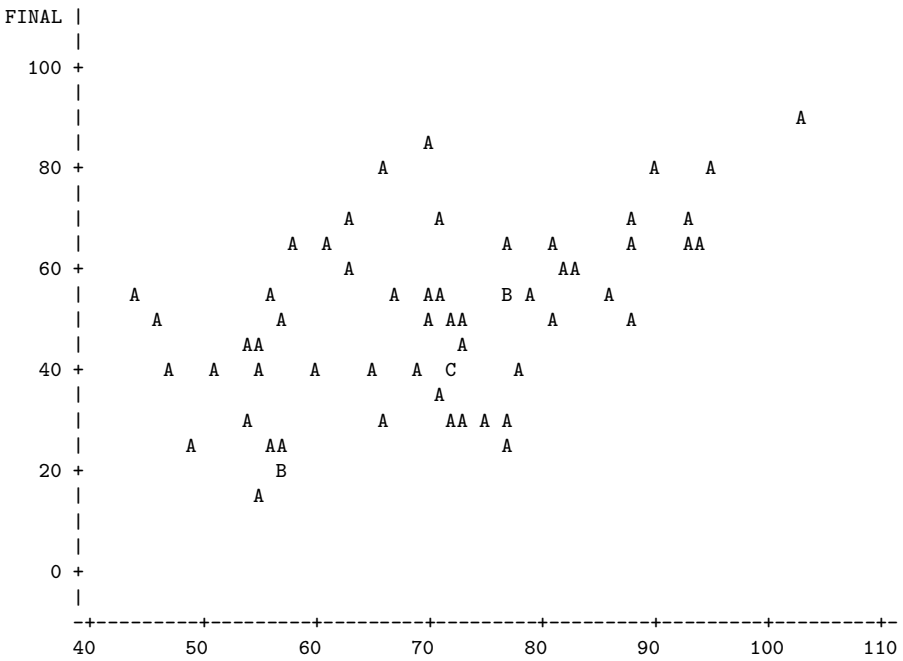
COMPAVE	5.000000	10.000000	Computer Average (drop lowest)

Correlation Analysis

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 62

	FINAL	MIDTERM	QUIZAVE	COMPAVE
FINAL	1.00000	0.51078	0.47127	0.14434
	0.0	0.0001	0.0001	0.2630
MIDTERM	0.51078	1.00000	0.59294	0.41277
	0.0001	0.0	0.0001	0.0009
QUIZAVE	0.47127	0.59294	1.00000	0.52649
Quiz Average (drop lowest)	0.0001	0.0001	0.0	0.0001
COMPAVE	0.14434	0.41277	0.52649	1.00000
Computer Average (drop lowest)	0.2630	0.0009	0.0001	0.0

Plot of FINAL*MIDTERM. Legend: A = 1 obs, B = 2 obs, etc.



Model: MODEL1
Dependent Variable: FINAL

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	4881.79529	4881.79529	21.180	0.0001
Error	60	13829.64019	230.49400		
C Total	61	18711.43548			

Root MSE	15.18203	R-square	0.2609
Dep Mean	49.46774	Adj R-sq	0.2486
C.V.	30.69077		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	3.375101	10.19938324	0.331	0.7419
MIDTERM	1	0.656651	0.14268372	4.602	0.0001

Multiple regression output was deleted.

2.2.7 SAS Example Two: The SENIC data

These data are from a disk that comes with Neter et al's [3] *Applied linear statistical models*. The acronym SENIC stands for "Study of Nosocomial Infection Control." "Nosocomial" means acquired in hospital. Sometimes, patients go to hospital with a broken leg or something, and catch a severe respiratory infection, presumably from other patients. The observations here are hospitals, and the dependent variable is `infrisk`, the probability of catching an infection while in hospital (multiplied by 100). The other variables are explained fairly well by the `labels` statement.

First we will look at the file `senic0.sas`. This is a very basic program that just reads the data and does frequency distributions of everything (even identification number; you don't want to print this!). The idea is that you start out this way, checking for data errors, and then gradually build up the program, adding labels, printing formats and new variables a little bit at a time. This makes it easier to catch your errors.

```
/* senic0.sas */
options linesize = 79;
data  simple;
      infile 'senic.dat';
      input   id stay age infrisk culratio xratio nbeds medschl
              region census nurses service;
proc freq;
      tables _all_;
```

Now suppose we discovered that the file has some weird missing value codes. The next version of the program might look like this.

```
/* senic0.1.sas */
options linesize = 79;
data  simple;
      infile 'senic.dat';
      input   id stay age infrisk culratio xratio nbeds medschl
              region census nurses service;

      /*** sas doesn't like numeric missing value codes.  a period  .  is
                      best for missing. however ....  ***/

      if stay eq 9999 then stay = .  ;
      if age eq 9999 then age = .  ;
      if xratio eq 9999 then xratio = .  ;
      if culratio eq 9999 then culratio = .  ;
      if infrisk = 999 then infrisk  = .  ;
      if nbeds = 9 then nbeds  = .  ;
      if medschl = 9 then medschl  = .  ;
      if region = 9 then region  = .  ;
      if census = 9 then census  = .  ;
      if service = 9 then service  = .  ;
```

```
      if nurses eq (0 or .999) then nurses = .  ;

proc freq;
      tables _all_;
```

The process continues. On the way, we switch to a version of the data file that has the data lined up in fixed columns, with blanks for missing values (a common situation). We wind up with a program called `senicread.sas`. Notice that is consists of just a `proc format` and a data step. There are no statistical procedures, except a `proc freq` that is commented out. This file will be read by programs that invoke statistical procedures, as you will see.

```
/*******  senicread.sas  Just reads and labels data  ***********/
title 'SENIC data';
options linesize=79;

proc format;  /* value labels used in data step below */
     value yesnofmt 1 = 'Yes'   2 = 'No' ;
     value regfmt  1 =  'Northeast'
                   2 =  'North Central'
                   3 =  'South'
                   4 =  'West' ;
     value acatfmt 1 = '53 & under' 2 = 'Over 53';

data senic;
     infile 'senic.raw' missover ;
            /*  in senic.raw, missing=blank  */
            /*  missover causes all blanks to be missing,
                even at the end of a line. */
     input
         #1  id       1-5
             stay     7-11
             age      13-16
             infrisk  18-20
             culratio 22-25
             xratio   27-31
             nbeds    33-35
             medschl  37
             region   39
             census   41-43
             nurses   45-47
             service  49-52 ;
     label id       = 'Hospital identification number'
           stay     = 'Av length of hospital stay, in days'
           age      = 'Average patient age'
           infrisk  = 'Prob of acquiring infection in hospital'
           culratio = '# cultures  / # no hosp acq infect'
           xratio   = '# x-rays / # no signs of pneumonia'
           nbeds    = 'Average # beds during study period'
           medschl  = 'Medical school affiliation'
           region   = 'Region of country (usa)'
           census   = 'Aver # patients in hospital per day'
           nurses   = 'Aver # nurses during study period'
           service  = '% of 35 potential facil. & services' ;
        /* associating variables with their value labels */
     format medschl yesnofmt.;
     format region  regfmt.;

        /***** recodes, computes & ifs  *****/
```

```
     if 0<age<=53 then agecat=1;
     else if age>53 then agecat=2;
     label   agecat = 'av patient age category';
     format agecat acatfmt.;

  /*  compute ad hoc index of hospital quality  */

     quality=(2*service+nurses+nbeds+10*culratio
                 +10*xratio-2*stay)/medschl;
     if (region eq 3) then quality=quality-100;
     label quality = 'jerry''s bogus hospital quality index';


/* Commented out

proc freq;
      tables _all_;
*/
```

Here are some comments.

- Notice that we are reading the variables from specified columns. This allows data to be packed into adjacent columns (some data files are like this), and also allows missing data to be represented by blanks. But it means that the data must be perfectly aligned into columns. Don't *assume* this is true just because you were told by someone who should know. Check!
- The `missover` option is highly recommended if missing values are represented by blanks.
- `if 0<age<=53` means "if $0 < \text{age} \leq 53$.'
- Age = 0 or negative would result in a missing value for `agecat`.
- a missing value for `xratio` (or any other variable in the formula) would result in a missing value for `quality`.
- The double quotation mark in the middle of the label for `quality` is how you get an apostrophe in a label.
- `tables _all_` in `proc freq`: The reserved name `_all_` means all the variables in the data set.

Here is a program that pulls in `senicread.sas` with a `%include` statement, and then does some statistical tests. Keeping the data definition in a separate file is often a good strategy, because most data analysis projects involve a substantial number of statistical procedures. It is common to have maybe twenty program files that carry out various analyses. You *could* have the data step at the beginning of each program, but what

happens when (inevitably) you want to make a change in the data step and re-run your analyses? You find yourself making the same change in twenty files. Probably you will forget to change some of them, and the result is a big mess. If you keep your data definition in just one place, you only have to edit it once, and a lot of problems are avoided.

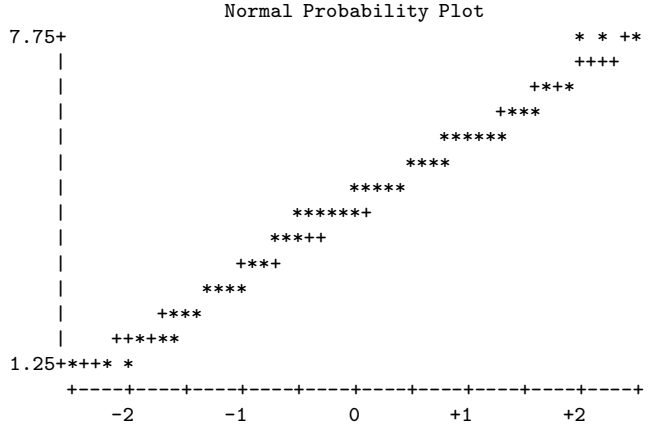
```
/****************** basicsenic.sas ***************/
/*               Basic stats on SENIC Data         */
/*************************************************/

%include 'senicread.sas';  /* senicread.sas reads data, etc.  */

proc univariate  plot normal ; /* Plots and a test for normality */
     title2 'Describe Quantitative Variables';
     var stay -- nbeds census nurses service;
     /* single dash only works with numbered lists, like item1-item50    */
proc freq;
     title2 'Frequency distributions of categorical variables';
     tables medschl region agecat;
proc chart;
     title2 'Vertical bar charts';
     vbar region medschl agecat /discrete ;
proc chart ;
     title2 'Pie chart';
     pie region/type=freq;

proc chart;
     title2 'Pseudo 3-d chart - just playing around';
     block region / sumvar=infrisk type=mean group=medschl discrete;

/* Now elementary tests */

proc freq;  /* use freq to do crosstabs */
     tables region*medschl / nocol nopercent expected chisq;
proc ttest;
     class medschl;
     var infrisk age ;
proc glm;  /* one-way anova */
     class region;
     model infrisk=region;
     means region/ snk scheffe;
proc plot;
     plot infrisk * nurses
          infrisk * nurses = medschl;
proc corr;
     var stay -- nbeds census nurses service;
proc glm;  /*  simple regression with glm*/
```

```
model infrisk=nurses;
```

The list file from this job is long, so we will just look at the `proc univariate` output for the dependent variable.

```
^L                                  SENIC data
6
                        Describe Quantitative Variables
                                                11:47 Friday, January 4, 2002

                             Univariate Procedure

Variable=INFRISK       Prob of acquiring infection in hospital

                                  Moments

                    N              113  Sum Wgts        113
                    Mean       4.354867  Sum           492.1
                    Std Dev    1.340908  Variance   1.798034
                    Skewness   -0.11976  Kurtosis   0.182355
                    USS         2344.41  CSS        201.3798
                    CV         30.79102  Std Mean   0.126142
                    T:Mean=0   34.52353  Pr>|T|       0.0001
                    Num ^= 0        113  Num > 0         113
                    M(Sign)        56.5  Pr>=|M|      0.0001
                    Sgn Rank     3220.5  Pr>=|S|      0.0001
                    W:Normal   0.970897  Pr<W         0.1280

                               Quantiles(Def=5)



                     100% Max       7.8       99%       7.7
                      75% Q3        5.2       95%       6.4
                      50% Med       4.4       90%       5.8
                      25% Q1        3.7       10%       2.6
                       0% Min       1.3        5%       1.8
                                                1%       1.3
                     Range          6.5
                     Q3-Q1          1.5
                     Mode           4.3

                                  Extremes

                       Lowest    Obs     Highest    Obs
                          1.3(      93)      6.5(      47)
                          1.3(      40)      6.6(     104)
                          1.4(     107)      7.6(      53)
                          1.6(       2)      7.7(      13)
                          1.7(      85)      7.8(      54)

^L                                  SENIC data
7
                        Describe Quantitative Variables
                                                11:47 Friday, January 4, 2002

                             Univariate Procedure

Variable=INFRISK       Prob of acquiring infection in hospital

                  Stem Leaf                     #             Boxplot
                     7 678                      3                0
                     7                                           |
                     6 56                       2                |
                     6 12334                    5                |
                     5 5555666777889           13                |
                     5 0000112233344           13             +-----+
                     4 5555555666778888999     19             |     |
                     4 0111122222333333344444  22             *--+--*
```

```
    3 5577778999               10             +-----+
    3 011244                    6                |
    2 5677899999               10                |
    2 0013                      4                |
    1 678                       3                |
    1 334                       3                0
      ----+----+----+----+--
```



2.2.8 SAS Reference Materials

This course is trying to teach you SAS by example, without full explanation, and certainly without discussion of all the options. If you need more detail, there are several approaches you can take. The most obvious is to consult the SAS manuals. The full set of manuals runs to over a dozen volumes, and most of them look like telephone directories. For a beginner, it is hard to know where to start. And even if you know where to look, the SAS manuals can be hard to read, because they assume you already understand the statistical procedures fairly thoroughly, and on a mathematical level. They are really written for professional statisticians. The SAS Institute also publishes a variety of manual-like books that are intended to be more instructional, most of them geared to specific topics (like *The SAS system for multiple regression* and *the SAS system for linear models*). These are a bit more readable, though it helps to have a real textbook on the topic to fill in the gaps.

A better place to start is a wonderful book by Cody and Smith [1] entitled *Applied statistics and the SAS programming language*. They do a really good job of presenting and documenting the language of the data step, and and they also cover a set of statistical procedures ranging from elementary to moderately advanced. If you had to own just one SAS book, this would be it.

If you consult *any* SAS book or manual (Cody and Smith's book included), you'll need to translate and filter out some details. First, you're advised to ignore anything about the SAS Display Manager. In this course, there are raw data file, program files, log files and list files; that's it.

Second, many of the examples you see in Cody and Smith's book and elsewhere will not have separate files for the raw data and the program. They include the raw data in the program file in the data step, after a `datalines` or `cards` statement. Here is an example from page 3 of [1].

```
data test;
```

```
    input subject 1-2 gender $ 4 exam1 6-8 exam2 10-12 hwgrade $ 14;
    datalines;
10 M  80  84 A
 7 M  85  89 A
 4 F  90  86 B
20 M  82  85 B
25 F  94  94 A
14 F  88  84 C
;
proc means data=test;
run;
```

Having the raw data and the SAS code together in one display is so attractive for small datasets that most textbook writers cannot resist it. But think how unpleasant it would be if you had 10,000 lines of data. The way we would do this example is to have the data file (named, say, `example1.dat`) in a separate file. The data file would look like this.

```
10 M  80  84 A
 7 M  85  89 A
 4 F  90  86 B
20 M  82  85 B
25 F  94  94 A
14 F  88  84 C
```

and the program file would look like this.

```
data test;
    infile 'example1.dat'; /* Read data from example1.dat */
    input subject 1-2 gender $ 4 Exam1 6-8 exam2 10-12 hwgrade $ 14;
proc means data=test;
```

Using this as an example, you should be able to translate any textbook example into the program-file data-file format used in this course.