Darlington, Chapter 1 pp 1-6 10/26

# CHAPTER 1

# BASIC CONCEPTS

## 1.1 STATISTICAL CONTROL

### 1.1.1 The Need for Control

If you have ever described a piece of research to a friend, it was probably not very long before you were asked a question like, "But did the researchers control for this?" If the research found a difference between the average salaries of men and women, did it control for differences in years of employment? If the research found differences among several ethnic groups in attitudes toward social welfare spending, did it control for income differences among the groups? If the research found that high-status female wolves have more pups on the average than low-status wolves, did it control for age differences among the wolves?

All these studies concern the relationship between an *independent variable* and a *dependent variable*. The study on salary differences concerns the relationship between the independent variable of gender and the dependent variable of salary. The study on welfare spending concerns the relationship between the independent variable of ethnicity and the dependent variable of attitude. The study on wolves concerns the relationship between the independent variable of status and the dependent variable of fertility. In each case there is a need to control a third variable; this third variable is called a *covariate*. The covariates for the three studies are, respectively, years of employment, income, and age.

Suppose you wanted to study these three relationships without worrying about covariates. You may be familiar with three very different statistical methods for analyzing these three problems. You may have studied the *t* test for testing questions like the sex difference in salaries, analysis of variance for questions like difference in average attitude among several ethnic groups, and the Pearson or rank-order correlation for questions like the relationship between status and number of pups. But in this book we will regard the differences among these three problems as minor in comparison with their similarities. The problems differ primarily in the type of independent variable. Gender is *dichotomous;* that is, there are two categories—male and female. Ethnicity is *multicategorical,* since there are several categories—the various ethnic groups in the study. Status is *numerical,* since there is a more or less continuous dimension from high status to low status. For our purposes, the differences among these three variable types are relatively minor. You should begin thinking of problems like these as basically similar, since all concern the relationship between an independent and a dependent variable. We shall return to this point in Secs. 3.2 and 10.1.

### 1.1.2 Five Methods of Control

You may already be somewhat familiar with four ways of controlling covariates: by *random assignment on the independent variable,* by *exclusion of cases,* by *manipulation of covariates,* and by *other types of randomization.* For instance, suppose you want to know whether driver training courses help students pass driving tests. One problem is that the students who take driver training courses may differ in various ways from those who do not. A second problem is that in a particular town, some testers may be easier than others. The driving schools may know which testers are easiest and encourage their students to take their tests when they know those testers are on duty.

You might control the first problem by using a list of applicants for driving courses, randomly choosing which of the applicants are allowed to take the course, and using the rejected applicants as the control group. This is *random assignment on the independent variable.* Or, if you find that more women take the courses than men, you might use a sample which is half female and half male for both the trained and the untrained groups. This would require discarding some available data, and is control by *exclusion of cases.* You might control the second problem by training testers to make them apply more uniform standards; that would be *manipulation of covariates.* Or you might control that problem by randomly altering the schedule different testers work, so that nobody would know which testers are on duty at a particular moment. That would not be random assignment on the independent variable, since you have not determined which applicants take the course; rather, it would be *other types of randomization.* This includes randomly assigning forms of the dependent variable (as in this example), choosing stimuli from a population of stimuli (for example, all common English adjectives), and manipulating the order of presentation of stimuli.

All these methods except exclusion of cases are types of *experimental control,* since they all require you to manipulate the situation in some way rather than merely observe it. These methods are often impractical or even

lysis of variance for
ethnic groups, and
he relationship be-
ll regard the differ-
parison with their
lependent variable.
—male and female.
ories—the various
e is a more or less
our purposes, the
minor. You should
nce all concern the
le. We shall return

ontrolling covari-
, by *exclusion of*
*ndomization*. For
ing courses help
s who take driver
do not. A second
asier than others.
d encourage their
e on duty.
licants for driving
owed to take the
). This is *random*
nore women take
f female and half
l require discard-
ou might control
y more uniform
ight control that
s work, so that
r moment. That
, since you have
would be *other*
ns of the depen-
lation of stimuli
ing the order of

of *experimental*
n in some way
ractical or even

impossible. For instance, you might not be allowed to decide which students take the driving course, or to train testers or alter their schedules. Or, if a covariate is worker seniority, as in one of our earlier examples, you cannot manipulate the covariate by telling workers how long to keep their jobs. In the same example, the independent variable is sex, and you cannot randomly decide that a particular worker will be male or female the way you can decide whether the worker will be in the experimental or control condition of an experiment. Even when experimental control is possible, the very exertion of control often intrudes the investigator into the situation in a way that disturbs subjects and alters results; ethologists and anthropologists are especially sensitive to such issues. Experimental control may be difficult even in laboratory studies on animals. Researchers may not be able to control how long a rat looks at a stimulus, but they are able to measure looking time.

Control by exclusion of cases avoids these difficulties, because you are manipulating data rather than subjects. But this method lowers sample size, and thus lowers the precision of estimates and the power of hypothesis tests.

A fifth method of controlling covariates—*statistical control*—is the topic of this book. It avoids the disadvantages of the previous four methods. No manipulation of subjects or conditions is required, and no data are excluded. Several terms mean the same thing: to *control* a covariate statistically means the same as to *adjust for* it or to *correct for* it, or to *hold constant* or to *partial out* the covariate.

Statistical control has limitations. Scientists may disagree on what variables need to be controlled—an investigator who has controlled age, income, and ethnicity may be criticized for failing to control education and family size. And because covariates must be measured to be controlled, they will be controlled inaccurately if they are measured inaccurately. We shall return to these and other problems in Chaps. 4 and 8. But because control of some covariates is almost always needed, and because the other four methods of control are so limited, statistical control is widely recognized as one of the most important statistical tools.

### 1.1.3 Examples of Statistical Control

The nature of statistical control can be illustrated by a simple fictitious example, though the precise methods used in this example are not those we shall emphasize later. In Holly City, 130 children attended a city-subsidized preschool program and 130 others did not. Later, all 260 children took a "school readiness test" on entering first grade. Of the 130 preschool children, only 60 scored above the median on the test; of the other 130 children, 70 scored above the median. In other words, the preschool children scored worse on the test than the others. These results are shown in the "Total" section of Table 1.1; A and B refer to scoring above and below the test median.

But when the children were divided into "middle class" and "working class," the results were as shown on the left and center of Table 1.1. We see that of the 40 middle-class children attending preschool, 30, or 75%, scored

**TABLE 1.1**
**Holly City**

| | Raw frequencies | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Middle | | | Working | | | Total | | |
| | A | B | Tot. | A | B | Tot. | A | B | Tot. |
| Preschool | 30 | 10 | 40 | 30 | 60 | 90 | 60 | 70 | 130 |
| Other | 60 | 30 | 90 | 10 | 30 | 40 | 70 | 60 | 130 |

above the median. There were 90 middle-class children not attending preschool, and 60, or 67%, of them scored above the median. These values of 75% and 67% are shown on the left in Table 1.2. Similar calculations based on the working-class and total tables yield the other figures in Table 1.2. This table shows clearly that within each level of socioeconomic status (SES), the preschool children outperform the other children, even though they appear to do worse than the other children in the "total" table. We have *held constant* or *controlled* or *partialed out* the covariate of SES.

When we perform a similar analysis for nearby Ivy City, we find the results in Table 1.3. When we inspect the total percentages, preschool appears to have a positive effect. But when we look within each SES group, no effect is found. Thus the "total" tables overstate the effect of preschool in Ivy City and understate it in Holly City. In these examples the independent variable is preschool attendance and the dependent variable is test score. In Holly City, we found a negative simple relationship between these two variables (those attending preschool scored lower on the test) but a positive partial relationship when SES was controlled. In Ivy City, we found a positive simple relationship but no partial relationship.

By examining the data more carefully, we can see what caused these paradoxical results. In Holly City, the 130 children attending preschool included 90 working-class children and 40 middle-class children, so 69% of the preschool attenders were working-class. But the 130 nonpreschool children included 90 middle-class children and 40 working-class children, so this group was only 31% working-class. Thus the test scores of the preschool group were lowered by the disproportionate number of working-class children in that group. This might have occurred if city-subsidized preschool programs had

**TABLE 1.2**
**Holly City**

| | Percentage scoring above the median | | |
|---|---|---|---|
| | Middle | Working | Total |
| Preschool | 75 | 33 | 46 |
| Other | 67 | 25 | 54 |

| Total | | |
|---|---|---|
| A | B | Tot. |
| 60 | 70 | 130 |
| 70 | 60 | 130 |

not attending pre-
ıese values of 75%
tions based on the
ɔle 1.2. This table
ıs (SES), the pre-
they appear to do
: *held constant* or

Ɔity, we find the
ɔreschool appears
;roup, no effect is
ol in Ivy City and
ıdent variable is
e. In Holly City,
variables (those
ırtial relationship
nple relationship

ıat caused these
ıg preschool in-
ı, so 69% of the
:school children
n, so this group
ıool group were
:hildren in that
l programs had

TABLE 1.3
**Ivy City**

| | Raw frequencies | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Middle | | | Working | | | Total | | |
| | A | B | Tot. | A | B | Tot. | A | B | Tot. |
| Preschool | 90 | 30 | 120 | 10 | 30 | 40 | 100 | 60 | 160 |
| Other | 30 | 10 | 40 | 30 | 90 | 120 | 60 | 100 | 160 |

| | Percentage scoring above the median | | |
|---|---|---|---|
| | Middle | Working | Total |
| Preschool | 75 | 25 | 62 |
| Other | 75 | 25 | 38 |

been established primarily in poorer neighborhoods. But in Ivy City this difference was in the opposite direction: the preschool group was 75% middle-class, while the nonpreschool group was only 25% middle-class; thus the test scores of the preschool group were raised by the disproportionate number of middle-class children. This might have occurred if parents had to pay for their children to attend preschool. In both cities the effects of preschool were seen more clearly by controlling for SES.

All three variables in this example were dichotomous—they had just two levels each. The independent variable of preschool attendance had two levels we called "preschool" and "other." The dependent variable of test score was dichotomized into those above and below the median. The covariate of socioeconomic status was also dichotomized. But any or all of the variables in this problem might have been numerical variables. Test scores might have ranged from 0 to 100, and SES might have been measured on a scale with many points. Even preschool attendance might have been numerical, if we scored the exact number of days each child had attended preschool. Changing some or all variables from dichotomous to numerical would change the details of analysis, but in its underlying logic the problem would remain the same. The use of numerical variables may be more complex, but it usually raises statistical power. Thus by dichotomizing SES and test scores in our examples above, we sacrificed power for simplicity.

Consider now a problem in which the dependent variable is numerical. At Swamp College, the dean calculated that among professors and instructors under 30 years of age, the average salary among males was $27,000 and the average salary among females was only $23,000. To see whether this difference might be attributed to different proportions of men and women who had completed the Ph.D., the dean made up the table given as Table 1.4.

If the dean had hoped that different rates of completion of the Ph.D. would explain the $4000 difference between men and women in average salary,

TABLE 1.4
**Average salaries at Swamp College, by sex and completion of Ph.D.**

| | Ph.D. completed | | |
|---|---|---|---|
| | Yes | No | Total |
| Men | \$30,000<br>$n = 10$ | \$26,000<br>$n = 30$ | \$27,000<br>$n = 40$ |
| Women | \$25,000<br>$n = 15$ | \$21,000<br>$n = 15$ | \$23,000<br>$n = 30$ |

that hope was frustrated. We see that men had completed the Ph.D. *less* often than women: 10 of 40 men, versus 15 of 30 women. The first column of the table shows that among instructors with a Ph.D., the difference in mean salaries between men and women is \$5000. The second column shows the same difference of \$5000 among instructors with no Ph.D. Therefore, in this artificial example, controlling for completion of the Ph.D. does not lower the difference between the mean salaries of men and women, but rather raises it from \$4000 to \$5000.

This example differs from the preschool example in its mechanical details; we are dealing with means rather than frequencies and proportions. But the underlying logic is the same. In the present case the independent variable is sex, the dependent variable is salary, and the covariate is educational level. Again, the partial relationship differs from the simple relationship, though this time both relationships have the same sign, since men always have higher salaries than women.

These examples are so simple that you may be wondering why a whole book is needed to discuss statistical control. But when the covariate is numerical, it may be that no two subjects have the same score on the covariate and so we cannot construct tables like those in the examples above. And we may want to control many covariates at once; the college dean might want to simultaneously control teaching ratings and other covariates as well as completion of the Ph.D. Also, we need methods for testing the significance of partial relationships. Other complexities are introduced later.

### 1.1.4 What You Should Know Already

This book assumes a working familiarity with the concepts of means and standard deviations, score distributions, samples and populations, random sampling, sampling distributions, null hypotheses, standard errors, statistical significance, power, confidence bands, one-tailed and two-tailed tests, summation, subscripts, and similar basic statistical terms and concepts. It refers