AP Statistics Summer Assignment

- *Introduction to Statistics & Data Analysis by Peck, Olsen, Devore. Chapter 2.*

We all know what it means for something to be random. Or do we? Many children's games rely on chance outcomes. Rolling dice, spinning spinners, and shuffling cards all select at random. Adult games use randomness as well, from card games to lotteries to Bingo. What's the most important aspect of the randomness in these games? It must be fair.

What is it about random selection that makes it seem fair? It's really two things. First, nobody can guess the outcome before it happens. Second, when we want things to be fair, usually some underlying set of outcomes will be equally likely (although in many games, some combinations of outcomes are more likely than others).

Randomness is not always what we might think of as "at random." Random outcomes have a lot of structure, especially when viewed in the long run. You can't predict how a fair coin will land on any single toss, but you're pretty confident that if you flipped it thousands of times you'd see about 50% heads. As we will see, randomness is an essential tool of Statistics. Statisticians don't think of randomness as the annoying tendency of things to be unpredictable or haphazard. Statisticians use randomness as a tool. In fact, without deliberately applying randomness, we couldn't do most of Statistics, and this book would stop right about here.[1]

But truly random values are surprisingly hard to get. Just to see how fair humans are at selecting, pick a number at random from the top of the next page. Go ahead. Turn the page, look at the numbers quickly, and pick a number at random.

Ready?

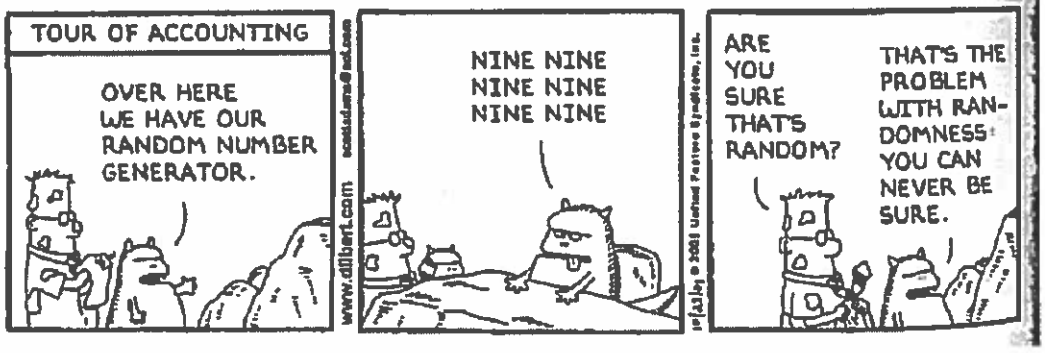
Go.

AP Statistics Summer Assignment

1 2 3 4

Did you pick 3? If so, you've got company. Almost 75% of all people pick the number 3. About 20% pick either 2 or 4. If you picked 1, well, consider yourself a little different. Only about 5% choose 1. Psychologists have proposed reasons for this phenomenon, but for us, it simply serves as a lesson that we've got to find a better way to choose things at random.

So how should we generate random numbers? It's surprisingly difficult to get random values even when they're equally likely. Computers have become a popular way to generate random numbers. Even though they often do much better than humans, computers can't generate truly random numbers either. Computers follow programs. Start a computer from the same place, and it will always follow exactly the same path. So numbers generated by a computer program are not truly random. Technically, "random" numbers generated this way are *pseudorandom* numbers. Pseudorandom values are generated in a fixed sequence, and because computers can represent only a finite number of distinct values, the sequence of pseudorandom numbers must eventually repeat itself. Fortunately, pseudorandom values are good enough for most purposes because they are virtually indistinguishable from truly random numbers.



We will study methods of randomization (ie. how to use your calculator as a random number generator) in the first unit of class.

AP Statistics Summer Assignment

When selecting a random sample, researchers can choose to do the sampling with or without replacement. **Sampling with replacement** means that after each successive item is selected for the sample, the item is "replaced" back into the population and may therefore be selected again at a later stage. In practice, sampling with replacement is rarely used. Instead, the more common method is to not allow the same item to be included in the sample more than once. After being included in the sample, an individual or object would not be considered for further selection. Sampling in this manner is called **sampling without replacement.**

DEFINITION

Sampling without replacement: Once an individual from the population is selected for inclusion in the sample, it may not be selected again in the sampling process. A sample selected without replacement includes n distinct individuals from the population.

Sampling with replacement: After an individual from the population is selected for inclusion in the sample and the corresponding data are recorded, the individual is placed back in the population and can be selected again in the sampling process. A sample selected with replacement might include any particular individual from the population more than once.

Although these two forms of sampling are different, when the sample size n is small relative to the population size, as is often the case, there is little practical difference between them. In practice, the two can be viewed as equivalent if the sample size is at most 10% of the population size.

It isn't sufficient to just draw a sample and start asking questions. We'll want our survey to be *valid*. A valid survey yields the information we are seeking about the population we are interested in. Before setting out to survey, ask yourself:

- What do I want to know?
- Am I asking the right respondents?
- Am I asking the right questions?
- What would I do with the answers if I had them; would they address the things I want to know?

These questions may sound obvious, but there are a number of pitfalls to avoid.

Know what you want to know. Before considering a survey, understand what you hope to learn and about whom you hope to learn it. Far too often, people decide to perform a survey without any clear idea of what they hope to learn.

Use the right frame. A valid survey obtains responses from the appropriate respondents. Be sure you have a suitable *sampling frame*. Have you identified the population of interest and sampled from it appropriately? A company might survey customers who returned warranty registration cards, a readily available sampling frame. But if the company wants to know how to make their product more attractive, the most important population is the customers who rejected their product in favor of one from a competitor.

Tune your instrument. It is often tempting to ask questions you don't really need, but beware—longer questionnaires yield fewer responses and thus a greater chance of nonresponse bias.

Ask specific rather than general questions. People are not very good at estimating their typical behavior, so it is better to ask "How many hours did you sleep last night?" than "How much do you usually sleep?" Sure, some responses will include some unusual events (My dog was sick; I was up all night.), but overall you'll get better data.

Ask for quantitative results when possible. "How many magazines did you read last week?" is better than "How much do you read: A lot, A moderate amount, A little, or None at all?"

Be careful in phrasing questions. A respondent may not understand the question—or may understand the question differently than the researcher intended it. ("Does anyone in your family belong to a union?" Do you mean just me, my spouse, and my children? Or does "family" include my father, my siblings, and my second cousin once removed? What about my grandfather, who is staying with us? I think he once belonged to the Autoworkers Union.) Respondents are unlikely (or may not have the opportunity) to ask for clarification. A question like "Do you approve of the recent actions of the Secretary of Labor?" is likely not to measure what you want if many re-

AP Statistics Summer Assignment

We draw samples because we can't work with the entire population, but we want the statistics we compute from a sample to reflect the corresponding parameters accurately. A sample that does this is said to be **representative.** A biased sampling methodology tends to over- or underestimate the parameter of interest.

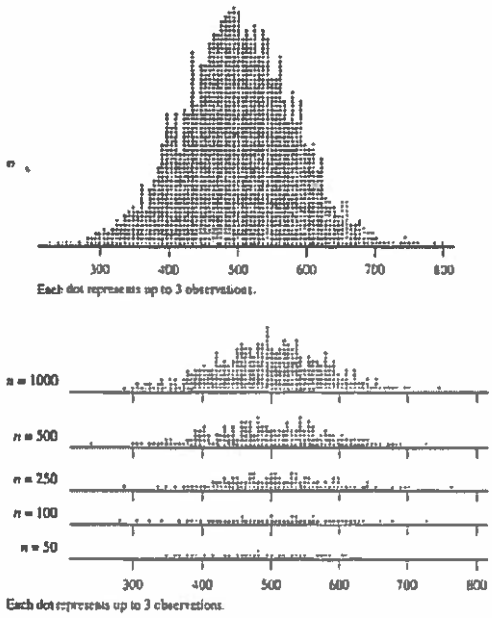
Here's a table summarizing the notation:

Name	Statistic	Parameter
Mean	$\bar{y}$	μ (mu, pronounced "meeoo," not "moo")
Standard deviation	s	σ (sigma)
Correlation	r	ρ (rho)
Regression coefficient	b	β (beta, pronounced "baytah"[7])
Proportion	$\hat{p}$	p (pronounced "pee"[8])

An Important Note Concerning Sample Size

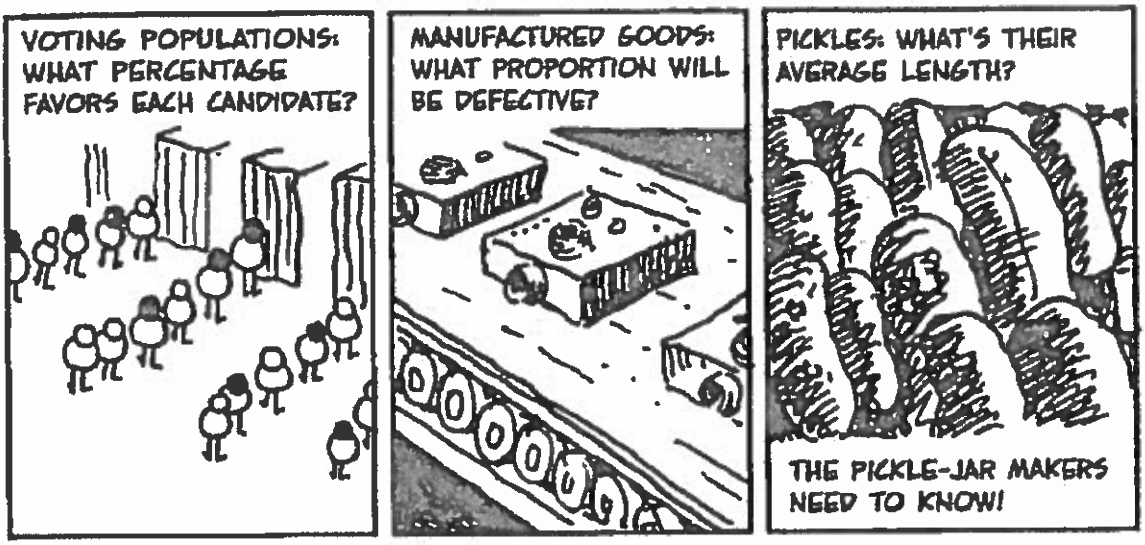
It is a common misconception that if the size of a sample is relatively small compared to the population size, the sample cannot possibly accurately reflect the population. Critics of polls often make statements such as, "There are 14.6 million registered voters in California. How can a sample of 1000 registered voters possibly reflect public opinion when only about 1 in every 14,000 people is included in the sample?" These critics do not understand the power of random selection!

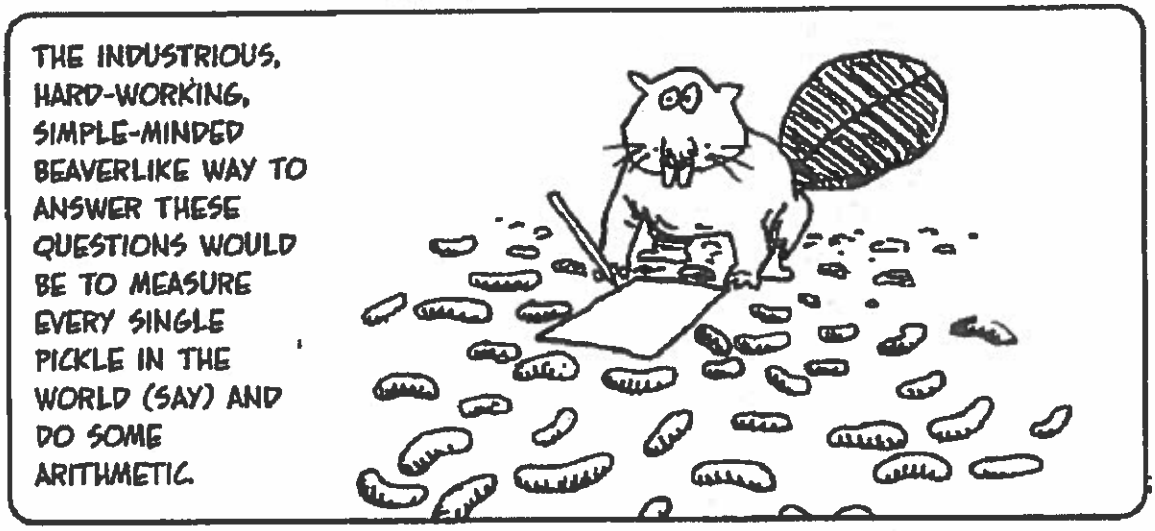
Consider a population consisting of 5000 applicants to a state university, and suppose that we are interested in math SAT scores for this population. A dotplot of the values in this population is shown in Figure 2.1(a). Figure 2.1(b) shows dotplots of the math SAT scores for individuals in five different random samples from the population, ranging in sample size from $n = 50$ to $n = 1000$. Notice that the samples tend to reflect the distribution of scores in the population. If we were interested in using the sample to estimate the population average or to say something about the variability in SAT scores, even the smallest of the samples ($n = 50$) pictured would provide reliable information. Although it is possible to obtain a simple random sample that does not do a reasonable job of representing the population, this is likely only when the sample size is very small, and unless the population itself is small, this risk does not depend on what fraction of the population is sampled. The random selection process allows us to be confident that the resulting sample adequately reflects the population, even when the sample consists of only a small fraction of the population.



AP Statistics Summer Assignment

THE PROBLEM WITH THE WORLD IS THAT THE COLLECTIONS OF STUFF IN IT ARE SO LARGE, IT'S HARD TO GET THE INFORMATION WE WANT:

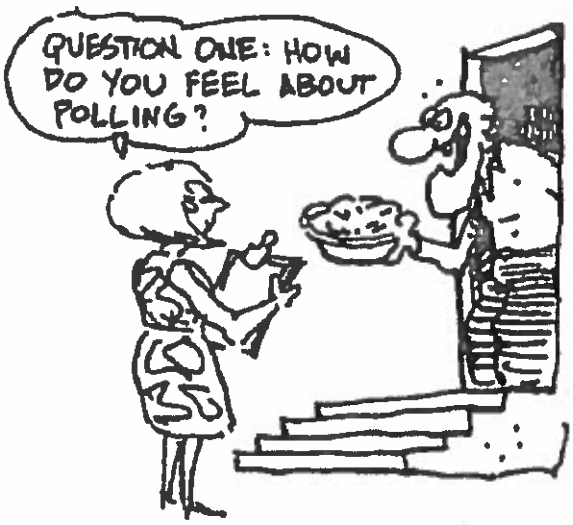




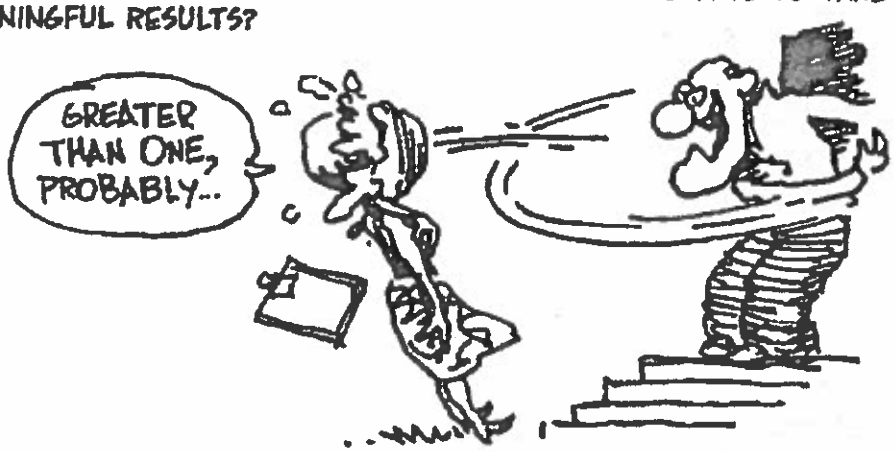
BUT WE AREN'T BEAVERS—WE'RE *STATISTICIANS!* WE'RE LOOKING FOR THE *EASY* WAY OUT...

AP Statistics Summer Assignment

OUR METHOD IS TO TAKE A *SAMPLE*... A RELATIVELY SMALL SUBSET OF THE TOTAL POPULATION, THE WAY POLLSTERS DO AT ELECTION TIME.

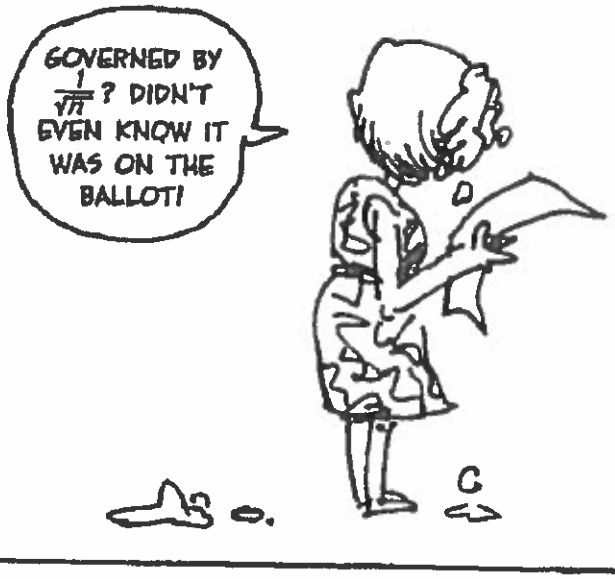


AN OBVIOUS QUESTION IS: HOW BIG A SAMPLE DO WE HAVE TO TAKE TO GET MEANINGFUL RESULTS?



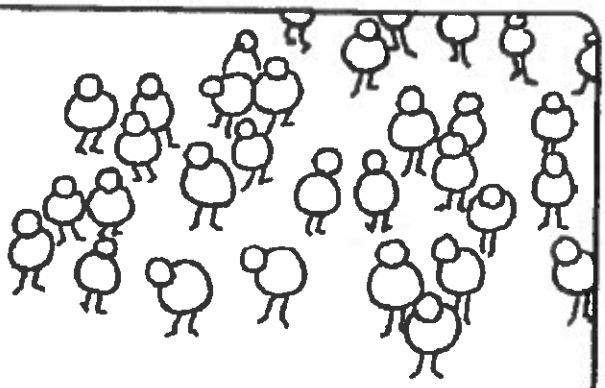
AND THE ANSWER, WHICH YOU SHOULD INSCRIBE IN YOUR BRAIN FOREVERMORE, WILL TURN OUT TO BE: IF n IS THE NUMBER OF ITEMS IN THE SAMPLE, THEN EVERYTHING IS GOVERNED BY

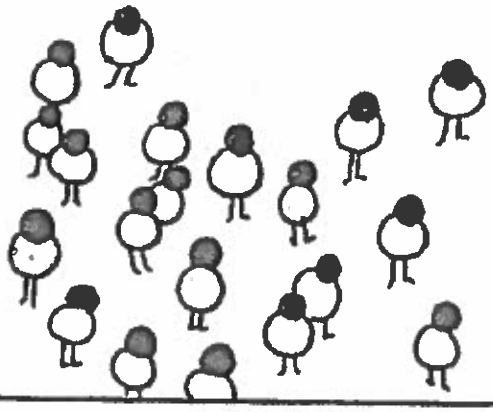
$$\frac{1}{\sqrt{n}}.$$



SAMPLING DESIGN

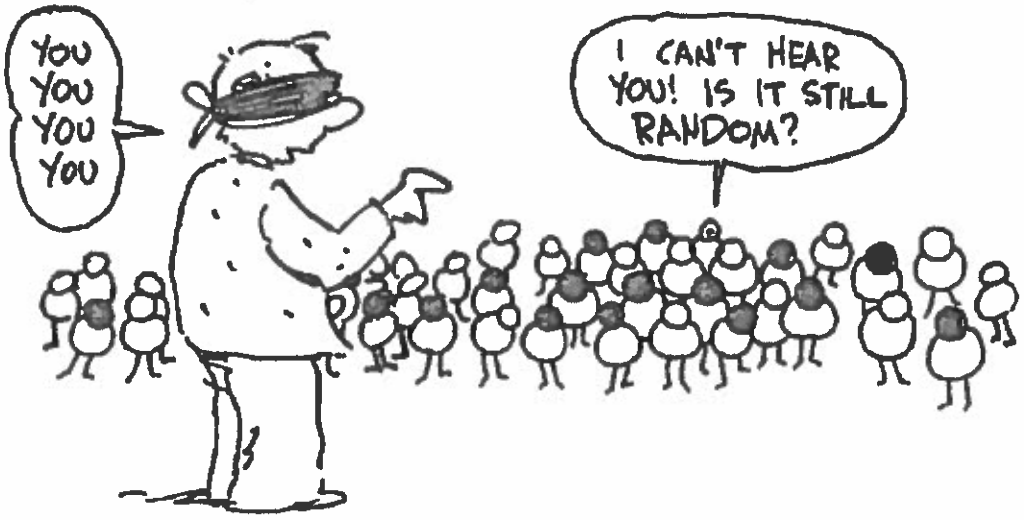
BEFORE DOING THE NUMBERS, WE SHOULD POINT OUT THAT THE *QUALITY* OF THE SAMPLE IS AS IMPORTANT AS ITS *SIZE*. HOW DO WE ASSURE OURSELVES THAT WE'RE CHOOSING A REPRESENTATIVE SAMPLE?





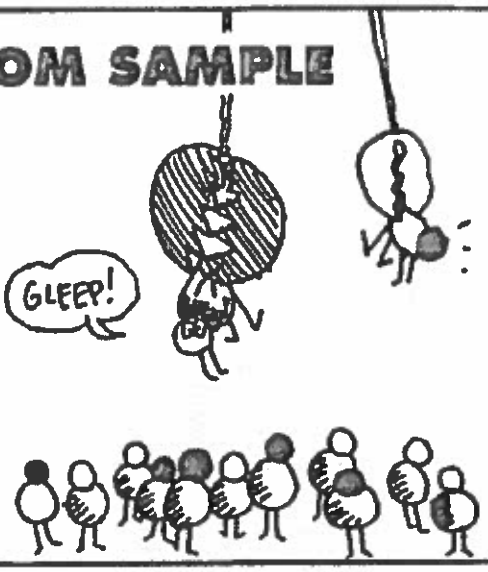
THE *SELECTION PROCESS ITSELF* IS CRITICAL. FOR EXAMPLE, A VOTER SURVEY THAT SYSTEMATICALLY EXCLUDED BLACK PEOPLE WOULD BE WORTHLESS, AND THERE ARE A HOST OF OTHER WAYS TO RUIN, OR BIAS, A SAMPLE.

NOT TO PROLONG THE MYSTERY, THE WAY TO GET STATISTICALLY DEPENDABLE RESULTS IS TO CHOOSE THE SAMPLE AT **random.**

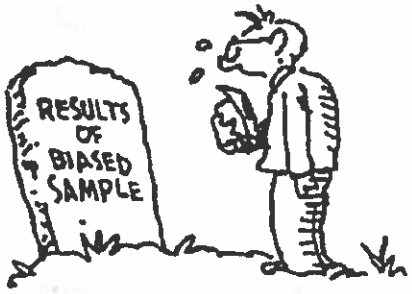


THE SIMPLE RANDOM SAMPLE

SUPPOSE WE HAVE A LARGE POPULATION OF OBJECTS AND A PROCEDURE FOR SELECTING n OF THEM. IF THE PROCEDURE ENSURES THAT *ALL POSSIBLE SAMPLES OF n OBJECTS ARE EQUALLY LIKELY*, THEN WE CALL THE PROCEDURE A **simple random sample.**



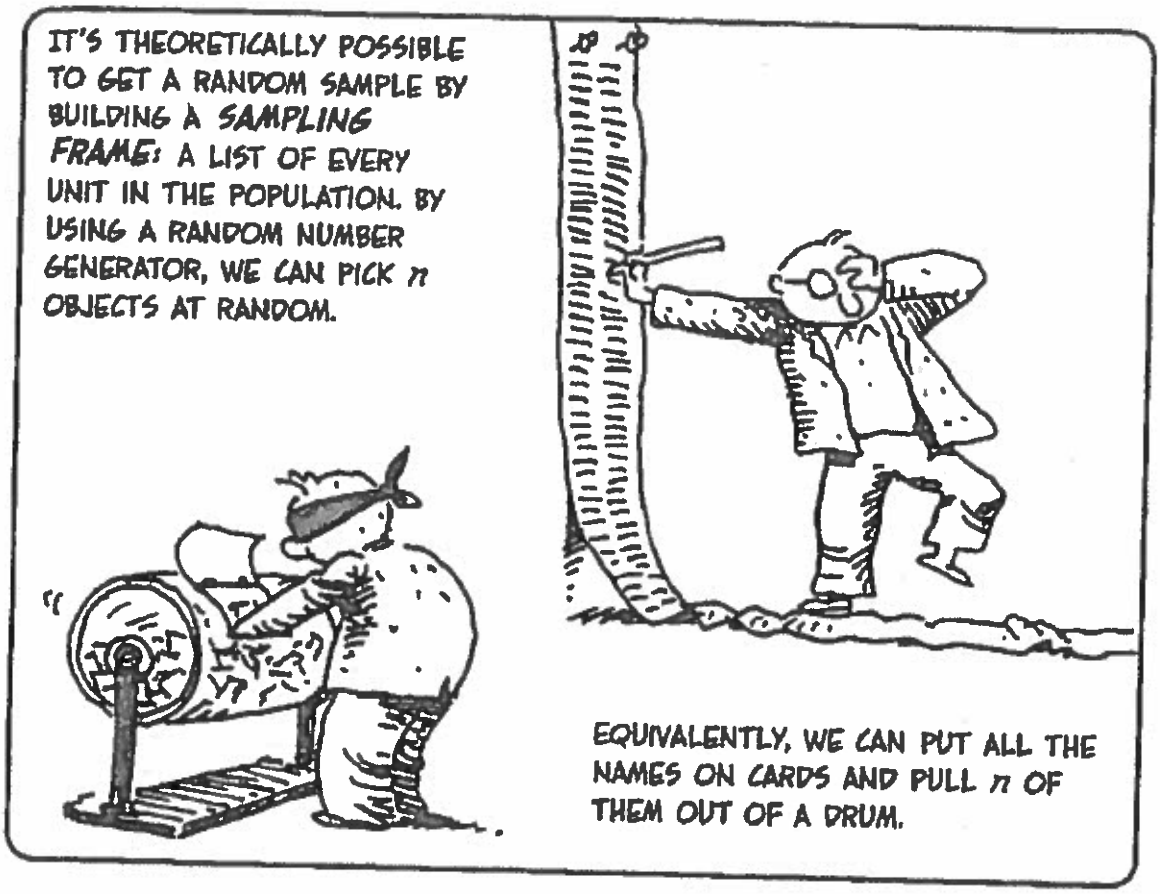
THE SIMPLE RANDOM SAMPLE HAS TWO PROPERTIES THAT MAKE IT THE STANDARD AGAINST WHICH WE MEASURE ALL OTHER METHODS:



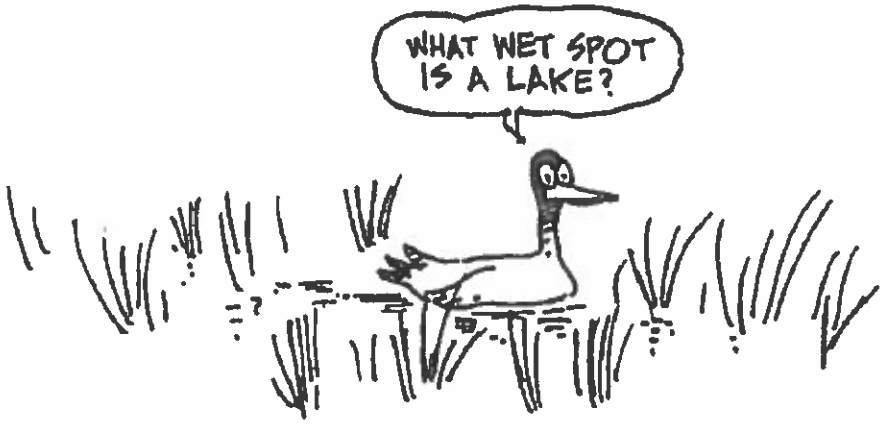
1) UNBIASED: EACH UNIT HAS THE SAME CHANCE OF BEING CHOSEN.

2) INDEPENDENCE: SELECTION OF ONE UNIT HAS NO INFLUENCE ON THE SELECTION OF OTHER UNITS.

UNFORTUNATELY, IN THE REAL WORLD, COMPLETELY UNBIASED, INDEPENDENT SAMPLES ARE HARD TO FIND. FOR INSTANCE, SURVEYING VOTERS BY RANDOMLY DIALING TELEPHONE NUMBERS IS BIASED: IT IGNORES VOTERS WITHOUT A TELEPHONE AND OVERSAMPLES PEOPLE WITH MORE THAN ONE NUMBER.



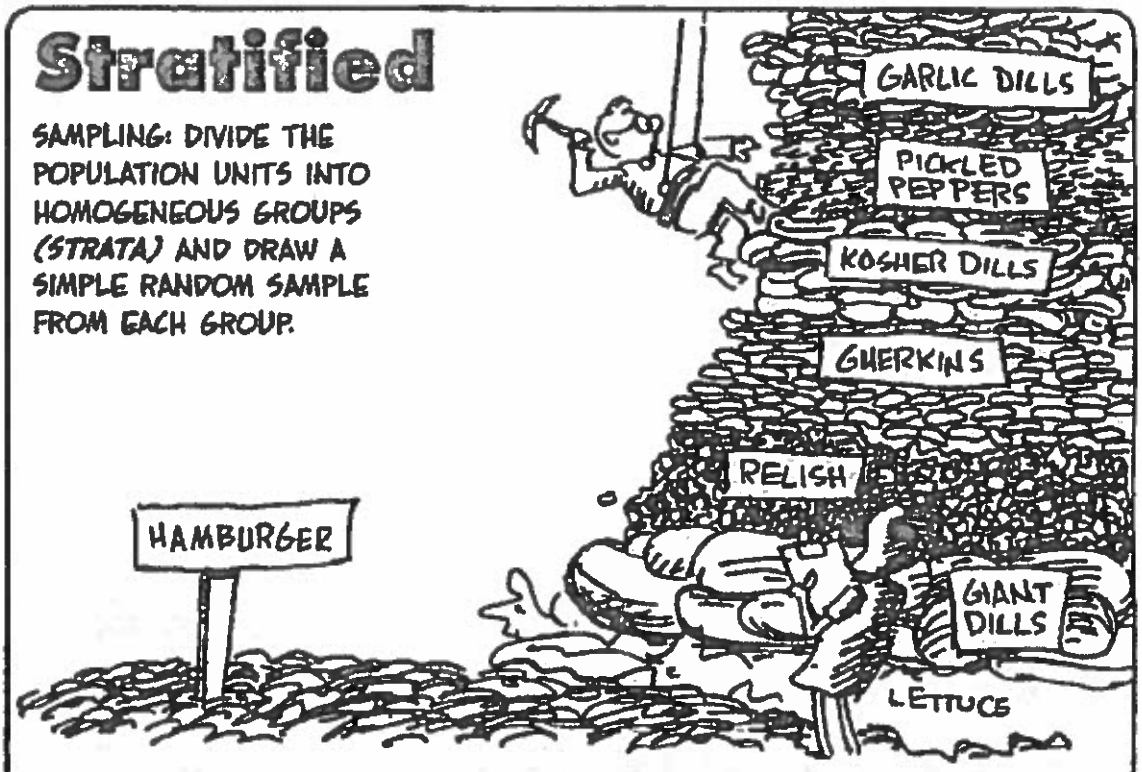
BUT THIS IS NOT ALWAYS EASY. MAKING THE FRAME MAY BE PROHIBITIVELY COSTLY, CONTROVERSIAL, OR EVEN IMPOSSIBLE. FOR EXAMPLE, AN E.P.A. WATER QUALITY STUDY NEEDED A SAMPLING FRAME OF LAKES IN THE U.S., SO THEN SOMEBODY HAS TO DECIDE:



ARE THERE OTHER WAYS TO SAMPLE THAT ARE MORE *EFFICIENT* AND *COST-EFFECTIVE* THAN A SIMPLE RANDOM SAMPLE? YES—IF YOU ALREADY KNOW SOMETHING ABOUT THE POPULATION. FOR INSTANCE...

Stratified

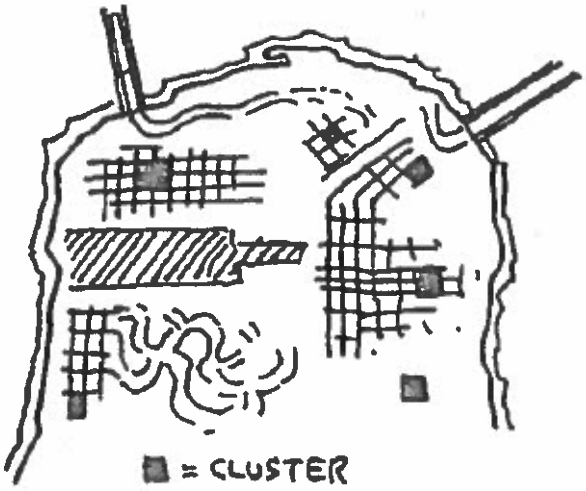
SAMPLING: DIVIDE THE POPULATION UNITS INTO HOMOGENEOUS GROUPS *(STRATA)* AND DRAW A SIMPLE RANDOM SAMPLE FROM EACH GROUP.



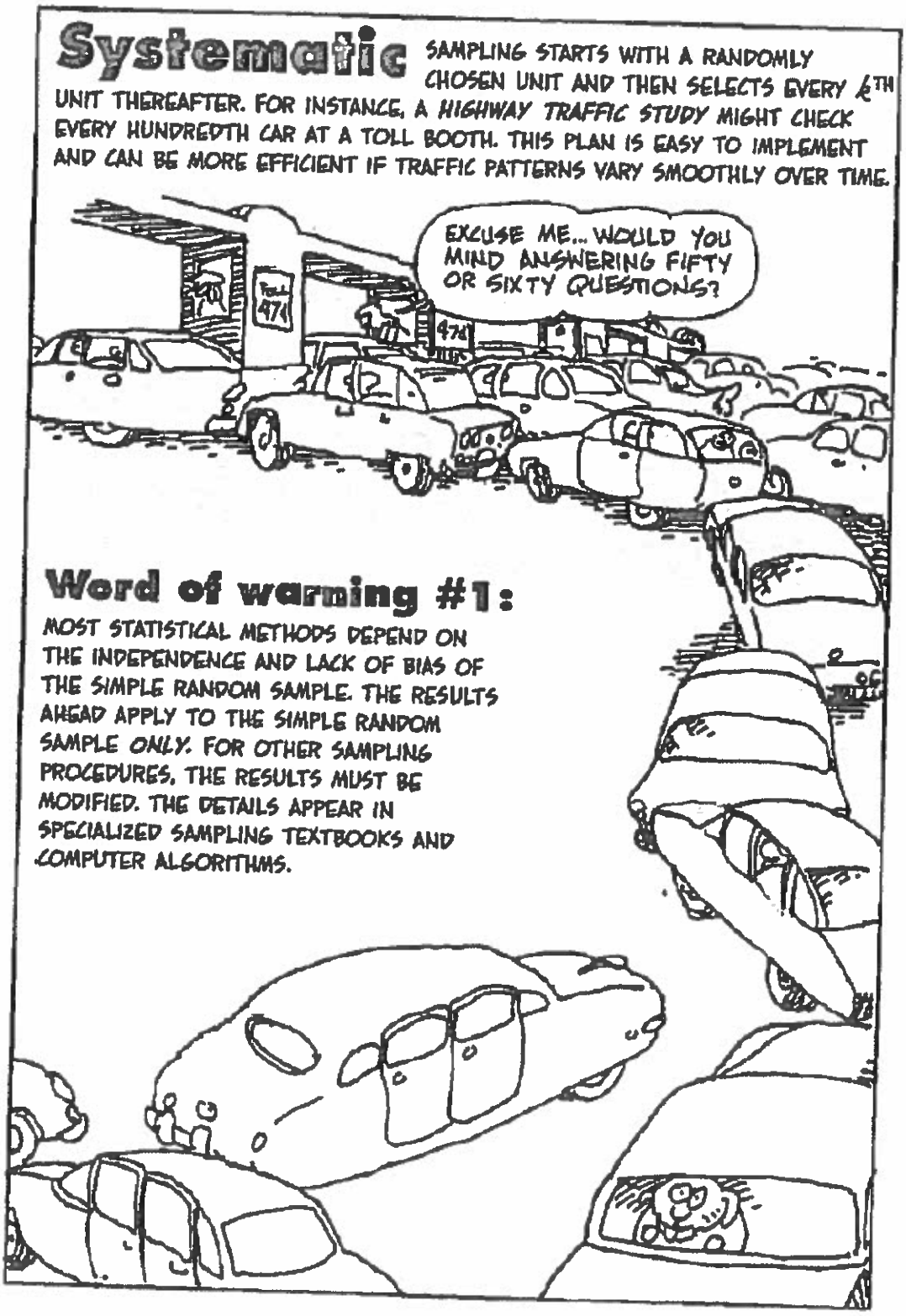
FOR EXAMPLE, THE POPULATION OF ALL *PICKLES* CAN BE STRATIFIED BY *TYPE OF PICKLE*. WITHIN EACH TYPE OR STRATUM, THE SIZE SHOULD BE LESS VARIABLE.

Cluster

SAMPLING GROUPS THE POPULATION INTO SMALL CLUSTERS, DRAWS A SIMPLE RANDOM SAMPLE OF CLUSTERS, AND OBSERVES EVERYTHING IN THE SAMPLED CLUSTERS. THIS CAN BE COST-EFFECTIVE IF TRAVEL COSTS BETWEEN RANDOMLY SAMPLED UNITS IS HIGH.



AN EXAMPLE IS A CITY HOUSING SURVEY WHICH DIVIDES A CITY INTO BLOCKS, RANDOMLY SAMPLES THE BLOCKS, AND LOOKS AT EVERY HOUSING UNIT IN EACH SAMPLED BLOCK.


Systematic
SAMPLING STARTS WITH A RANDOMLY CHOSEN UNIT AND THEN SELECTS EVERY k^{TH} UNIT THEREAFTER. FOR INSTANCE, A HIGHWAY TRAFFIC STUDY MIGHT CHECK EVERY HUNDREDTH CAR AT A TOLL BOOTH. THIS PLAN IS EASY TO IMPLEMENT AND CAN BE MORE EFFICIENT IF TRAFFIC PATTERNS VARY SMOOTHLY OVER TIME.
EXCUSE ME... WOULD YOU MIND ANSWERING FIFTY OR SIXTY QUESTIONS?
Word of warning #1:
MOST STATISTICAL METHODS DEPEND ON THE INDEPENDENCE AND LACK OF BIAS OF THE SIMPLE RANDOM SAMPLE. THE RESULTS AHEAD APPLY TO THE SIMPLE RANDOM SAMPLE ONLY. FOR OTHER SAMPLING PROCEDURES, THE RESULTS MUST BE MODIFIED. THE DETAILS APPEAR IN SPECIALIZED SAMPLING TEXTBOOKS AND COMPUTER ALGORITHMS.

Word of warning #2:

WITHOUT RANDOMIZED DESIGN, THERE CAN BE NO DEPENDABLE STATISTICAL ANALYSIS, NO MATTER HOW IT IS MODIFIED. THE BEAUTY OF RANDOM SAMPLING IS THAT IT "STATISTICALLY GUARANTEES" THE ACCURACY OF THE SURVEY.

A COMMONLY USED METHOD IS ESPECIALLY PRONE TO BIAS: IT'S CALLED AN **opportunity** SAMPLE. AVOIDING ALL THE BOTHER OF DESIGNING A PROCEDURE, THE OPPORTUNITY SAMPLER JUST GRABS THE FIRST n POPULATION UNITS TO COME ALONG.

A CLASSIC EXAMPLE IS SHERE HITE'S BOOK, *WOMEN AND LOVE*. 100,000 QUESTIONNAIRES WENT TO WOMEN'S ORGANIZATIONS (AN *OPPORTUNITY SAMPLE*), ONLY 4.5% WERE FILLED OUT AND RETURNED (*RESPONSE BIAS*). SO HER "RESULTS" WERE BASED ON A SAMPLE OF WOMEN WHO WERE HIGHLY MOTIVATED TO ANSWER THE SURVEY'S QUESTIONS, FOR WHATEVER REASON.

AP Statistics Summer Assignment

DEFINITION: Bias

The design of a statistical study shows bias if it systematically favors certain outcomes.

AP EXAM TIP If you're asked to describe how the design of a study leads to bias, you're expected to identify the direction of the bias. Suppose you were asked, "Explain how using a convenience sample of students in your statistics class to estimate the proportion of all high school students who own a graphing calculator could result in bias." You might respond, "This sample would probably include a much higher proportion of students with a graphing calculator than in the population at large because a graphing calculator is required for the statistics class. That is, this method would probably lead to an overestimate of the actual population proportion."

Sample Badly with Volunteers

One of the most common dangerous sampling methods is a voluntary response sample. In a voluntary response sample, a large group of individuals is invited to respond, and all who do respond are counted. This method is used by call-in shows, 900 numbers, Internet polls, and letters written to members of Congress. Voluntary response samples are almost always biased, and so conclusions drawn from them are almost always wrong.

It's often hard to define the sampling frame of a voluntary response study. Practically, the frames are groups such as Internet users who frequent a particular Web site or those who happen to be watching a particular TV show at the moment. But these sampling frames don't correspond to interesting populations.

Even within the sampling frame, voluntary response samples are often biased toward those with strong opinions or those who are strongly motivated. People with very negative opinions tend to respond more often than those with equally strong positive opinions. The sample is not representative, even though every individual in the population may have been offered the chance to respond. The resulting voluntary response bias invalidates the survey.

If you had it to do over again, would you have children? Ann Landers, the advice columnist, asked parents this question. The overwhelming majority—70% of the more than 10,000 people who wrote in—said no, kids weren't worth it. A more carefully designed survey later showed that about 90% of parents actually are happy with their decision to have children. What accounts for the striking difference in these two results? What parents do you think are most likely to respond to the original question?

Sample Badly, but Conveniently

Another sampling method that doesn't work is convenience sampling. As the name suggests, in convenience sampling we simply include the individuals who are convenient for us to sample. Unfortunately, this group may not be representative of the population. A recent survey of 437 potential home buyers in Orange County, California, found, among other things, that

> *All but 2 percent of the buyers have at least one computer at home, and 62 percent have two or more. Of those with a computer, 99 percent are connected to the Internet (Jennifer Hieger, "Portrait of Homebuyer Household: 2 Kids and a PC," Orange County Register, 27 July 2001).*

Later in the article, we learn that the survey was conducted via the Internet! That was a convenient way to collect data and surely easier than drawing a simple random sample, but perhaps home builders shouldn't conclude from this study that *every* family has a computer and an Internet connection.

Many surveys conducted at shopping malls suffer from the same problem. People in shopping malls are not necessarily representative of the population of interest. Mall shoppers tend to be more affluent and include a larger percentage of teenagers and retirees than the population at large. To make matters worse, survey interviewers tend to select individuals who look "safe," or easy to interview.

Sample from a Bad Sampling Frame

An SRS from an incomplete sampling frame introduces bias because the individuals included may differ from the ones not in the frame. People in prison, homeless people, students, and long-term travelers are all likely to be missed. In telephone surveys, people who have only cell phones or who use VOIP Internet phones are often missing from the sampling frame.

Undercoverage

Many survey designs suffer from undercoverage, in which some portion of the population is not sampled at all or has a smaller representation in the sample than it has in the population. Undercoverage can arise for a number of reasons, but it's always a potential source of bias.

Telephone surveys are usually conducted when you are likely to be home, interrupting your dinner. If you eat out often, you may be less likely to be surveyed, a possible source of undercoverage.

AP Statistics Summer Assignment

DEFINITION

A study is an observational study if the investigator observes characteristics of a sample selected from one or more existing populations. The goal of an observational study is usually to draw conclusions about the corresponding population or about differences between two or more populations. In a well-designed observational study, the sample is selected in a way that is designed to produce a sample that is representative of the population.

A study is an experiment if the investigator observes how a response variable behaves when one or more explanatory variables, also called factors, are manipulated. The usual goal of an experiment is to determine the effect of the manipulated explanatory variables (factors) on the response variable. In a well-designed experiment, the composition of the groups that will be exposed to different experimental conditions is determined by random assignment.

The type of conclusion that can be drawn from a statistical study depends on the study design. Both observational studies and experiments can be used to compare groups, but in an experiment the researcher controls who is in which group, whereas this is not the case in an observational study. This seemingly small difference is critical when it comes to drawing conclusions based on data from the study.

A well-designed experiment can result in data that provide evidence for a cause-and-effect relationship. This is an important difference between an observational study and an experiment. In an observational study, it is impossible to draw clear cause-and-effect conclusions because we cannot rule out the possibility that the observed effect is due to some variable other than the explanatory variable being studied. Such variables are called confounding variables.

DEFINITION

An experiment is a study in which one or more explanatory variables are manipulated in order to observe the effect on a response variable.

The explanatory variables are those variables that have values that are controlled by the experimenter. Explanatory variables are also called factors.

The response variable is a variable that is not controlled by the experimenter and that is measured as part of the experiment.

An experimental condition is any particular combination of values for the explanatory variables. Experimental conditions are also called treatments.

Is it *ever* possible to get convincing evidence of a cause-and-effect relationship? Well, yes it is, but we would have to take a different approach. We could take a group of third graders, randomly assign half to take music lessons, and forbid the other half to do so. Then we could compare their grades several years later. This kind of study design is called an experiment.

An experiment requires a random assignment of subjects to treatments. Only an experiment can justify a claim like "Music lessons cause higher grades." Questions such as "Does taking vitamin C reduce the chance of getting a cold?" and "Does working with computers improve performance in Statistics class?" and "Is this drug a safe and effective treatment for that disease?" require a designed experiment to establish cause and effect.

Experiments study the relationship between two or more variables. An experimenter must identify at least one explanatory variable, called a factor, to manipulate and at least one response variable to measure. What distinguishes an experiment from other types of investigation is that the experimenter actively and deliberately manipulates the factors to control the details of the possible treatments, and assigns the subjects to those treatments *at random*. The experimenter then observes the response variable and *compares* responses for different groups of subjects who have been treated differently. For example, we might design an experiment to see whether the amount of sleep and exercise you get affects your performance.

The individuals on whom or which we experiment are known by a variety of terms. Humans who are experimented on are commonly called subjects or participants. Other individuals (rats, days, petri dishes of bacteria) are commonly referred to by the more generic term experimental unit. When we recruit subjects for our sleep deprivation experiment by advertising in Statistics class, we'll probably have better luck if we invite them to be participants than if we advertise that we need experimental units.

The specific values that the experimenter chooses for a factor are called the levels of the factor. We might assign our participants to sleep for 4, 6, or 8 hours. Often there are several factors at a variety of levels. (Our subjects will also be assigned to a treadmill for 0 or 30 minutes.) The combination of specific levels from all the factors that an experimental unit receives is known as its treatment. (Our subjects could have any one of six different treatments—three sleep levels, each at two exercise levels.)

How should we assign our participants to these treatments? Some students prefer 4 hours of sleep, while others need 8. Some exercise regularly; others are couch potatoes. Should we let the students choose the treatments they'd prefer? No. That would not be a good idea. To have any hope of drawing a fair conclusion, we must assign our participants to their treatments *at random*.

It may be obvious to you that we shouldn't let the students choose the treatment they'd prefer, but the need for random assignment is a lesson that was once hard for some to accept. For example, physicians might naturally prefer to assign patients to the therapy that they think best rather than have a random element such as a coin flip determine the treatment. But we've known for more than a century that for the results of an experiment to be valid, we must use deliberate randomization.

An Experiment:
Manipulates the factor levels to create treatments.
Randomly assigns subjects to these treatment levels.
Compares the responses of the subject groups across treatment levels.

There are two kinds of gardeners. Some water frequently, making sure that the plants are never dry. Others let Mother Nature take her course and leave the watering to her. The makers of OptiGro want to ensure that their product will work under a wide variety of watering conditions. Maybe we should include the amount of watering as part of our experiment. Can we study a second factor at the same time and still learn as much about fertilizer?

We now have two factors (fertilizer at three levels and irrigation at two levels). We combine them in all possible ways to yield six treatments:

	No Fertilizer	Half Fertilizer	Full Fertilizer
No Added Water	1	2	3
Daily Watering	4	5	6

If we allocate the original 12 plants, the experiment now assigns 2 plants to each of these six treatments at random. This experiment is a **completely randomized two-factor experiment** because any plant could end up assigned at random to any of the six treatments (and we have two factors).

Key Concepts in Experimental Design

Random Assignment

Random assignment (of subjects to treatments or of treatments to trials) to ensure that the experiment does not systematically favor one experimental condition (treatment) over another.

Blocking

Using extraneous variables to create groups (blocks) that are similar. All experimental conditions (treatments) are then tried in each block.

Direct Control

Holding extraneous variables constant so that their effects are not confounded with those of the experimental conditions (treatments).

Replication

Ensuring that there is an adequate number of observations for each experimental condition.

Blocking can be considered under the "umbrella" of controlling the experiment.

AP Statistics Summer Assignment

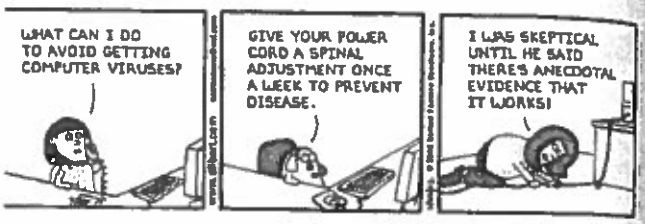
Control. We control sources of variation other than the factors we are testing by making conditions as similar as possible for all treatment groups. For human subjects, we try to treat them alike. However, there is always a question of degree and practicality. Controlling extraneous sources of variation reduces the variability of the responses, making it easier to detect differences among the treatment groups.

Making generalizations from the experiment to other levels of the controlled factor can be risky. For example, suppose we test two laundry detergents and carefully control the water temperature at 180°F. This would reduce the variation in our results due to water temperature, but what could we say about the detergents' performance in cold water? Not much. It would be hard to justify extrapolating the results to other temperatures.

Although we control both experimental factors and other sources of variation, we think of them very differently. We control a factor by assigning subjects to different factor levels because we want to see how the response will change at those different levels. We control other sources of variation to *prevent* them from changing and affecting the response variable.

Randomize. As in sample surveys, randomization allows us to equalize the effects of unknown or uncontrollable sources of variation. It does not eliminate the effects of these sources, but it should spread them out across the treatment levels so that we can see past them. If experimental units were not assigned to treatments at random, we would not be able to use the powerful methods of Statistics to draw conclusions from an experiment. Assigning subjects to treatments at random reduces bias due to uncontrolled sources of variation. Randomization protects us even from effects we didn't know about. There's an adage that says "control what you can, and randomize the rest."

Replicate. Two kinds of replication show up in comparative experiments. First, we should apply each treatment to a number of subjects. Only with such replication can we estimate the variability of responses. If we have not assessed the variation, the experiment is not complete. The outcome of an experiment on a single subject is an anecdote, not data.



A second kind of replication shows up when the experimental units are not a representative sample from the population of interest. We may believe that what is true of the students in Psych 101 who volunteered for the sleep experiment is true of all humans, but we'll feel more confident if our results for the experiment are *replicated* in another part of the country, with people of different ages, and at different times of the year. Replication of an entire experiment with the controlled sources of variation at different levels is an essential step in science.

Block. The ability of randomizing to equalize variation across treatment groups works best in the long run. For example, if we're allocating players to two 6-player soccer teams from a pool of 12 children, we might do so at random to equalize the talent. But what if there were two 12-year-olds and ten 6-year-olds in the group? Randomizing may place both 12-year-olds on the same team. In the long run, if we did this over and over, it would all equalize. But wouldn't it be better to assign one 12-year-old to each group (at random) and five 6-year-olds to each team (at random)? By doing this, we would improve fairness in the short run. This approach makes the division more fair by recognizing the variation in *age* and allocating the players at random *within* each age level. When we do this, we call the variable *age* a blocking variable. The levels of *age* are called blocks.

Sometimes, attributes of the experimental units that we are not studying and that we can't control may nevertheless affect the outcomes of an experiment. If we group similar individuals together and then randomize within each of these blocks, we can remove much of the variability due to the difference among the blocks. Blocking is an important compromise between randomization and control. However, unlike the first three principles, blocking is not *required* in an experimental design.

Use of a Control Group

If the purpose of an experiment is to determine whether some treatment has an effect, it is important to include an experimental group that does not receive the treatment. Such a group is called a **control group**. The use of a control group allows the experimenter to assess how the response variable behaves when the treatment is not used. This provides a baseline against which the treatment groups can be compared to determine whether the treatment had an effect.

Single-Blind and Double-Blind Experiments

Because people often have their own personal beliefs about the effectiveness of various treatments, it is desirable to conduct experiments in such a way that subjects do not know what treatment they are receiving. For example, in an experiment comparing four different doses of a medication for relief of headache pain, someone who knows that he is receiving the medication at its highest dose may be subconsciously influenced to report a greater degree of headache pain reduction. By ensuring that subjects are not aware of which treatment they receive, we can prevent the subjects' personal perceptions from influencing the response.

An experiment in which subjects do not know what treatment they have received is described as *single-blind*. Of course, not all experiments can be made single-blind. For example, in an experiment to compare the effect of two different types of exercise on blood pressure, it is not possible for participants to be unaware of whether they are in the swimming group or the jogging group! However, when it is possible, "blinding" the subjects in an experiment is generally a good strategy.

In some experiments, someone other than the subject is responsible for measuring the response. To ensure that the person measuring the response does not let personal beliefs influence the way in which the response is recorded, the researchers should make sure that the measurer does not know which treatment was given to any particular individual. For example, in a medical experiment to determine whether a new vaccine reduces the risk of getting the flu, doctors must decide whether a particular individual who is not feeling well actually has the flu or some other unrelated illness. If the doctor knew that a participant with flu-like symptoms had received the new flu vaccine, she might be less likely to determine that the participant had the flu and more likely to interpret the symptoms as being the result of some other illness.

There are two ways in which blinding might occur in an experiment. One involves blinding the subjects, and the other involves blinding the individuals who measure the response. If subjects do not know which treatment was received *and* those measuring the response do not know which treatment was given to which subject, the experiment is described as **double-blind**. If only one of the two types of blinding is present, the experiment is single-blind.

DEFINITION

A placebo is something that is identical (in appearance, taste, feel, etc.) to the treatment received by the treatment group, except that it contains no active ingredients.

A "fake" treatment that looks just like the treatments being tested is called a placebo. Placebos are the best way to blind subjects from knowing whether they are receiving the treatment or not. One common version of a placebo in drug testing is a "sugar pill." Especially when psychological attitude can affect the results, control group subjects treated with a placebo may show an improvement.

The fact is that subjects treated with a placebo sometimes improve. It's not unusual for 20% or more of subjects given a placebo to report reduction in pain, improved movement, or greater alertness, or even to demonstrate improved health or performance. This placebo effect highlights both the importance of effective blinding and the importance of comparing treatments with a control. Placebo controls are so effective that you should use them as an essential tool for blinding whenever possible.

The best experiments are usually

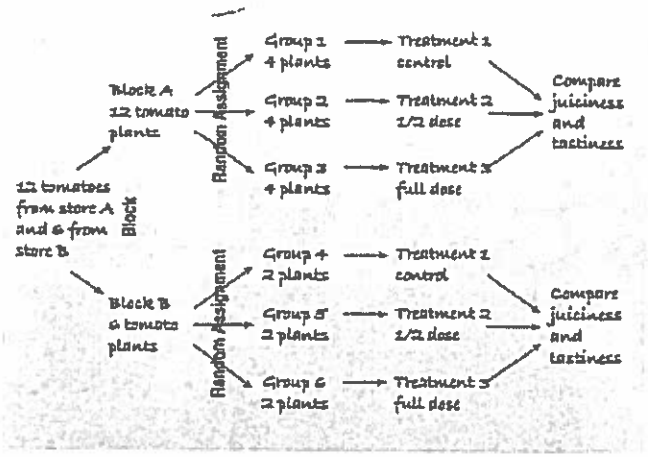
- randomized.
- comparative.
- double-blind.
- placebo-controlled.

AP Statistics Summer Assignment

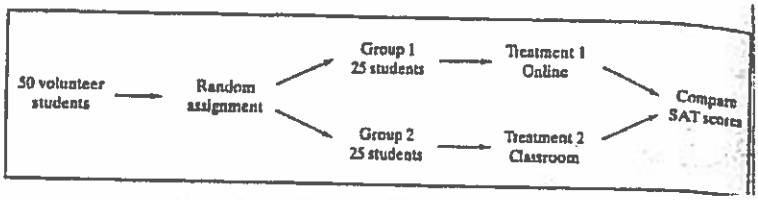
Because stores may vary in the care they give plants or in the sources of their seeds, the plants from either store are likely to be more like each other than they are like the plants from the other store. When groups of experimental units are similar, it's often a good idea to gather them together into blocks. By blocking, we isolate the variability attributable to the differences between the blocks, so that we can see the differences caused by the treatments more clearly. Here, we would define the plants from each store to be a block. The randomization is introduced when we randomly assign treatments within each block.

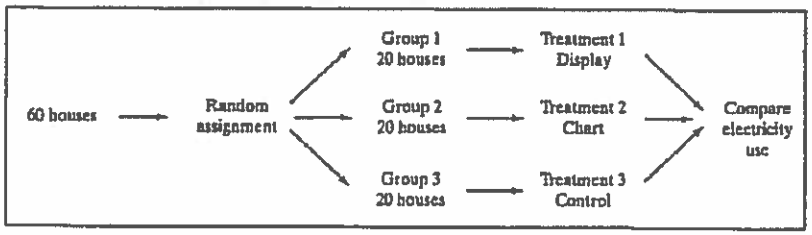
In a completely randomized design, each of the 18 plants would have an equal chance to land in each of the three treatment groups. But we realize that the store may have an effect. To isolate the store effect, we block on store by assigning the plants from each store to treatments at random. So we now have six treatment groups, three for each block. Within each block, we'll randomly assign the same number of plants to each of the three treatments. The experiment is still fair because each treatment is still applied (at random) to the same number of plants and to the same proportion from each store: 4 from store A and 2 from store B. Because the randomization occurs only within the blocks (plants from one store cannot be assigned to treatment groups for the other), we call this a randomized block design.

In effect, we conduct two parallel experiments, one for tomatoes from each store, and then combine the results. The picture tells the story:

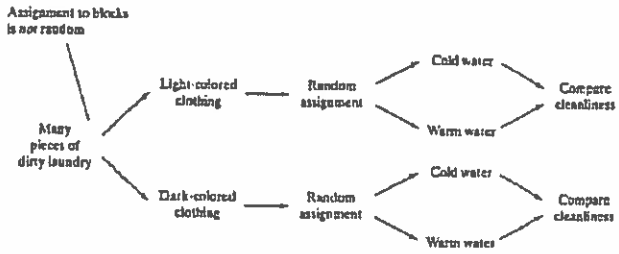


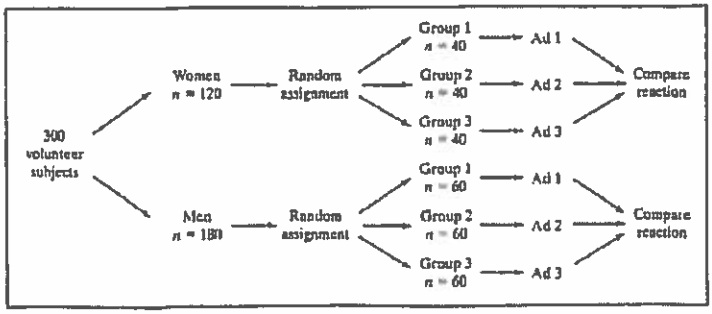
Completely Randomized Design Examples





Blocked Design Examples





AP Statistics Summer Assignment

Summer 2018

Matched Pairs Design

A common type of randomized block design for comparing two treatments is a matched pairs design. The idea is to create blocks by matching pairs of similar experimental units. Then we can use chance to decide which member of a pair gets the first treatment. The other subject in that pair gets the other treatment. That is, the random assignment of subjects to treatments is done within each matched pair. Just as with other forms of blocking, matching helps reduce the effect of variation among the experimental units.

Sometimes each "pair" in a matched pairs design consists of just one experimental unit that gets both treatments one after the other. In that case, each experimental unit serves as its own control. The *order* of the treatments can influence the response, so we randomize the order for each experimental unit.

Try this for fun-sies ☺

ACTIVITY *Get your heart beating*

MATERIALS: Clock or stopwatch

Are standing pulse rates generally higher than sitting pulse rates? In this Activity, you will perform two experiments to try to answer this question.

1. *Completely randomized design* For the first experiment, you'll randomly assign half of the students in your class to stand and the other half to sit. You can use the hat method, Table D, or technology to carry out the random assignment. Once the two treatment groups have been formed, students should stand or sit as required. Then they should measure their pulses for one minute. Have the subjects in each group record their data on the board.

2. *Matched pairs design* In a matched pairs design, each student should receive both treatments in a random order. Since you already sat or stood in Step 1, you just need to do the opposite now. As before, everyone should measure their pulses for one minute after completing the treatment (that is, once they are standing or sitting). Have all the subjects record their data (both measurements) in a chart on the board.

In some cases, it isn't practical or even ethical to do an experiment. Consider these important questions:

- Does texting while driving increase the risk of having an accident?
- Does going to church regularly help people live longer?
- Does smoking cause lung cancer?

To answer these cause-and-effect questions, we just need to perform a randomized comparative experiment. Unfortunately, we can't randomly assign people to text while driving or to attend church or to smoke cigarettes. The best data we have about these and many other cause-and-effect questions come from observational studies.

It is sometimes possible to build a strong case for causation in the absence of experiments. The evidence that smoking causes lung cancer is about as strong as nonexperimental evidence can be.

EXAMPLE

Does Smoking Cause Lung Cancer?

Living with observational studies

Doctors had long observed that most lung cancer patients were smokers. Comparison of smokers and similar nonsmokers showed a very strong association between smoking and death from lung cancer. Could the association be due to a lurking variable? Is there some genetic factor that makes people both more likely to get addicted to nicotine and to develop lung cancer? If so, then smoking and lung cancer would be strongly associated even if smoking had no direct effect on the lungs. Or maybe confounding is to blame. It might be that smokers live unhealthy lives in other ways (diet, alcohol, lack of exercise) and that some other habit confounded with smoking is a cause of lung cancer. How were these objections overcome?

Data Ethics*

Medical professionals are taught to follow the basic principle "First, do no harm." Shouldn't those who carry out statistical studies follow the same principle? Most reasonable people think so. But this may not always be as simple as it sounds. Decide whether you think each of the following studies is ethical or unethical.

- A promising new drug has been developed for treating cancer in humans. Before giving the drug to human subjects, researchers want to administer the drug to animals to see if there are any potentially serious side effects.
- Are companies discriminating against some individuals in the hiring process? To find out, researchers prepare several equivalent résumés for fictitious job applicants, with the only difference being the gender of the applicant. They send the fake résumés to companies advertising positions and keep track of the number of males and females who are contacted for interviews.
- In a medical study of a new drug for migraine sufferers, volunteer subjects are randomly assigned to two groups. Members of the first group are given a placebo pill. Subjects in the second group are given the new drug. None of the subjects knows whether they are taking a placebo or the active drug. Neither do any of the physicians who are interacting with the subjects.
- Will people try to stop someone from driving drunk? A television news program hires an actor to play a drunk driver and uses a hidden camera to record the behavior of individuals who encounter the driver.

The most complex issues of data ethics arise when we collect data from people. The ethical difficulties are more severe for experiments that impose some treatment on people than for sample surveys that simply gather information. Trials of new medical treatments, for example, can do harm as well as good to their subjects. Here are some basic standards of data ethics that must be obeyed by all studies that gather data from human subjects, both observational studies and experiments.