2 RESEARCH METHODOLOGY

In the early 1980s there was a flurry of articles in the popular press that reported the supposed hazards of video display terminals (VDTs) — those TV-like terminals connected to large computers or sitting on top of microcomputers. These purported adverse effects ranged from relatively mild ones like fatigue to more serious ones that affected pregnant women, such as stillbirths, miscarriages, and congenital fetal abnormalities. One newspaper reported that four of seven women who worked with VDTs gave birth to children with defects. This news story created a considerable stir and was cited in a Canadian task-force report on hazards in the workplace. Since anywhere between 1 million and 7 million people in North America use these terminals on a daily basis, they would represent a major health hazard were these reports true.

The task of the epidemiologist in this situation is twofold: (1) to determine if there is indeed an increased risk to the fetus caused by the mother working with VDTs; and (2) if so, to determine what the magnitude of that risk is. In this section we explore some of the possible research designs that could be used to answer these questions. We begin with the basic elements of research design; then discuss various factors, called **threats to validity**, that may lead us to draw erroneous conclusions from the data; and then show how the different design elements can be combined into various types of studies to minimize these threats to validity.

When discussing the different types of sampling strategies, biases, designs, and other elements our aim is not to be comprehensive; any such compendium is always incomplete, since the number of types is based solely on the imagination and inventiveness of the researcher. Rather, we mention some of the more common varieties of each of these factors to illustrate how they can be combined in various ways to address different issues.

DESIGN ELEMENTS

EXPERIMENTAL OR OBSERVATIONAL STUDIES

In **experimental** studies the intervention is under the control of the researcher. For example, the research team may determine (by random allocation) which subjects receive a novel treatment and which ones get traditional (or no) treatment, when an intervention is carried out in a community, or how much of a new drug each patient is given. The aim is to determine how changes in the **independent variable** (the one under the researcher's control) affect some outcome (the **dependent variable**). By controlling the timing or amount of the intervention, or which subjects get it and which ones do not, the chances are minimized that other factors outside of the researcher's control could have affected the results.

By contrast, the researcher does not control the intervention in **observational** studies, but rather observes the effects of an experiment in nature. It would be both unethical and impractical, for example, to expose some people to cigarette smoke or putative occupational carcinogens deliberately for 20 years to determine their effects. However, by choice or chance, some people have been exposed, so it is possible to draw some tentative conclusions based on observation of these subjects and, if possible, control subjects.

Most well-designed studies of a new treatment are experimental, in that the research team determines which subjects receive the new drug or intervention and which ones receive traditional treatment or a placebo. Almost all studies that involve exposure to harmful agents or that try to trace the natural history of a disorder are observational. However, these general rules naturally have exceptions. For example, if VDTs were being introduced gradually into the workplace, so that there were fewer terminals than eligible workers and there was no hard evidence of any adverse effects, women could be randomly assigned to work with them or remain using typewriters. However, this may be difficult to do because of practical considerations, and an observational type of study may be more realistic. (Needless to say, the researcher cannot control which women become pregnant. The last one who tried was hauled up on morality charges.)

NUMBER OF OBSERVATIONS

The simplest research design would involve looking at or measuring the outcome only once. In many cases, such as when the outcome is either present or absent or when the *timing* of the outcome is of minor interest, one observation may be all that is necessary. For example, if the question is whether working at a VDT results in a higher incidence of stillbirths, miscarriages, or congenital abnormalities, then we could record these outcomes 9 months after conception for this group of women and for an appropriate control group. The outcome is recorded only on a single occasion.

However, if we were interested in the *time course* of an outcome, one observation is not sufficient. To use a different example, Bagby and his colleagues looked at the effects of a new mental health act introduced toward the end of 1978 on the proportion of psychiatric patients who were involuntarily admitted to hospital (Fig. 2-1). The graph shows a dramatic decline in this type of admission following the new, more restrictive legislation. If the analysis had stopped at this point, it's likely that people would have come to the erroneous conclusion that the new act resulted in a reduction in the proportion of people being admitted to psychiatric wards on an involuntary basis. Multiple observations over time, however, show a different picture, that is, a gradual return to a level even higher than those of the 7 years preceding the new law. So, not only do multiple observations tell us something different than a single look does, they also reveal something about the "natural history" of the legislation; there was a gradual return to the previous mode of practice as psychiatrists learned to live with the new law.

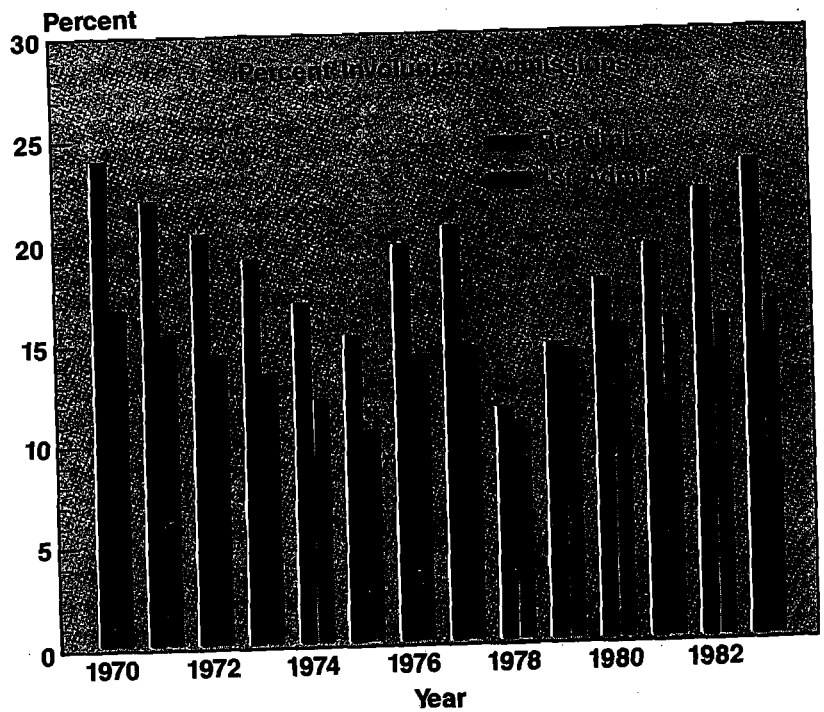


Figure 2-1 The proportion of psychiatric patients involuntarily admitted to hospital before and after the new mental health act of 1978.

DIRECTION OF DATA GATHERING

Data can be gathered in one of two ways: (1) looking *forward* and getting new data after the start of the study, or (2) looking *backward*, and using data that have already been collected. Specific names are used for each of these strategies. Studies that involve gathering data after the study has begun are called **prospective**; in **retrospective** studies the data have already been recorded for other reasons at some time in the past. The advantage of prospective data collection is that the nature of the data, the definitions of symptoms, the method by which the data are recorded, and other factors can be worked out ahead of time and are constant over the course of the trial. In retrospective studies definitions of symptoms or diseases may have been modified over time, units of measurement may have changed, and old methods for diagnosis may have been replaced, thereby resulting in more variability in the data. Perhaps the greatest advantage of prospective studies is that they allow us to determine the *directionality* of events (i.e., what occurred first and what happened later). As we'll see in *Assessing Causation*, this is necessary (but not sufficient) if we want to be able to say anything about causation. Information of this sort is far more difficult (some would say impossible) to obtain accurately in retrospective studies.

Doing the study retrospectively would involve identifying all women who were pregnant and worked with VDTs at least 9 months ago, and then either interviewing them or reviewing their hospital charts to determine the outcome of the pregnancy. This is advantageous because the study could be done relatively quickly, but it suffers from a few risks: the type of terminals may have changed, it may be difficult to establish how much time the women spent in front of the VDTs, and hospital documentation of all possible birth defects may be difficult to acquire (e.g., miscarriages may not have been recorded in hospital records.) A prospective study would enter women into the trial only if they became pregnant after the start date. Although the researcher could now record all the relevant information with greater accuracy, the study might have to continue for a few years until enough women became pregnant to allow analysis of the results.

The term "prospective" should *not* be used to describe trials in which historical data are gathered after a diagnosis or exposure that occurred some time in the past. For example, if we gather hospital utilization data from 1945 to the present on people who witnessed the A-bomb tests in Nevada, the data would still be retrospective, although the hospitalizations occurred after the exposure. Even though the subjects were followed forward in time, the data involve events that happened before *now*, and so the study would be called retrospective (Fig. 2-2).

A few authors have tried to clarify this confusion in nomenclature by introducing terms such as "retrolective," "prolective," or "retrospective-prospective." Laudable as this goal is, we feel that these neologisms have only further obfuscated the sufficiently murky picture.

X = When subject is enrolled in study

→ = Direction of data gathering

Figure 2-2 Prospective versus retrospective studies.

COMPARISON GROUPS

Keeping with our study of women who worked with video display terminals, we could easily derive prevalence figures for each of the outcomes of interest (stillbirths, miscarriages, and congenital abnormalities), but the meaning of these numbers would be unclear. The major reason is that women who do *not* work in front of VDTs also experience these adverse effects.

So, now the question has become somewhat more complicated: Do women who work with VDTs have these outcomes *at a higher rate* than women who do not work in front of terminals? This means that we now need a group against which we can compare our prevalence results to determine if they are higher or not.

There are two major types of comparison or control groups: **historical** and **concurrent**. In the former case we would compare our results with data that already exist from previous studies (e.g., a large survey of the prevalence of miscarriages, stillbirths, and congenital abnormalities in the general population). If such data do not exist or if they are suspect for one reason or another, the researchers must gather information from a control group concurrently; in essence, the researchers have at least two groups in the study.

When good historical control groups exist, they can save a considerable amount of time, effort, and expense. Unfortunately most historical control groups are compromised for some reason. The primary reason is that factors in the environment, such as clinical policies, may have changed since the data were originally gathered. For example, not too long ago very few infants under 2,500 g survived, whereas now it is not uncommon for neonatologists to save kids who weigh in under 1 kg. So, if infant mortality were one of our endpoints, it may appear that women who work with VDTs have a *lower* infant mortality rate than the historical controls. Conversely, it may be expected that infants who are born weighing 800 g or less may have more abnormalities than kids who were born weighing 2,500 g or more. So, the overall prevalence of birth defects may be increasing over time. The result is that this outcome may look poor when compared with a historical control, irrespective of any effect VDTs may have. The lesson is that when a historical control is used, we have to be quite certain that nothing has changed in the interim that could affect its comparability with the group we are looking at now.

On rare occasions a control group may not be necessary at all. To quote Bradford Hill, "If we survey the deaths of infants in the first month of life and find that so many are caused by dropping the baby on its head on the kitchen floor I am not myself convinced that we need controls to satisfy us that this is a bad habit." The classic case of a study where a control group was unnecessary was the use of streptomycin for tuberculous meningitis; without treatment the disease was universally fatal, so that any improvement in survival was significant. Fortunately or unfortunately, such examples are rare.

SAMPLING

Needless to say the most accurate information about the incidence of adverse outcomes in pregnancy caused by working with VDTs would be gained if we could gather data from all women who had ever worked in front of these terminals at some point during their pregnancy. Just as obviously, however, this would be impractical; there may be hundreds of thousands of such women scattered over most of the globe. Practical considerations dictate that we could follow up only a small proportion of these women, and if we select them appropriately, our estimates won't be too far off. (However, the famous prediction in 1936 that Alf Landon would decisively beat FDR must serve as a constant reminder that "appropriately" isn't all that easy to define — much to Roosevelt's relief.) In this section we discuss various ways in which we could go about choosing the group or groups we will include in our study.

BASIC TERMINOLOGY

Population All of the people to whom the results should be applicable constitute the populations. In this example the population would consist of all females who worked at VDTs at some time during their pregnancy (Fig. 2-3). (Note that "population" does not refer to all the people in the world; just to those who have a specific disorder, were exposed to some agent, or underwent some procedure.)

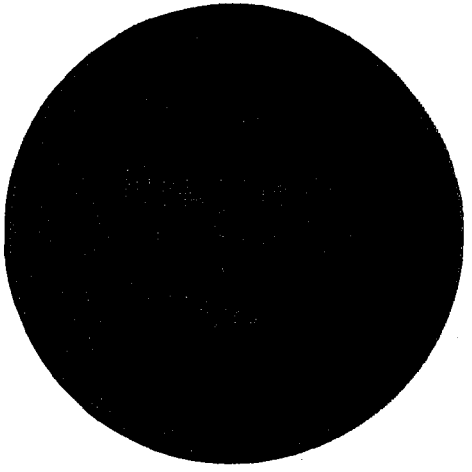


Figure 2-3 Population.

Sample In most cases the population is quite large, and it is impractical to study all people. We limit our study to a subset of the population; this smaller group is called the sample (Fig. 2-4).

Cohort Originally, cohort referred to a group of people born in the same year. Nowadays it has the broader, if less precise, meaning of a group of people who share some attribute. For instance, all people who began working at a specific job within a given time period can be referred to as a cohort, as can all people who entered the study at a certain time.

PROBABILITY SAMPLING

Probability sampling refers to a number of different strategies used to choose a sample. The term comes from the procedure used; every person in the population has a fixed and known probability of being selected to be part of the sample. For a number of reasons most studies try to use one or more of these strategies if at all possible.

The primary reason is that this method allows the investigator to generalize the results from the sample to the population, which is usually the major reason for doing a study. Second, it can tell the researcher the margin of error that could be expected from these estimates, that is, how far off the estimates can be. We see this in the reporting of polls, which often have a line stating that the results are accurate to within plus or minus 4 percent. In a related vein most statistical tests are based on the assumption of some sort of random sampling. When probability sampling is not used, we shouldn't use these tests (although that has never stopped people from doing so), and the ability to generalize the results from the sample to the population is questionable. (This is in contrast to the view of one politician who trusted letters he received more than polls and complained that the latter were "only" random.)

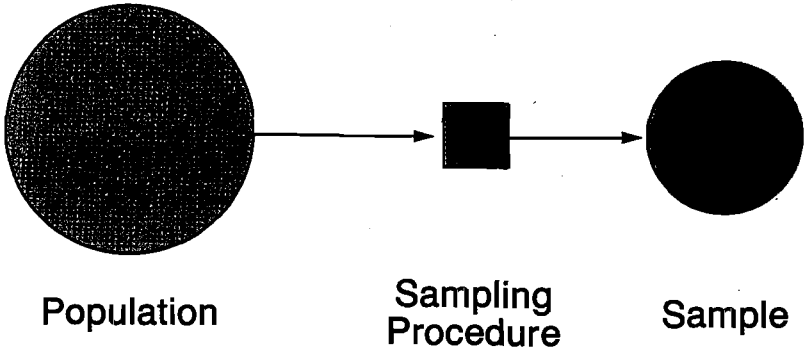


Figure 2-4 The sample is a subset of the population.

RANDOM SAMPLING

In **random sampling** (sometimes called "strictly random sampling" to differentiate it from the other varieties) each subject in the population has an equal chance of being chosen for the study. As we've mentioned, this approach maximizes the likelihood that the results of the study can be generalized to the entire population.

Random sampling is most often used in survey research (Fig. 2-5). Nearly all towns and cities have lists of taxpayers (for obvious reasons) or of street and house addresses. This makes it relatively simple for the researcher to select people, or at least dwellings, at random. These days about 98 percent of people have telephones, so it is also quite easy to draw a random sample from municipal or telephone lists, or from dialing digits at random.

Once we move out of the realm of surveys of the general population, however, it often becomes impossible to draw a pure random sample. We would have to know, for example, every company that used VDTs and all of the pregnant women at each business who had ever worked with VDTs in order to select people randomly for the study. More often we choose one or a number of businesses and hope that the use of VDTs within them is representative of companies in general, and that the women who work there are representative of female workers in other companies. We would then randomly select people within those companies for our study.

The same situation exists even for experimentally based studies. The hospital where a new treatment is tried out is not really chosen at random; it is most likely selected on the basis of convenience (e.g., the investigator works there or knows someone there who owes him a favor). The assumption is made that it is representative of hospitals in general, and that the randomly selected patients from that hospital are representative of the general population of patients with that condition. Unfortunately, these assumptions are not always correct and result in many of the various types of selection biases, which we discuss starting on p. 33 in *Threats to Validity*.

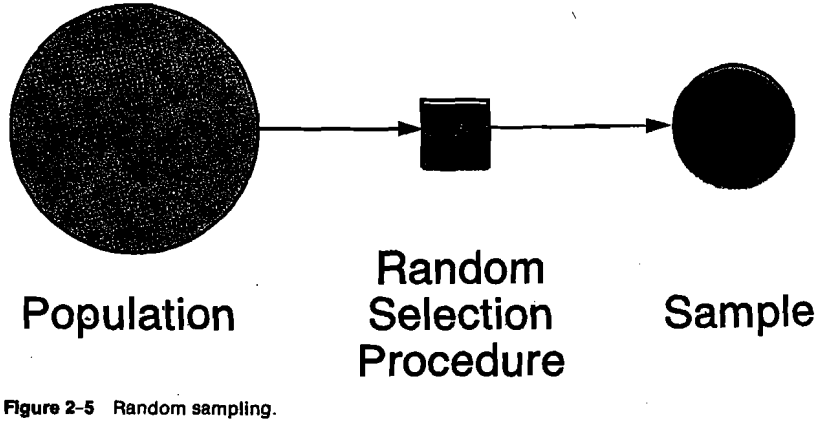


Figure 2-5 Random sampling.

STRATIFIED RANDOM SAMPLING

There are some circumstances in which we may wish to deviate from strictly random sampling. One major reason is that, with random sampling, we may end up with too few people in one subgroup or another. For instance, if we thought that the teratogenicity of VDTs was related to the number of previous pregnancies, random sampling might result in very few women who had three or more children before working on the terminals; the sample would be too small to allow us to analyze the effects of parity. Similarly, we may want to have equal numbers of women in each age category to maximize the power of our statistical tests.

Conversely, we may want to ensure that our sample is equivalent to the general population in terms of a few key variables, such as age at first pregnancy or number of children (it's obviously not necessary to match for sex). Random sampling ensures this matching in the long run with large enough samples, but not necessarily in our particular study, especially if there are fewer than 1,000 subjects. By chance, we could over- or under-sample people from a particular age or parity group.

To achieve these goals, we divide the key variables into various levels, or **strata**. For instance, we can divide age into 10-year increments, or parity into one kid, two kids, and three or more (Fig. 2-6). Then subjects are selected randomly from the stratum into which they fall. If toward the end of the study we have enough women who have had one or two children, but not three or more, we would restrict entry into the study to only this latter group. Since we know how our strata deviate from a strictly random sample, we can correct for this during the analyses when we derive the prevalence figures.

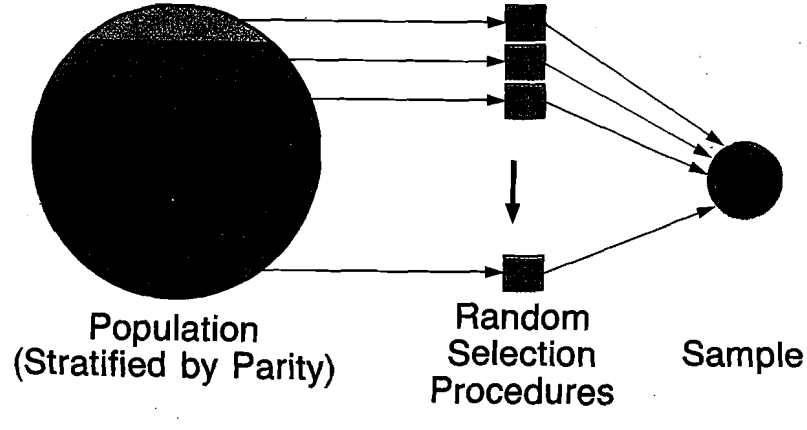


Figure 2-6 Stratified random samplings.

CLUSTER SAMPLING

In some designs it is impractical to assign individual subjects to the various groups. For example, in the Burlington Randomized Trial, nurse practitioners were placed in the offices of some family physicians to see whether they could reduce the cost of primary care without adversely affecting its quality. Outcome was measured at the level of the individual patient. However, since most families tend to use the same family doc, it would have been unfeasible to allocate randomly members of the same family to different practices. In this case each family was considered to be a **cluster**, and the unit of randomization was the family rather than the individual (Fig. 2-7).

However, the two, three, or more people in the same household cannot be considered to be independent of one another in terms of health status; they share the same diet, environment, and likely have similar attitudes toward exercise or other behaviors that affect health. Consequently, the husband's health is probably more correlated with his wife's than it is with that of another randomly chosen person.

Since the outcomes are correlated to some degree across people (who are usually considered to be independent in the usual statistical tests), studies that use cluster sampling usually need larger sample sizes than investigations in which the subjects are truly independent. How much larger the sample size has to be depends on the average number of people in the cluster, as well as on how strongly the variables are correlated within members of the cluster.

HAPHAZARD SAMPLING

In a **haphazard sample**, which is also called a "sample of convenience," subjects are selected on the basis of their availability, or in any other nonrandom way. For example, a researcher can interview people who pass

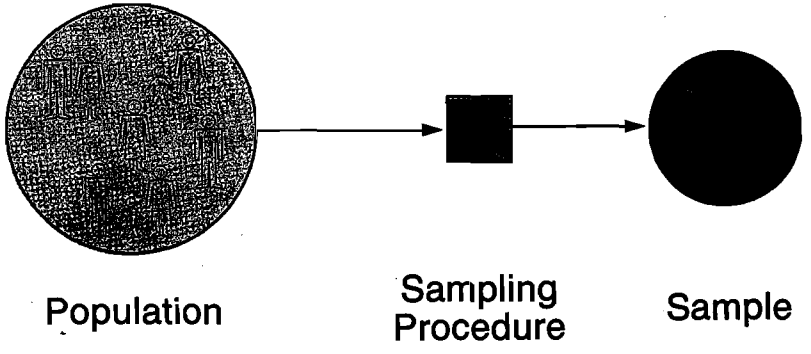


Figure 2-7 Cluster sampling.

a certain street corner or take blood samples from the research assistants who work in his or her laboratory. There is always the very real danger that this is a biased, nonrepresentative sample. During the day, housewives, shift workers, or the unemployed are more likely to be walking around outside than are people who work 9 to 5, and the location of the specific corner may differentially favor people from one social class over another. (On Wall Street in New York and Bay Street in Toronto you were more likely to find Yuppies in 1986, and the unemployed in 1987.) Similarly, those working in a lab may be healthier, brighter, or disproportionately female compared with the population of interest.

Unfortunately, newscasters rely on just this sort of haphazard, "person in the street" interview to find out (often erroneously) what the people "really" think about some issue. Politicians, who rely on letters they receive, fall prey to the same trap; those who are concerned enough to write are not representative of the electorate in general. Lest we as researchers develop undue pride about our avoidance of such egregious errors as are committed by those who are untrained in the strict disciplines of science, two examples may suffice to remind us of our fallibility. Mueller and his colleagues developed a test for plasma unesterified fatty acid to be used for patients with neoplastic disease. Their 30 normal subjects were "members of the professional staff . . . or hospitalized normal volunteers." The sampling for this test may have been a marked improvement over another test, which studied hemolysate prothrombin consumption time; the authors gave no indication at all regarding how many normal blood samples were used, much less where they came from. To assume that these samples were randomly selected, and hence representative of normal people, requires a leap of faith that we, at least, cannot make.

SUBJECT ALLOCATION

As we have noted, in experimental studies, whether the person receives a treatment or some other intervention is under the control of the researcher. Just as subjects can be *selected* for the study in various ways, they can be assigned or **allocated** to the different groups in a number of ways.

Sometimes these two steps are combined; as subjects are selected from the population, they are assigned to groups. In other instances the two steps are explicitly differentiated; a sample is derived, and then a separate procedure is used to allocate the subjects to the various groups. However, it is important to be aware of these two steps because, many times, the first step (subject selection) is only implicit in the study. For example, while patients in a hospital can be randomly allocated to receive conventional therapy or a new treatment, there is actually an initial stage that may not have been acknowledged, namely, the selection of the hospitals where the study was carried out. In many instances this initial selection procedure was not random.

Unfortunately, the similarity of terms used to describe subject selection and allocation can lead to considerable confusion for the uninitiated or unwary reader, and offers an area of potential mischief for unscrupulous researchers (a group that fortunately does not include epidemiologists — often). In the above example the sample was randomly assigned to the treatment groups, but it was selected haphazardly. Describing the procedure as randomized, without adequately delineating the somewhat suspect origins of the sample, can be misleading.

RANDOMIZED ALLOCATION

With random allocation, all subjects in the sample have the same probability of being assigned to the experimental or to the control groups. (This is not the same as a specific subject having an equal probability of being assigned to the groups; for design reasons, one group may be deliberately larger than the other, so the probability of ending up in that group is higher. However, the probability would be the same for all subjects.) This ensures that in the long run (i.e., with a large number of subjects) any underlying factors that may affect the outcome are equivalent for each group.

The subjects are allocated to groups by a **randomization device** or **scheme**. If there are only two groups that are equal in size, this can be accomplished by a simple coin toss: if heads, then the first group, or if tails, the other group. However, it is more common to use a table of random numbers, which can be found in most introductory statistics books. These tables consist of many numbers, often listed in groups of five for the sake of readability, which are generated in a completely random fashion. An example of a small portion of a table of random numbers would look something like this:

92778	07201	92632	93521	18235
83855	98335	11980	90040	22843
85527	62908	55960	80310	46765
34606	20883	66096	23610	00765
37375	68228	49966	20361	57424
81839	59252	91022	94233	93928
67018	85005	03174	89887	94262

To assign subjects to two groups, the table is entered at random; if the first number is odd, for example, the subject is allocated to Group A, and if it is even, to Group B. The second subject is assigned in the same way on the basis of the next number in the table; "next" can mean moving your finger right, left, up, or down. When there are three groups, the subject is assigned to the first group if the number is 1, 2, or 3; to the second group if the number is 4, 5, or 6; and to the last group if the number is between 7 and 9. If a zero is encountered, it is simply ignored and the next nonzero number is used. Groups of unequal sizes can be created in the same way. If Group A is to be twice the size of B, then numbers 1 through 6 can be used to allot subjects to Group A, and 7 to 9 to Group B.

Now that you've mastered the arcane art of using tables of random numbers, the good news is that you probably won't need to do it, since most computers can produce random numbers quite easily. There are a number of programs that capitalize on this and produce lists of random assignments according to your specifications — equal numbers in all groups, one group twice the size of the others, and so on. However, they're based on the same principles as those of the random number table, so your mental effort was not in vain.

BLOCK RANDOMIZATION

Block randomization is a modification of random allocation, in which subjects are allocated in small blocks that usually consist of two to four times the number of groups (Fig. 2-8). If there are three groups, then the block size is often six, nine, or 12 subjects.

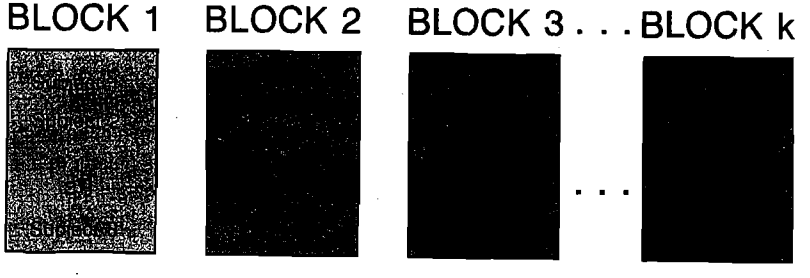


Figure 2-8 Allocation of subjects into blocks.

The subjects in the first block are randomly assigned so that there are equal numbers in each group (or, if the groups are not to be equal, they are assigned in proportion to the size of each group). The subjects in the succeeding blocks are then randomized in turn, until the final sample size is achieved (Fig. 2-9).

Block randomization ensures that, even if the study ends prematurely, there will be nearly equal numbers in all groups. With simple randomization it is possible to have a "run" of subjects assigned to one group; if the study ends at this point, an imbalance could result that would tend to reduce the efficiency of most statistical tests.

STRATIFIED ALLOCATION

The aim of **stratified allocation** is slightly different from that of stratified selection. In the selection phase stratification is used to ensure that the sample has certain desired characteristics. These characteristics may demand that the sample (1) matches the population on certain key variables, (2) includes sufficient numbers of subjects in all strata to permit subanalyses, or (3) has a normal distribution. The purpose of stratified allocations is more simple; it ensures that the groups do not differ too significantly on the stratification variables.

Stratified allocation is done when it is believed that the stratification variables may affect the outcome. If the groups are not balanced, any difference in outcome may result from these "nuisance" variables rather than from our intervention. For instance, if response to treatment is related to the patient's age, we do *not* want the experimental and control groups to differ on this factor.

For logistic reasons it is often impractical to have more than two or three stratifying variables, unless the available population is very large in relation to the sample size. Variables for stratification are chosen on the basis of their potential to affect the outcome. For example, since we felt that response to treatment was related to age but not to sex, only the former variable should be considered as a stratifying variable. If both age and duration of illness affect the outcome, but only one can be used as a stratification variable because of sample size limitations, the one that is more strongly associated with the outcome would be the variable to choose.

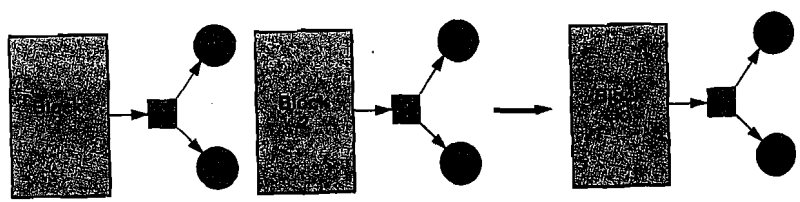


Figure 2-9 Block randomization.

MINIMIZATION

Minimization is a relatively recent and sophisticated method of assigning subjects to groups, and is used when there are many variables on which they should be matched. To keep matters simple, let's assume that we want to match the groups on only two variables: age and parity. The first few subjects are assigned to the groups by simple randomization. When a new person comes along, she is tentatively placed into each group in turn, and we compute what the mean age and parity level would be if she were in that group. The group to which she is ultimately assigned is based on minimizing the age and parity differences among the groups.

Because of the number of calculations required before a subject can be finally assigned (the number of variables multiplied by the number of groups), this method is unfeasible without a computer. Also, the criterion for "minimum differences between the groups" is somewhat arbitrary; how do we trade off an imbalance among the groups based on sex with differences based on age? For these reasons minimization is not yet widely used, although it may become more common in the future.

NONRANDOM (HAPHAZARD) ALLOCATION

Nonrandom allocation refers to situations in which subjects end up in the various groups by some manner other than having been randomly assigned. Let's assume that we wanted to compare the mean Apgar scores of kids whose mothers worked with VDTs against a group of kids whose mothers did not use display terminals. While we could *select* mothers at random from these two groups, the *allocation* would not have been random; they would have selected themselves to work or not work with the terminals.

The difficulty here is that there may be other factors on which these two groups of people differ. Some factors to be taken into consideration include the following:

1. Working women may be healthier than women in general (see the discussion on subject selection biases in *Threats to Validity*)
2. They may be working because they are poorer than other women (or become richer because they are working), and therefore provide a different prenatal environment
3. Even if we match for working status, those who have been chosen to be moved from typewriters to computers may be the brighter women.

In brief, the investigator has no control over factors that may, on the one hand, determine group membership and, on the other hand, affect the outcome.

The problem is even more acute in therapy trials. Clinical factors, which are also related to outcome, may have dictated whether the patient received medical or surgical treatment for his or her condition, or was given one drug

rather than another. So, simply comparing the success rates of these haphazardly selected groups may lead to erroneous results, because we conclude that the difference between the groups was caused by the intervention rather than by the factors that originally placed the subject in one group rather than in the other.

MATCHING

The term **matching** can have two meanings: one applies at the level of the individual subject and the other describes the general strategy for selecting a control group.

Matching at the individual level means that a pair of experimental and control subjects are chosen to be as similar as possible in terms of certain key variables, such as age, sex, race, socioeconomic status, number of hospital admissions, or diagnosis. A person from the smaller subject pool is often chosen first (e.g., if there are fewer "exposed" than "nonexposed" people in a case-control design, the pool of potential experimental subjects is smaller than that of the controls). Then a subject from the other pool is selected and matched as closely as possible on the key characteristics. The larger the ratio of potential subjects to the desired number to be chosen, the more matching variables can be used. If there are not too many people to choose from, the number of matching variables must be reduced or the criteria for similarity are relaxed (e.g., matching for age within plus or minus 10 years rather than within 5). Matching results in the two groups being as similar as possible on these key variables.

At the level of the group, matching refers to selecting a control group that has certain characteristics as an aggregate. For example, subjects in this control group can be (1) patients at the same hospital, but with a different diagnosis; (2) drawn from the same community; or (3) working at similar jobs. Control subjects, however, are not matched to experimental subjects on a one-to-one basis.

The purpose of matching on certain variables is to eliminate the effect of those variables on group differences. If the two groups are matched on age, for example, any difference in outcome between the groups cannot result from this factor. The downside is that matching prevents us from examining at some later point the effect of age on the outcome. The moral is to match only when you're certain that you aren't ruling out examination of an association in which you may later be interested.

Groups are **undermatched** if they differ on some variable that is related to the outcome. The effect of undermatching is that group differences at the end may be caused by the variables that aren't matched. So, there is a fine line between overmatching, and thus being unable to explore potentially interesting relationships, and undermatching, which may cause your results to be explained by some extraneous variable.

THREATS TO VALIDITY

The purpose of any study is to tell us what is "really" happening in the world: Does streptokinase reduce cardiac mortality? What causes sudden infant death syndrome? Did the swine flu vaccination program do more good than harm? We hope that the results from our sample can be generalized to the population at large, so that our findings also hold true for similar people. Consequently it is disconcerting, at the least, to find different studies coming to opposite conclusions.

The major reason for these differences is that all studies have flaws involving (1) the definition of the disorder or phenomenon of interest, (2) the selection of the subjects, or (3) the design or execution of the study itself. Cook and Campbell call these flaws **threats to validity**. In this discussion we examine some of the more common ones, and see how they can affect the interpretation of the results. In *Measurement* we discuss those forms of bias that affect eliciting and recording information.

SUBJECT SELECTION BIASES

Subject selection biases involve a host of factors that may result in the subjects in the sample being unrepresentative of the population. We've already discussed one class of selection bias—nonrandom sampling. However, even with the best of sampling strategies, nature (human and otherwise) conspires against us in many ways. Sackett compiled a list of various biases, 57 at last count, and even this is probably incomplete. To keep life simple, we can think of two major types of subject selection biases: who gets *invited* to participate in a study, and who *accepts*. We cannot even attempt to provide a complete catalog of these two classes of factors; rather, the following three examples of invitational bias (healthy worker, incidence-prevalence, and Berkson's) and one of acceptance bias (volunteer) are illustrative only. We hope these examples help to enlighten and warn the reader of where things can go wrong.

HEALTHY WORKER BIAS

Randon sampling does not help us if the group from which the sample is drawn is unrepresentative of the population to which we want to generalize. For example, comparing the outcome of pregnancies of women who work with VDTs with those of a group of women chosen at random may open the researcher up to the **healthy worker** bias; that is, people who work are, as a group, healthier than the population as a whole. The entire adult population consists of those people who are working, those who are able to work but do not for one reason or another, and those who cannot work because of health problems. Any group of workers, by definition, does not include this last category of people that tends to lower the overall health status of the population. This selection bias operates even more strongly when the job

applicants have to pass a physical examination, as in the Armed Forces or for certain labor-intensive occupations. Seltzer and Jablon, for example, found lower morbidity rates among people discharged from the Army than among people of similar ages in the general population. This effect was seen even 23 years after the men had been discharged. (Some have hypothesized that this is caused by Army food killing off the less fit before they can be discharged.)

The effects of this bias are to (1) make any sample drawn from a group of workers appear healthier than the general population; (2) make the standardized mortality rate (see *Measurement*) less than 1:1 when workers are compared with the general population; and (3) make the proportional mortality rate (see *Measurement*) for occupational hazards greater than 1.0 because of "borrowing" (i.e., if they are dying less from heart disease, they must be dying more from something else).

INCIDENCE-PREVALENCE (NEYMAN) BIAS

If a group is investigated a significant amount of time after the people have been exposed to a putative cause or after the disorder has developed, those who have died and those who have recovered will be missed. This is known as the **incidence-prevalence** bias or the **Neyman** bias. For example, a cross-sectional look at depressed patients in hospital misses those in whom the depression culminated in suicide or resolved itself. Similarly, a study of cardiac patients in a tertiary care hospital does not include (1) those who died before reaching hospital and (2) those whose myocardial infarction was not sufficiently severe to warrant transfer to a specialized facility.

As another example, even the latest version of the *Diagnostic and Statistical Manual of Mental Disorders* (1987) is somewhat pessimistic regarding the long-term prognosis in schizophrenia. However, this pessimism may be unwarranted, and may be based on the fact that most "natural history" studies use patients who are in hospital at a given time. Follow-up studies with patients who have been admitted for the first time, which are much less susceptible to the Neyman bias than cross-sectional ones, give a very different picture; according to these follow-up studies, the majority of patients — anywhere between 60 and 80 percent — go on to lead productive lives outside the hospital.

The effects of the Neyman bias can be in two different directions. Missing those who died before they could be included in the study makes the disorder look less severe, since the outcome is generally more positive than had all patients been included. Conversely, missing those who have already gotten better makes the outcome look grimmer. The net effect is often unknowable, and depends on the relative proportions of patients in the three groups (i.e., studied, died, and improved).

BERKSON'S BIAS

Berkson's bias is the spurious association found between some characteristic and a disease, and it results from admission rates to hospital (or any other setting where the study is carried out) being different for those persons (1) with the disease, (2) without the disease, and (3) with the characteristic. For example (Table 2-1), assume that in the general population there is no relationship at all between vaginal bleeding (the characteristic) and endometrial cancer (the disease).

Let us further assume that 10 percent of patients with endometrial cancer have vaginal bleeding and 10 percent of patients with other cancers have bleeding. If the probability of being admitted to hospital because of vaginal bleeding is 70 percent, if it's 10 percent because of endometrial cancer, and if it's 50 percent because of other forms of cancer, then:

1. Of the 100 patients with vaginal bleeding and endometrial cancer (cell A), 10 will be admitted because of endometrial cancer (i.e., 10 percent). Of the remaining 90 patients in cell A, 63 (70 percent) will be admitted because of vaginal bleeding, so that a total of 73 women will be admitted with endometrial cancer and bleeding.
2. Of the 100 patients with vaginal bleeding and other forms of cancer (cell B), 50 will be admitted because of the other cancers. Of the remaining 50, 35 (again, 70 percent) will be admitted because of vaginal bleeding, so that in total 85 will be admitted with bleeding and other cancers.
3. Of the 900 patients with endometrial cancer and no bleeding (cell C), 90 (again 10 percent) will be admitted because of endometrial cancer.

TABLE 2-1 Association Between Endometrial Cancer and Vaginal Bleeding

Vaginal Bleeding	Type of Cancer: Endometr.	Type of Cancer: Other	
Yes	100 (A)	100 (B)	200
No	900 (C)	900 (D)	1,800
	1,000	1,000	2,000

4. Of the 900 patients with other forms of cancer and no bleeding (cell D), 450 will be admitted because of the other cancers.

Table 2-2 shows the graphic results of these different admission rates. Now it appears that 44.8 percent of patients with endometrial cancer have vaginal bleeding, whereas only 15.9 percent of patients with other forms of cancer have vaginal bleeding. This apparent (and false) association is the result of different hospitalization rates for endometrial and other cancers and for vaginal bleeding. Thus, Berkson's bias comes into play whenever we sample from a setting in which there are different rates of admission for different disorders.

VOLUNTEER BIAS

To be ethical, most studies allow patients to refuse to participate. Thus the results are predicated to some degree on the assumption that those who do not volunteer are similar to those who do. However, there is now ample evidence to show that this is not the case and that volunteers differ systematically from nonvolunteers.

For example, the National Diet-Heart Study found that, compared with nonvolunteers, volunteers more frequently (1) were nonsmokers, (2) were concerned about health matters, (3) had a higher level of education, (4) were employed in professional and skilled jobs, (5) were Protestant or Jewish, (6) were living in households with children, and (7) were active in community affairs.

In one arm of the Coronary Drug Project the 5-year mortality rate for compliers (those who took 80 percent or more of their medication) was

TABLE 2-2 Results Caused by Different Hospitalization Rates for Characteristic (Bleeding) and Disease (Cancer)

		Type of Cancer		
		Endometr.	Other	
Vaginal Bleeding	Yes	73 (A)	85 (B)	158
	No	90 (C)	450 (D)	540
		163	535	698

15.1 percent. It was almost twice as high among noncompliers (28.2 percent), even though the "medication" they were complying with was a placebo! Although all subjects were volunteers, those who complied with the treatment regimen were apparently a different breed from those who did not comply.

Even for those who participate in a trial, a type of volunteer bias may operate. The incidence of inactive tuberculosis was lower among volunteers who appeared early during a mass screening than among those who appeared later, whereas the opposite trend was noted for pneumoconiosis.

HAWTHORNE EFFECT

According to legend, worker productivity improved at the Hawthorne plant of the Western Electric Company not only when the illumination was increased, but also later when it was decreased. The reason for this was supposed to be the attention paid to the workers by the researchers and not the lighting itself. Although later studies showed that the increase in productivity likely resulted from other factors, the term **Hawthorne effect** has remained to explain the phenomenon that occurs when a subject's performance changes simply because he or she is being studied (some have referred to this as the psychological equivalent of the Heisenberg Uncertainty Principle).

For example, Frank reported that the introduction of a research project onto a hospital ward was "followed by considerable behavioral improvement in the patients," even though no medication or special treatments were involved. He felt that the most likely explanation was that "participation in the project raised the general level of interest of the treatment staff, and the patients responded favorably to this."

To counteract the Hawthorne effect it is often necessary to use an *attention control* group, which is treated exactly the same as the experimental group except for the active treatment. For example, studies of psychotherapy often employ a control group that meets with the therapist as frequently and for the same duration as does the treatment group, but the content of the session is not supposed to be therapeutic. In drug trials the control group receives a placebo, which usually involves taking the same number of pills at the same time of day as the experimental subjects.

BLINDING

One effect of the attention control group we just discussed is to **blind** the subject and perhaps the experimenter. A person is considered blind if he or she is unaware of the group to which a subject belongs. If only the subject is unaware but the experimenter knows, the study is called single blind. If both the subject and the researcher do not know, the study is labeled double blind. (Some people have proposed the term triple blind for the occasions when the subject and evaluator are blind, and the pharmacist has lost the key that tells who got the drug and who got the placebo. However, this is more a threat to the pharmacist's life than to validity.)

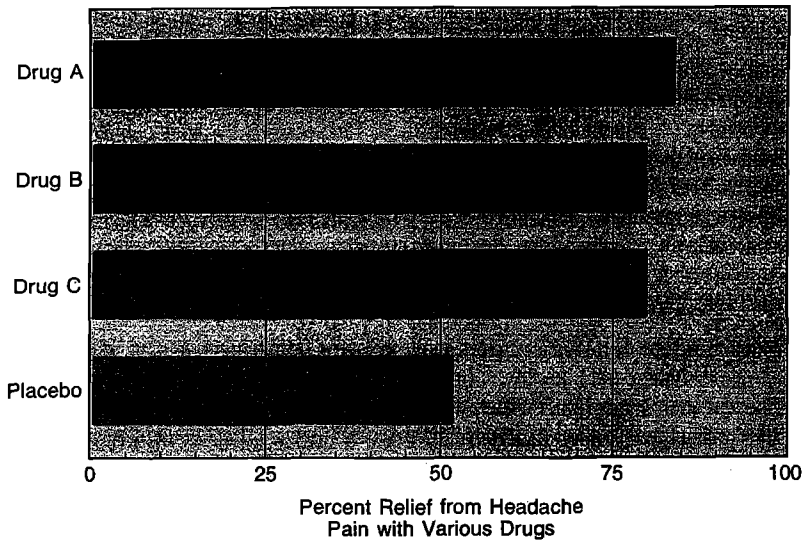


Figure 2-10 Results of this study show the placebo effect. In this case more than 50 percent of subjects on placebo experienced relief of headache pain.

The purpose of blinding is to prevent various biases from affecting the results. Subjects may show a *placebo effect* if they know they are receiving an active agent, or may not show it if they think they are not receiving the new drug. With single blinding, both groups should show an equivalent reaction. The magnitude of the placebo effect should not be underestimated (indeed, it's what kept medicine alive for a few millennia). The results of one typical study, shown in Figure 2-10, indicate that more than 50 percent of patients experienced relief of headache pain from placebos.

If the clinicians (or evaluators) were aware of group membership, they could be more alert or attentive to signs of improvement. Likewise, clinicians who know that a disease should be present may be more diligent when looking for it (*diagnostic suspicion bias*). Rosenthal conducted a series of studies that showed that what a researcher expects to find in an experiment affects what does occur, irrespective of whether the subjects are humans or rats.

CONFOUNDING

Confounding is the illusory association between two variables when in fact no such association exists. It is caused by a third variable (the "confounder"), which is correlated with the first two. For example, Table 2-3 shows bifocal use (needed or not) and nocturnal enuresis (present or absent) in a group of 200 patients.

TABLE 2-3 Relationship Between the Need for Bifocals and Nocturnal Enuresis

Bifocals Needed	Nocturnal Enuresis: Present	Nocturnal Enuresis: Absent	
Yes	17	83	100
No	8	92	100
	25	175	200

The odds ratio is 1.93, which indicates that persons who need bifocals are twice as likely to have enuresis as those who don't need bifocals. (This may be related to the supposed link between masturbation and blindness.)

However, a closer look at these data shows that there are actually two age groups involved (Table 2-4). For each age group there is no association between bifocal use and enuresis. In those under age 60, 5 percent of bifocal users are enuretic (1 of 20 subjects), as are 5 percent of nonusers (4 of 80 subjects). For those over age 60, 20 percent are enuretic, irrespective of bifocal use. The confounder here is age; bifocal users are more apt to be over age 60, which is also the group that has the higher rate of enuresis. (Fig. 2-11)

TABLE 2-4 No Association Between Bifocal Need and Nocturnal Enuresis When Subjects are Divided by Age

Bifocals Needed	Under 60: Present	Under 60: Absent		Over 60: Present	Over 60: Absent	
Yes	1	19	20	16	64	80
No	4	76	80	4	16	20
	5	95	100	20	80	100

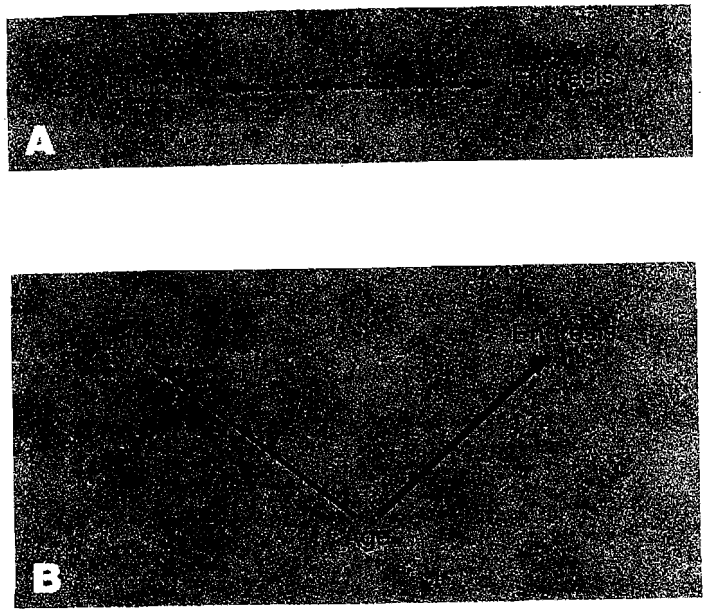


Figure 2-11 *A*, When unaware of the confounder, it appears that there is a direct association between enuresis and bifocals. *B*, There is a direct association between age (confounder) and bifocals and between age and enuresis.

CONTAMINATION

In studies in which one group receives the experimental treatment and another group gets either conventional treatment or a placebo, the validity of the results is predicated on the *purity* of the groups. If some subjects in the control group receive the new treatment, both groups will improve to some degree (assuming that the treatment works). Thus, differences between the groups are diminished or even eliminated. This condition is referred to as **contamination.**

Contamination is a particular problem when a medication used in a study is also available over the counter or as an ingredient in other compounds (e.g., aspirin), or when it can be prescribed by family physicians who are unaware (or have forgotten) that certain drugs should not be given to some of their patients. However, contamination is not limited to drug trials; it can occur with any form of intervention, such as respite care for those taking care of demented elderly, psychotherapy, and similar maneuvers in which subjects in the control group receive some form of the treatment.

In cohort and case-control studies contamination is caused by misclassification, that is, assigning exposed subjects to the nonexposed group or vice versa. This is often caused by errors in recall by the subjects.

The effect of contamination is to reduce differences between the treated and untreated groups. This may lead us to draw the erroneous conclusion that the intervention is of limited or no use.

COINTERVENTION

Cointervention refers to subjects in a study receiving therapies other than those given as part of the experiment that affect the outcome of interest. For example, some subjects in a study that compares the effectiveness of various nonsteroidal anti-inflammatory drugs for arthritis could be given other drugs by another physician, be enrolled in a program using transcutaneous stimulation, or might be taking over-the-counter aspirin.

Cointervention differs from contamination in two ways: (1) the intervention and (2) the groups that are affected. First, contamination refers to the control group receiving the experimental intervention, whereas cointervention refers to some treatment other than the one under investigation. Second, all groups in a study can be witting or unwitting recipients of a certain cointervention, but only the control group can be contaminated.

Although all groups can be subject to cointervention, it is a particular danger when the control subjects do not improve or even deteriorate on placebo. If any other clinician is involved in the case and unaware of the study, he or she may prescribe other treatments to help the person, thereby minimizing differences between the groups. If subjects in all groups receive other therapies, then it becomes almost impossible to determine if the results are caused by the treatment under study, by the cointervention, or by both.

REGRESSION TOWARD THE MEAN

Regression toward the mean refers to the phenomenon whereby groups of subjects that are chosen because of their extreme score on any variable will have scores that are less extreme and closer to the mean value when they are retested. The reason is that any test result we observe—some serum value, a decision based on an x-ray, or a score on a paper and pencil test—is comprised of two parts: the *true* score and an *error* score. Written out in the form of an equation, we say:

$$\text{Observed Score} = \text{True Score} \pm \text{Error Component}$$

There are many sources of error (see the discussion in *Measurement with Continuous Variables* on reliability), including variations in the machine, biologic variation within the subject, motivation, fatigue, and recording error. The assumption is that this error component is random, sometimes adding to the true score and sometimes diminishing it. We can never see the true score, only the observed score.

When we select a group because of its extreme scores (either very high or very low), we are including two types of persons: (1) those whose true scores are extreme and (2) those whose true scores do *not* fall in the extreme range, but the error component added to the true score has placed them in

the extreme region. Similarly, we have excluded persons whose true scores are extreme, but whose observed scores are below the cut-off level. For example, let's assume that we're using a test with a mean of 50, and a score of 70 or above identifies the most extreme 2 percent of the sample, which is the group we want to include in our study. We've shown the true score plus or minus the error component for the 10 subjects whose observed scores are over 70 and for a few of the other subjects (Fig. 2-12). Thus we have biased our sample to include an overrepresentation of people who have error scores in the direction *away from* the mean. Since the error component is random, when these people are retested only half of them will have error scores away from the mean (keeping them in the extreme range), and half will have error scores that move the observed score closer to the mean. On the whole, the group average on the second testing will be closer to the mean than on the first testing.

In practical terms this means that if we select a group of subjects because they appear abnormal on some test (that is, their score differs from the mean) and do *nothing* to them, they will seem to improve (move closer to the mean) when they are retested. So, if we had intervened, it would be impossible to know if the improvement was caused by us or simply by regression effects.

The magnitude of this effect is inversely related to the reliability of the test; the less reliable the test is, the greater the regression effect. The reason is that reliability expresses the relative contributions of the true score and the error scores, so that an unreliable test has a large error component (see the discussion on reliability in *Measurement with Continuous Variables* for more detail).

Regression toward the mean can be minimized in two ways: (1) by increasing the reliability of the test, and (2) by testing each subject at least twice and requiring all the tests to be extreme before he or she is included in the study. This is often done in hypertension trials in which the person has to have three consecutive abnormal readings before being called hypertensive.

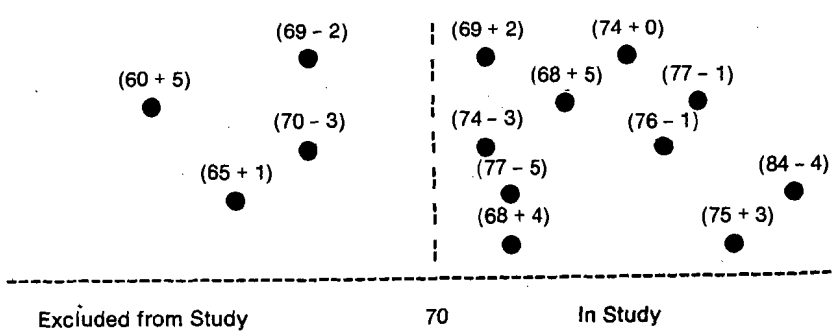


Figure 2-12 True score ± error component for 10 subjects with observed scores over 70, and four subjects with observed scores under 70.

COHORT EFFECTS

Nowadays a **cohort** refers to a group of people chosen because they share some common characteristic (e.g., employment in a specific job, or exposure to a given agent). Previously, however, cohort was used in a narrower sense and meant a group of similar age, the members of which having in common only their year of birth. Cohorts of this type have been extremely useful in elucidating many epidemiologic findings, such as increases in longevity and height over time. A danger arises when one attempts to attribute a *causal* factor to differences among age cohorts, since one cohort differs from another on many variables other than age.

For example, studies done in the 1940s and 1950s tended to show a decline in intelligence over the age of 50 by comparing various age cohorts on a standardized test. Subsequently longitudinal studies have demonstrated that, while performance on timed tasks does decrease, scores on other tests either remain stable or actually increase with age (we can all now breathe a sigh of relief). The problem with the original studies was that not only were the older subjects more advanced in years than the younger ones, they were also exposed to a very different educational and cultural environment, which accounted for most of the differences among the cohorts and hence for most of the apparent decline.

ECOLOGIC FALLACY

Ecologic studies attempt to demonstrate a relationship between two variables, such as suicide rate and religion, by using aggregate data. These are data about groups of people rather than individuals. For example, we can look at the rates of lung cancer per 100,000 individuals in a number of cities, and see if these are correlated with pollution levels.

While this technique is very inexpensive and has at times led to useful findings, there is one major problem — there is no guarantee that those people who developed lung cancer were the same ones who were exposed to the pollution. That is, it is possible (although unlikely) that pollution is unrelated to cancer of the lung, but that pollution is caused by large factories. We know that cigarette smoking is related to social class, and that factory workers smoke more heavily than the general population. So, it may be that pollution is simply a marker for heavy smoking, and it is the smoking that is producing cancer.

The ecologic fallacy was nicely demonstrated by Robinson, who showed that there was a strong relationship (r=0.62) between literacy rates and the proportion of non-native born people; that is, regions with the largest number of immigrants had the lowest rates of illiteracy. Since most immigrants had relatively little education, especially in the 1930s when the data were collected, this seems to fly in the face of common sense. However, the individual correlation between literacy and foreign birth was -0.12, which is lower in magnitude (correlations based on individuals are almost always lower than ecologic correlations) and in the reverse direction.

The explanation is that immigrants usually settle in large cities, which have high rates of literacy, rather than in rural areas where literacy rates are lower. Thus, *areas* with low rates of illiteracy have a high proportion of immigrants, but illiteracy and immigrant status are correlated (albeit weakly) within the *individual*.

EPIDEMIOLOGIC RESEARCH STRATEGIES

The hallmark of a scientific theory is that its hypotheses are capable of being disproved. This does not always require experiments under the control of the researcher; astronomers haven't yet figured out how to experimentally induce stars to form or evolve. However, when experimental studies can be done, they can provide powerful tests of hypotheses that are not feasible when we have to rely solely on observations of naturally occurring events.

Over the years many different study designs have been developed to deal with the multitude of research questions that have been asked. We cannot begin to describe all of these methods here; entire books have been written on just this one area. Rather we have chosen the six designs that are used most frequently. The first four (cross-sectional, ecologic, cohort, and case-control) are commonly referred to as **descriptive** or **analytic** designs. These are most appropriate when, for one reason or another, experimental control over the independent variable is not feasible. This would include, for instance, exposure to potentially harmful agents (e.g., cigarette smoke), situations in which there may be a long interval between exposure and outcome (such as diethylstilbestrol use and vaginal cancer in female offspring), or when our state of knowledge (or rather, ignorance) doesn't yet allow us to state whether there is an effect that's worth following up with a more expensive trial.

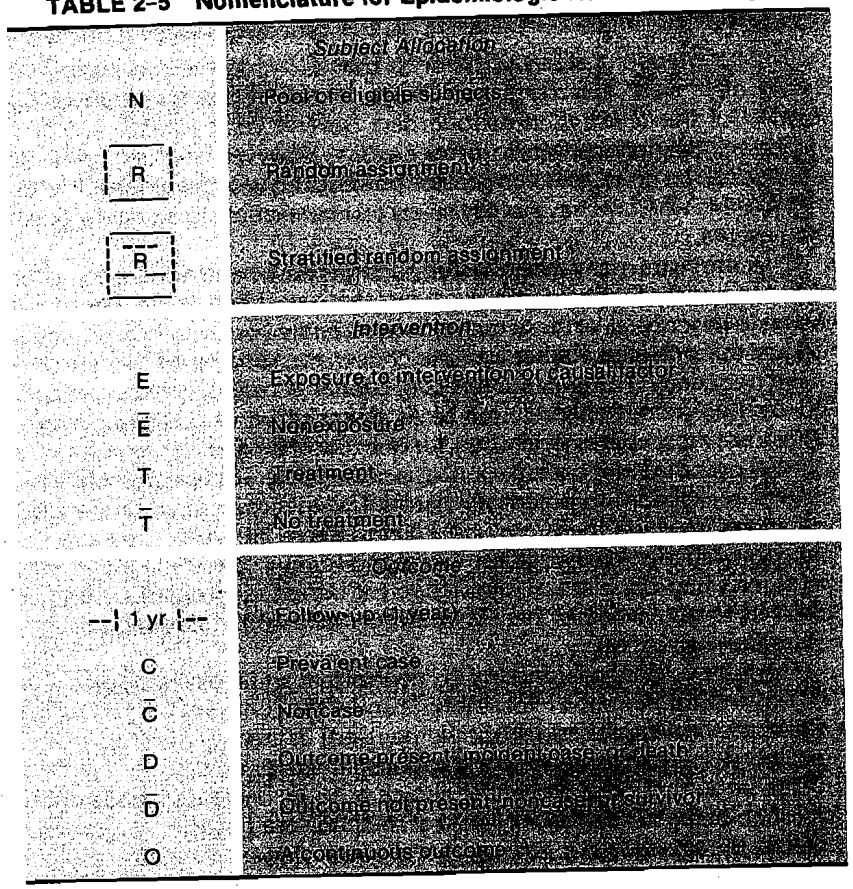
The last two designs (randomized control trial and cross-over) are called **experimental**, since the intervention is under the control of the researcher. These methods are used (or should be used) in therapy trials, since their results are least susceptible to the various threats to validity.

The important point is that the choice of study design depends on the question being asked. Usually several methods are possible, and we may look for the strongest (i.e., the one that allows the fewest alternative explanations for the results). However, we may instead opt for a "quick and dirty" design, even if it isn't the optimal one, simply to see if there is anything worth looking into at greater expense.

NOMENCLATURE

Table 2-5 is based on the nomenclature introduced by Kleinbaum, Kupper, and Morgenstern and modified by the Department of Clinical Epidemiology and Biostatistics at McMaster University.

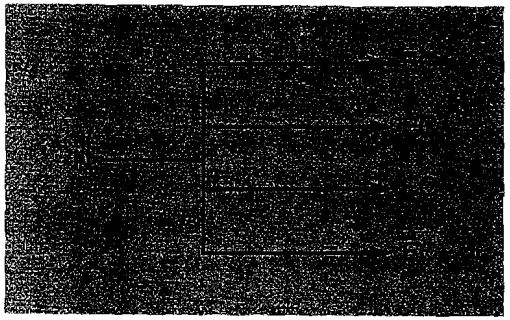
TABLE 2-5 Nomenclature for Epidemiologic Research Strategies



DESCRIPTIVE AND ANALYTIC STRATEGIES

CROSS-SECTIONAL SURVEY

Design



Example

A group of women (N) are interviewed to determine (1) their use of video display terminals (E) and (2) whether they had a miscarriage (C).

Major Features

Exposure and caseness are determined simultaneously.

Advantages

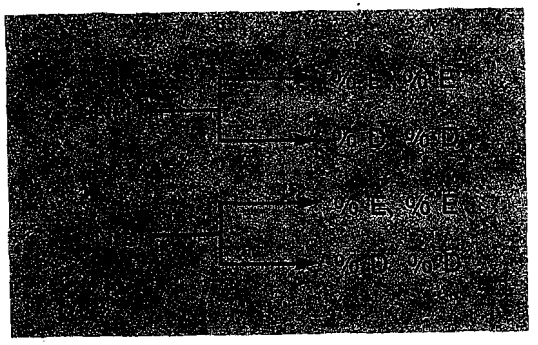
1. This design is relatively inexpensive and simple to carry out because no follow-up is required.
2. No one is exposed to the putative causal agent because of the study, or denied a potentially beneficial therapy.

Disadvantages

1. A cross-sectional design can establish association, but it is impossible to determine causation, since exposure and caseness are determined at the same time.
2. It is impossible to ensure that confounders are equally distributed among the groups.
3. Often either exposure or caseness or both depend upon recall, which is fallible.
4. This design is susceptible to the Neyman bias, that is, cases with early deaths and those in which evidence of exposure has disappeared are missed.
5. The groups could end up having very different sample sizes, resulting in a loss of statistical efficiency.

ECOLOGIC STUDY

Design



Example

Ecologic studies are used quite often in cancer research, in which the rates of cancer of different organs are examined by geographic area (county, province, or state). This has led to some fruitful hypotheses regarding the association between cancer of the esophagus and diet in Eastern Europe and China, for instance.

Major Features

The group, usually defined geographically, is the unit of analysis, and the data are most often already available.

Advantages

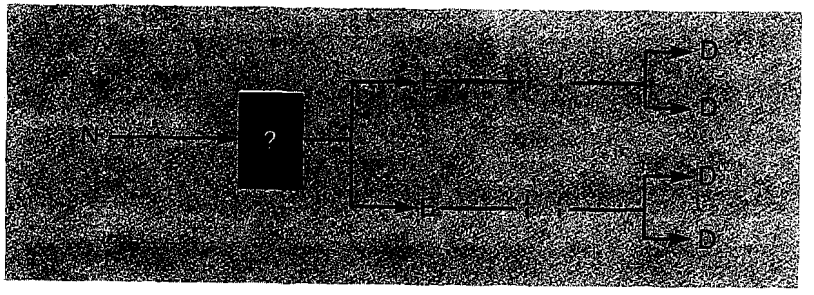
1. Data are usually available, so this type of study is quite inexpensive.

Disadvantages

1. We know how many people were exposed within each group and how many have the outcome, but *not* how many exposed people have the outcome. That is, it is quite possible that the outcome occurred in unexposed people and the variables are not related (see the discussion in *Threats to Validity* on ecologic fallacy).
2. Correlations from ecologic studies are usually much higher than in studies where both variables are gathered on the same individuals.

COHORT STUDY

Design



Example

A group of women who used VDTs during their pregnancy and a second group who did not use them are followed to determine the rates of miscarriages, stillbirths, and congenital abnormalities.

Major Features

Exposure to the putative causal agent or treatment is *not* under the researcher's control. Subjects are divided into exposed (or treated) and nonexposed (or untreated) groups on the basis of past history. The design can be prospective (following the groups forward in time from the present), or retrospective (choosing groups that were formed some time in the past, and then following them forward from that time to the present).

Advantages

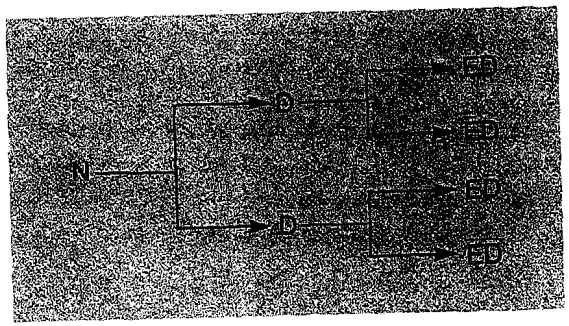
1. Treatment is not withheld from subjects and they are not artificially subjected to potential hazards.
2. Subjects can be matched for possible confounders.
3. When the design is prospective, eligibility criteria and outcome assessments can be standardized.
4. It is administratively easier and less costly than an RCT.
5. It can establish the timing and directionality of events.

Disadvantages

1. It may be difficult to obtain controls if therapy is popular or if most people have been exposed.
2. Exposure may be related to some other unknown factor that is correlated with the outcome (confounding).
3. Blindness among subjects and assessors may be difficult to achieve.
4. It is expensive to do well.
5. It may violate some statistical tests based on the assumption of randomization.
6. For rare disorders, large sample sizes or follow-up periods are necessary.

CASE-CONTROL STUDY

Design



Example

The mothers of children born with (D) and without ($\bar{D}$) birth defects are interviewed to determine whether or not they used video display terminals during their pregnancy.

Major Features

The groups are identified on the basis of the *outcome* (e.g., birth defects), and the search for exposure (to video display terminals) is retrospective.

Advantages

1. It can be done relatively quickly and inexpensively.
2. It may be the only feasible method for very rare disorders, or for situations in which there is a long lag between exposure and outcome.
3. It usually requires fewer subjects than cross-sectional studies.

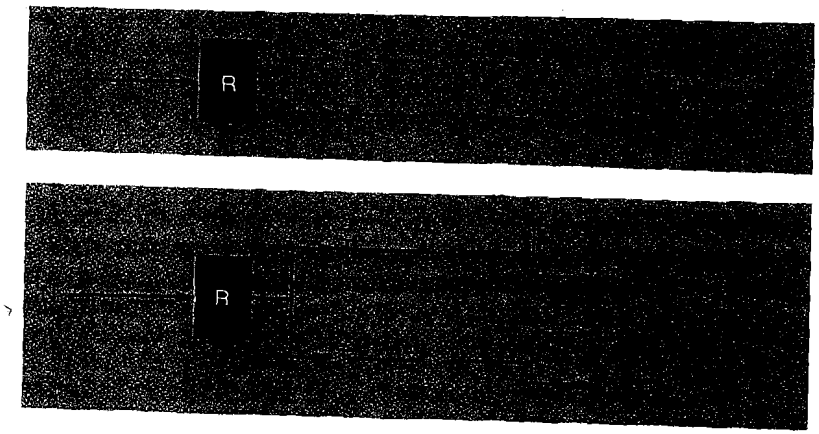
Disadvantages

1. It relies on recall or records to determine exposure, and both are notoriously inaccurate.
2. The groups may be confounded, that is, exposure may have been caused by some other factor that is correlated with the outcome (e.g., income, area of residence, age).
3. It may be difficult to select and then find an appropriate control group.
4. If the index group is aware of the hypothesis, there is the possibility of recall bias.

EXPERIMENTAL DESIGNS

RANDOMIZED CONTROLLED TRIAL

Design



Example

Hemiplegic stroke patients currently receiving physiotherapy are randomly assigned to receive or not receive transcutaneous stimulation. After 3 months, they are compared on walking speed (continuous outcome) and presence or absence of footdrop (discrete outcome).

Major Features

Subject allocation to treatments or exposure is under the control of the experimenter.

Advantages

1. Groups are likely more comparable because confounding variables are probably balanced.
2. There is a greater likelihood that patients, staff, and assessors can be blinded.
3. Most statistical tests rest on the assumption of random allocation.

Disadvantages

1. These trials are expensive in terms of time and money.
2. Those who volunteer may not be representative of all patients.
3. A potentially effective treatment is withheld from some subjects, or some may be exposed to a possibly dangerous one.

CROSS-OVER DESIGN

Design

Example

Patients are randomly allocated to receive carbamazepine (CBZ) to control their manic-depressive disorder or a placebo. After 4 weeks they are given a placebo until all the drug is out of their system. Then those who had been given CBZ are given placebo for 4 weeks, and those given placebo are given the active drug.

Major Features

Randomization is under the researcher's control; all patients receive both the active treatment and the placebo (or control treatment).

Advantages

1. Subjects serve as their own controls, thereby reducing error variance. Consequently, fewer subjects generally are needed than for RCTs.
2. All subjects receive the treatment at least for some period.
3. Statistical tests assuming randomization can be used.
4. Blindness of patients, staff, and assessors can be maintained.

Disadvantages

1. Subjects who responded to the treatment are taken off it and given placebo (or the alternative treatment).
2. The wash-out period with some drugs can be quite lengthy, during which time the patients are often given placebos.
3. It cannot be used if the treatment has any permanent effects (e.g., educational programs, physiotherapy, behavior therapy).

C.R.A.P. DETECTORS

C.R.A.P. DETECTOR II-1

Question In one of the seminal books on the etiology of homosexuality Bieber and his associates derived their sample by mailing three copies of a questionnaire to fellow members of a New York-based psychoanalytic society. The analysts filled them out for any homosexual patients they had in therapy. If the psychiatrist had fewer than three such patients in treatment, he or she was to fill out the remaining questionnaires on male heterosexual patients; the heterosexual subjects constituted the control group. What are the problems with this sampling strategy?

Answer Unfortunately a listing of all the problems would fill a book thicker than this one. First, persons who elect to go into psychoanalysis are not representative of the general population. Obviously, those who are happy with their lives never spend time on the analytic couch. Second, those who are unhappy but poor must settle for less comfortable and less expensive chairs, or get no help at all. Finally, leaving the choice of which patients to include up to the individual analysts opens the door to a host of biases; it is doubtful whether the sample would include patients who didn't improve or who didn't match the psychoanalytic sample.

C.R.A.P. DETECTOR II-2

Question Those who disapprove of social assistance programs state that welfare fosters dependence, and encourages people to behave in ways that enable them to remain on assistance for a long time. The opponents buttress their arguments with surveys showing that, at any one time, the majority of welfare recipients have been on it for extended periods. How much can we trust these data?

Answer This is a nice example of the incidence-prevalence bias. Figure 2-13A shows the proportion of women who have ever received Aid for Families with Dependent Children (AFDC), and how long they were on it. Of these women, 30 percent were on AFDC for only 1 or 2 years, and 70 percent received it for less than 8 years. However, if the investigators had done a cross-sectional survey that asked women currently on AFDC how long they had been on it, a very different picture would emerge. Now, as Figure 2-13B shows, the vast majority (65 percent) have been getting benefits for more than 7 years. The problem is that long-term recipients are more likely to be picked up in a one-time survey than short-term recipients who had been on AFDC in the past, but were not at the time of the survey.

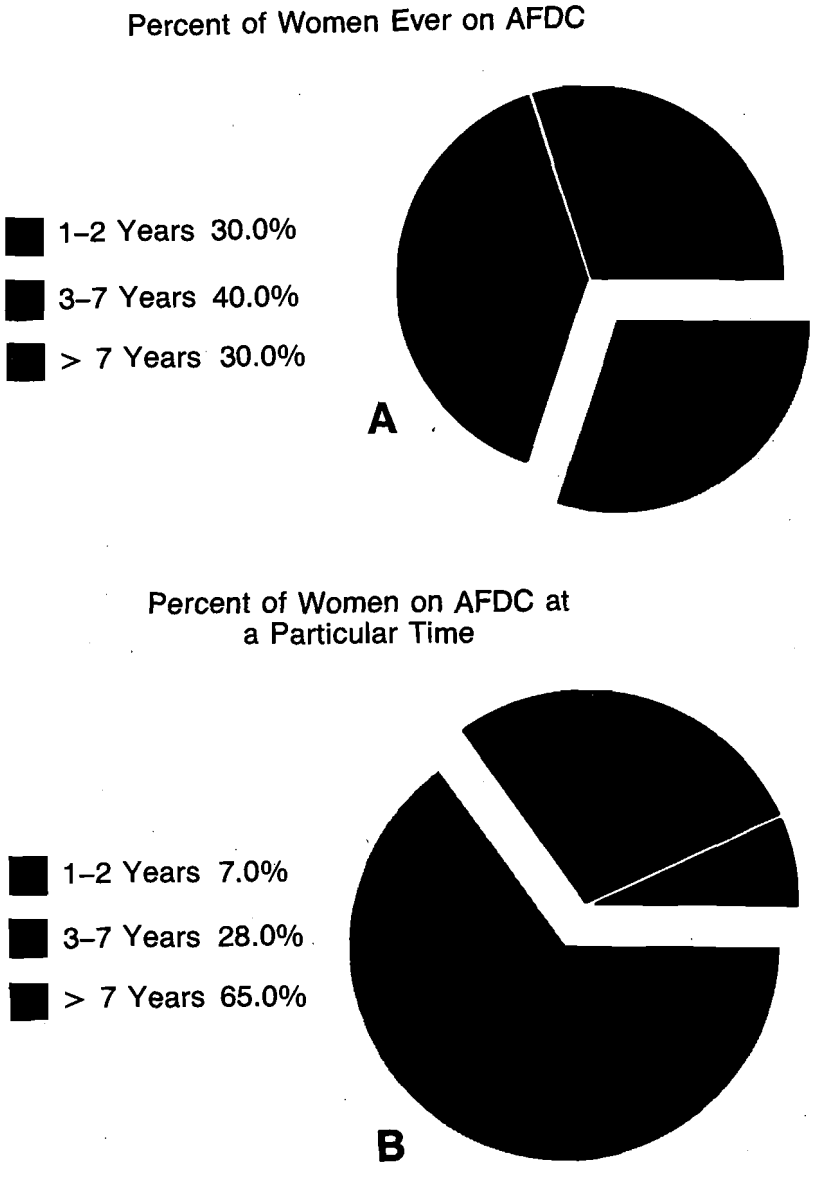


Figure 2-13 *A*, Percentage of women who have ever received AFDC and length of time they received it. *B*, Percentage of women who received AFDC at a particular time and the length of time they had been receiving it.

C.R.A.P. DETECTOR II-3

Question Schroeder, among others, concluded that there was a relationship between water hardness and cardiovascular disease. Specifically, he found a correlation of -0.56, which indicated that states with the softest water had the highest death rates for heart disease. Should you be worried if you live in an area with soft water?

Answer Schroeder's study used data aggregated at the level of states, and as such it was susceptible to the ecologic fallacy. Comstock followed up this finding by gathering data on individuals, and found no relationship between cardiovascular disease and trace elements in water. So, what holds at the level of the community or state may not obtain for the individual.

C.R.A.P. DETECTOR II-4

Question According to Ederer the 10 top batters in the American League in 1968 had a mean batting average of .414, and the 10 worst batted an average of .083, in the first week of play. As can be seen in Figure 2-14, by the second week both groups were batting in the low .200s. Does this mean that the good batters suddenly got worse, and the bad batters mysteriously got better?

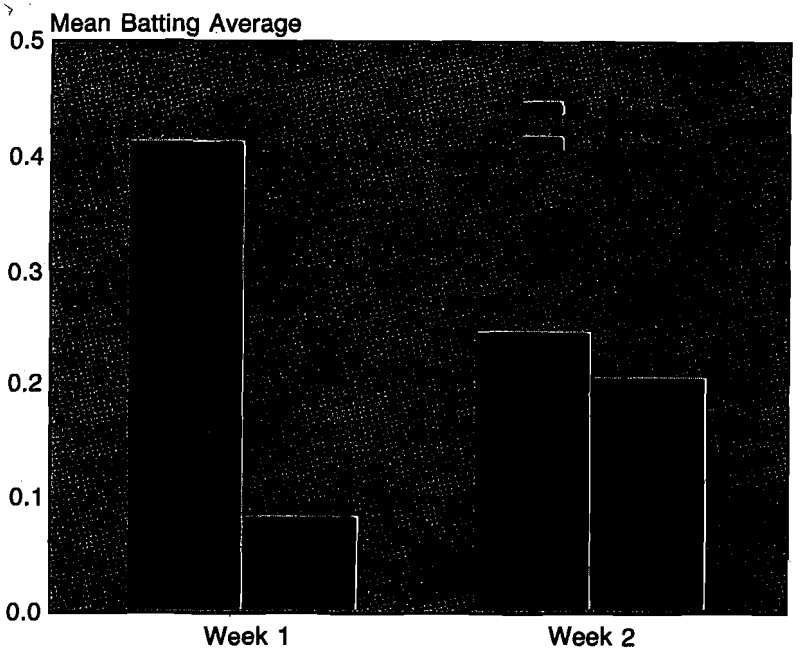


Figure 2-14 *Mean batting average of 10 best and 10 worst batters.*

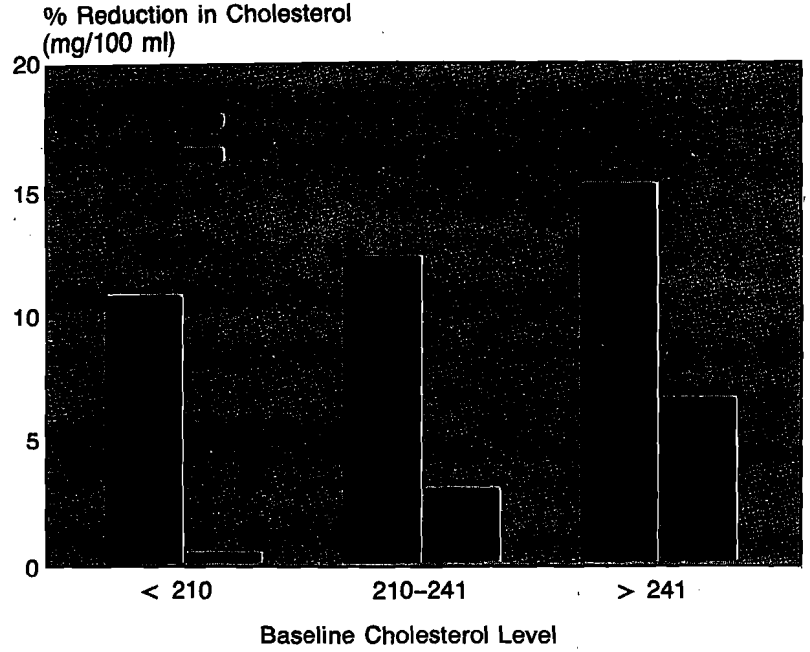


Figure 2-15 Example of regression toward the mean effect: the higher the baseline serum cholesterol level, the greater the subsequent improvement, regardless of diet.

Answer This is an example of the regression toward the mean effect; on the average, persons chosen because they are above the mean on one occasion tend to "regress" down toward it at a second measurement period, and those below the mean regress upward. Ederer showed the same effect with serum cholesterol (Fig. 2-15): the higher the baseline level, the greater the later "improvement," whether the subjects had been on a cholesterol-lowering diet or a control diet.

REFERENCES

Chenier NM. Reproductive hazards at work. Ottawa: Canadian Advisory Council on the Status of Women, 1982.

DESIGN ELEMENTS

Number of Observations

Bagby RM, Silverman I, Ryan DP, Dickens SE. Effects of mental health legislative reforms in Ontario. Can Psychol 1987; 28:21-29.

Comparison Groups

Hill AB. Statistical methods in clinical and preventive medicine. Edinburgh: Livingstone, 1962:365.

Hinshaw HC, Feldman WH, Pfuetze KH. Treatment of tuberculosis with streptomycin: summary of observations on 100 cases. JAMA 1946; 132:778-782.

SAMPLING

Cluster Sampling

Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. Am J Epidemiol 1981; 114:906-914.

Spitzer WO, Sackett DL, Sibley JC, Roberts RS, Gent M, Kergin DJ, Hackett BC, Olynich A. The Burlington Randomized Trial of the nurse practitioner. N Engl J Med 1974; 290:251-256.

Haphazard Sampling

Mueller PS, Watkin DM. Plasma unesterified fatty acid concentrations in neoplastic disease. J Lab Clin Med 1961; 57:95-108.

Quick AJ. Hemolysate prothrombin consumption time: a new test for thromboplastinogenic coagulation defects. J Lab Clin Med 1961; 57:290-299.

THREATS TO VALIDITY

Cook, TD, Campbell DT. Quasi-experimentation: design issues for field settings. Chicago: Rand McNally, 1979.

Subject Selection Biases

Sackett DL. Bias in analytic research. J Chronic Dis 1979; 32:51-63.

Healthy Worker Bias

Seltzer CC, Jablon S. Effects of selection on mortality. Am J Epidemiol 1974; 100:367-372.

Incidence–Prevalence Bias

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 3rd ed. revised. Washington: APA, 1987.

Harding CM, Zubin J, Strauss JS. Chronicity in schizophrenia: fact, partial fact, or artifact? Hosp Community Psychiatry 1987; 38:477-486.

Volunteer Bias

American Heart Association. The National Diet-Heart Study: final report. American Heart Association Monograph No. 18. New York: AHA, 1980.

Cochrane AL. The application of scientific methods to industrial and social medicine. In: Morris JN, ed. Uses of epidemiology. Baltimore: Williams & Wilkins, 1964.

Coronary Drug Project Research Group. Influence of adherence to treatment and response to cholesterol on mortality in the Coronary Drug Project. N Engl J Med 1980; 303:1038-1041.

Hawthorne Effect

Bramel D, Friend R. Hawthorne, the myth of the docile worker, and class bias in psychology. Am Psychol 1981; 36:867-878.
Frank JD. Persuasion and healing. Baltimore: Johns Hopkins Press, 1961.
Parsons HM. What happened at Hawthorne? Science 1974; 183:922-932.

Blinding

Beecher HK. The powerful placebo. JAMA 1955; 159:1602-1606.
Rosenthal R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1966.

Cohort Effects

Horn JL, Donaldson G. On the myth of intellectual decline in adulthood. Am Psychol 1976; 31:701-719.

Ecologic Fallacy

Robinson WS. Ecological correlations and the behavior of individuals. Am Sociol Rev 1950; 15:351-357.

EPIDEMIOLOGIC RESEARCH STRATEGIES

Nomenclature

Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. Belmont, CA: Lifetime Learning Publications, 1982.

C.R.A.P. DETECTORS

Bieber I, Dain HJ, Dince PR, et al. Homosexuality. New York: Basic Books, 1962.
Comstock GW. Fatal arteriosclerotic heart disease, water hardness at home, and socioeconomic characteristics. Am J Epidemiol 1971; 94:1-10.
Duncan GJ, Hill MS, Hoffman SD. Welfare dependence within and across generations. Science 1988; 239:467-471.
Ederer F. Serum cholesterol changes: effects of diet and regression toward the mean. J Chronic Dis 1972; 25: 277-289.
Schroeder HA. Relationship between mortality from cardiovascular disease and treated water supplies: variations in states and 163 largest municipalities of the United States. JAMA 1960; 172:1902-1908.

TO READ FURTHER

SAMPLING

Abramson JH. Survey methods in community medicine. Edinburgh: Churchill-Livingstone, 1974.
Levy P, Lemeshow S. Sampling for health professionals. Belmont, CA: Lifetime Learning Publications, 1980.

THREATS TO VALIDITY

Ederer F. Patient bias, investigator bias and the double-masked procedure in clinical trials. Am J Med 1975; 58:295-299.
Morgenstern H. Uses of ecologic analysis in epidemiological research. Am J Public Health 1982; 72:1336-1344.
Rosenthal R, Rosnow RL. The volunteer subject. New York: Wiley, 1975.
Sackett DL. Bias in analytic research. J Chronic Dis 1979; 32:51-63.
Walter SD. Cause-deleted proportional mortality analysis and the healthy worker effect. Stat Med 1986; 5:61-71.

EPIDEMIOLOGIC RESEARCH STRATEGIES

Cook TD, Campbell DT. Quasi-experimentation: design issues for field settings. Chicago: Rand McNally, 1979.
Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. Belmont, CA: Lifetime Learning Publications, 1982.