Chapter 1

Introduction

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). There are n cases, where n is the sample size.

- 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

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

Response Variable: 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 response variable and B is the explanatory 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. 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 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 relative frequency 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[1].

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

We will say that the explanatory and response variables are **unrelated** if the *conditional distribution of the response variable is identical for each value of the explanatory variable*[2]. That is, the relative frequency histogram of the response variable does not depend on the value of the explanatory variable. If the distribution of the response variable does depend on the value of the explanatory variable, we will describe the two variables as **related**. All this vocabulary applies to sample as well as population data-sets[3].

[1] Since the area under such a curve equals one (remember, it's a *relative* frequency histogram), the smooth curve is a probability density function.

[2] As a technical note, suppose that X and Y are both continuous. Then the definition of "unrelated" says $f(y|x) = f(y)$, which is equivalent to $f(x, y) = f(x)f(y)$. This is the definition of independence. So the proposed definition of "unrelated" is a way of smuggling the idea of statistical independence into this non-technical discussion. I *said* I was going to put the mathematical digressions in footnotes.

[3] A population dataset may be entirely hypothetical. For example, if a collection of cancer-prone laboratory mice are given an anti-cancer vaccine, one might pretend that those mice are a random sample from a population of all cancer-prone mice receiving the vaccine – but of course there is no such population.

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

Sample Question 1.2.1 *Give an example of a study with two categorical explanatory variables, one quantitative explanatory 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 explanatory variables are Sex and Immigration Status (Citizen, Permanent Resident or Visa), and the quantitative explanatory 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 explanatory and response 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 explanatory variable and response variable 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 explanatory variable and the response variable in the population. In this case we will conclude there really *is* a relationship.

Of course we seldom or never know for sure what is happening in the entire population. So when we reject a null hypothesis, we may be right or wrong. Sometimes, the null hypothesis is true (nothing is going on) and we mistakenly reject it; this is called a **Type One Error**. It is also possible that the null hypothesis is false (there really is a relationship between explanatory and response variable in the population) but we fail to

reject it. This is called a **Type Two Error**. This numbering expresses the philosophy that false knowledge is a really bad thing – it's the Number One kind of mistake you can make.

The probability of correctly rejecting the null hypothesis – that is, the probability of discovering something that really is present, is one minus the probability of a Type Two error. This is called the **Power** of a statistical test. Clearly, more power is a good thing. But there is a tradeoff between power and Type One error, so that it is impossible for any statistical test to simultaneously minimize the chances of Type One error and maximize the power. The accepted solution is to insist that the Type One error probability be no more than some small value (the significance level – 0.05 for us), and use the test that has the greatest power subject to this constraint. An important part of theoretical statistics is concerned with proving that certain significance tests that have the best power, and the tests that are used in practice tend to be the winners of this contest.

If you think about it for a moment, you will realize that most of the time, even a test with good overall power will not have exactly the same power in every situation. The two main principles are:

- The stronger the relationship between variables in the population, the greater the power.

- The larger the sample size, the greater the power.

These two principles may be combined to yield a method for choosing a sample size based on power, before any data have been collected. You choose a strength of relationship that you want to detect, ideally one that is just barely strong enough to be scientifically meaningful. Then you choose a (fairly high) probability with which you want to be able to detect it. Next, you pick a sample size and calculate the power – not difficult, in this age of computers. It will almost certainly be too low, though it may be higher than you need if you have started with a huge sample size. So you increase (or decrease) the sample size, and calculate the power again. Continue until you have located the smallest sample size that gives you the power you want for the strength of relationship you have chosen. This is not the only rational way to choose sample size, but it is one of the two standard ones.[4] Examples will be given later.

Closely related to significance tests are **confidence intervals**. A confidence interval corresponds to a pair of numbers calculated from the sample data, a lower confidence limit and an upper confidence limit. The confidence limits are chosen so that the probability of the interval containing some parameter (or *function* of the parameters, like a difference between population means) equals a large value, say 0.95. Such a confidence interval would be called a "ninety-five percent confidence interval." The connection between tests and confidence intervals is that a two tailed t-test or Z-test will be significant at the 0.05 level if and only if the 95% confidence interval does not contain zero.

[4]The other standard way is to choose the sample size so that a chosen confidence interval will have at most some specified width.

1.2.2 Should You *Accept* the Null Hypothesis?

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 explanatory and response variables are unrelated. Instead, we can say only that there is insufficient evidence to conclude that there is a relationship. A good reference is Fisher's masterpiece, *Statistical methods for research workers* [9], which had its first edition in 1925, and its 14th and last edition in 1970, eight years after Fisher's death.

In some courses, Fisher's advice is given as an absolute rule. Students are told that one *never* accepts the null hypothesis. But in other courses, if the null hypothesis is not rejected, then it is accepted without further question. Who is right? This is the echo of a very old quarrel between Fisher, who is responsible for the concept of hypothesis testing more or less as we know it, and the team of Jerzy Neyman and Egon Pearson, who came along a bit later and cleaned up Fisher's method, putting it on a firm decision-theoretic basis. The *decision* in question is between the null hypothesis and the alternative hypothesis, period. According to Neyman and Pearson, you have to pick one of them, based on the data. Refusal to decide is not an option.

During their lifetimes, Fisher fought bitterly with Neyman and Pearson. To Neyman and Pearson, Fisher was creative but mathematically unsophisticated. To Fisher, Neyman and Pearson were good mathematicians, but they were missing the point, because science does not proceed by simple yes or no decisions made in isolation from one another. Today, Neyman-Pearson theory usually dominates in theoretical research and theoretical courses, while Fisher's approach dominates in applications and applied courses. One might think that because this is an applied course, we'll just side with Fisher. But it's a bit trickier than that.

In the typical data analysis project, the first step is to assemble the data file and check it for errors. Then, the usual practice is to carry out a variety of statistical tests to get a preliminary idea of how the variables are related to each other. This phase can be automated (as in stepwise regression) or not, but in general you try a lot of tests, and if a potential explanatory variable is not significantly related to the response variable in question, you usually just drop it and look elsewhere. That is, the null hypothesis is freely accepted, and the Neyman-Pearson approach seems to govern this most applied of statistical pursuits.

You can't fault this; scientists must explore their data, and statistical testing is a good way to do it. But it is helpful to distinguish between *exploratory* and *confirmatory* statistical analysis. In an exploratory analysis, the researcher carries out a large number of tests in an attempt to understand how the variables are related to one another. Various statistical models are employed, variables may be defined and re-defined several times, and the sample may be subdivided in various ways. Anything reasonable may be (and should be) attempted. Numerous null hypotheses may be tentatively rejected, and numerous others may be tentatively accepted. Properly speaking, the product of an exploratory analysis is hypotheses, not conclusions. It is rare for all the details of an exploratory analysis to be given in writing, though it is good practice to keep a record of what has been tried.

In a confirmatory analysis, a more limited number of tests are carried out with the intention of coming to firm conclusions.[5] The results of confirmatory analyses *are* often written up, because communication of results is in many ways the most important phase of any investigation. It is clear that acceptance of the null hypothesis is a standard feature of good exploratory analysis, even if it is not recognized as such. The argument between Fisher and Neyman-Pearson is whether the null hypothesis should be accepted in confirmatory analysis.

First of all, it's clear that Fisher is right in a way. Suppose you wish to compare two methods of teaching the piano. You randomly assign three students to one method and two students to the other. After some reasonable period of time, you compare ratings of their performance, using a two-sample t test or something. Suppose the results are not statistically significant. Does it make sense to conclude that the two methods are equally effective? Obviously not; the sample size is so small that we probably don't have enough power to detect even a fairly large effect.

But Neyman and Pearson do not give up, even in this situation. They say that if one had to choose based just on this tiny data set, the conclusion of no effect would be the rational choice. Meanwhile, Fisher is going crazy. Who would decide anything based on such inadequate evidence? He does not know whether to laugh at them or tear his hair out, so he does both, in public. On their side, Neyman and Pearson are irritated by Fisher's unwillingness (or inability) to appreciate that when statistical tests emerge as mathematical consequences of a general theory, this is better than just making them up out of thin air.

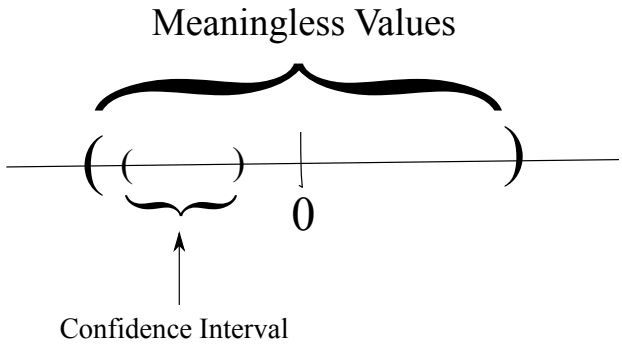
Fisher wins this round, but it's not over. The trouble with his approach is that it *never* allows one to conclude that the null hypothesis is true. But sometimes, experimental treatments just don't do anything, and it is of scientific and practical importance to be able to say so. For example, medical researchers frequently conclude that drugs don't work. On what basis are they drawing these conclusions? On what basis *should* they draw such conclusions?

Unfortunately, though there are clear conventional standards for deciding when a relationship is present, there is much less agreement on how to decide that one is absent. In medical research, scientists often get away with such claims based only on the fact that a test fails to attain statistical significance. Then, if the sample size is not unusually small, nobody objects. It seems to depend on the editor of the journal.

There are a couple of reasonable suggestions about how to be more systematic (need references here). Both methods stop short of allowing you to conclude that a relationship is completely absent. Instead, they focus on deciding that the relationship between ex-

[5] Ideally, exploratory and confirmatory analyses should be carried out on different data sets, possibly by randomly splitting the data into exploratory and confirmatory sub-samples. But this is only feasible when data are not too expensive or time-consuming to collect. In practice, researchers often explore their data thoroughly, and then report the most interesting results as if they were a confirmatory analysis. This practice is almost guaranteed to inflate the probability of Type One error, so it is wise to treat the results of most scientific investigations as tentative until they have been independently replicated. In any case, it is useful to distinguish *conceptually* between exploratory and confirmatory analysis, even though the pure forms may seen only rarely in practice.

Figure 1.1: A relationship that is significant but too weak to be meaningful.



planatory variable and response variable is so weak that it does not matter, if it exists at all.

One approach is based on power. Suppose you have selected the sample size so that that there is a high probability (maybe 95%) of detecting a relationship that is just barely meaningful (of course, if the relationship in the population happens to be stronger, the probability of detecting it will be even higher). Then, if the test is non-significant, you conclude that the relationship is not strong enough to be meaningful.

Another approach is based on confidence intervals. Again, you need to be able to specify what's scientifically or perhaps clinically meaningful, in terms of the population parameters. You construct a confidence interval for the quantity in question (for example a difference between means). If the 95% confidence interval lies entirely within a range of values that is scientifically meaningless, you conclude that the relationship is not strong enough to be meaningful.

These two reasonable methods need not yield the same conclusion for a given data set; the confidence interval approach allows a relationship to be deemed negligible even though it is statistically significant, while the power approach does not. Figure 1.1 shows how this can happen. Notice that the 95% confidence interval is entirely within the range of values deemed too small to be meaningful. But the confidence interval does not contain zero, so $p < 0.05$. Any time the true parameter value is in the non-meaningful range but is not exactly zero, a configuration like this is guaranteed to occur if the sample size is large enough.

Unfortunately, both the power method and the confidence interval method typically require a very large sample to conclude that a relationship is (virtually) absent. So it often happens that an important test is non-significant, but the power for detecting a marginal effect was fairly low, and the confidence interval includes both zero *and* values that are not trivial. In this situation, the best we can do is follow Fisher's advice, and

say that the data do not provide sufficient evidence to conclude that the explanatory and response variables are related.

Frequently, one has to write for a non-technical audience, and an important part of this course is to express conclusions in plain, non-technical language — language that is understandable to someone with no statistical training, but at the same time acceptable to experts. Suppose you need to state conclusions, and the results are not statistically significant. Most of your primary audience has no statistical background, so you need to speak in clear, non-statistical language. But *some* of the audience (maybe including the technical staff of your main audience) will be very disturbed if you seem to be accepting the null hypothesis; they can make a lot of trouble. How do you finesse this?

Here are some statements that are acceptable. It's good not to use exactly the same phrase over and over.

- The data do not provide evidence that the treatment has any effect.
- There was no meaningful connection between ...
- The results were consistent with no treatment effect.
- The results were consistent with no association between astrological sign and personality type.
- The small differences in average taste ratings could have been due to sampling error.
- The small differences in average taste ratings were within the range of sampling error.

The nice thing about using this kind of language is that it communicates clearly to non-experts, but it lets the experts read between the lines and see that you are aware of the technical (philosophic) issue, and that you are being careful. There are many, many more examples in Moore and McCabe's *Introduction to the practice of statistics* [15]. This introductory text is simple and non-technical on the surface, but written with all the theoretical complexities clearly in mind and under control. The result is a book that satisfies both the absolute beginner and the professional statistician — quite an accomplishment.

1.2.3 The Format of the Data File is Important!

If you're the person who will be doing the statistical analysis for a research study, there is an initial period where you are learning the objectives of the study and how the data are going to be collected. For example, perhaps participants are going to watch some commercials and then fill out a questionnaire. From the very beginning, you should be thinking about what the cases are, what the explanatory and response variables are, checking whether determining the relationships between explanatory and response variables will satisfy the objectives of the research, and deciding what statistical tests to

employ. All this applies whether you are helping plan the study, or (more likely, if you are a statistician) you are being brought in only after the data have already been collected.

Many scientific questions can be answered by determining whether explanatory variables and response variables are related. This makes it helpful to arrange data files in the row-by-column format suggested at the beginning of this chapter. Again, rows are usually cases, and columns are usually variables. But most data do not automatically come in this format unless a knowledgeable person has arranged it that way.

Data Analysis Hint 1 *If a data set is not already in a row-by-column format with rows corresponding to cases and columns corresponding to variables, you should put it in this format yourself, or get someone else to do it.*

Statistical software (including SAS) mostly expects data to be arranged this way, so Hint 1 is partly a matter of convenience. But there's more to it than that. You might be surprised how much a good data format can support good research design. For example, it is common for people who are very smart in other ways to record data over time at considerable effort and expense, but to change the data that are recoded or the way they are recorded throughout the course of the study. As a result, almost nothing is comparable, and most of the effort is wasted. An investigator who is thinking in terms of variables and cases is less likely to make this blunder.

The row-by-column format forces you to know how many cases there are, and which data come from the same case. Also, thinking in terms of variables helps you decide whether two different variables are intended as measures of the same thing at different times, or as quantities that are completely different.

On the other hand, you should keep your mind open. It is possible that for some studies and certain advanced statistical models, a different structure of the data file could be better. But I have never seen an example that applies to real data. In my experience, when data are recorded in a format other than the one advocated here, it is a sign of *lack* of sophistication on the part of the researchers.

So in the next section, please pay attention to the format of the data files. Bear in mind, though, that these are all *elementary* tests, with one explanatory variable and one response variable. Almost all real data sets have more than two variables.

1.2.4 Standard elementary significance tests

We will now consider some of the most common elementary statistical methods; these are covered in most introductory statistics courses. There is always just one explanatory variable and one response variable. For each test, 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 explanatory variable?
3. In your example, what is the response variable?
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 response 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 explanatory 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 response 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. The matched t-test takes this lack of independence into account 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 explanatory variable if a case (unit of analysis, subject, participant in the study) contributes a value of the response variable for each value of the explanatory variable in question. A variable on which there are repeated measures is sometimes called a **within-cases** (or within-subjects) variable. When this language is being spoken, variables on which there are not repeated measures are called **between-cases**. In a within-cases design, each case serves as its own control. When the correlations among data from the same case are substantial, a within-cases design can have higher power than a between-cases design.

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 explanatory 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 response 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 explanatory 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 response variable (or variables)?*

Answer to Sample Question 1.2.14 *There are 2 response 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 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 too many. The point is that order of presentation has to be unrelated to any manipulated explanatory variable.

Two (and higher) way Analysis of Variance

Extension of One-Way ANOVA to allow assessment of the joint relationship of several categorical explanatory variables to one quantitative response variable that is assumed normal within treatment combinations. Tests for interactions between explanatory variables are possible. An interaction means that the relationship of one explanatory variable to the response variable *depends* on the value of another explanatory variable. This method is not really elementary, because there is more than one explanatory variable.

Crosstabs and chi-squared tests

Cross-tabulations (Crosstabs) are joint frequency distribution of two categorical variables. One can be considered an explanatory variable, the other a response variable if you like. In any case (even when the explanatory variable 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 yourself for publishing the results. It's totally unclear which is the explanatory variable and which is the response variable, 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 explanatory variable (or variables)?*

Answer to Sample Question 1.2.18 *Gender*

Sample Question 1.2.19 *In your example, what is the response 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 explanatory 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.

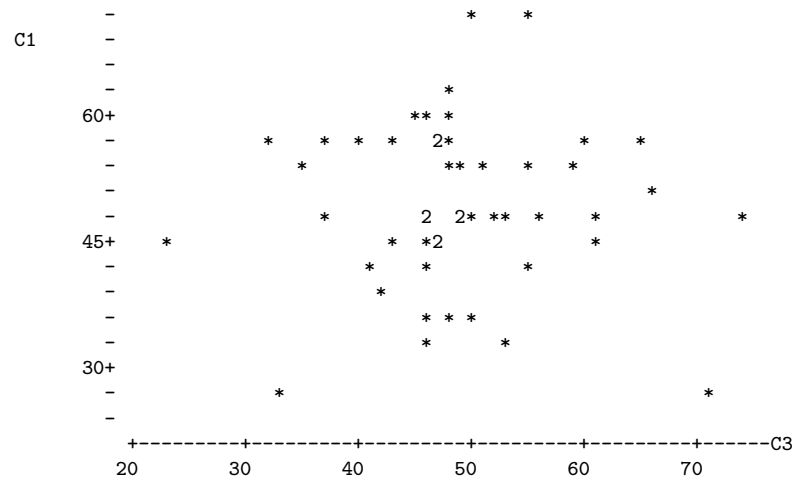
Often, the points in a scatterplot cluster around a straight line. The correlation coefficient (Pearson's r) expresses how close the points are to the line.

Here are some properties of the correlation coefficient r:

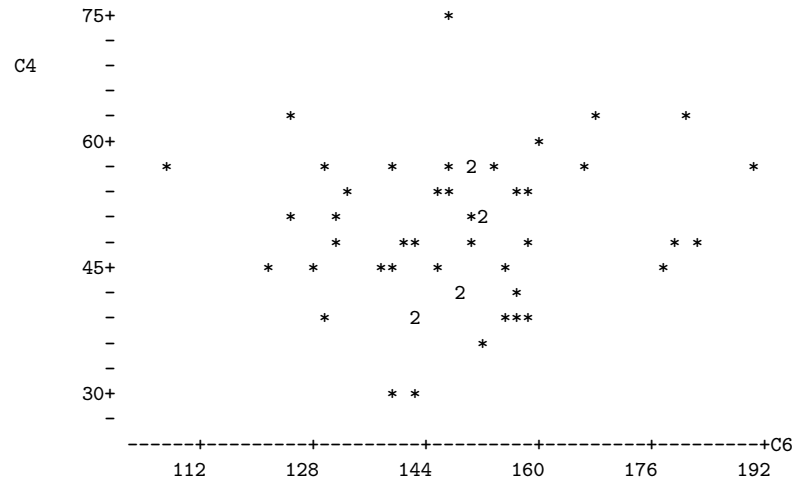
- $-1 \leq r \leq 1$
- $r = +1$ indicates a perfect positive linear relationship. All the points are exactly on a line with a positive slope.
- $r = -1$ indicates a perfect negative linear relationship. All the points are exactly on a line with a negative slope.
- $r = 0$ means no *linear* relationship (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.

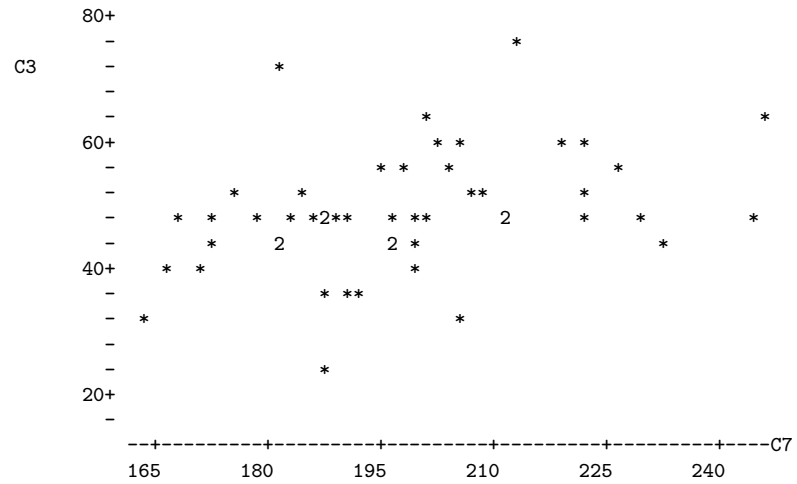
Here are some examples of scatterplots and the associated correlation coefficients. The number 2 on a plot means that two points are on top of each other, or at least too close to be distinguished in this crude line printer graphic.



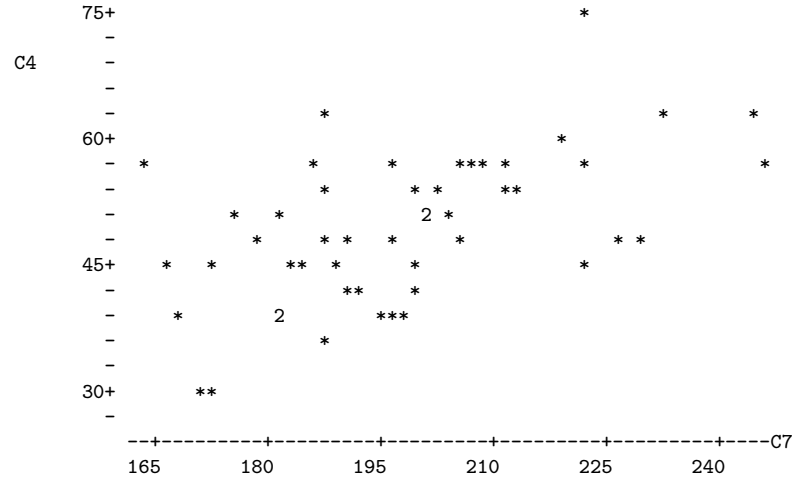
Correlation of C1 and C3 = 0.004



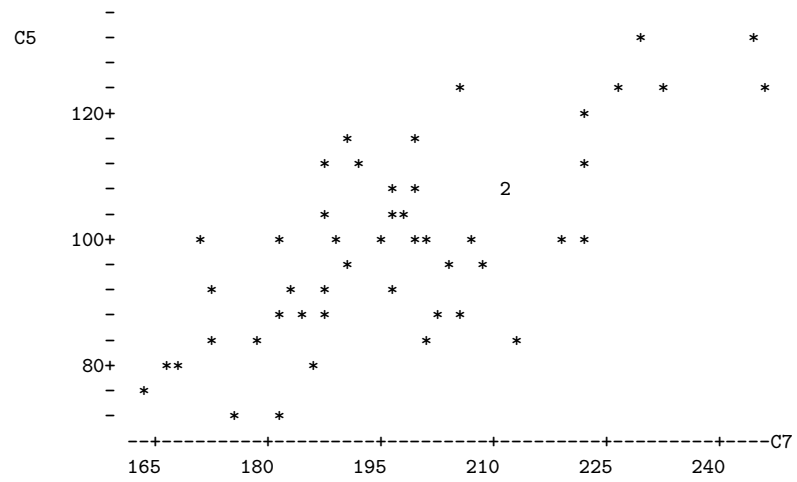
Correlation of C4 and C6 = 0.112



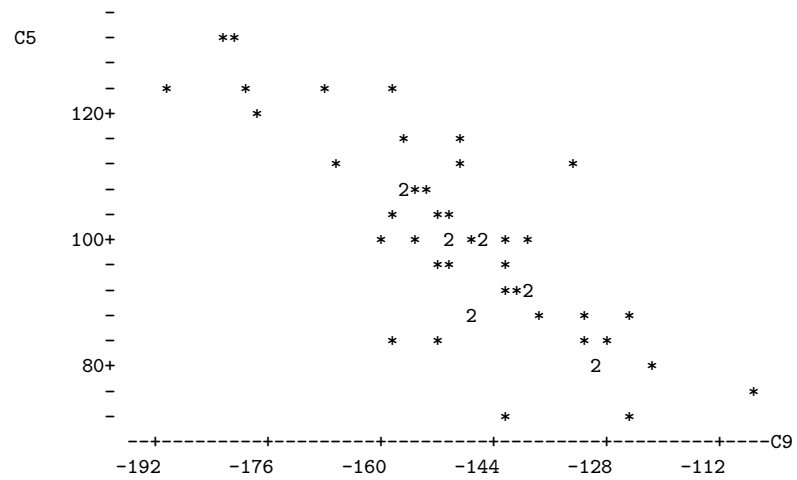
Correlation of C3 and C7 = 0.368



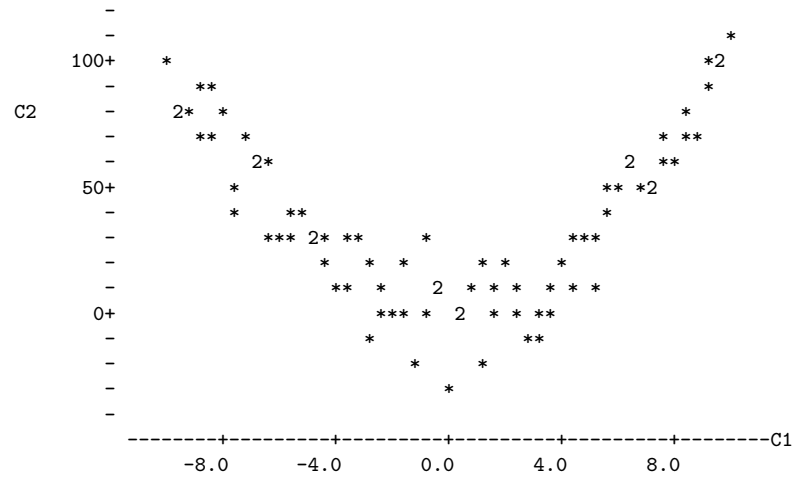
Correlation of C4 and C7 = 0.547



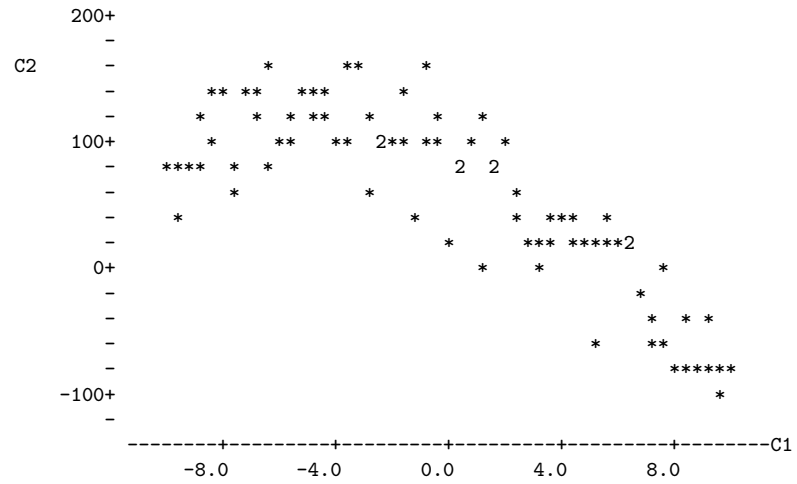
Correlation of C5 and C7 = 0.733



Correlation of C5 and C9 = -0.822



Correlation of C1 and C2 = 0.025




```
Correlation of C1 and C2 = -0.811
```

Simple Regression One explanatory 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 explanatory variable, the response variable has a normal distribution with the same variance (variance does not depend on value of the response variable). 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 explanatory 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 will be given later.

1.3 Experimental versus observational studies

Why might someone want to predict a response variable from an explanatory 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 explanatory 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 explanatory variable *cause* variation in the response variable? Who cares?

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

In all these discussions of causality, it is important to understand what the term does *not* mean. If we say that smoking cigarettes causes lung cancer, it does not mean that you will get lung cancer if and only if you smoke cigarettes. It means that smoking *contributes* to the *chances* that you will get cancer. So when we say "cause," we really mean "contributing factor." And it is almost always one contributing factor among many.

Now here are some 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 explanatory variable, that might be related to both the explanatory variable and the response 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 explanatory variable (or variables). An **observational study** is one in which the values of the explanatory 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 explanatory *variables* as experimental (i.e., randomly assigned, manipulated) or observed.

In an experimental study, there is no way the response variable could be causing the explanatory variable, because values of the explanatory variable 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 explanatory variable, 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 explanatory 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 explanatory variable and one categorical manipulated explanatory variable. Make*

the study multivariate, with one response variable consisting of unordered categories and two quantitative response variables.

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 response variable is whether the patient is alive or not 5 years after the study begins. The quantitative response 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 explanatory variable *causes* a response 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 to use at home. The response 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 response 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 response variable 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* [19]. 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* [20].

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 explanatory 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* [7] 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 [18], but the general idea is that when people know they are participating 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 consideration, 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.