6. Research Design

A. Scientific Method
B. Measurement
C. Basics of Data Collection
D. Sampling Bias
E. Experimental Designs
F. Causation
G. Exercises

Scientific Method

by David M. Lane

Prerequisites

- none

This section contains a brief discussion of the most important principles of the scientific method. A thorough treatment of the philosophy of science is beyond the scope of this work.

One of the hallmarks of the scientific method is that it depends on empirical data. To be a proper scientific investigation, the data must be collected systematically. However, scientific investigation does not necessarily require experimentation in the sense of manipulating variables and observing the results. Observational studies in the fields of astronomy, developmental psychology, and ethology are common and provide valuable scientific information.

Theories and explanations are very important in science. Theories in science can never be proved since one can never be 100% certain that a new empirical finding inconsistent with the theory will never be found.

Scientific theories must be potentially disconfirmable. If a theory can accommodate all possible results then it is not a scientific theory. Therefore, a scientific theory should lead to testable hypotheses. If a hypothesis is disconfirmed, then the theory from which the hypothesis was deduced is incorrect. For example, the secondary reinforcement theory of attachment states that an infant becomes attached to its parent by means of a pairing of the parent with a primary reinforcer (food). It is through this "secondary reinforcement" that the child-parent bond forms. The secondary reinforcement theory has been disconfirmed by numerous experiments. Perhaps the most notable is one in which infant monkeys were fed by a surrogate wire mother while a surrogate cloth mother was available. The infant monkeys formed no attachment to the wire monkeys and frequently clung to the cloth surrogate mothers (Harlow, 1958).

If a hypothesis derived from a theory is confirmed, then the theory has survived a test and it becomes more useful and better thought of by the researchers in the field. A theory is not confirmed when correct hypotheses are derived from it.

A key difference between scientific explanations and faith-based explanations is simply that faith-based explanations are based on faith and do not

need to be testable. This does not mean that an explanation that cannot be tested is incorrect in some cosmic sense. It just means that it is not a scientific explanation.

The method of investigation in which a hypothesis is developed from a theory and then confirmed or disconfirmed involves deductive reasoning. However, deductive reasoning does not explain where the theory came from in the first place. In general, a theory is developed by a scientist who is aware of many empirical findings on a topic of interest. Then, through a generally poorly understood process called "induction," the scientist develops a way to explain all or most of the findings within a relatively simple framework or theory.

An important attribute of a good scientific theory is that it is parsimonious. That is, that it is simple in the sense that it uses relatively few constructs to explain many empirical findings. A theory that is so complex that it has as many assumptions as it has predictions is not very valuable.

Although strictly speaking, disconfirming an hypothesis deduced from a theory disconfirms the theory, it rarely leads to the abandonment of the theory. Instead, the theory will probably be modified to accommodate the inconsistent finding. If the theory has to be modified over and over to accommodate new findings, the theory generally becomes less and less parsimonious. This can lead to discontent with the theory and the search for a new theory. If a new theory is developed that can explain the same facts in a more parsimonious way, then the new theory will eventually supersede the old theory.

Measurement

by David M. Lane

Prerequisites

- Values of Pearson's Correlation
- Variance Sum Law
- Chapter 3: Measures of Variability

Learning Objectives

1. Define reliability
2. Describe reliability in terms of true scores and error
3. Compute reliability from the true score and error variance
4. Define the standard error of measurement and state why it is valuable
5. State the effect of test length on reliability
6. Distinguish between reliability and validity
7. Define three types of validity
8. State the how reliability determines the upper limit to validity

The measurement of psychological attributes such as self-esteem can be complex. A good measurement scale should be both reliable and valid. These concepts will be discussed in turn.

Reliability

The notion of reliability revolves around whether you would get at least approximately the same result if you measure something twice with the same measurement instrument. A common way to define reliability is the correlation between parallel forms of a test. Letting "test" represent a parallel form of the test, the symbol $r_{test,test}$ is used to denote the reliability of the test.

True Scores and Error

Assume you wish to measure a person's mean response time to the onset of a stimulus. For simplicity, assume that there is no learning over tests which, of

course, is not really true. The person is given 1,000 trials on the task and you obtain the response time on each trial.

The mean response time over the 1,000 trials can be thought of as the person's "true" score, or at least a very good approximation of it. Theoretically, the true score is the mean that would be approached as the number of trials increases indefinitely.

An individual response time can be thought of as being composed of two parts: the true score and the error of measurement. Thus if the person's true score were 345 and their response on one of the trials were 358, then the error of measurement would be 13. Similarly, if the response time were 340, the error of measurement would be -5.

Now consider the more realistic example of a class of students taking a 100-point true/false exam. Let's assume that each student knows the answer to some of the questions and has no idea about the other questions. For the sake of simplicity, we are assuming there is no partial knowledge of any of the answers and for a given question a student either knows the answer or guesses. Finally, assume the test is scored such that a student receives one point for a correct answer and loses a point for an incorrect answer. In this example, a student's true score is the number of questions they know the answer to and their error score is their score on the questions they guessed on. For example, assume a student knew 90 of the answers and guessed correctly on 7 of the remaining 10 (and therefore incorrectly on 3). Their true score would be 90 since that is the number of answers they knew. Their error score would be 7 - 3 = 4 and therefore their actual test score would be 90 + 4.

Every test score can be thought of as the sum of two independent components, the true score and the error score. This can be written as:

$$y_{test} = y_{true} + y_{error}$$

The following expression follows directly from the Variance Sum Law:

$$\sigma^2_{Test} = \sigma^2_{True} + \sigma^2_{error}$$

Reliability in Terms of True Scores and Error

It can be shown that the reliability of a test, $r_{test,test}$, is the ratio of true-score variance to test-score variance. This can be written as:

$$r_{test,test} = \frac{\sigma^2_{True}}{\sigma^2_{Test}} = \frac{\sigma^2_{True}}{\sigma^2_{True} + \sigma^2_{error}}$$

It is important to understand the implications of the role the variance of true scores plays in the definition of reliability: If a test were given in two populations for which the variance of the true scores differed, the reliability of the test would be higher in the population with the higher true-score variance. Therefore, reliability is not a property of a test per se but the reliability of a test in a given population.

Assessing Error of Measurement

The reliability of a test does not show directly how close the test scores are to the true scores. That is, it does not reveal how much a person's test score would vary across parallel forms of the test. By definition, the mean over a large number of parallel tests would be the true score. The standard deviation of a person's test scores would indicate how much the test scores vary from the true score. This standard deviation is called the standard error of measurement. In practice, it is not practical to give a test over and over to the same person and/or assume that there are no practice effects. Instead, the following formula is used to estimate the standard error of measurement.

$$S_{measurement} = S_{test}\sqrt{1 - r_{test,test}}$$

where $s_{measurement}$ is the standard error of measurement, s_{test} is the standard deviation of the test scores, and $r_{test,test}$ is the reliability of the test. Taking the extremes, if the reliability is 0, then the standard error of measurement is equal to the standard deviation of the test; if the reliability is perfect (1.0) then the standard error of measurement is 0.

Increasing Reliability

It is important to make measures as reliable as is practically possible. Suppose an investigator is studying the relationship between spatial ability and a set of other variables. The higher the reliability of the test of spatial ability, the higher the correlations will be. Similarly, if an experimenter seeks to determine whether a particular exercise regiment decreases blood pressure, the higher the reliability of

the measure of blood pressure, the more sensitive the experiment. More precisely, the higher the reliability the higher the power of the experiment. Power is covered in detail in Chapter 13. Finally, if a test is being used to select students for college admission or employees for jobs, the higher the reliability of the test the stronger will be the relationship to the criterion.

Two basic ways of increasing reliability are (1) to improve the quality of the items and (2) to increase the number of items. Items that are either too easy so that almost everyone gets them correct or too difficult so that almost no one gets them correct are not good items: they provide very little information. In most contexts, items which about half the people get correct are the best (other things being equal).

Items that do not correlate with other items can usually be improved. Sometimes the item is confusing or ambiguous.

Increasing the number of items increases reliability in the manner shown by the following formula:

$$r_{new,new} = \frac{kr_{test,test}}{1 + (k-1)r_{test,test}}$$

where k is the factor by which the test length is increased, $r_{new,new}$ is the reliability of the new longer test, and $r_{test,test}$ is the current reliability. For example, if a test with 50 items has a reliability of .70 then the reliability of a test that is 1.5 times longer (75 items) would be calculated as follows

$$r_{new,new} = \frac{(1.5)(0.70)}{1 + (1.5-1)(0.70)}$$

which equals 0.78. Thus increasing the number of items from 50 to 75 would increase the reliability from 0.70 to 0.78.

It is important to note that this formula assumes the new items have the same characteristics as the old items. Obviously adding poor items would not increase the reliability as expected and might even decrease the reliability.

Validity

The validity of a test refers to whether the test measures what it is supposed to measure. The three most common types of validity are face validity, empirical validity, and construct validity. We consider these types of validity below.

Face Validity

A test's face validity refers to whether the test appears to measure what it is supposed to measure. That is, does the test "on its face" appear to measure what it is supposed to be measuring. An Asian history test consisting of a series of questions about Asian history would have high face validity. If the test included primarily questions about American history then it would have little or no face validity as a test of Asian history.

Predictive Validity

Predictive validity (sometimes called empirical validity) refers to a test's ability to predict a relevant behavior. For example, the main way in which SAT tests are validated is by their ability to predict college grades. Thus, to the extent these tests are successful at predicting college grades they are said to possess predictive validity.

Construct Validity

Construct validity is more difficult to define. In general, a test has construct validity if its pattern of correlations with other measures is in line with the construct it is purporting to measure. Construct validity can be established by showing a test has both convergent and divergent validity. A test has convergent validity if it correlates with other tests that are also measures of the construct in question. Divergent validity is established by showing the test does not correlate highly with tests of other constructs. Of course, some constructs may overlap so the establishment of convergent and divergent validity can be complex.

To take an example, suppose one wished to establish the construct validity of a new test of spatial ability. Convergent and divergent validity could be established by showing the test correlates relatively highly with other measures of spatial ability but less highly with tests of verbal ability or social intelligence.

Reliability and Predictive Validity

The reliability of a test limits the size of the correlation between the test and other measures. In general, the correlation of a test with another measure will be lower than the test's reliability. After all, how could a test correlate with something else as high as it correlates with a parallel form of itself? Theoretically it is possible for a test to correlate as high as the square root of the reliability with another measure. For example, if a test has a reliability of 0.81 then it could correlate as high as 0.90 with another measure. This could happen if the other measure were a perfectly reliable test of the same construct as the test in question. In practice, this is very unlikely.

A correlation above the upper limit set by reliabilities can act as a red flag. For example, Vul, Harris, Winkielman, and Paschler (2009) found that in many studies the correlations between various fMRI activation patterns and personality measures were higher than their reliabilities would allow. A careful examination of these studies revealed serious flaws in the way the data were analyzed.

Basics of Data Collection

by Heidi Zeimer

Prerequisites

- None

Learning Objectives

1. Describe how a variable such as height should be recorded
2. Choose a good response scale for a questionnaire

Most statistical analyses require that your data be in numerical rather than verbal form (you can't punch letters into your calculator). Therefore, data collected in verbal form must be coded so that it is represented by numbers. To illustrate, consider the data in Table 1.

Table 1. Example Data

Student Name	Hair Color	Gender	Major	Height	Computer Experience
Norma	Brown	Female	Psychology	5'4"	Lots
Amber	Blonde	Female	Social Science	5'7"	Very little
Paul	Blonde	Male	History	6'1"	Moderate
Christopher	Black	Male	Biology	5'10"	Lots
Sonya	Brown	Female	Psychology	5'4"	Little

Can you conduct statistical analyses on the above data or must you re-code it in some way? For example, how would you go about computing the average height of the 5 students. You cannot enter students' heights in their current form into a statistical program -- the computer would probably give you an error message because it does not understand notation such as 5'4". One solution is to change all the numbers to inches. So, 5'4" becomes (5 x 12) + 4 = 64, and 6'1" becomes (6 x 12) + 1 = 73, and so forth. In this way, you are converting height in feet and inches to simply height in inches. From there, it is very easy to ask a statistical program to calculate the mean height in inches for the 5 students.

You may ask, “Why not simply ask subjects to write their height in inches in the first place?” Well, the number one rule of data collection is to ask for information in such a way as it will be most accurately reported. Most people know their height in feet and inches and cannot quickly and accurately convert it into inches “on the fly.” So, in order to preserve data accuracy, it is best for researchers to make the necessary conversions.

Let’s take another example. Suppose you wanted to calculate the mean amount of computer experience for the five students shown in Table 1. One way would be to convert the verbal descriptions to numbers as shown in Table 2. Thus, “Very Little” would be converted to “1” and “Little” would be converted to “2.”

Table 2. Conversion of verbal descriptions to numbers

1	2	3	4	5
Very Little	Little	Moderate	Lots	Very Lots

Measurement Examples

Example #1: How much information should I record?

Say you are volunteering at a track meet at your college, and your job is to record each runner’s time as they pass the finish line for each race. Their times are shown in large red numbers on a digital clock with eight digits to the right of the decimal point, and you are told to record the entire number in your tablet. Thinking eight decimal places is a bit excessive, you only record runners’ times to one decimal place. The track meet begins, and runner number one finishes with a time of 22.93219780 seconds. You dutifully record her time in your tablet, but only to one decimal place, that is 22.9. Race number two finishes and you record 32.7 for the winning runner. The fastest time in Race number three is 25.6. Race number four winning time is 22.9, Race number five is…. But wait! You suddenly realize your mistake; you now have a tie between runner one and runner four for the title of Fastest Overall Runner! You should have recorded more information from the digital clock -- that information is now lost, and you cannot go back in time and record running times to more decimal places.

The point is that you should think very carefully about the scales and specificity of information needed in your research before you begin collecting data. If you believe you might need additional information later but are not sure,

measure it; you can always decide to not use some of the data, or "collapse" your data down to lower scales if you wish, but you cannot expand your data set to include more information after the fact. In this example, you probably would not need to record eight digits to the right of the decimal point. But recording only one decimal digit is clearly too few.

Example #2

Pretend for a moment that you are teaching five children in middle school (yikes!), and you are trying to convince them that they must study more in order to earn better grades. To prove your point, you decide to collect actual data from their recent math exams, and, toward this end, you develop a questionnaire to measure their study time and subsequent grades. You might develop a questionnaire which looks like the following:

1. Please write your name: _____
2. Please indicate how much you studied for this math exam:
   a lot……moderate……..little
3. Please circle the grade you received on the math exam:
   A B C D F

Given the above questionnaire, your obtained data might look like the following:

Name	Amount Studied	Grade
John	Little	C
Sally	Moderate	B
Alexander	Lots	A
Linda	Moderate	A
Thomas	Little	B

Eyeballing the data, it seems as if the children who studied more received better grades, but it's difficult to tell. "Little," "lots," and "B," are imprecise, qualitative terms. You could get more precise information by asking specifically how many hours they studied and their exact score on the exam. The data then might look as follows:

Name	Hours studied	% Correct

John	5	71
Sally	9	83
Alexander	13	97
Linda	12	91
Thomas	7	85

Of course, this assumes the students would know how many hours they studied. Rather than trust the students' memories, you might ask them to keep a log of their study time as they study.

Sampling Bias

by David M. Lane

Prerequisites

- Inferential Statistics (including sampling)

Learning Objectives

1. Recognize sampling bias
2. Distinguish among self-selection bias, undercoverage bias, and survivorship bias

Descriptions of various types of sampling such as *simple random sampling* and *stratified random sampling* are covered in the inferential statistics section of Chapter 1. This section discusses various types of sampling biases including self-selection bias and survivorship bias. Examples of other sampling biases that are not easily categorized will also be given.

It is important to keep in mind that sampling bias refers to the method of sampling, not the sample itself. There is no guarantee that random sampling will result in a sample representative of the population just as not every sample obtained using a biased sampling method will be greatly non-representative of the population.

Self-Selection Bias

Imagine that a university newspaper ran an ad asking for students to volunteer for a study in which intimate details of their sex lives would be discussed. Clearly the sample of students who would volunteer for such a study would not be representative of the students at the university. Similarly, an online survey about computer use is likely to attract people more interested in technology than is typical. In both of these examples, people who "self-select" themselves for the experiment are likely to differ in important ways from the population the experimenter wishes to draw conclusions about. Many of the admittedly "non-scientific" polls taken on television or web sites suffer greatly from self-selection bias.

A self-selection bias can result when the non-random component occurs after the potential subject has enlisted in the experiment. Considering again the hypothetical experiment in which subjects are to be asked intimate details of their sex lives, assume that the subjects did not know what the experiment was going to be about until they showed up. Many of the subjects would then likely leave the experiment resulting in a biased sample.

Undercoverage Bias

A common type of sampling bias is to sample too few observations from a segment of the population. A commonly-cited example of undercoverage is the poll taken by the Literary Digest in 1936 that indicated that Landon would win an election against Roosevelt by a large margin when, in fact, it was Roosevelt who won by a large margin. A common explanation is that poorer people were undercovered because they were less likely to have telephones and that this group was more likely to support Roosevelt.

A detailed analysis by Squire (1988) showed that it was not just an undercoverage bias that resulted in the faulty prediction of the election results. He concluded that, in addition to the undercoverage described above, there was a nonresponse bias (a form of self-selection bias) such that those favoring Landon were more likely to return their survey than were those favoring Roosevelt.

Survivorship Bias

Survivorship bias occurs when the observations recorded at the end of the investigation are a non-random set of those present at the beginning of the investigation. Gains in stock funds is an area in which survivorship bias often plays a role. The basic problem is that poorly-performing funds are often either eliminated or merged into other funds. Suppose one considers a sample of stock funds that exist in the present and then calculates the mean 10-year appreciation of those funds. Can these results be validly generalized to other stock funds of the same type? The problem is that the poorly-performing stock funds that are not still in existence (did not survive for 10 years) are not included. Therefore, there is a bias toward selecting better-performing funds. There is good evidence that this survivorship bias is substantial (Malkiel, 1995).

In World War II, the statistician Abraham Wald analyzed the distribution of hits from anti-aircraft fire on aircraft returning from missions. The idea was that

this information would be useful for deciding where to place extra armor. A naive approach would be to put armor at locations that were frequently hit to reduce the damage there. However, this would ignore the survivorship bias occurring because only a subset of aircraft return. Wald's approach was the opposite: if there were few hits in a certain location on returning planes, then hits in that location were likely to bring a plane down. Therefore, he recommended that locations without hits on the returning planes should be given extra armor. A detailed and mathematical description of Wald's work can be found in Mangel and Samaniego (1984.)

Experimental Designs

by David M. Lane

Prerequisites

- Chapter 1: Variables

Learning Objectives

1. Distinguish between between-subject and within-subject designs
2. State the advantages of within-subject designs
3. Define "multi-factor design" and "factorial design"
4. Identify the levels of a variable in an experimental design
5. Describe when counterbalancing is used

There are many ways an experiment can be designed. For example, subjects can all be tested under each of the treatment conditions or a different group of subjects can be used for each treatment. An experiment might have just one *independent variable* or it might have several. This section describes basic experimental designs and their advantages and disadvantages.

Between-Subjects Designs

In a *between-subjects* design, the various experimental treatments are given to different groups of subjects. For example, in the "*Teacher Ratings*" case study, subjects were randomly divided into two groups. Subjects were all told they were going to see a video of an instructor's lecture after which they would rate the quality of the lecture. The groups differed in that the subjects in one group were told that prior teaching evaluations indicated that the instructor was charismatic whereas subjects in the other group were told that the evaluations indicated the instructor was punitive. In this experiment, the *independent variable* is "Condition" and has two levels (charismatic teacher and punitive teacher). It is a *between-subjects* variable because different subjects were used for the two levels of the independent variable: subjects were in either the "charismatic teacher" or the "punitive teacher" condition. Thus the comparison of the charismatic-teacher condition with the punitive-teacher condition is a comparison between the subjects in one condition with the subjects in the other condition.

The two conditions were treated exactly the same except for the instructions they received. Therefore, it would appear that any difference between conditions should be attributed to the treatments themselves. However, this ignores the possibility of chance differences between the groups. That is, by chance, the raters in one condition might have, on average, been more lenient than the raters in the other condition. Randomly assigning subjects to treatments ensures that all differences between conditions are chance differences; it does not ensure there will be no differences. The key question, then, is how to distinguish real differences from chance differences. The field of inferential statistics answers just this question. The inferential statistics applicable to testing the difference between the means of the two conditions covered in Chapter 12. Analyzing the data from this experiment reveals that the ratings in the charismatic-teacher condition were higher than those in the punitive-teacher condition. Using *inferential statistics*, it can be calculated that the probability of finding a difference as large or larger than the one obtained if the treatment had no effect is only 0.018. Therefore it seems likely that the treatment had an effect and it is not the case that all differences were chance differences.

Independent variables often have several levels. For example, in the "Smiles and Leniency" case study, the independent variable is "type of smile" and there are four levels of this independent variable: (1) false smile, (2) felt smile, (3) miserable smile, and (4) a neutral control. Keep in mind that although there are four levels, there is only one independent variable. Designs with more than one independent variable are considered next.

Multi-Factor Between-Subject Designs

In the "*Bias Against Associates of the Obese*" experiment, the qualifications of potential job applicants were judged. Each applicant was accompanied by an associate. The experiment had two independent variables: the weight of the associate (obese or average) and the applicant's relationship to the associate (girl friend or acquaintance). This design can be described as an Associate's Weight (2) x Associate's Relationship (2) *factorial design*. The numbers in parentheses represent the number of levels of the independent variable. The design was a factorial design because all four combinations of associate's weight and associate's relationship were included. The dependent variable was a rating of the applicant's qualifications (on a 9-point scale).

If two separate experiments had been conducted, one to test the effect of Associate's Weight and one to test the effect of Associate's Relationship then there would be no way to assess whether the effect of Associate's Weight depended on the Associate's Relationship. One might imagine that the Associate's Weight would have a larger effect if the associate were a girl friend rather than merely an acquaintance. A factorial design allows this question to be addressed. When the effect of one variable does differ depending on the level of the other variable then it is said that there is an *interaction* between the variables.

Factorial designs can have three or more independent variables. In order to be a between-subjects design there must be a separate group of subjects for each combination of the levels of the independent variables.

Within-Subjects Designs

A within-subjects design differs from a between-subjects design in that the same subjects perform at all levels of the *independent variable*. For example consider the "*ADHD Treatment*" case study. In this experiment, subjects diagnosed as having attention deficit disorder were each tested on a delay of gratification task after receiving methylphenidate (MPH). All subjects were tested four times, once after receiving one of the four doses. Since each subject was tested under each of the four levels of the independent variable "dose," the design is a *within-subjects design* and dose is a *within-subjects variable*. Within-subjects designs are sometimes called *repeated-measures designs*.

Counterbalancing

In a within-subject design it is important not to *confound* the order in which a task is performed with the experimental treatment. For example, consider the problem that would have occurred if, in the ADHD study, every subject had received the doses in the same order starting with the lowest and continuing to the highest. It is not unlikely that experience with the delay of gratification task would have an effect. If practice on this task leads to better performance, then it would appear that higher doses caused the better performance when, in fact, it was the practice that caused the better performance.

One way to address this problem is to *counterbalance* the order of presentations. In other words, subjects would be given the doses in different orders

in such a way that each dose was given in each sequential position an equal number of times. An example of counterbalancing is shown in Table 1.

Table 1. Counterbalanced order for four subjects.

Subject	0 mg/kg	.15 mg/kg	.30 mg/kg	.60 mg/kg
1	First	Second	Third	Fourth
2	Second	Third	Fourth	First
3	Third	Fourth	First	Second
4	Fourth	First	Second	Third

It should be kept in mind that counterbalancing is not a satisfactory solution if there are complex dependencies between which treatment precedes which and the dependent variable. In these cases, it is usually better to use a between-subjects design than a within-subjects design.

Advantage of Within-Subjects Designs

An advantage of within-subjects designs is that individual differences in subjects' overall levels of performance are controlled. This is important because subjects invariably will differ greatly from one another. In an experiment on problem solving, some subjects will be better than others regardless of the condition they are in. Similarly, in a study of blood pressure some subjects will have higher blood pressure than others regardless of the condition. Within-subjects designs control these individual differences by comparing the scores of a subject in one condition to the scores of the same subject in other conditions. In this sense each subject serves as his or her own control. This typically gives within-subjects designs considerably more *power* than between-subjects designs. That is, this makes within-subjects designs more able to detect an effect of the independent variable than are between-subjects designs.

Within-subjects designs are often called “repeated-measures” designs since repeated measurements are taken for each subject. Similarly, a within-subject variable can be called a repeated-measures factor.

Complex Designs

Designs can contain combinations of between-subject and within-subject variables. For example, the "*Weapons and Aggression*" case study has one between-subject variable (gender) and two within-subject variables (the type of priming word and the type of word to be responded to).

Causation

by David M. Lane

Prerequisites

- Chapter 1: What are Statistics
- Chapter 3: Measures of Variability
- Chapter 4: Pearson's Correlation
- Chapter 6: Experimental Designs

Learning Objectives

1. Explain how experimentation allows causal inferences
2. Explain the role of unmeasured variables
3. Explain the "third-variable" problem
4. Explain how causation can be inferred in non-experimental designs

The concept of causation is a complex one in the philosophy of science. Since a full coverage of this topic is well beyond the scope of this text, we focus on two specific topics: (1) the establishment of causation in experiments and (2) the establishment of causation in non-experimental designs.

Establishing Causation in Experiments

Consider a simple experiment in which subjects are *sampled randomly* from a *population* and then *assigned randomly* to either the experimental group or the control group. Assume the condition means on the *dependent variable* differed. Does this mean the treatment caused the difference?

To make this discussion more concrete, assume that the experimental group received a drug for insomnia, the control group received a placebo, and the dependent variable was the number of minutes the subject slept that night. An obvious obstacle to inferring causality is that there are many unmeasured variables that affect how many hours someone sleeps. Among them are how much stress the person is under, physiological and genetic factors, how much caffeine they consumed, how much sleep they got the night before, etc. Perhaps differences between the groups on these factors are responsible for the difference in the number of minutes slept.

At first blush it might seem that the random assignment eliminates differences in unmeasured variables. However, this is not the case. Random

assignment ensures that differences on unmeasured variables are chance differences. It does not ensure that there are no differences. Perhaps, by chance, many subjects in the control group were under high stress and this stress made it more difficult to fall asleep. The fact that the greater stress in the control group was due to chance does not mean it could not be responsible for the difference between the control and the experimental groups. In other words, the observed difference in "minutes slept" could have been due to a chance difference between the control group and the experimental group rather than due to the drug's effect.

This problem seems intractable since, by definition, it is impossible to measure an "unmeasured variable" just as it is impossible to measure and control all variables that affect the dependent variable. However, although it is impossible to assess the effect of any single unmeasured variable, it is possible to assess the combined effects of all unmeasured variables. Since everyone in a given condition is treated the same in the experiment, differences in their scores on the dependent variable must be due to the unmeasured variables. Therefore, a measure of the differences among the subjects within a condition is a measure of the sum total of the effects of the unmeasured variables. The most common measure of differences is the variance. By using the within-condition variance to assess the effects of unmeasured variables, statistical methods determine the probability that these unmeasured variables could produce a difference between conditions as large or larger than the difference obtained in the experiment. If that probability is low, then it is inferred (that's why they call it *inferential statistics*) that the treatment had an effect and that the differences are not entirely due to chance. Of course, there is always some nonzero probability that the difference occurred by chance so total certainty is not a possibility.

Causation in Non-Experimental Designs

It is almost a cliché that correlation does not mean causation. The main fallacy in inferring causation from correlation is called the "*third-variable problem*" and means that a third variable is responsible for the correlation between two other variables. An excellent example used by Li (1975) to illustrate this point is the positive correlation in Taiwan in the 1970's between the use of contraception and the number of electric appliances in one's house. Of course, using contraception does not induce you to buy electrical appliances or vice versa. Instead, the third variable of education level affects both.

Does the possibility of a third-variable problem make it impossible to draw causal inferences without doing an experiment? One approach is to simply assume that you do not have a third-variable problem. This approach, although common, is not very satisfactory. However, be aware that the assumption of no third-variable problem may be hidden behind a complex causal model that contains sophisticated and elegant mathematics.

A better though, admittedly more difficult approach, is to find converging evidence. This was the approach taken to conclude that smoking causes cancer. The analysis included converging evidence from retrospective studies, prospective studies, lab studies with animals, and theoretical understandings of cancer causes.

A second problem is determining the direction of causality. A correlation between two variables does not indicate which variable is causing which. For example, Reinhart and Rogoff (2010) found a strong correlation between public debt and GDP growth. Although some have argued that public debt slows growth, most evidence supports the alternative that slow growth increases public debt.

Statistical Literacy

by David M. Lane

Prerequisites

- Chapter 6: Causation

A low level of HDL have long been known to be a risk factor for heart disease. Taking niacin has been shown to increase HDL levels and has been recommended for patients with low levels of HDL. The assumption of this recommendation is that niacin causes HDL to increase thus causing a lower risk for heart disease.

What do you think?

What experimental design involving niacin would test whether the relationship between HDL and heart disease is causal?

> You could randomly assign patients with low levels of HDL to a condition in which they received niacin or to one in which they did not. A finding that niacin increased HDL without decreasing heart disease would cast doubt on the causal relationship. This is exactly what was found in a study conducted by the NIH. See the description of the results here.

References

Harlow, H. (1958) The nature of love. *American Psychologist*, *13*, 673-685.

Li, C. (1975) *Path analysis: A primer*. Boxwood Press, Pacific Grove. CA .

Malkiel, B. G. (1995) Returns from investing in equity mutual funds 1971 to 1991. *The Journal of Finance, 50*, 549-572.

Mangel, M. & Samaniego, F. J. (1984) Abraham Wald's work on aircraft survivability. *Journal of the American Statistical Association, 79,* 259-267.

Reinhart, C. M. and Rogoff, K. S. (2010). Growth in a Time of Debt. Working Paper 15639, National Bureau of Economic Research, http://www.nber.org/papers/w15639

Squire, P. (1988) Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly, 52*, 125-133.

Vul, E., Harris, C., Winkielman, P., & Paschler, H. (2009) Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, *4*, 274-290.

Exercises

1. To be a scientific theory, the theory must be potentially _____.
2. What is the difference between a faith-based explanation and a scientific explanation?
3. What does it mean for a theory to be parsimonious?
4. Define reliability in terms of parallel forms.
5. Define true score.
6. What is the reliability if the true score variance is 80 and the test score variance is 100?
7. What statistic relates to how close a score on one test will be to a score on a parallel form?
8. What is the effect of test length on the reliability of a test?
9. Distinguish between predictive validity and construct validity.
10. What is the theoretical maximum correlation of a test with a criterion if the test has a reliability of .81?
11. An experiment solicits subjects to participate in a highly stressful experiment. What type of sampling bias is likely to occur?
12. Give an example of survivorship bias not presented in this text.
13. Distinguish "between-subject" variables from "within-subjects" variables.
14. Of the variables "gender" and "trials," which is likely to be a between-subjects variable and which a within-subjects variable?
15. Define interaction.
16. What is counterbalancing used for?
17. How does randomization deal with the problem of pre-existing differences between groups?
18. Give an example of the "third-variable problem" other than those in this text.