Draft Lecture Notes for Methods of Applied Statistics (STA442H/1008H)

Jerry Brunner

January 8, 2002

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. 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 sizeable 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 `S`. 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 `S` 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 randon number generation for statisitcal 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 statisticl texts use the term **qualitative** to mean categorical. When an anthropologist uses the word “qualitative,” however, it usually means “non-quantitative.”

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 A is thought to cause or influence B. Sometimes it can go either way or it's not clear. 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 about 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 almost the entire 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 variable are **unrelated** if the *conditional distribution of the dependent variable in the population 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**.

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 variale 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 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.

Here is another way to talk about p-values and significance testing. *The p-value is the probability of getting our results (or better) just by chance.* If p is small enough (we will use) 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 is a relationship between the independent and dependent, and we will say our results are "statistically significant."

If $p > .05$, we will not conclude anything. All we can say is that there is no evidence of a relationship between the independent variable and the dependent variable.

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

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, meaning that* ***the same or related (non-independent) units contribute a value of the DV for each value of the IV.***

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."*

22	18
40	34
20	10
⋮	⋮

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 quesionnaire.*

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*

1	2	4.75	5.31
2	4	2.91	6.82
⋮	⋮	⋮	⋮
80	2	5.12	4.06

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 16 pieces of data for each bank.*

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 presentaion 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 chisquared 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 co-ordinates 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.

- $-1 \leq r \leq 1$
- $r = +1$ 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 beteen them is linear. As usual, the assumption of independent observations is always important.

Here are some examples of scatterplots and the associated correlation coefficients.

```
MTB > plot c1 c3

         -                               *    *
 C1      -
         -
         -                             *
       60+                          ** *
         -             *    *  *  *   2*           *    *
         -                *            ** *   *   *
         -                                               *
         -                  *        2  2* **  *    *            *
       45+    *                   *  *2             *
         -                      *    *        *
         -                       *
         -                           * * *
         -                           *      *
       30+
         -              *                                     *
         -
           +---------+---------+---------+---------+---------+------C3
          20        30        40        50        60        70

MTB > corr c1 c3
Correlation of C1 and C3 = 0.004
```

```
MTB > plot c4 c6

       75+                             *
         -
 C4      -
         -
         -               *                          *       *
       60+                                     *
         -    *             *     *    * 2 *       *              *
         -                    *       **     **
         -               *   *           *2
         -                   *     **    *    *            * *
       45+             *   *     **   *     *             *
         -                              2    *
         -                  *       2       ***
         -                                *
         -
       30+                        * *
         -
           ------+---------+---------+---------+---------+---------+C6
               112       128       144       160       176       192

MTB > corr c4 c6
Correlation of C4 and C6 = 0.112
```

```
MTB > plot c3 c7

       80+
         -                                   *
 C3      -              *
         -
         -                           *                             *
       60+                            * *        * *
         -                       * *   *              *
         -         *     *              **        *
         -    *  *   *  * *2**   * **      2      *    *         *
         -        *     2         2 *                     *
       40+    *  *                  *
         -                  * **
         -  *                           *
         -
         -                  *
       20+
         -
           --+---------+---------+---------+---------+---------+----C7
           165       180       195       210       225       240

MTB > correlation between c3 and c7 please
Correlation of C3 and C7 = 0.368
```

```
MTB > plot c4 c7

       75+                                         *
         -
 C4      -
         -
         -                  *                             *       *
       60+                                       *
         -  *              *      *     *** *      *               *
         -                  *       * *     **
         -          *   *            2 *
         -            *     * *   *     *             * *
       45+    *   *      **  *      *              *
         -                    **    *
         -     *        2        ***
         -                  *
         -
       30+       **
         -
           --+---------+---------+---------+---------+---------+----C7
           165       180       195       210       225       240

MTB > corr c4 c7
Correlation of C4 and C7 = 0.547
```

```
MTB > plot c5 c7

         -
 C5      -                                              *         *
         -
         -                              *             *   *        *
      120+                                         *
         -                    *     *
         -                  *  *                   *
         -                        * *       2
         -                  *     **
      100+       *      *    *   *  **   *       * *
         -                    *        *  *
         -        *      *  *     *
         -              * * *         * *
         -        *   *              *       *
       80+    **           *
         -  *
         -          *   *
           --+---------+---------+---------+---------+---------+----C7
           165       180       195       210       225       240


MTB > corr c5 c7
Correlation of C5 and C7 = 0.733


MTB > plot c5 c9

         -
 C5      -         **
         -
         -    *      *      *     *
      120+            *
         -                         *    *
         -                   *          *         *
         -                         2**
         -                        *   **
      100+                       *  *  2 *2 * *
         -                            **    *
         -                                  **2
         -                               2     *   *   *
         -                        *   *            * *
       80+                                          2    *
         -                                                        *
         -                                  *          *
           --+---------+---------+---------+---------+---------+----C9
          -192      -176      -160      -144      -128      -112

MTB > corr c5 c9
Correlation of C5 and C9 = -0.822
```

```
MTB > plot c2 c1

         -
         -                                                      *
      100+    *                                               *2
         -       **                                           *
 C2      -     2*  *                                        *
         -       **  *                                    * **
         -            2*                               2  **
       50+          *                                ** *2
         -          *    **                          *
         -             *** 2* **     *            ***
         -                  *   *  *      * *    *
         -                   **  *    2  * * *  * * *
        0+                       *** *  2  * * **
         -                      *             **
         -                          *     *
         -                             *
         -
           --------+---------+---------+---------+---------+--------C1
                -8.0      -4.0       0.0       4.0       8.0

MTB > corr c1 c2
Correlation of C1 and C2 = 0.025

      200+
         -
 C2      -             *      **     *
         -        ** **   ***      *
         -       *    *  * **   *     *   *
      100+        *     **   **  2** **  *  *
         -    ****  *  *                2  2
         -          *           *            *
         -     *                    *        *  ***  *
         -                             *      *** *****2
        0+                                *    *          *
         -                                              *
         -                                               *  * *
         -                                          *    **
         -                                                 ******
     -100+                                                     *
         -
           --------+---------+---------+---------+---------+--------C1
                -8.0      -4.0       0.0       4.0       8.0

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.

Choosing an Elementary Technique

Make a table in lecture.

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 ex-*

perimental, 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 qualtitative dependent variables are systolic and dyastolic 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.

Placebo Effects

Experimenter Expectancy

Internal and external validity

Quasi-experimental designs