Introduction to Applied Statistics (Draft)

Jerry Brunner

January 11, 2007

Chapter 1

Introduction

This course is about using statistical methods to draw conclusions from real data. It is deliberately non-mathematical, relying on translations of statistical theory into English. For the most part, formulas are avoided. While this involves some loss of precision, it also makes the course accessible to students from non-statistical disciplines (particularly graduate students and advanced undergraduates on their way to graduate school) who need to use statistics in their research. Even for students with strong training in theoretical statistics, the use of plain English can help reveal the connections between theory and applications, while also suggesting a useful way to communicate with non-statisticians.

We will avoid mathematics, but we will not avoid computers. Learning to apply statistical methods to real data involves actually doing it, and the use of software is not optional. Furthermore, we will *not* employ "user-friendly" menu-driven statistical programs. Why?

- It's just too easy to poke around in the menus trying different things, produce some results that seem reasonable, and then two weeks later be unable to say exactly what one did.

- Real data sets tend to be large and complex, and most statistical analyses involve a sizable number of operations. If you discover a tiny mistake after you produce your results, you don't want to go back and repeat two hours of menu selections and mouse clicks, with one tiny variation.

- If you need to analyze a data set that is similar to one you have analyzed in the past, it's a lot easier to edit a program than to remember a collection of menu selections from last year.

Don't worry! The word "program" does *not* mean we are going to write programs in some true programming language like C or Java. We'll use statistical software in which most of the actual statistical procedures have already been written by experts; usually, all we have to do is invoke them by using high-level commands.

The statistical packages we will use in this course are `SAS` and `R`. These packages are command-oriented rather than menu-oriented, and are very powerful. They are industrial strength tools, and will be illustrated in an industrial strength environment — `unix`. This is mostly for local convenience. There are Windows versions of both `SAS` and `R` that work just as well as the unix versions, except for very big jobs.

Applied Statistics really refers to two related enterprises. The first might be more accurately termed "Applications of Statistics," and consists of the appropriate application of standard general techniques. The second enterprise is the development of specialized techniques that are designed specifically for the data at hand. The difference is like buying your clothes from Walmart versus sewing them yourself (or going to a tailor). In this course, we will do both. We'll maintain the non-mathematical nature of the course in the second half by substituting computing power and random number generation for statistical theory.

1.1 Vocabulary of data analysis

We start with a **data file**. Think of it as a rectangular array of numbers, with the rows representing **cases** (units of analysis, observations, subjects, replicates) and the columns representing **variables** (pieces of information available for each case).

- A physical data file might have several lines of data per case, but you can imagine them listed on a single long line.

- Data that are *not* available for a particular case (for example because a subject fails to answer a question, or because a piece of measuring equipment breaks down) will be represented by missing value codes.

Missing value codes allow observations with missing information to be automatically excluded from a computation.

- Variables can be **quantitative** (representing amount of something) or **categorical**. In the latter case the "numbers" are codes representing category membership. Categories may be **ordered** (small vs. medium vs. large) or **unordered** (green vs. blue vs. yellow). When a quantitative variable reflects measurement on a scale capable of very fine gradation, it is sometimes described as **continuous**. Some statistical texts use the term **qualitative** to mean categorical. When an anthropologist uses the word "qualitative," however, it usually refers to ethnographic or case study research in which data are not explicitly assembled into a data file.

Another very important way to classify variables is

Independent Variable (IV): Predictor = X (actually $X_i, i = 1, \ldots, n$)

Dependent Variable (DV): Predicted = Y (actually $Y_i, i = 1, \ldots, n$)

Example: X = weight of car in kilograms, Y = fuel efficiency in litres per kilometer

Sample Question 1.1.1 *Why isn't it the other way around?*

Answer to Sample Question 1.1.1 *Since weight of a car is a factor that probably influences fuel efficiency, it's more natural to think of predicting fuel efficiency from weight.*

The general principle is that if it's more natural to think of predicting A from B, then A is the dependent variable and B is the independent variable. This will usually be the case when B is thought to cause or influence A. Sometimes it can go either way or it's not clear. But usually it's easy to decide.

Sample Question 1.1.2 *Is it possible for a variable to be both quantitative and categorical? Answer Yes or No, and either give an example or explain why not.*

Answer to Sample Question 1.1.2 *Yes. For example, the number of cars owned by a person or family.*

In some fields, you may hear about **nominal, ordinal, interval** and **ratio** variables, or variables measured using "scales of measurement" with those names. Ratio means the scale of measurement has a true zero point, so that a value of 4 represents twice as much as 2. An interval scale means that the difference (interval) between 3 and 4 means the same thing as the difference between 9 and 10, but zero does not necessarily mean absence of the thing being measured. The usual examples are shoe size and ring size. In ordinal measurement, all you can tell is that 6 is less than 7, not how much more. Measurement on a nominal scale consists of the assignment of unordered categories. For example, citizenship is measured on a nominal scale.

It is usually claimed that one should calculate means (and therefore, for example, do multiple regression) only with interval and ratio data; it's usually acknowledged that people do it all the time with ordinal data, but they really shouldn't. And it is obviously crazy to calculate a mean on numbers representing unordered categories. Or is it?

Sample Question 1.1.3 *Give an example in which it's meaningful to calculate the mean of a variable measured on a nominal scale.*

Answer to Sample Question 1.1.3 *Code males as zero and females as one. The mean is the proportion of females.*

It's not obvious, but actually all this talk about what you should and shouldn't do with data measured on these scales does not have anything to do with *statistical* assumptions. That is, it's not about the mathematical details of any statistical model. Rather, it's a set of guidelines for what statistical model one ought to adopt. Are the guidelines reasonable? It's better to postpone further discussion until after we have seen some details of multiple regression.

1.2 Statistical significance

We will often pretend that our data represent a **random sample** from some **population**. We will carry out formal procedures for making inferences about this (usually fictitious) population, and then use them as a basis for drawing conclusions from the data.

Why do we do all this pretending? As a formal way of filtering out things that happen just by coincidence. The human brain is organized to find *meaning* in what it perceives, and it will find apparent meaning even in a sequence of random numbers. The main purpose of testing for statistical significance is to protect Science against this. Even when the data do not fully satisfy the assumptions of the statistical procedure being used (for example, the data are not really a random sample) significance testing can be a useful as a way of restraining scientists from filling the scientific literature with random garbage. This is such an important goal that we will spend a substantial part of the course on significance testing.

1.2.1 Definitions

Numbers that can be calculated from sample data are called **statistics**. Numbers that could be calculated if we knew the whole population are called **parameters**. Usually parameters are represented by Greek letters such as α, β and γ, while statistics are represented by ordinary letters such as a, b, c. Statistical inference consists of making decisions about parameters based on the values of statistics.

The **distribution** of a variable corresponds roughly to a histogram of the values of the variable. In a large population for a variable taking on many values, such a histogram will be indistinguishable from a smooth curve.

For each value x of the independent variable X, in principle there is a separate distribution of the dependent variable Y. This is called the **conditional distribution** of Y given $X = x$.

We will say that the independent and dependent variables are **unrelated** if the *conditional distribution of the dependent variable is identical for each value of the independent variable*. That is, the histogram of the dependent variable does not depend on the value of the independent variable. If the distribution of the dependent variable does depend on the value of the independent variable, we will describe the two variables as **related**. All this applies to sample as well as population data-sets (a population dataset may be entirely hypothetical).

Most research questions involve more than one independent variable. It is also common to have more than one dependent variable. When there is one dependent variable, the analysis is called **univariate**. When more than one dependent variable is being considered simultaneously, the analysis is called **multivariate**.

Sample Question 1.2.1 *Give an example of a study with two categorical independent variables, one quantitative independent variable, and two quantitative dependent variables.*

Answer to Sample Question 1.2.1 *In a study of success in university, the subjects are first-year university students. The categorical independent variables are Sex and Immigration Status (Citizen, Permanent Resident or Visa), and the quantitative independent variable is family income. The dependent variables are cumulative Grade Point Average at the end of first year, and number of credits completed in first year.*

Many problems in data analysis reduce to asking whether one or more variables are related – not in the actual data, but in some hypothetical population from which the data are assumed to have been sampled. The reasoning goes like this. Suppose that the independent and dependent variables are actually unrelated *in the population*. If this **null hypothesis** is true, what is the probability of obtaining a *sample* relationship between the variables that is as strong or stronger than the one we have observed? If the probability is small (say, $p < 0.05$), then we describe the sample relationship as **statistically significant**, and it is socially acceptable to discuss the results. In particular, there is some chance of having the results taken seriously enough to publish in a scientific journal.

The number 0.05 is called the **significance level**. In principle, the exact value of the significance level is arbitrary as long as it is fairly small, but scientific practice has calcified around a suggestion of R. A. Fisher (in whose honour the F-test is named), and the 0.05 level is an absolute rule in many journals in the social and biological sciences.

We will willingly conform to this convention. We conform *willingly* because we understand that scientists can be highly motivated to get their results into print, even if those "results" are just trends that could easily be random noise. To restrain these people from filling the scientific literature with random garbage, we need a clear rule.

For those who like precision, the formal definition of a p-value is this. It is the minimum significance level α at which the null hypothesis (of no relationship between IV and DV in the population) can be rejected.

Here is another useful way to talk about p-values. *The p-value is the probability of getting our results (or better) just by chance.* If p is small enough, then the data are very unlikely to have arisen by chance, assuming there is

really no relationship between the independent variable and the dependent variable in the population. In this case we will conclude there really is a relationship between the independent variable and the dependent variable.

What should we do if $p > .05$? Fisher suggested that we should not conclude anything. In particular, he suggested that we should *not* conclude that the independent and dependent variables are unrelated. Instead, we can say that there is insufficient evidence evidence of a relationship between the independent variable and the dependent variable. A good reference is Fisher's masterpiece, *Statistical methods for research workers* [5], which had its first edition in 1925, and its 14th and last edition in 1970, eight years after Fisher's death.

The trouble with Fisher's formulation is that it never allows us to conclude that the null hypothesis is true. But sometimes, experimental treatments just don't do anything, and it is of scientific and practical importance to be able to say so. For example, medical researchers frequently conclude that drugs don't work. On what basis are they drawing these conclusions? On what basis *should* they draw such conclusions? We will get back to this important issue later. For now, let us agree that if a test is not significant, then we certainly can agree with Fisher that there is not enough evidence to conclude that the independent and dependent variables are related. As for concluding that the variables are *not* related, we don't yet have a formal rule.

1.2.2 Standard elementary significance tests

We will now consider some of the most common elementary statistical methods. For each one, you should be able to answer the following questions.

1. Make up your own original example of a study in which the technique could be used.
2. In your example, what is the independent variable (or variables)?
3. In your example, what is the dependent variable (or variables)?
4. Indicate how the data file would be set up.

Independent observations One assumption shared by most standard methods is that of *"independent observations."* The meaning of the assumption is this. Observations 13 and 14 are independent if and only if the conditional distribution of observation 14 given observation 13 is the same for each possible value observation 13. For example if the observations are temperatures on consecutive days, this would not hold. If the dependent variable is score on a homework assignment and students copy from each other, the observations will not be independent.

When significance testing is carried out under the assumption that observations are independent but really they are not, results that are actually due to chance will often be detected as significant with probability considerably greater than 0.05. This is sometimes called the problem of *inflated* n. In other words, you are pretending you have more separate pieces of information than you really do. When observations cannot safely be assumed independent, this should be taken into account in the statistical analysis. We will return to this point again and again.

Independent (two-sample) t-test

This is a test for whether the means of two independent groups are different. Assumptions are independent observations, normality within groups, equal variances. For large samples normality does not matter. For large samples with nearly equal sample sizes, equal variance assumption does not matter. The assumption of independent observations is always important.

Sample Question 1.2.2 *Make up your own original example of a study in which a two-sample t-test could be used.*

Answer to Sample Question 1.2.2 *An agricultural scientist is interested in comparing two types of fertilizer for potatoes. Fifteen small plots of ground receive fertilizer A and fifteen receive fertilizer B. Crop yield for each plot in pounds of potatoes harvested is recorded.*

Sample Question 1.2.3 *In your example, what is the independent variable (or variables)?*

Answer to Sample Question 1.2.3 *Fertilizer, a binary variable taking the values A and B.*

Sample Question 1.2.4 *In your example, what is the dependent variable (or variables)?*

Answer to Sample Question 1.2.4 *Crop yield in pounds.*

Sample Question 1.2.5 *Indicate how the data file might be set up.*

Answer to Sample Question 1.2.5

A	13.1
A	11.3
⋮	⋮
B	12.2
⋮	⋮

Matched (paired) t-test

Again comparing two means, but from paired observations. Pairs of observations come from the same case (subject, unit of analysis), and presumably are non-independent. Again, the data from a given pair are not really separate pieces of information, and if you pretend they are, then you are pretending to have more accurate estimation of population parameters — and a more sensitive test — than you really do. The probability of getting results that are statistically significant will be greater than 0.05, even if nothing is going on.

In a matched t-test, this problem is taken care of by computing a difference for each pair, reducing the volume of data (and the apparent sample size) by half. This is our first example of a *repeated measures* analysis. Here is a general definition. We will say that there are **repeated measures** on an independent variable if a case (unit of analysis, subject, participant in the study) contributes a value of the dependent variable for each value of the independent variable in question. A variable on which there are repeated measures is sometimes called a **within-subjects** variable. When this language is being spoken, variables on which there are not repeated measures are called **between-subjects**.

The assumptions of the matched t-test are that the differences represent independent observations from a normal population. For large samples, normality does not matter. The assumption that different cases represent independent observations is always important.

Sample Question 1.2.6 *Make up your own original example of a study in which a matched t-test could be used.*

Answer to Sample Question 1.2.6 *Before and after a 6-week treatment, participants in a quit-smoking program were asked "On the average, how many cigarettes do you smoke each day?"*

Sample Question 1.2.7 *In your example, what is the independent variable (or variables)?*

Answer to Sample Question 1.2.7 *Presence versus absence of the program, a binary variable taking the values "Absent" or "Present" (or maybe "Before" and "After"). We can say there are* repeated measures *on this factor, or that it is a* within-subjects *factor.*

Sample Question 1.2.8 *In your example, what is the dependent variable (or variables)?*

Answer to Sample Question 1.2.8 *Reported number of cigarettes smoked per day.*

Sample Question 1.2.9 *Indicate how the data file might be set up.*

Answer to Sample Question 1.2.9 *The first column is "Before," and the second column is "After."*

$$\begin{array}{cc} 22 & 18 \\ 40 & 34 \\ 20 & 10 \\ \vdots & \vdots \end{array}$$

One-way Analysis of Variance

Extension of the independent t-test to two or more groups. Same assumptions, everything. $F = t^2$ for two groups.

Sample Question 1.2.10 *Make up your own original example of a study in which a one-way analysis of variance could be used.*

Answer to Sample Question 1.2.10 *Eighty branches of a large bank were chosen to participate in a study of the effect of music on tellers' work behaviour. Twenty branches were randomly assigned to each of the following 4 conditions. 1=No music, 2=Elevator music, 3=Rap music, 4=Individual choice (headphones). Average customer satisfaction and worker satisfaction were assessed for each bank branch, using a standard questionnaire.*

Sample Question 1.2.11 *In your example, what are the cases?*

Answer to Sample Question 1.2.11 *Branches, not people answering the questionnaire.*

Sample Question 1.2.12 *Why do it that way?*

Answer to Sample Question 1.2.12 *To avoid serious potential problems with independent observations within branches. The group of interacting people within social setting is the natural unit of analysis, like an organism.*

Sample Question 1.2.13 *In your example, what is the independent variable (or variables)?*

Answer to Sample Question 1.2.13 *Type of music, a categorical variable taking on 4 values.*

Sample Question 1.2.14 *In your example, what is the dependent variable (or variables)?*

Answer to Sample Question 1.2.14 *There are 2 dependent variables, average customer satisfaction and average worker satisfaction. If they were analyzed simultaneously the analysis would be multivariate (and not elementary).*

Sample Question 1.2.15 *Indicate how the data file might be set up.*

Answer to Sample Question 1.2.15 *The columns correspond to Branch, Type of Music, Customer Satisfaction and Worker Satisfaction*

$$\begin{array}{cccc} 1 & 2 & 4.75 & 5.31 \\ 2 & 4 & 2.91 & 6.82 \\ \vdots & \vdots & \vdots & \vdots \\ 80 & 2 & 5.12 & 4.06 \end{array}$$

Sample Question 1.2.16 *How could this be made into a repeated measures study?*

Answer to Sample Question 1.2.16 *Let each branch experience each of the 4 music conditions in a random order (or better, use only 72 branches, with 3 branches receiving each of the 24 orders). There would then be 10 pieces of data for each bank: Branch, Order (a number from 1 to 24), and customer satisfaction and worker satisfaction for each of the 4 conditions.*

Including all orders of presentation in each experimental condition is an example of **counterbalancing** — that is, presenting stimuli in such a way that order of presentation is unrelated to experimental condition. That way, the effects of the treatments are not confused with fatigue or practice effects (on the part of the experimenter as well as the subjects). In counterbalancing, it is often not feasible to include *all* possible orders of presentation it each experimental condition, because sometimes there are two many. The point is that order of presentation has to be unrelated to any manipulated independent variable.

Two (and higher) way Analysis of Variance

Extension of One-Way ANOVA to allow assessment of the joint relationship of several categorical independent variables to one quantitative dependent variable that is assumed normal within treatment combinations. Tests for interactions between IVs are possible. An interaction means that the relationship of one independent variable to the dependent variable *depends* on the value of another independent variable. More on this later.

Crosstabs and chi-squared tests

Cross-tabulations (Crosstabs) are joint frequency distribution of two categorical variables. One can be considered an IV, the other a DV if you like. In any case (even when the IV is manipulated in a true experimental study) we will test for significance using the *chi-squared test of independence*. Assumption is independent observations are drawn from a multinomial distribution. Violation of the independence assumption is common and very serious.

Sample Question 1.2.17 *Make up your own original example of a study in which this technique could be used.*

Answer to Sample Question 1.2.17 *For each of the prisoners in a Toronto jail, record the race of the offender and the race of the victim. This is illegal; you could go to jail for publishing the results. It's totally unclear which is the IV and which is the DV, so I'll make up another example.*

For each of the graduating students from a university, record main field of study and and gender of the student (male or female).

Sample Question 1.2.18 *In your example, what is the independent variable (or variables)?*

Answer to Sample Question 1.2.18 *Gender*

Sample Question 1.2.19 *In your example, what is the dependent variable (or variables)?*

Answer to Sample Question 1.2.19 *Main field of study (many numeric codes).*

Sample Question 1.2.20 *Indicate how the data file would be set up.*

Answer to Sample Question 1.2.20 *The first column is Gender (0=Male, 1=F). The second column is Field.*

$$\begin{array}{cc} 1 & 2 \\ 0 & 14 \\ 0 & 9 \\ \vdots & \vdots \end{array}$$

Correlation and Simple Regression

Correlation Start with a **scatterplot** showing the association between two (quantitative, usually continuous) variables. A scatterplot is a set of Cartesian coordinates with a dot or other symbol showing the location of each (x, y) pair. If one of the variables is clearly the independent variable, it's traditional to put it on the x axis. There are n points on the scatterplot, where n is the number of cases in the data file.

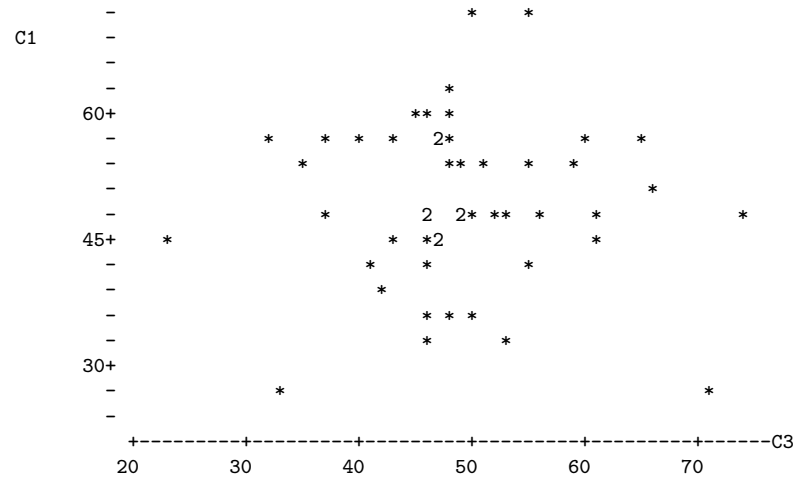
Often, the points in a scatterplot cluster around a straight line. The correlation coefficient (Pearson's r) expresses the extent to which the points cluster tightly around a straight line.

Here are some properties of the correlation coefficient r:

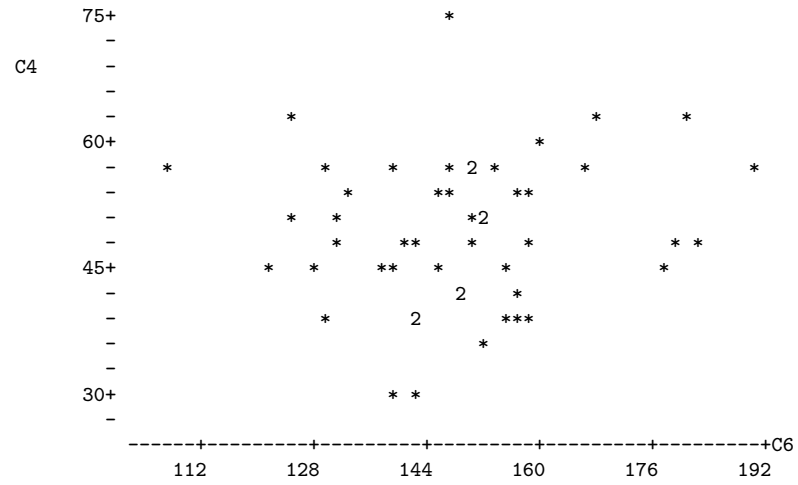
- $-1 \leq r \leq 1$
- $r = +1$ indicates a perfect positive linear relationship. All the points are exactly on a line with a positive slope.
- $r = -1$ indicates a perfect negative linear relationship. All the points are exactly on a line with a negative slope.
- $r = 0$ means no *linear* relationship (curve possible)
- r^2 represents explained variation, reduction in (squared) error of prediction. For example, the correlation between scores on the Scholastic Aptitude Test (SAT) and first-year grade point average (GPA) is around +0.50, so we say that SAT scores explain around 25% of the variation in first-year GPA.

The test of significance for Pearson's r assumes a bivariate normal distribution for the two variables; this means that the only possible relationship between them is linear. As usual, the assumption of independent observations is always important.

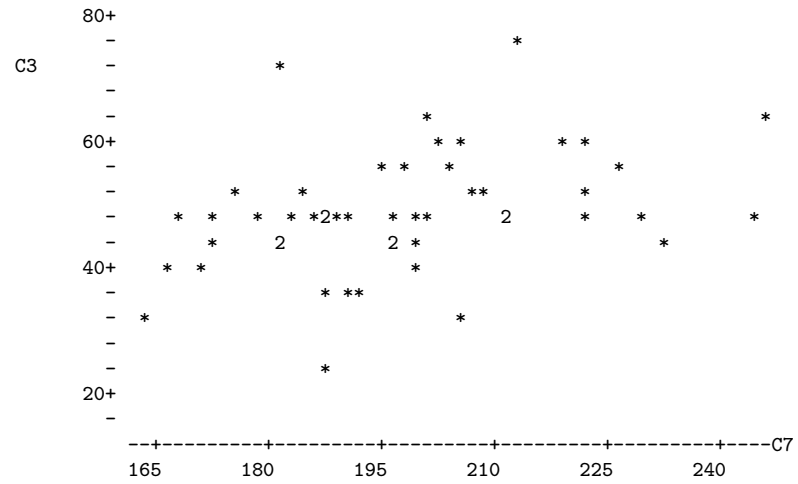
Here are some examples of scatterplots and the associated correlation coefficients. The number 2 on a plot means that two points are on top of each other, or at least too close to be distinguished in this crude line printer graphic.



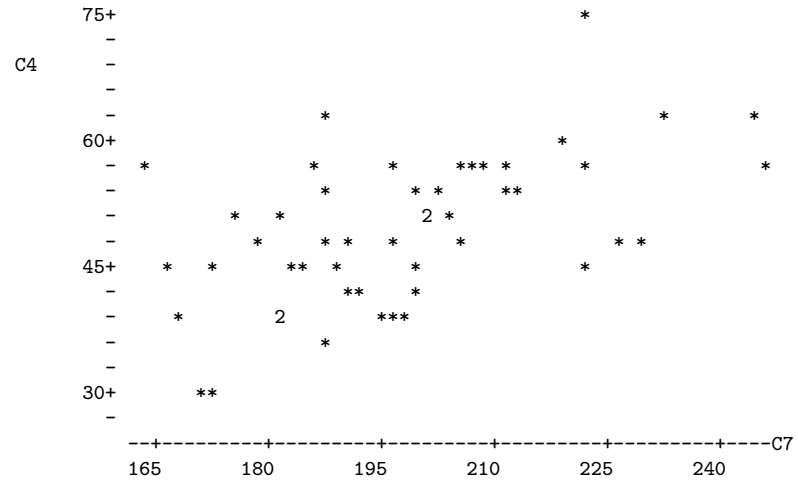
Correlation of C1 and C3 = 0.004



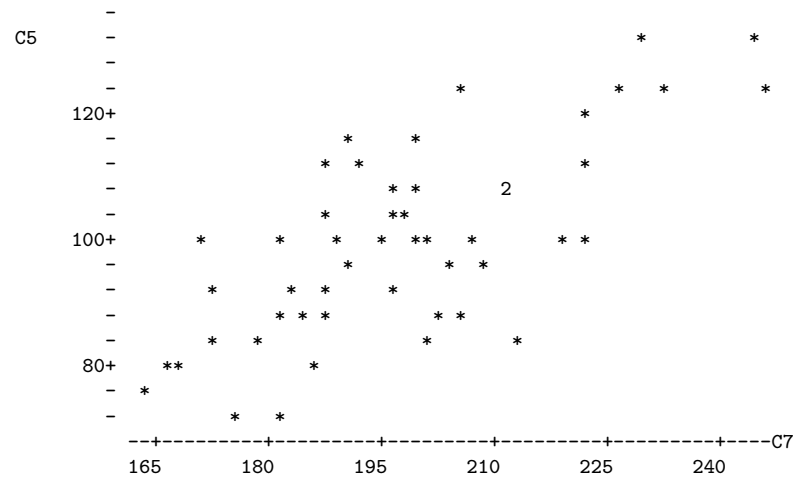
Correlation of C4 and C6 = 0.112



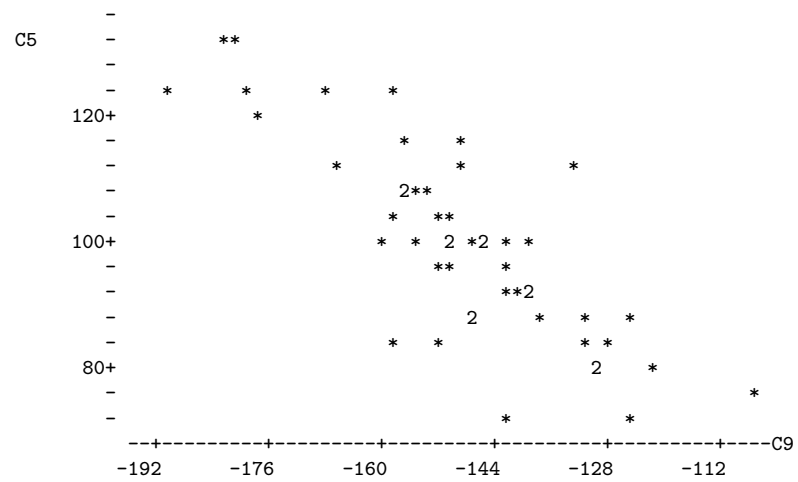
Correlation of C3 and C7 = 0.368



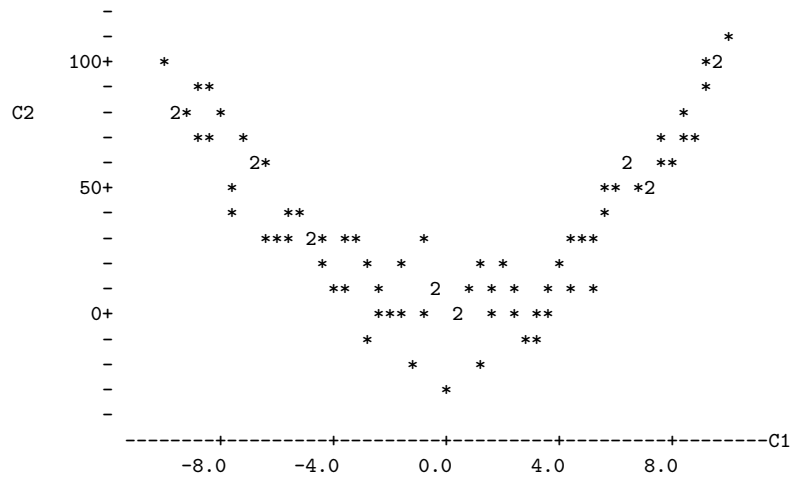
Correlation of C4 and C7 = 0.547



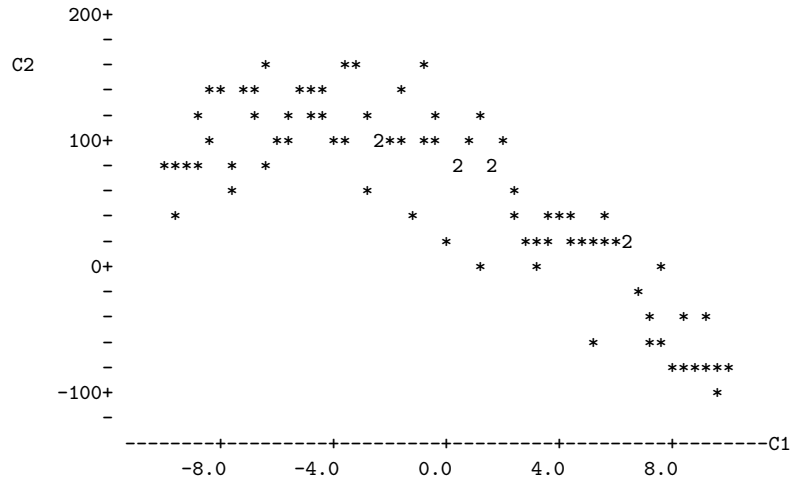
Correlation of C5 and C7 = 0.733



Correlation of C5 and C9 = -0.822



Correlation of C1 and C2 = 0.025



Correlation of C1 and C2 = -0.811

Simple Regression One independent variable, one dependent. In the usual examples both are quantitative (continuous). We fit a **least-squares** line to the cloud of points in a scatterplot. The least-squares line is the unique line that minimizes the sum of squared vertical distances between the line and the points in the scatterplot. That is, it minimizes the total

(squared) error of prediction.

Denoting the slope of the least-squares line by b_1 and the intercept of the least-squares line by b_0,

$$b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \overline{Y} - b_1\overline{X}.$$

That is, the slope of the least squares has the same sign as the correlation coefficient, and equals zero if and only if the correlation coefficient is zero.

Usually, you want to test whether the slope is zero. This is the same as testing whether the correlation is zero, and mercifully yields the same p-value. Assumptions are independent observations (again) and that within levels of the IV, the DV has a normal distribution with the same variance (variance does not depend on value of the DV). Robustness properties are similar to those of the 2-sample t-test. The assumption of independent observations is always important.

Multiple Regression

Regression with several independent variables at once; we're fitting a (hyper) plane rather than a line. Multiple regression is very flexible; all the other techniques mentioned above (except the chi-squared test) are special cases of multiple regression. More details later.

1.3 Experimental versus observational studies

Why might someone want to predict a dependent variable from an independent variable? There are two main reasons.

- There may be a practical reason for prediction. For example, a company might wish to predict who will buy a product, in order to maximize the productivity of its sales force. Or, an insurance company might wish to predict who will make a claim, or a university computer centre might wish to predict the length of time a type of hard drive will last before failing. In each of these cases, there will be some independent variables that are to be used for prediction, and although the people doing the study may be curious and may have some ideas

about how things might turn out and why, they don't really care why it works, as long as they can predict with some accuracy. Does variation in the IV *cause* variation in the DV? Who cares?

- This may be science (of some variety). The goal may be to understand how the world works — in particular, to understand the dependent variable. In this case, most likely we are implicitly or explicitly thinking of a causal relationship between the IV and DV. Think of attitude similarity and interpersonal attraction ....

Sample Question 1.3.1 *A study finds that high school students who have a computer at home get higher grades on average than students who do not. Does this mean that parents who can afford it should buy a computer to enhance their children's chances of academic success?*

Here is an answer that gets **zero** points. "Yes, with a computer the student can become computer literate, which is a necessity in our competitive and increasingly technological society. Also the student can use the computer to produce nice looking reports (neatness counts!), and obtain valuable information on the World Wide Web." **ZERO**.

The problem with this answer is that while it makes some fairly reasonable points, it is based on personal opinion, and fails to address the real question, which is "**Does this mean** ..." Here is an answer that gets full marks.

Answer to Sample Question 1.3.1 *Not necessarily. While it is possible that some students are doing better academically and therefore getting into university because of their computers, it is also possible that their parents have enough money to buy them a computer, and also have enough money to pay for their education. It may be that an academically able student who is more likely to go to university will want a computer more, and therefore be more likely to get one somehow. Therefore, the study does not provide good evidence that a computer at home will enhance chances of academic success.*

Note that in this answer, the *focus is on whether the study provides good evidence* for the conclusion, not whether the conclusion is reasonable on other grounds. And the answer gives *specific alternative explanations* for the results as a way of criticizing the study. If you think about it, suggesting plausible alternative explanations is a very damaging thing to say about any empirical study, because you are pointing out that the investigators expended

a huge amount of time and energy, but didn't establish anything conclusive. Also, suggesting alternative explanations is extremely valuable, because that is how research designs get improved and knowledge advances.

Now here are the general principles. If X and Y are measured at roughly the same time, X could be causing Y, Y could be causing X, or there might be some third variable (or collection of variables) that is causing both X and Y. Therefore we say that "Correlation does not necessarily imply causation." Here, by correlation we mean association (lack of independence) between variables. It is not limited to situations where you would compute a correlation coefficient.

A **confounding variable** is a variable not included as an independent variable, that might be related to both the independent variable and the dependent variable – and that might therefore create a seeming relationship between them where none actually exists, or might even hide a relationship that is present. Some books also call this a "lurking variable." You are responsible for the vocabulary "confounding variable."

An **experimental study** is one in which cases are randomly assigned to the different values of an independent variable (or variables). An **observational study** is one in which the values of the independent variables are not randomly assigned, but merely observed.

Some studies are purely observational, some are purely experimental, and many are mixed. It's not really standard terminology, but in this course we will describe independent *variables* as experimental (i.e., randomly assigned, manipulated) or observed.

In an experimental study, there is no way the dependent variable could be causing the independent variable, because values of the IV are assigned by the experimenter. Also, it can be shown (using the Law of Large Numbers) that when units of observation are randomly assigned to values of an IV, all potential confounding variables are cancelled out as the sample size increases. This is very wonderful. You don't even have to know what they are!

Sample Question 1.3.2 *Is it possible for a continuous variable to be experimental, that is, randomly assigned?*

Answer to Sample Question 1.3.2 *Sure. In a drug study, let one of the independent variables consist of n equally spaced dosage levels spanning some range of interest, where n is the sample size. Randomly assign one participant to each dosage level.*

Sample Question 1.3.3 *Give an original example of a study with one quantitative observed independent variable and one categorical manipulated independent variable. Make the study multivariate, with one dependent variable consisting of unordered categories and two quantitative dependent variables. categorical*

Answer to Sample Question 1.3.3 *Stroke patients in a drug study are randomly assigned to either a standard blood pressure drug or one of three experimental blood pressure drugs. The categorical dependent variable is whether the patient is alive or not 5 years after the study begins. The quantitative dependent variables are systolic and diastolic blood pressure one week after beginning drug treatment.*

In practice, of course there would be a lot more variables; but it's still a good answer.

Because of possible confounding variables, only an experimental study can provide good evidence that an independent variable *causes* a dependent variable. Words like effect, affect, leads to etc. imply claims of causality and are only justified for experimental studies.

Sample Question 1.3.4 *Design a study that could provide good evidence of a causal relationship between having a computer at home and academic success.*

Answer to Sample Question 1.3.4 *High school students without computers enter a lottery. The winners (50% of the sample) get a computer and modem to use at home. The dependent variable is whether or not the student enters university.*

Sample Question 1.3.5 *Is there a problem with independent observations here? Can you fix it?*

Answer to Sample Question 1.3.5 *Oops. Yes. Students who win may be talking to each other, sharing software, etc.. Actually, the losers will be communicating too. Therefore their behaviour is non-independent and standard significance tests will be invalid. One solution is to hold the lottery in n separate schools, with one winner in each school. If the dependent variable were GPA, we could do a matched t-test comparing the performance of the winner to the average performance of the losers.*

Sample Question 1.3.6 *What if the DV is going to university or not?*

Answer to Sample Question 1.3.6 *We are getting into deep water here. Here is how I would do it. In each school, give a score of "1" to each student who goes to university, and a "0" to each student who does not. Again, compare the scores of the winners to the average scores of the losers in each school using a matched t-test. Note that the mean difference that is to be compared with zero here is the mean difference in probability of going to university, between students who get a computer to use and those who do not. While the differences for each school will not be normally distributed, the central limit theorem tells us that the mean difference will be approximately normal if there are more than about 20 schools, so the t-test is valid. In fact, the t-test is conservative, because the tails of the t distribution are heavier than those of the standard normal. This answer is actually beyond the scope of the present course.*

Artifacts and Compromises

Random assignment to experimental conditions will take care of confounding variables, but only if it is done right. It is amazingly easy for for confounding variables to sneak back into a true experimental study through defects in the procedure. For example, suppose you are interested in studying the roles of men and women in our society, and you have a 50-item questionnaire that (you hope) will measure traditional sex role attitudes on a scale from 0 = Very Non-traditional to 50 = Very Traditional. However, you suspect that the details of how the questionnaire is administered could have a strong influence on the results. In particular, the sex of the person administering the questionnaire and how he or she is dressed could be important.

Your subjects are university students, who must participate in your study in order to fulfill a course requirement in Introductory Psychology. You randomly assign your subjects to one of four experimental conditions: Female research assistant casually dressed, Female research assistant formally dressed, Male research assistant casually dressed, or Male research assistant formally dressed. Subjects in each experimental condition are instructed to report to a classroom at a particular time, and they fill out the questionnaire sitting all together.

This is an appealing procedure from the standpoint of data collection, because it is fast and easy. However, it is so flawed that it may be a complete waste of time to do the study at all. Here's why. Because subjects are

run in four batches, an unknown number of confounding variables may have crept back into the study. To name a few, subjects in different experimental conditions will be run at different times of day or different days of the week. Suppose subjects in the the male formally dressed condition fill out the questionnaire at 8 in the morning. Then *all* the subjects in that condition are exposed to the stress and fatigue of getting up early, as well as the treatment to which they have been randomly assigned.

There's more, of course. Presumably there are just two research assistants, one male and one female. So there can be order effects; at the very least, the lab assistant will be more practiced the second time he or she administers the questionnaire. And, though the research assistants will surely try to administer the questionnaire in a standard way, do you really believe that their body language, facial expressions and tone of voice will be identical both times?

Of course, the research assistants know what condition the subjects are in, they know the hypotheses of the study, and they probably have a strong desire to please the boss — the investigator (professor or whatever) who is directing this turkey, uh, excuse me, I mean this research. Therefore, their behaviour could easily be slanted (perhaps unconsciously so) to produce the hypothesized effects.

This kind phenomenon is well-documented. It's called *experimenter expectancy*. Experimenters find what they expect to find. If they are led to believe that certain mice are very intelligent, then those mice will do better on all kinds of learning tasks, even though in fact the mice were randomly assigned to be labeled as "intelligent." This kind of thing applies all the way down to flatworms. The classic reference is Robert Rosenthal's *Experimenter expectancy in behavioral research* [9]. Naturally, the expectancy phenomenon applies to teachers and students in a classroom setting, where it is called *teacher expectancy*. The reference for this is Rosenthal and Jacobson's *Pygmalion in the classroom* [10].

It is wrong (and complacent) to believe that expectancy effects are confined to psychological research. In medicine, *placebo effects* are well-documented. Patients who are given an inert substance like a sugar pill do better than patients who are not, provided that they (or their doctors) believe that they are getting medicine that works. Is it the patients' expectancies that matter, or the doctors'? Probably both. The standard solution, and the *only* acceptable solution in clinical trials of new drugs, is the so called *double blind*, in which subjects are randomly assigned to receive either the drug or a placebo,

and neither the patient nor the doctor knows which it is. This is the gold standard. Accept no substitutes.

Until now, we have been discussing threats to the *Internal Validity* of research. A study has good internal validity if it's designed to eliminate the influence of confounding variables, so one can be reasonably sure that the observed effects really are being produced by the independent variables of interest. But there's also *External Validity*. External validity refers to how well the phenomena outside the laboratory or data-collection situation are being represented by the study. For example, well-controlled, double-blind taste tests indicated that the Coca-cola company had a recipe that consumers liked better than the traditional one. But attempts to market "New" Coke were an epic disaster. There was just more going on in the real world of soft drink consumption than in the artificial laboratory setting of a taste test. Cook and Campbell's *Quasi-experimentation* [3] contains an excellent discussion of internal versus external validity.

In Industrial-Organizational psychology, we have the *Hawthorne Effect*, which takes its name from the Hawthorne plant of General Electric, where some influential studies of worker productivity were carried out in the 1930's. The basic idea is that when workers know that they are part of a study, almost anything you do will increase productivity. Make the lights brighter? Productivity increases. Make the lights dimmer? Productivity increases. This is how the Hawthorne Effect is usually described. The actual details of the studies and their findings are more complex [8], but the general idea is that when people know they are particpating in a study, they tend to feel more valued, and act accordingly. In this respect, the fact that the subjects know that a study is being carried can introduce a serious distortion into the way things work, and make the results unrepresentative of what normally happens.

Medical research on non-human animals is always at least subject to discussion on grounds of external validity, as is almost any laboratory research in Psychology. Do you know why the blood vessels running away from the heart are called "arteries?" It's because they were initially thought to contain air. Why? Because medical researchers were basing their conclusions entirely on dissections of dead bodies. In live bodies, the arteries are full of blood.

Generally speaking, the controlled environments that lead to the best internal validity also produce the greatest threats to external validity. Is a given laboratory setup capturing the essence of the phenomena under con-

sideration, or is it artificial and irrelevant? It's usually hard to tell. The best way to make an informed judgement is to compare laboratory studies and field studies that are trying to answer the same questions. The laboratory studies usually have better internal validity, and the field studies usually have better external validity. When the results are consistent, we feel more comfortable.

Chapter 2

First set of tools: SAS running under unix (including linux)

The SAS language is the same regardless of what hardware you use or what operating system is running on the hardware. SAS programs are simple text files that can be transported from one machine to another with minimal difficulty. In this course, everything will be illustrated with SAS running under the unix operating system, but it's not a problem even if the next place you go only has PCs. The adjustment to SAS-PC should be fast and fairly painless.

2.1 Unix

Unix is a line-oriented operating system. Well, there's X-windows (a graphical shell that runs on top of unix), but we won't bother with it. Basically, you type a command, press Enter, and unix does something for (or to) you. It may help to think of unix as DOS on steroids, if you remember DOS. The table below has all the unix commands you will need for this course. Throughout, *fname* stands for the name of a file.

A Minimal Set of unix Commands

exit Logs you off the system: ALWAYS log off before leaving!

passwd Lets you change your password. Recommended.

man *command name* Online help: explains *command name*, (like `man sort`).

ls Lists names of the files in your directory.

less *fname* Displays fname on screen, one page at a time. Spacebar for next page, `q` to quit.

lpr *fname* Prints hard copy on a laser printer. lpr stands for line printer. These physical devices no longer exist in most installations.

rm *fname* Removes *fname*, erasing it forever.

cp *fname1 fname2* Makes a copy of *fname1*. The new copy is named *fname2*.

mv *fname1 fname2* Moves (renames) *fname1*

emacs *fname* Starts the `emacs` text editor, editing *fname* (can be new file).

R Gets you into the `R` implementation of the S environment.

sas *fname* Executes `SAS` commands in the file *fname*`.sas`, yielding *fname*`.log` and (if no fatal errors) *fname*`.lst`.

ps Shows active processes

kill -9 # Kills process (job) number `#`. Sometimes you must do this when you can't log off because there are stopped jobs. Use `ps` to see the job numbers.

mail yourname@yourisp.com < *fname* Email a file to yourself. Very handy for getting files to your home computer for printing.

curl *URL* > *fname* A *URL* is a Web address. This command is intended to help you get a copy of the source code of Web pages. But when the web page contains just a data file, as it sometimes does in this course, this is a great way to get a copy of the data. Copy the *URL* from your browser. `curl http://fisher.utstat.toronto.edu/~brunner/429s07/code_n_data/drp.dat > drp.dat`

This really is a minimal set of commands. The unix operating system is extremely powerful, and has an enormous number of commands. You can't really see the power from the minimal set above, but you can see the main drawback from the standpoint of the new user. Commands tend to be terse, consisting of just a few keystrokes. They make sense once you are familiar with them (like `ls` for listing the files in a directory, or `rm` for remove), but they are hard to guess. The `man` command (short for manual) gives very accurate information, but you have to know the name of the command before you can use `man` to find out about it.

Just for future reference, here are a few more commands that you may find useful, or otherwise appealing.

A Few More unix Commands

mkdir *dirname* Makes a new sub-directory (like a folder) named *dirname*. You can have sub-directories within sub-directories; it's a good way to organize your work.

cp *fname dirname* Copies the file *fname* into the directory *dirname*.

cd *dirname* Short for Change Directory. Takes you to the sub-directory *dirname*.

cd .. Moves you up a directory level.

cd Moves you to your main directory from wherever you are.

ls > *fname* Sends the output of the `ls` command to the file *fname* instead of to the screen.

cat *fname* Lists the whole file on your screen, not one page at a time. It goes by very fast, but usually you can scroll back up to see the entire file, if it's not too long.

cat *fname1 fname2* > *fname3* Concatenates *fname1* and *fname2* (sticks them together) and re-directs the output to *fname3*

grep ERROR cartoon1.log Searches for the string ERROR in the file cartoon1.log. Echos each line containing the string. Silent if ERROR does not occur. Case sensitive.

alias chk "grep ERROR *.log ; grep WARN *.log" Makes a new command called `chk`. It checks for the string ERROR and the string WARN in any log file.

cal Displays a calendar for this month

cal 1 3002 Displays a calendar for January 3002.

unset noclobber Are you tired of being asked if you really want to remove or overwrite a file?

rm *fname1 fname2* Remove both

rm -f *fname* Remove without asking for confirmation, this time only.

alias rm "rm -f" `rm` now means `rm -f`.

rm -r *dirname* Remove the directory, and everything in it recursively.

R –vanilla < *fname1* > *fname2* Execute the S language commands in *fname1*, sending output to *fname2*. Run in "plain vanilla" mode.

Printing files at home This is a question that always comes up. Almost surely, the printer connected to your printer at home is not directly connected to the university network. If you want to do something like print your SAS output at home, you have to transfer the file on the unix machine to the hard drive of your home computer, and print it from there. One way is to use some kind of `ftp` (file transfer protocol) tool to get the file in question onto your hard drive. For short files, you can also use the `less` or `cat` command to list the file on your screen, select it with your mouse, copy it, paste it to a word processing document, and print it from there. It is a good idea to use a fixed-width font like Courier, and not the Times or Times Roman font. Everything will be lined up better.

Perhaps easiest of all is to email yourself the file. This is illustrated in the first set of unix commands. To repeat, `mail yourname@yourisp.com < fname`.

2.2 Introduction to SAS

SAS stands for "Statistical Analysis System." Even though it runs on PCs 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 the interface still basically file and text oriented, not graphical.

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 `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 `.log`. 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. In this text, all the examples will be run from the unix command line (`terminal`). In my view, this way is simplest and also the best way to start. Also, it is by far the easiest way to use SAS from home, assuming that SAS is running on a remote server and not your home computer.

The following illustrates a simple SAS run from the command line (using an application called `terminal` in some unix and linux environments). Initially, there are only two files in the 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 `appsrv01.srv` (the name of the unix machine used to produce the examples), followed by a `>` sign.

```
appsrv01.srv> ls
drp.dat         reading.sas
appsrv01.srv> sas reading
appsrv01.srv> 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 at least one 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 here, 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 tandard` 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 most noticeable exception is `proc format`, which is used to provide labels for the values of categorical 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 set 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 `less` 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. This is 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 and end 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 `less` command or the `cat` command, 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 that the name of the SAS data set being created is "reading." The names of data sets

are arbitrary, but you should make them informative.

- `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.

```
appsrv01.srv> cat reading.log
1                                                          The SAS System
                          11:40 Thursday, September 2, 2007

NOTE: Copyright (c) 1999-2001 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.2 (TS2M0)
      Licensed to UNIVERSITY OF TORONTO/COMPUTING & COMMUNICATIONS, Site 00089870
NOTE: This session is executing on the Linux 2.6.8.1-smp-athlon-bk platform.



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: SAS initialization used:
      real time           0.08 seconds
      cpu time            0.01 seconds

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
8          data reading;
```

```
9               infile 'drp.dat';
10              input group $ score;
11              label group = 'Get Directed Reading Programme?'
12                    score = 'Degree of Reading Power Test Score';

NOTE: The infile 'drp.dat' is:
      File Name=/homes/students/u0/stats/brunner/drp.dat,
      Owner Name=brunner,Group Name=stats,
      Access Permission=rw-r-----,
      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           0.01 seconds
      cpu time            0.01 seconds


13         proc ttest;
14              class group;
15              var score;
NOTE: There were 44 observations read from the data set WORK.READING.
NOTE: The PROCEDURE TTEST printed page 1.
NOTE: PROCEDURE TTEST used:
      real time           0.08 seconds
      cpu time            0.01 seconds


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

2                                The SAS System
                                              11:40 Thursday, September 2, 2007

NOTE: The SAS System used:
      real time           0.24 seconds
      cpu time            0.03 seconds
```

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 statistical output — the t-test we want, and also a bunch of other stuff, 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 things 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.

```
appsrv01.srv> cat reading.lst
```

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

              More & McCabe (1993) textbook t-test Example 7.8              1
                                          11:40 Thursday, September 2, 2007

                            The TTEST Procedure

                               Statistics

                                  Lower CL          Upper CL  Lower CL
  Variable  group          N       Mean    Mean      Mean   Std Dev  Std Dev

  score     Control       23     34.106  41.522    48.937    13.263   17.149
  score     Treatmen      21     46.466  51.476    56.487    8.4213   11.007
  score     Diff (1-2)           -18.82  -9.954    -1.091    11.998   14.551
```

Statistics

Variable	group	Upper CL Std Dev	Std Err	Minimum	Maximum
score	Control	24.271	3.5758	10	85
score	Treatmen	15.895	2.402	24	71
score	Diff (1-2)	18.495	4.3919		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
score	Pooled	Equal	42	-2.27	0.0286
score	Satterthwaite	Unequal	37.9	-2.31	0.0264

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
score	Folded F	22	20	2.43	0.0507

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

- The first part of the output is labelled "Statistics," containing confidence intervals and some descriptive statistics. There are three rows to this display: one for the Control group, one for the Treatment group, and one for the difference between groups.
  - `Variable: score` The first column, labelled "Variable," tells you what the dependent variable is – particularly useful if you have more than one.
  - `group` The independent variable. Underneath are the values of the independent variable in the first two rows. The third row is for the difference, computed as Control minus Treatment (1-2).

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.

- `N` The third column gives sample sizes; n=23 for the control group, and n=21 for the treatment group.
- The next three columns contain means and their associated 95% confidence intervals. The middle column has the means. For the Control group, the sample mean score on the DRP test was 41.522; for the Treatment group, the sample mean was 51.476. The difference between means is -9.954 = 41.522-51.476 (One minus Two).

  To the left of the mean, labelled "Lower CL Mean," is the lower confidence limit of the 95% confidence interval for the population mean. Thus, the 95% confidence interval for the Treatment mean is from 46.466 to 56.487, and the 95% confidence interval for the *difference* between means is from -18.82 to -1.091. The fact that this last interval does not contain zero means that the usual two-tailed t-test will be statistically significantly at the 0.05 level. Thre is a lovely consistency between the classical tests and confidence intervals.
- The next three columns give confidence intervals for the standard deviations. We have the lower confidence limit, the standard deviation, and the upper confidence limit in the continuation below.
- Then we get standard errors (estimated standard deviations of the sample mean or difference between means), and finally the minimum and maximum for each group.

- Then finally, under "T-Tests," we get what we want – a t-test for the difference between the means of the Control and Treatment groups. Two tests are given; as usual there seems to be more output than

we were expecting or wishing for. Probably we were looking for the first one, using the Pooled Method. This is the traditional test, which assumes equal population variances, and therefore is based on a *Pooled* estimate of the common within-groups standard deviation.

- The value of the test statistic is `-2.27`.
- 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. The question "What do we conclude from this study?" will always be asked. For now, 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. Later we will take up the possibility of concluding that the effect we are testing is actually *absent* (or at least trivially small).

Many students, even when they have been warned, respond to the "what do you conclude" question 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 Satterthwaite method 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 [6], or lots of other elementary texts, 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 a test for Equality of Variances. This "Folded" F 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* [6]. 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 .... 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. Here are the first few lines.

```
appsrv01.srv>  less statclass1.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  .
1  1  10  6  7  9  8  8  5  7  10  9  10  9  5  6  4  8  10  57  52
0  1  0  0  9  9  10  5  2  2  8  7  7  10  10  6  3  7  10  49  .
0  1  10  9  5  8  9  8  5  6  8  7  5  6  10  6  5  9  9  77  64
0  1  10  8  6  8  9  5  3  6  9  9  6  9  10  6  5  7  10  65  42
1  1  10  5  6  7  10  4  6  0  10  9  10  9  10  6  7  8  10  73  .
0  1  9  0  4  6  10  5  3  3  10  8  10  5  10  10  9  9  10  71  37

⋮
```

Notice the periods at the ends of lines 5, 7 and 10. The period is the SAS *missing value code*. These people did not show up for the final exam. They may have taken a makeup exam, but if so their scores did not make it into this data file. When a case has a missing value recorded for a variable, SAS automatically excludes that case from any statistical calculation involving the variable. If a new variable is being created based on the value of a variable with a missing value, the new variable will usually have a missing value for that case too.

Here is the SAS program textttstatmarks1.sas. It reads and labels the data, and then does a variety of significance tests. They are all elementary except the last one, which illustrates testing for one set of independent variables controlling for another set in multiple regression.

```
appsrv01.srv>  cat statmarks1.sas

                  /* statmarks1.sas */
options linesize=79 noovp formdlim='_';
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 'statclass1.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';
     mark2 = round(mark);
     /* Bump up at grade boundaries */
     if mark2=89 then mark2=90;
     if mark2=79 then mark2=80;
     if mark2=69 then mark2=70;
     if mark2=59 then mark2=60;
     /* Assign letter grade */
     if mark2=. then grade='Incomplete';
        else if mark2 ge 90 then grade = 'A';
        else if 80 le mark2 le 89 then grade='B';
        else if 70 le mark2 le 79 then grade='C';
        else if 60 le mark2 le 69 then grade='D';
        else grade='F';
```

```
     format sex sexfmt.;       /* Associates sex & ethnic    */
     format ethnic ethfmt.;    /* with formats defined above */

/* Now the proc steps */


proc freq;
      title3 'Frequency distributions of the categorical variables';
      tables sex ethnic grade;

proc means n mean std;
     title3 'Means and SDs of quantitative variables';
     var quiz1 -- mark;        /*  single dash only works with numbered
                                  lists, like quiz1-quiz8    */
proc ttest;
     title3 'Independent t-test';
     class sex;
     var mark;
proc means n mean std t;
     title3 'Matched t-test: Quiz 1 versus 8';
     var quiz1 quiz8 diff;
proc glm;
     title3 'One-way anova';
     class ethnic;
     model mark = ethnic;
     means ethnic;
     means ethnic / Tukey Bon Scheffe;
proc freq;
     title3 'Chi-squared Test of Independence';
     tables sex*ethnic sex*grade ethnic*grade / chisq;
proc freq; /*  Added after seeing warning from chisq test above */
     title3 'Chi-squared Test of Independence: Version 2';
     tables sex*ethnic grade*(sex ethnic) / norow nopercent chisq expected;
proc corr;
     title3 'Correlation Matrix';
     var final midterm quizave compave;
proc plot;
     title3 'Scatterplot';
```

```
     plot final*midterm; /* Really should do all combinations */
proc reg;
     title3 'Simple regression';
     model final=midterm;

/* Predict final exam score from midterm, quiz & computer */
proc reg simple;
     title3 'Multiple Regression';
     model final = midterm quizave compave / ss1;
     smalstuf: test quizave = 0, compave = 0;
run;

/* Note that the final run statement is not needed when
   running SAS from the unix command line. */
```

Noteworthy features of this program include

- `options`: Already discussed in connection with `reading.sas`.
- `title2`: Subtitle
- `proc format`: This is a non-statistical procedure – a rarity in the SAS language. It is the way SAS takes care of labelling categorical variables when the categories are coded as numbers. `proc format` defines *printing formats*. For any variable associated with the printing format named `sexfmt`, any time it would print the value "0" (in a table or something) it instead prints the string "Male." The associations between variables and printing formats are accomplished in the `format` statement at the end of the data step. The names of formats have a period at the end to distinguish them from variable names. Of course formats must be defined before they can be associated with variables. This is why `proc format` precedes the data step.
- `quiz1-quiz8`: One may refer to a *range* of variables ending with consecutive numbers using a minus sign. In the `input` statement, a range can be defined (named) this way. It saves typing and is easy to read.
- Creating new variables with assignment statements. The variables `quizave`, `compave` and `mark` are not in the original data file. They are created here, and they are appended to the end of the SAS data

set in oder of creation. Variables like this should never be in the raw data file.

Data Analysis Hint 3 *When variables are exact mathematical functions of other variables, always create them in the data step rather than including them in the raw data file. It saves data entry, and makes the data file smaller and easier to read. If you want to try out a different definition of the variable, it's easy to change a few statements in the data step.*

- `sum(of quiz1-quiz8)`: Without the word "of," the minus sign is ambiguous. In the SAS language, `sum(quiz1-quiz8)` is the sum of a single number, the difference between `quiz1` and `quiz8`.
- `format sex sexfmt.;` Associates the variable `sex` with its printing format. In questionnaire studies where a large number of items have the same potential responses (like a scale from 1 = Strongly Agree to 7=Strongly Disagree), it is common to associate a long list of variables with a single printing format.
- `quiz1 -- mark` in the first `proc means`: A double dash refers to a list of variables *in the order of their creation* in the `data` step. Single dashes are for numerical order, while double dashes are for order of creation; it's very handy.
- Title inside a procedure labels just that procedure.
- `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 the IV ethnic is categorical.
  - `model` Dependent variable(s) = independent variable(s)
  - `means ethnic`: Mean of `mark` separately for each value of `ethnic`.
  - `means ethnic / Tukey Bon Scheffe`: Post hoc tests (multiple comparisons, probing, follow-ups). Used if the overall F-test is significant, to see which means are different from which other means.

- `chisq` option on `proc freq`: Gives a large collection of chisquare tests. The first one is the familiar Pearson chisquare test of independence (the one comparing observed and expected frequencies).

- `tables sex*ethnic / norow nopercent chisq expected;` In this second version of the crosstab produced `proc freq`, we suppress the row and total percentages, and look at the expected frequencies because SAS warned us that some of them were too small. SAS issues a warning if any expected frequency is below 5; this is the old-fashioned rule of thumb. But it has been known for some time that Type I eror rates are affected mostly by expected frequencies smaller than one, not five — so I wanted to take a look.

- `proc corr` After `var`, list the variables you want to see in a correlation matrix.

- `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.

statmarks1.lst

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

           Grades from STA3000 at Roosevelt University:  Fall, 1957           1
                          Illustrate Elementary Tests
             Frequency distributions of the categorical variables
                                                10:10 Sunday, September 5, 2007

                               The FREQ Procedure

                                               Cumulative    Cumulative
             sex    Frequency     Percent     Frequency      Percent
          -----------------------------------------------------------
          Male            39       62.90            39        62.90
          Female          23       37.10            62       100.00
```

Apparent ethnic background (ancestry)

ethnic	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Chinese	41	66.13	41	66.13
European	15	24.19	56	90.32
Other	6	9.68	62	100.00

grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	3	4.84	3	4.84
B	6	9.68	9	14.52
C	18	29.03	27	43.55
D	21	33.87	48	77.42
F	10	16.13	58	93.55
Incomplete	4	6.45	62	100.00

Grades from STA3000 at Roosevelt University: Fall, 1957 2
Illustrate Elementary Tests
Means and SDs of quantitative variables
10:10 Sunday, September 5, 2007

The MEANS Procedure

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		58	50.3103448	17.2496701
quizave	Quiz Average (drop lowest)	62	7.6751152	1.1266917
compave	Computer Average (drop lowest)	62	8.8346774	1.1204997
mark	Final Mark	58	68.4830049	10.3902874

Illustrate Elementary Tests

Independent t-test

10:10 Sunday, September 5, 2007

The TTEST Procedure

Statistics

Variable	sex	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev
mark	Male	36	65.604	68.57	71.535	7.1093	8.7653
mark	Female	22	62.647	68.341	74.036	9.8809	12.843
mark	Diff (1-2)		-5.454	0.2284	5.9108	8.8495	10.482

Statistics

Variable	sex	Upper CL Std Dev	Std Err	Minimum	Maximum
mark	Male	11.434	1.4609	54.057	89.932
mark	Female	18.354	2.7382	48.482	95.457
mark	Diff (1-2)	12.859	2.8366		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
mark	Pooled	Equal	56	0.08	0.9361
mark	Satterthwaite	Unequal	33.1	0.07	0.9418

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
mark	Folded F	21	35	2.15	0.0443

Illustrate Elementary Tests

Matched t-test: Quiz 1 versus 8

10:10 Sunday, September 5, 2007

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	t Value
quiz1		62	9.0967742	2.2739413	31.50
quiz8		62	6.3064516	2.3787909	20.87
diff	Quiz 8 minus Quiz 1	62	-2.7903226	3.1578011	-6.96

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

          Grades from STA3000 at Roosevelt University:  Fall, 1957          5
                        Illustrate Elementary Tests
                              One-way anova  10:10 Sunday, September 5, 2007

                               The GLM Procedure

                            Class Level Information

                 Class         Levels    Values

                 ethnic             3    Chinese European Other


                           Number of observations    62

NOTE: Due to missing values, only 58 observations can be used in this analysis.

---------------------------------------------------------------------------

          Grades from STA3000 at Roosevelt University:  Fall, 1957          6
                        Illustrate Elementary Tests
                              One-way anova  10:10 Sunday, September 5, 2007

                               The GLM Procedure

Dependent Variable: mark   Final Mark

                                     Sum of
Source                       DF        Squares    Mean Square   F Value   Pr > F

Model                         2    1238.960134     619.480067      6.93   0.0021

Error                        55    4914.649951      89.357272

Corrected Total              57    6153.610084


                R-Square     Coeff Var      Root MSE     mark Mean

                0.201339      13.80328      9.452898      68.48300


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

ethnic                        2    1238.960134     619.480067      6.93   0.0021


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

ethnic                        2    1238.960134     619.480067      6.93   0.0021

---------------------------------------------------------------------------

          Grades from STA3000 at Roosevelt University:  Fall, 1957          7
                        Illustrate Elementary Tests
```

One-way anova 10:10 Sunday, September 5, 2007

The GLM Procedure

Level of ethnic	N	mark Mean	mark Std Dev
Chinese	37	65.2688224	7.9262171
European	15	76.0142857	11.2351562
Other	6	69.4755952	13.3097753

Illustrate Elementary Tests
One-way anova 10:10 Sunday, September 5, 2007

The GLM Procedure

Tukey's Studentized Range (HSD) Test for mark

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

Alpha	0.05
Error Degrees of Freedom	55
Error Mean Square	89.35727
Critical Value of Studentized Range	3.40649

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

ethnic Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
European - Other	6.539	-4.460	17.538	
European - Chinese	10.745	3.776	17.715	***
Other - European	-6.539	-17.538	4.460	
Other - Chinese	4.207	-5.814	14.228	
Chinese - European	-10.745	-17.715	-3.776	***
Chinese - Other	-4.207	-14.228	5.814	

Illustrate Elementary Tests
One-way anova 10:10 Sunday, September 5, 2007

The GLM Procedure

Bonferroni (Dunn) t Tests for mark

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

```
                    Alpha                          0.05
                    Error Degrees of Freedom         55
                    Error Mean Square          89.35727
                    Critical Value of t         2.46941


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


                                Difference
                                  Between     Simultaneous 95%
          ethnic                    Means    Confidence Limits
        Comparison

        European - Other             6.539      -4.737   17.814
        European - Chinese          10.745       3.600   17.891  ***
        Other    - European         -6.539     -17.814    4.737
        Other    - Chinese           4.207      -6.067   14.480
        Chinese  - European        -10.745     -17.891   -3.600  ***
        Chinese  - Other            -4.207     -14.480    6.067
```

```
                      Illustrate Elementary Tests
                             One-way anova  10:10 Sunday, September 5, 2007

                           The GLM Procedure

                         Scheffe's Test for mark

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


                    Alpha                          0.05
                    Error Degrees of Freedom         55
                    Error Mean Square          89.35727
                    Critical Value of F         3.16499


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


                                Difference
                                  Between     Simultaneous 95%
          ethnic                    Means    Confidence Limits
        Comparison

        European - Other             6.539      -4.950   18.027
        European - Chinese          10.745       3.466   18.025  ***
        Other    - European         -6.539     -18.027    4.950
        Other    - Chinese           4.207      -6.260   14.674
        Chinese  - European        -10.745     -18.025   -3.466  ***
        Chinese  - Other            -4.207     -14.674    6.260
```

The FREQ Procedure

Table of sex by ethnic

sex ethnic(Apparent ethnic background (ancestry))

Frequency Percent Row Pct Col Pct	Chinese	European	Other	Total
Male	27 43.55 69.23 65.85	7 11.29 17.95 46.67	5 8.06 12.82 83.33	39 62.90
Female	14 22.58 60.87 34.15	8 12.90 34.78 53.33	1 1.61 4.35 16.67	23 37.10
Total	41 66.13	15 24.19	6 9.68	62 100.00

Statistics for Table of sex by ethnic

Statistic	DF	Value	Prob
Chi-Square	2	2.9208	0.2321
Likelihood Ratio Chi-Square	2	2.9956	0.2236
Mantel-Haenszel Chi-Square	1	0.0000	0.9949
Phi Coefficient		0.2170	
Contingency Coefficient		0.2121	
Cramer's V		0.2170	

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

Sample Size = 62

The FREQ Procedure

Table of sex by grade

sex	grade						
Frequency| Percent | Row Pct | Col Pct	A	B	C	D	F	Incomplete	Total
Male	1	3	13	14	5	3	39
	1.61	4.84	20.97	22.58	8.06	4.84	62.90
	2.56	7.69	33.33	35.90	12.82	7.69	
	33.33	50.00	72.22	66.67	50.00	75.00	
Female	2	3	5	7	5	1	23
	3.23	4.84	8.06	11.29	8.06	1.61	37.10
	8.70	13.04	21.74	30.43	21.74	4.35	
	66.67	50.00	27.78	33.33	50.00	25.00	
Total	3	6	18	21	10	4	62
	4.84	9.68	29.03	33.87	16.13	6.45	100.00

Statistics for Table of sex by grade

Statistic	DF	Value	Prob
Chi-Square	5	3.3139	0.6517
Likelihood Ratio Chi-Square	5	3.2717	0.6582
Mantel-Haenszel Chi-Square	1	0.2342	0.6284
Phi Coefficient		0.2312	
Contingency Coefficient		0.2253	
Cramer's V		0.2312	

WARNING: 58% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 62

The FREQ Procedure

Table of ethnic by grade

ethnic(Apparent ethnic background (ancestry)) grade

Frequency / Percent / Row Pct / Col Pct	A	B	C	D	F	Incomplete	Total
Chinese	0	2	11	17	7	4	41
	0.00	3.23	17.74	27.42	11.29	6.45	66.13
	0.00	4.88	26.83	41.46	17.07	9.76	
	0.00	33.33	61.11	80.95	70.00	100.00	
European	2	4	5	3	1	0	15
	3.23	6.45	8.06	4.84	1.61	0.00	24.19
	13.33	26.67	33.33	20.00	6.67	0.00	
	66.67	66.67	27.78	14.29	10.00	0.00	
Other	1	0	2	1	2	0	6
	1.61	0.00	3.23	1.61	3.23	0.00	9.68
	16.67	0.00	33.33	16.67	33.33	0.00	
	33.33	0.00	11.11	4.76	20.00	0.00	
Total	3	6	18	21	10	4	62
	4.84	9.68	29.03	33.87	16.13	6.45	100.00

Statistics for Table of ethnic by grade

Statistic	DF	Value	Prob
Chi-Square	10	18.2676	0.0506
Likelihood Ratio Chi-Square	10	19.6338	0.0329
Mantel-Haenszel Chi-Square	1	5.6222	0.0177
Phi Coefficient		0.5428	
Contingency Coefficient		0.4771	
Cramer's V		0.3838	

WARNING: 78% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 62

The FREQ Procedure

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	2.9208	0.2321
Likelihood Ratio Chi-Square	2	2.9956	0.2236
Mantel-Haenszel Chi-Square	1	0.0000	0.9949
Phi Coefficient		0.2170	
Contingency Coefficient		0.2121	
Cramer's V		0.2170	

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

Sample Size = 62

The FREQ Procedure

Table of grade by sex

grade sex

Frequency / Expected / Col Pct	Male	Female	Total
A	1 1.8871 2.56	2 1.1129 8.70	3
B	3 3.7742 7.69	3 2.2258 13.04	6
C	13 11.323 33.33	5 6.6774 21.74	18
D	14 13.21 35.90	7 7.7903 30.43	21
F	5 6.2903 12.82	5 3.7097 21.74	10
Incomplete	3 2.5161 7.69	1 1.4839 4.35	4
Total	39	23	62

Statistics for Table of grade by sex

Statistic	DF	Value	Prob
Chi-Square	5	3.3139	0.6517
Likelihood Ratio Chi-Square	5	3.2717	0.6582
Mantel-Haenszel Chi-Square	1	0.2342	0.6284
Phi Coefficient		0.2312	
Contingency Coefficient		0.2253	
Cramer's V		0.2312	

WARNING: 58% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 62

The FREQ Procedure

Table of grade by ethnic

grade ethnic(Apparent ethnic background (ancestry))

Frequency / Expected / Col Pct	Chinese	European	Other	Total
A	0	2	1	3
	1.9839	0.7258	0.2903	
	0.00	13.33	16.67	
B	2	4	0	6
	3.9677	1.4516	0.5806	
	4.88	26.67	0.00	
C	11	5	2	18
	11.903	4.3548	1.7419	
	26.83	33.33	33.33	
D	17	3	1	21
	13.887	5.0806	2.0323	
	41.46	20.00	16.67	
F	7	1	2	10
	6.6129	2.4194	0.9677	
	17.07	6.67	33.33	
Incomplete	4	0	0	4
	2.6452	0.9677	0.3871	
	9.76	0.00	0.00	
Total	41	15	6	62

Statistics for Table of grade by ethnic

Statistic	DF	Value	Prob
Chi-Square	10	18.2676	0.0506
Likelihood Ratio Chi-Square	10	19.6338	0.0329
Mantel-Haenszel Chi-Square	1	5.6222	0.0177
Phi Coefficient		0.5428	
Contingency Coefficient		0.4771	
Cramer's V		0.3838	

WARNING: 78% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 62

The CORR Procedure

4 Variables: final midterm quizave compave

Simple Statistics

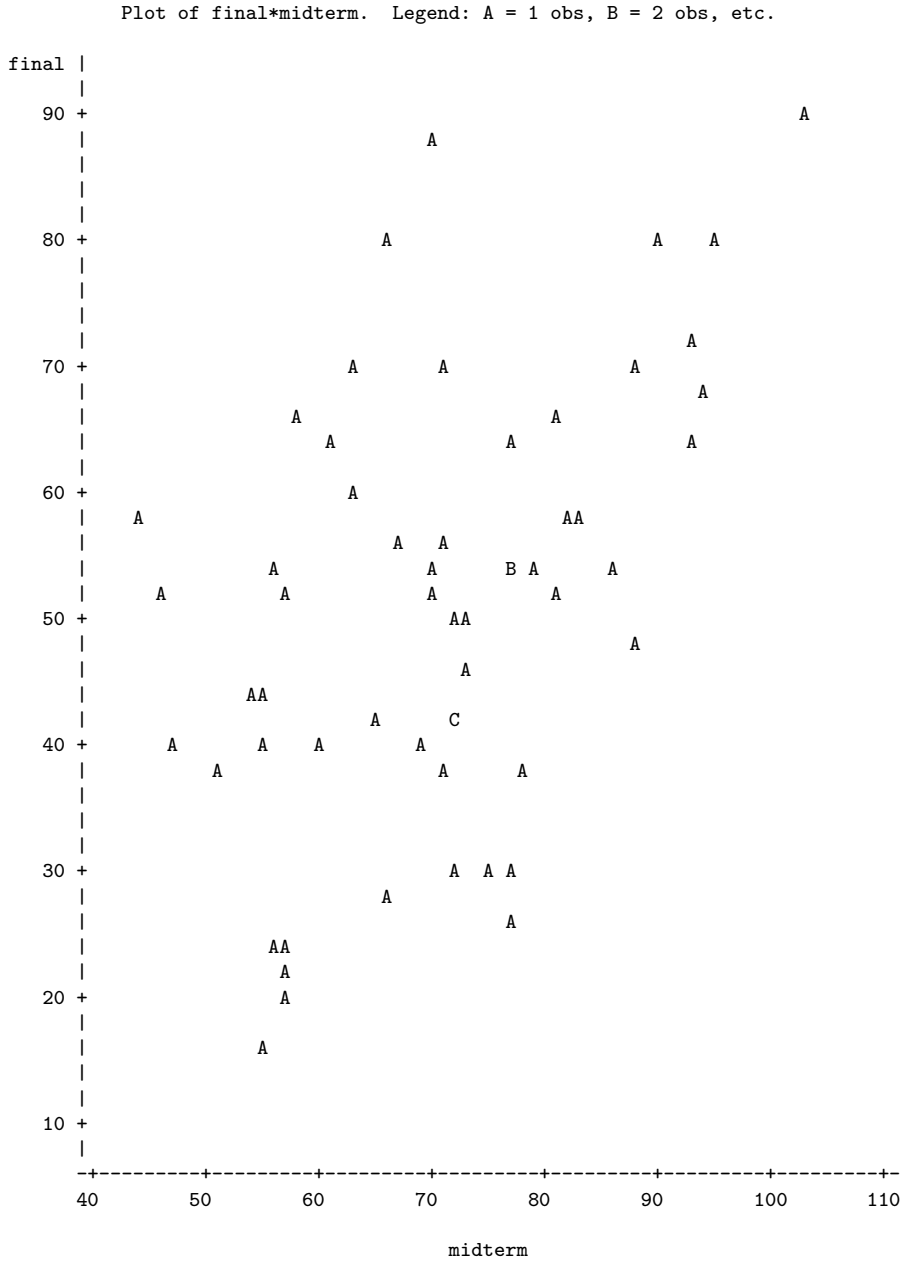
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
final	58	50.31034	17.24967	2918	15.00000	89.00000
midterm	62	70.19355	13.62356	4352	44.00000	103.00000
quizave	62	7.67512	1.12669	475.85714	4.57143	9.71429
compave	62	8.83468	1.12050	547.75000	5.00000	10.00000

Simple Statistics

Variable	Label
final	
midterm	
quizave	Quiz Average (drop lowest)
compave	Computer Average (drop lowest)

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	final	midterm	quizave	compave
final	1.00000	0.47963	0.41871	0.06060
		0.0001	0.0011	0.6513
	58	58	58	58
midterm	0.47963	1.00000	0.59294	0.41277
	0.0001		<.0001	0.0009
	58	62	62	62
quizave	0.41871	0.59294	1.00000	0.52649
Quiz Average (drop lowest)	0.0011	<.0001		<.0001
	58	62	62	62
compave	0.06060	0.41277	0.52649	1.00000
Computer Average (drop lowest)	0.6513	0.0009	<.0001	
	58	62	62	62



NOTE: 4 obs had missing values.

Simple regression
10:10 Sunday, September 5, 2007

The REG Procedure
Model: MODEL1
Dependent Variable: final

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3901.64751	3901.64751	16.73	0.0001
Error	56	13059	233.19226		
Corrected Total	57	16960			

Root MSE	15.27063	R-Square	0.2300
Dependent Mean	50.31034	Adj R-Sq	0.2163
Coeff Var	30.35287		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.88931	10.80304	0.64	0.5263
midterm	1	0.61605	0.15061	4.09	0.0001

Illustrate Elementary Tests
Multiple Regression
10:10 Sunday, September 5, 2007

The REG Procedure

Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	58.00000	1.00000	58.00000	0	0
midterm	4088.00000	70.48276	298414	180.35935	13.42979
quizave	451.57143	7.78571	3576.51020	1.06498	1.03198
compave	515.50000	8.88793	4641.50000	1.04862	1.02402
final	2918.00000	50.31034	163766	297.55112	17.24967

Descriptive Statistics

Variable	Label
Intercept	Intercept
midterm	
quizave	Quiz Average (drop lowest)

compave Computer Average (drop lowest)
final

The REG Procedure
Model: MODEL1
Dependent Variable: final

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4995.04770	1665.01590	7.51	0.0003
Error	54	11965	221.58085		
Corrected Total	57	16960			

Root MSE	14.88559	R-Square	0.2945
Dependent Mean	50.31034	Adj R-Sq	0.2553
Coeff Var	29.58754		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	9.01839	19.02591
midterm		1	0.50057	0.18178
quizave	Quiz Average (drop lowest)	1	4.80199	2.46469
compave	Computer Average (drop lowest)	1	-3.53028	2.17562

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|	Type I SS
Intercept	Intercept	1	0.47	0.6374	146806
midterm		1	2.75	0.0080	3901.64751
quizave	Quiz Average (drop lowest)	1	1.95	0.0566	509.97483
compave	Computer Average (drop lowest)	1	-1.62	0.1105	583.42537

The REG Procedure
Model: MODEL1

```
   Test smalstuf Results for Dependent Variable final

                                  Mean
Source             DF         Square    F Value    Pr > F

Numerator           2      546.70010       2.47    0.0943
Denominator        54      221.58085
```

Data in fixed columns When the data values have at least one space between them, the variables are recorded in the same order for each case, and missing values are indicated by periods, the default version of the `input` statement (list input) does the job perfectly. It is a bonus that the variables need not always be separated by the same number of spaces for each case. Also, there can be more than one line of data for each case, and in fact there need not even be the same number of data lines for all the cases, just as long as there are the same number of variables.

Another common situation is for the data to be lined up in fixed columns, with blanks for missing values. Sometimes, especially when there are many variables, the data are *packed* together, without spaces between values. For example, the Minnesota Multiphasic Personality Inventory (MMPI) consists of over 300 questions, all to be answered True or False. It would be quite natural to code 1=True and 0=False, and pack the data together. There would still be quite a few data lines for each case.

Here is the beginning of the file `statclass2.dat`. It is the same as `statclass1.dat`, except that the data are packed together. Most of the blanks occur because two columns are reserved for the marks on quizzes and computer assignments, because 10 out of 10 is possible. Three columns are reserved for the midterm and final scores, because 100% is possible. For all variables, missing values are represented by blanks. That is, if the field occupied by a variable is completely blank, it's a missing value.

```
appsrv01.srv>  less statclass2.dat
12 9 1 7 8 4 3 5 2 6101010 5 0 0 0 0  55 43
021010 5 910 8 6 81010 8 9 9 9 91010  66 79
121010 5101010 9 8101010101010 91010  94 67
121010 8 910 710 9101010 91010 91010  81 65
0110 1 0 0 8 6 5 210 9 0 010 6 0 5 0  54
1110 6 7 9 8 8 5 710 910 9 5 6 4 810  57 52
01 0 0 9 910 5 2 2 8 7 71010 6 3 710  49
```

```
0110 9 5 8 9 8 5 6 8 7 5 610 6 5 9 9  77 64
0110 8 6 8 9 5 3 6 9 9 6 910 6 5 710  65 42
1110 5 6 710 4 6 010 910 910 6 7 810  73
01 9 0 4 610 5 3 310 810 51010 9 910  71 37

⋮
```

Now we will take a look at `statread.sas`. It contains just the `proc format` and the `data` step; There are no statistical procedures. This file will be read by programs that invoke statistical procedures, as you will see.

```
                              /* statread.sas
Read the statclass data in fixed format, define and label variables. Use
with     %include 'statread.sas';   */

options linesize=79 noovp formdlim='_';
title 'Grades from STA3000 at Roosevelt University:  Fall, 1957';

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 'statclass2.dat' missover;
     input (sex ethnic) (1.)
           (quiz1-quiz8 comp1-comp9) (2.)
           (midterm final) (3.);
     /* 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';
   mark2 = round(mark);
   /* Bump up at grade boundaries */
   if mark2=89 then mark2=90;
   if mark2=79 then mark2=80;
   if mark2=69 then mark2=70;
   if mark2=59 then mark2=60;
   /* Assign letter grade */
   if mark2=. then grade='Incomplete';
      else if mark2 ge 90 then grade = 'A';
      else if 80 le mark2 le 89 then grade='B';
      else if 70 le mark2 le 79 then grade='C';
      else if 60 le mark2 le 69 then grade='D';
      else grade='F';
   format sex sexfmt.;       /* Associates sex & ethnic    */
   format ethnic ethfmt.;    /* with formats defined above */

/*************************************************************/
```

The data step in `statread.sas` differs from the one in `statmarks1.sas` in only two respects. First, the `missover` option on the infile statement causes blanks to be read as missing values even if they occur at the end of a line and the line just ends rather than being filled in with space characters. That is, such lines are shorter than the others in the file, and when SAS `over`-reads the end of the line, it sets all the variables it would have read to missing. This is what we want, so you should always use the `missover` option when missing values are represented by blanks.

The other difference between this data step and the one in `statmarks1.sas` is in the `input` statement. Here, we are using *formatted* input. `sex` and `ethnic` each occupy 1 column. `quiz1-quiz8` and `comp1-comp9` each occupy 2 columns. `midterm` and `final` each occupy 3 columns. You can supply a list of formats for each list of variables in parentheses, but if the number of formats is less than the number of variables, they are re-used. That's what's happening in the present case.

The program `statread.sas` reads and defines the data, but it requests no statistical output; `statdescribe.sas` pulls in `statread.sas` using a `%include` statement, and produces basic descriptive statistics. Significance tests would be produced by other short programs.

Keeping the data definition in a separate file and using `%include` (the only part of the powerful *SAS macro language* presented here) 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 in many cases the data step is long. And, 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.

```
                                /* statdescribe.sas */
%include 'statread.sas';
title2 'Basic Descriptive Statistics';

proc freq;
      title3 'Frequency distributions of the categorical variables';
      tables sex ethnic grade;

proc means n mean std;
     title3 'Means and SDs of quantitative variables';
     var quiz1 -- mark2;        /*  single dash only works with numbered
                                   lists, like quiz1-quiz8    */

proc univariate normal; /* the normal option gives a test for normality */
     title3 'Detailed look at mark and bumped mark (mark2)';
     var mark mark2;
```

2.2.7 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 [2] 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. Here is the main case. 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 [2].

```
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.

Chapter 3

More Than One Independent Variable at a time

The standard elementary tests typically involve one independent variable and one dependent variable. Now we will see why this can make them very misleading. The lesson you should take away from this discussion is that when important variables are ignored in a statistical analysis — particularly in an observational study — the result can be that we draw incorrect conclusions from the data. Potential confounding variables really need to be included in the analysis.

3.1 The chi-squared test of independence

In order to make sure the central example in this chapter is clear, it may be helpful to give a bit more background on the common Pearson chi-square test of independence. As stated earlier, the chi-square test of independence is for judging whether two categorical variables are related or not. It is based upon a *cross-tabulation*, or *joint frequency distribution* of the two variables. For example, suppose that in the statclass data, we are interested in the relationship between sex and apparent ethnic background. If the ratio of females to males depended upon ethnic background, this could reflect an interesting cultural difference in sex roles with respect to men and women going to university (or at least, taking Statistics classes). In `statmarks1.sas`, we did this test and obtained a chisquare statistic of 2.92 (df=2, $p = 0.2321$), which is not statistically significant. Now we'll do it just a bit differently to

illustrate the details. First, here is the program `ethsex.sas`.

```
/* ethsex.sas */
%include 'statread.sas';
title2 'Sex by Ethnic';
proc freq;
     tables sex*ethnic / chisq norow nocol nopercent expected;
```

And here is the output.

The FREQ Procedure

Table of sex by ethnic

sex ethnic(Apparent ethnic background (ancestry))

Frequency| Expected	Chinese	European	Other	Total
Male	27	7	5	39
	25.79	9.4355	3.7742	
Female	14	8	1	23
	15.21	5.5645	2.2258	
Total	41	15	6	62

Statistics for Table of sex by ethnic

Statistic	DF	Value	Prob
Chi-Square	2	2.9208	0.2321
Likelihood Ratio Chi-Square	2	2.9956	0.2236
Mantel-Haenszel Chi-Square	1	0.0000	0.9949
Phi Coefficient		0.2170	
Contingency Coefficient		0.2121	
Cramer's V		0.2170	

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

Sample Size = 62

In each cell of the table, we have an observed frequency and an expected frequency. The expected frequency is the frequency one would expect by chance if the two variables were completely unrelated.[1] If the observed frequencies are different enough from the expected frequencies, one would tend to disbelieve the null hypothesis that the two variables are unrelated. But how should one measure the difference, and what is the meaning of different "enough?"

The Pearson chi-square statistic (named after Karl Pearson, a famous racist, uh, I mean statistician) is defined by

$$\chi^2 = \sum_{\text{cells}} \frac{(f_o - f_e)^2}{f_e}, \tag{3.1}$$

where f_o refers to the observed frequence, f_e refers to expected frequency, and as indicated, the sum is over all the cells in the table.

If the two variables are really independent, then as the total sample size increases, the probability distribution of this statistic approaches a chisquare with degrees of freedom equal to (Number of rows - 1)$\times$(Number of columns - 1). Again, this is an approximate, large-sample result, one that obtains exactly only in the limit as the sample size approaches infinity. A traditional "rule of thumb" is that the approximation is okay if no expected frequency is less than five. This is why SAS gave us a warning.

More recent research suggests that to avoid inflated Type I error (false significance at a rate greater than 0.05), all you need is for no expected frequency to be less than one. You can see from formula (3.1) why an expected frequency less than one would be a problem. Division by a number close to zero can yield a very large quantity even when the observer and expected frequencies are fairly close, and the so-called chisquare value will be seriously inflated.

Anyway, The p-value for the chisquare test is the upper tail area, the area under the chi-square curve beyond the observed value of the test statistic. In the example from the statclass data, the test was not significant and we conclude nothing.

[1]The formula for the expected frequency in a given cell is (row total) $\times$ (column total)/(sample size). This follows from the definition of independent events given in introductory probability: the events A and B are independent if $P(A \cap B) = P(A)P(B)$. But this is too much detail, and we're not going there.

3.2 The Berkeley Graduate Admissions data

Now we're going to look at another example, one that should surprise you. In the 1970's the University of California at Berkeley was accused of discriminating against women in graduate admissions. Data from a large number of applicants are available. The three variables we will consider are sex of the person applying for graduate study, department to which the person applied, and whether or not the person was admitted. First, we will look at the table of sex by admission.

Table of sex by admit

sex	admit		
Frequency|			
Row Pct	No	Yes	Total
Male	1493	1198	2691
	55.48	44.52	
Female	1278	557	1835
	69.65	30.35	
Total	2771	1755	4526

The FREQ Procedure

Statistics for Table of sex by admit

Statistic	DF	Value	Prob
Chi-Square	1	92.2053	<.0001

It certainly looks suspicious. Roughly forty-five percent of the male applicants were admitted, compared to thirty percent of the female applicants. This difference in percentages (equivalent to the relationship between variables here) is highly significant; with $n = 4526$, the p-value is very close to zero.

3.3 Controlling for a variable by subdivision

However, things look different when we take into account the department to which the person applied. Think of a *three-dimensional* table in which the rows are sex, the columns are admission, and the third dimension (call it layers) is department. Such tables are easy to generate with SAS and other statistical packages.

The three-dimensional table is displayed by printing each layer on a separate page, along with test statistics (if requested) for each sub-table. This is equivalent to dividing the cases into sub-samples, and doing the chisquare test separately for each sub-sample. A useful way to talk about this is to say that that we are *controlling* for the third variable; that is, we are looking at the relationship between the other two variables with the third variable held constant. We will have more to say about controlling for collections of independent variables when we get to regression.

Here are the six sub-tables of sex by admit, one for each department, with a brief comment after each table. The SAS output is edited a bit to save paper.

```
      Table 1 of sex by admit
      Controlling for dept=A

sex       admit

Frequency|
Row Pct  |No      |Yes     |  Total
---------+--------+--------+
Male     |    313 |    512 |    825
         |  37.94 |  62.06 |
---------+--------+--------+
Female   |     19 |     89 |    108
         |  17.59 |  82.41 |
---------+--------+--------+
Total         332      601      933


Statistics for Table 1 of sex by admit
       Controlling for dept=A
```

Statistic	DF	Value	Prob
Chi-Square	1	17.2480	<.0001

For department A, 62% of the male applicants were admitted, while 82% of the female applicants were admitted. That is, women were *more* likely to get in than men. This is a *reversal* of the relationship that is observed when the data for all departments are pooled!

Table 2 of sex by admit
Controlling for dept=B

sex admit

Frequency| Row Pct	No	Yes	Total
Male	207	353	560
	36.96	63.04	
Female	8	17	25
	32.00	68.00	
Total	215	370	585

Statistics for Table 2 of sex by admit
Controlling for dept=B

Statistic	DF	Value	Prob
Chi-Square	1	0.2537	0.6145

For department B, women were somewhat more likely to be admitted (another reversal), but it's not statistically significant.

Table 3 of sex by admit
Controlling for dept=C

sex	admit		
Frequency|			
Row Pct	No	Yes	Total
Male	205	120	325
	63.08	36.92	
Female	391	202	593
	65.94	34.06	
Total	596	322	918

Statistics for Table 3 of sex by admit
Controlling for dept=C

Statistic	DF	Value	Prob
Chi-Square	1	0.7535	0.3854

For department C, men were slightly more likely to be admitted, but the 3% difference is much smaller than for the pooled data. Again, it's not statistically significant.

Table 4 of sex by admit
Controlling for dept=D

sex	admit		
Frequency|			
Row Pct	No	Yes	Total
Male	279	138	417
	66.91	33.09	
Female	244	131	375
	65.07	34.93	
Total	523	269	792

Statistics for Table 4 of sex by admit
Controlling for dept=D

Statistic	DF	Value	Prob
Chi-Square	1	0.2980	0.5852

For department D, women were a bit more likely to be admitted (a reversal), but it's far from statistically significant. Now department E:

Table 5 of sex by admit
Controlling for dept=E

sex	admit		
Frequency|			
Row Pct	No	Yes	Total
Male	138	53	191
	72.25	27.75	
Female	299	94	393
	76.08	23.92	
Total	437	147	584

Statistics for Table 5 of sex by admit
Controlling for dept=E

Statistic	DF	Value	Prob
Chi-Square	1	1.0011	0.3171

This time it's a non-significant tendency for men to get in more. Finally, department F:

```
        Table 6 of sex by admit
        Controlling for dept=F

sex       admit

Frequency|
Row Pct  |No      |Yes     |  Total
---------+--------+--------+
Male     |    351 |     22 |    373
         |  94.10 |   5.90 |
---------+--------+--------+
Female   |    317 |     24 |    341
         |  92.96 |   7.04 |
---------+--------+--------+
Total         668       46      714

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      0.3841    0.5354
```

For department F, women were slightly more likely to get in, but once again it's not significant.

So in summary, the pooled data show that men were more likely to be admitted to graduate study. But when take into account the department to which the student is applying, there is a significant relationship between sex and admission for only one department, and in that department, women are more likely to be accepted.

How could this happen? I generated two-way tables of sex by department and department by admit; both relationships were highly significant. Instead of displaying the SAS output, I have assembled some numbers from these two tables. The same thing could be accomplished with SAS `proc tabulate`, but it's too much trouble, so I did it by hand.

Now it is clear. The two departments with the lowest percentages of female applicants (A and B) also had the highest overall percentage of applicants accepted, while the department with the highest percentage of female applicants (E) also had the second-lowest overall percentage of applicants accepted. That is, the departments most popular with men were easiest to get into, and those most popular with women were more difficult. Clearly,

Table 3.1: Percentage of female applicants and overall percentage of applicants accepted for six departments

Department	Percent applicants female	Percentage applicants accepted
A	11.58%	64.42%
B	4.27	63.25
C	64.60	35.08
D	47.35	33.96
E	67.29	25.17
F	47.76	6.44

this produced the overall tendency for men to be admitted more than women.

By the way, does this mean that the University of California at Berkeley was *not* discriminating against women? By no means. Why does a department admit very few applicants relative to the number who apply? Because they do not have enough professors and other resources to offer more classes. This implies that the departments popular with men were getting more resources, relative to the level of interest measured by number of applicants. Why? Maybe because men were running the show. The "show," by the way definitely includes the U. S. military, which funds a lot of engineering and similar stuff at big American universities.

The Berkeley data, a classic example of *Simpson's paradox*, illustrate the following uncomfortable fact about observational studies. When you include a new variable in an analysis, the results you have could get weaker, they could get stronger, or they could reverse direction — all depending upon the inter-relations of the independent variables. Basically, if an observational study does not include every potential confounding variable you can think of, there is going to be trouble.

Now, the distinguishing feature of the "elementary" tests is that they all involve one independent variable and one dependent variable. Consequently, they can be *extremely* misleading when applied to the data from observational studies, and are best used as tools for preliminary exploration.

Pooling the chi-square tests When using sub-tables to control for a categorical independent variable, it is helpful to have a single test that allows you to answer a question like this: If you control for variable A, is B related

to C? For the chi-square test of independence, it's quite easy. Under the null hypothesis that B is unrelated to C for each value of A, the test statistics for the sub-tables are independent chisquare random variables. Therefore, there sum is also chisquare, with degrees of freedom equal to the sum of degrees of freedom for the sub-tables.

In the Berkeley example, we have a pooled chisquare value of

$$17.2480 + 0.2537 + 0.7535 + 0.2980 + 1.0011 + 0.3841 = 19.9384$$

with 6 degrees of freedom. Using any statistics text (except this one), we can look up the critical value at the 0.05 significance level. It's 12.59; since $19.9 > 12.59$, the pooled test is significant at the 0.05 level. To get a p-value for our pooled chisquare test, we can use SAS. See the program in the next section.

In summary, we need to use statistical methods that incorporate more than one independent variable at the same time; multiple regression is the central example. But even with advanced statistical tools, the most important thing in any study is to collect the right data in the first place. Looking at it the right way is critical too, but no statistical analysis can compensate for having the wrong data.

For more detail on the Berkeley data, see the article in *Science* by Bickel Hammel and O'Connell [1]. For the principle of adding chisquare values and adding degrees of freedom from sub-tables, a good reference is Feinberg's *The analysis of cross-classified categorical data* [4].

3.4 The SAS program

Here is the program `berkeley.sas`. It has several features that you have not seen yet, so a discussion follows the listing of the program.

```
/************************** berkeley.sas ********************************/
options linesize=79 pagesize=35 noovp formdlim='_';
title 'Berkeley Graduate Admissions Data: ';

proc format;
     value sexfmt 1 = 'Female' 0 = 'Male';
     value ynfmt 1 = 'Yes'  0 = 'No';
data berkley;
     input  line sex dept $ admit count;       %$
     format sex sexfmt.; format admit ynfmt.;
     datalines;
   1     0      A      1    512
   2     0      B      1    353
   3     0      C      1    120
   4     0      D      1    138
   5     0      E      1     53
   6     0      F      1     22
   7     1      A      1     89
   8     1      B      1     17
   9     1      C      1    202
  10     1      D      1    131
  11     1      E      1     94
  12     1      F      1     24
  13     0      A      0    313
  14     0      B      0    207
  15     0      C      0    205
  16     0      D      0    279
  17     0      E      0    138
  18     0      F      0    351
  19     1      A      0     19
  20     1      B      0      8
  21     1      C      0    391
  22     1      D      0    244
  23     1      E      0    299
  24     1      F      0    317
;
```

```
proc freq;
     tables sex*admit / nopercent nocol chisq;
     tables dept*sex / nopercent nocol chisq;
     tables dept*admit / nopercent nocol chisq;
     tables dept*sex*admit / nopercent nocol chisq;
     weight count;

/* Get p-value */
proc iml;
     x = 19.9384;
     pval = 1-probchi(x,6);
     print "Chisquare = " x "df=6, p = " pval;
```

The first unusual feature of `berkeley.sas` is in spite of my recommendations to the contrary, the data are in the program itself rather than in a separate file. The data are in the data step, following the `datalines` command and ending with a semicolon. You can always do this, but usually it's a bad idea. Here, it's a good idea. This is why.

I did not have access to a raw data file, only a 2 by 6 by 2 table of sex by department by admission. So I created a data set with just 24 lines, even though there are 4526 cases. Each line of the data set has values for the three variables, and also a variable called `count`, which is simply the observed cell frequency for that combination of sex, department and admission. Then, using the `weight` statement in `proc freq`, I "weighted" each of the 24 cases in the data file by `count`, essentially multiplying the sample size by count for each case.

The advantages are several. First, such a data set is easy to create from published tables, and is much less trouble than a raw data file with thousands of cases. Second, the data file is so short that it makes sense to put it in the data set for portability and ease of reference. Finally, this is the way you can get the data from published tables (which may not include any significance tests at all) into SAS, where you can compute any statistics you want, including sophisticated log-linear modelling analyses.

The last `tables` statement in the `proc freq` gives us the three-dimensional table. For a two-dimensional table, the first variable you mention will correspond to rows and the second will correspond to columns. For higher-dimensional tables, the second-to-last variable mentioned is rows, the last is

columns, and combinations of the variables listed first are the control variables for which sub-tables are produced.

Finally, the `iml` in `proc iml` stands for "Interactive Matrix Language," and you can use it to perform useful calculations in a syntax that is very similar to standard matrix algebra notation; this can be very convenient when formulas you want to compute are in that notation. Here, we're just using it to calculate the area under the curve of the chisquare density with 6 degrees of freedom, beyond the observed test statistic of 19.9384. The `probchi` function is the cumulative distribution function of the chisquare distribution; the second argument (6 in this case) is the degrees of freedom. `probchi(`x`,6)` gives the area under the curve between zero and x, and `1-probchi(`x`,6)` gives the tail area above x – that is, the p-value.

Summary The example of the Berkeley graduate admissions data teaches us that if potential confounding variables are not cancelled out by random assignment to experimental conditions, they need to be explicitly included in a statistical analysis. Otherwise, the results can be very misleading. In the Berkeley example, first we ignored department and there was a relationship between sex and admission that was statistically significant in one direction. Then, when we *controlled* for department — that is, when we took it into account — the relationship was either significant in the opposite direction, or it was not significant (depending on which department).

We also saw how to pool chi-square values and degrees of freedom by adding over sub-tables, obtaining a useful test of whether two categorical variables are related, while controlling for one or more other categorical variables. This is something SAS will not do for you, but it's easy to do with `proc freq` output and a calculator.

Bibliography

[1] Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, **187**, 398-403.

[2] Cody, R. P. and Smith, J. K. (1991). *Applied statistics and the SAS programming language.* (4th Edition) Upper Saddle River, New Jersey: Prentice-Hall.

[3] Cook, T. D. and Campbell, D. T. (1979). *Quasi-experimentation : design and analysis issues for field settings*. New York: Rand McNally.

[4] Feinberg, S. (1977) *The analysis of cross-classified categorical data.* Cambridge, Massachusetts: MIT Press.

[5] Fisher, R. A. (1925) *Statistical methods for research workers.* London: Oliver and Boyd.

[6] Moore, D. S. and McCabe, G. P. (1993). *Introduction to the practice of statistics.* New York: W. H. Freeman.

[7] Neter, J., Kutner, M. H., Nachhtscheim, C. J. and Wasserman, W. (1996) *Applied linear statistical models.* (4th Edition) Toronto: Irwin.

[8] Roethlisberger, F. J. (1941). *Management and morale*. Cambridge, Mass.: Harvard University Press.

[9] Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Croft.

[10] Rosenthal, R. and Jacobson, L. (1968). *Pygmalion in the classroom: teacher expectation and pupils' intellectual development.* New York: Holt, Rinehart and Winston.