CollegeBoard
connect to college success™

AP® Statistics Module

Sampling and Experimentation: Planning and Conducting a Study

connect to college success™
www.collegeboard.com

The College Board: Connecting Students to College Success

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 4,700 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three and a half million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

Equity Policy Statement

The College Board and the Advanced Placement Program encourage teachers, AP Coordinators, and school administrators to make equitable access a guiding principle for their AP programs. The College Board is committed to the principle that all students deserve an opportunity to participate in rigorous and academically challenging courses and programs. All students who are willing to accept the challenge of a rigorous academic curriculum should be considered for admission to AP courses. The Board encourages the elimination of barriers that restrict access to AP courses for students from ethnic, racial, and socioeconomic groups that have been traditionally underrepresented in the AP Program. Schools should make every effort to ensure that their AP classes reflect the diversity of their student population.

For more information about equity and access in principle and practice, please send an email to apequity@collegeboard.org.

For further information, visit apcentral.collegeboard.com.

Table of Contents

Acknowledgments

Early in Victor Hugo's *The Hunchback of Notre Dame* one of a group of travelers, stumbling upon a possibly less-than-polished play in progress, comments: "What paltry scribbler wrote this rhapsody?"

In clear contradistinction to the unknown playwrights of Hugo's imaginary rhapsody, it is my privilege to identify and acknowledge the scribblers who worked mightily on this real one: Peter Flanagan-Hyde, Roxy Peck, and Dick Scheaffer. It would be difficult to field a writing team with more knowledge of statistics and the AP Statistics program. Over and above their knowledge is the wisdom that can only come from years of classroom teaching and the creative crafting, reworking, and finally perfecting of lessons for students who are beginning their journey of learning statistics.

Every writer knows that works of consequence do not spring forth in one sitting. Each new draft, like the fabled phoenix, springs forth from the ashes of the previous one. No more important force drives the improvement of a written work than the feedback of colleagues; our writing team benefited greatly from colleagues who gave careful reading and generously offered lengthy and detailed suggestions and perfecting amendments. In the earliest stages of writing, Floyd Bullard, Gloria Barrett, and Beth Chance kindly allocated much of their time, combining their own experiences as accomplished writers with careful readings of the evolving manuscript to make it better with each draft.

Jessica Utts and Ken Koehler were the first to read the "complete" manuscript. Their expert comments about technical statistical issues—and excellent wordsmithing suggestions—ensured that yet another and better phoenix would spring forth. Brad Hartlaub and Linda Young were present at both the creation and the denouement, ensuring the fidelity of the final product to the original vision of the AP Statistics Development Committee.

Finally, Susan Kornstein, who has piloted so many publications at the College Board, was ever present and ever helpful during the long journey to completion of this work. A steadier hand at the helm can scarcely be imagined.

Chris Olsen
Cedar Rapids Community Schools
Cedar Rapids, Iowa

From the AP Statistics Development Committee

The AP Statistics Development Committee has reviewed students' performances on the AP Statistics Exam over several years. On the free-response portion of the exam, the students' scores on questions covering content area II, planning and conducting a study, exhibit considerable variability—while some students are doing well in this area, others are having problems. The Committee has discussed possible steps for improving scores in this content area.

Textbooks used in AP Statistics courses often offer an introductory discussion of planning a study, which may be skipped over or simply not revisited later in the course. The AP Statistics Development Committee suggests that teachers present basic concepts of planning a study as early as possible (perhaps even earlier than where the initial discussion in the textbook occurs). These concepts often represent ways of thinking about studies that are very different from what a student has experienced in the past. Some of the ideas are complex and cannot be fully grasped in one, or even a few, sessions. Presenting the ideas early and then returning to them frequently throughout the course will help students acquire the depth of understanding they need to perform well on the AP Exam.

If at all possible, students should actually plan and conduct a study. The data from such a study should be analyzed and the conclusions fully discussed. If the study is conducted early in the school year, an initial analysis may be based solely on graphs and observations made from those graphs. The data can be revisited as the students learn about estimation and hypothesis testing. Consideration should be given to whether the results can be applied to a larger population and to whether causal conclusions can be drawn. Students should always interpret the results and state their conclusions in the context of the study.

Published accounts of studies are a valuable classroom resource. The students should first attempt to discern how a study was actually conducted. Newspapers often provide only sketchy descriptions of designs, and the students may be asked to propose one or more plausible designs consistent with what is reported. When two or more designs have been suggested, students should explore the strengths and weaknesses of each. We also recommend following up with a discussion of appropriate methods of analysis.

We hope that this publication will be a valuable resource for teachers. It emphasizes the concepts that underlie survey sampling and the design of studies, and numerous examples are provided. These concepts are often new to AP Statistics teachers and yet are fundamental to statistical understanding.

Results from studies are reported daily. To become informed consumers, students need to gain a basic understanding of study design and the inferences that can be legitimately drawn. AP Statistics provides an excellent opportunity for students to learn these concepts, and this publication, along with its suggested activities, should be a great help in this quest.

2005-2006 AP Statistics Development Committee

Committee Chair
Kenneth J. Koehler, Iowa State University; Ames, Iowa

Committee Members
Gloria Barrett, North Carolina School of Science and Math; Durham, North Carolina
Beth Chance, California Polytechnic State University; San Luis Obispo, California
Kathy Fritz, Plano West Senior High School; Plano, Texas
Josh Tabor, Wilson High School; Hacienda Heights, California
Calvin Williams, Clemson University; Clemson, South Carolina

Chief Reader
Brad Hartlaub, Kenyon College; Gambier, Ohio

Introduction to "Sampling and Experimentation"

Chris Olsen
Cedar Rapids Community Schools
Cedar Rapids, Iowa

AP Statistics, developed in the last decade of the twentieth century, brings us into the twenty-first. In the introduction to the *AP® Statistics Teacher's Guide*, two past presidents of the American Statistical Association, David S. Moore and Richard L. Scheaffer, capture the importance and the vision of AP Statistics:

> The intent of the AP Statistics curriculum is to offer a modern introduction to statistics that is equal to the best college courses both in intellectual content and in its alignment with the contemporary practice of statistics. This is an ambitious goal, yet the first years of AP Statistics Examinations demonstrate that it is realistic.

Over a thousand mathematics teachers have responded to the challenge of delivering on this "ambitious goal"! Many have dusted off their old statistics books, taken advantage of College Board workshops, and even signed up for another statistics course or two.

While questions and discussions on the College Board's AP Statistics Electronic Discussion Group, at workshops, and during the AP Statistics Reading cover many topics in the course, teachers are particularly interested in the area of "Sampling and Experimentation: Planning and Conducting a Study."

Why Is the Topic of Planning and Conducting Studies Problematic?

It is not particularly surprising that the topic of planning studies should be somewhat unfamiliar to mathematics teachers, even to those who have taken more than one statistics course. For undergraduate mathematics majors, the first—or even the second—statistics course probably focused on data analysis and inference; if their statistics course(s) were calculus based, probability and random variables may have been well covered. Planning studies was apparently to be learned in more advanced courses, or possibly in "tool" method courses with syllabi specific to disciplines such as engineering or psychology—or perhaps science classes. Therefore, many undergraduate majors in mathematics did not experience, in classes or elsewhere, the planning of scientific studies.

This dearth of preparation led to a heightened concern by mathematics teachers as they prepared to teach AP Statistics and confronted the area of study design. Here is the introduction to the *AP® Statistics Teacher's Guide*'s description of the role of study design:

> Designs for data collection, especially through sampling and experiments, have always been at the heart of statistical practice. The fundamental ideas are not mathematical, which perhaps explains why data collection, along with [exploratory data analysis], was slighted in [many elementary statistics texts]. Yet the most serious flaws in statistical studies are almost always weaknesses in study design, and the harshest controversies in which statistics plays a role almost always concern issues where well-designed studies are lacking or, for practical or ethical reasons, not possible.

The heart of statistical practice? Not mathematical? Clearly, those who planned to teach AP Statistics were in for some serious preparation. Unfortunately, the new emphasis on planning and conducting studies did not arrive complete with easily accessible materials for the new teachers. Excellent texts on sampling, experimental design, and research methods exist, but their target audiences are usually current and future professional statisticians and researchers. While these textbooks are detailed and thorough enough for college courses and even future teachers, they do not necessarily speak to the needs of the current high-school mathematics teacher. Thankfully, recent statistics books, written with an awareness of the AP syllabus, have included introductory discussions of sampling and experimental design. But even the best book for AP Statistics students cannot be expected to present them with the topic and simultaneously provide a grounding for teachers sufficiently solid to respond to those "tough" student questions.

The Scope and Purpose of This Publication—and a Small Warning

The purpose of this publication is to provide a resource for veteran as well as less-experienced AP Statistics teachers. Beginning AP Statistics teachers may see the terminology and procedures of experimental design as an incoherent list of mantras, such as, "Every experiment must have a control group and a placebo," "Random assignment to treatments is an absolute necessity," and so on. Experienced AP Statistics teachers may have questions about planning studies, or may wish to extend their knowledge beyond what is presented in the textbook. Seasoned veterans will have different, perhaps more-specific questions than the others about the enterprise of planning studies. Answers to these more advanced questions are not found in elementary statistics books but rather appear to be cleverly tucked away in advanced statistics books intended for the future professional statistician.

We hope to introduce the idea of a scientific research study and to provide a coherent foundation regarding current scientific methodology and how its concepts apply to experimental studies. This background in science and philosophy might seem only tenuously related to the classroom experience of teaching statistics. However, the modern experiment did not suddenly appear on the scientific scene; it is the fruit of at least four

centuries of thought about how humans can effectively and efficiently create objective knowledge. With both beginning and experienced AP Statistics teachers in mind, struggling to understand the "big picture" of planning and conducting a study, we will provide the logical framework around which today's scientific experiment is designed, as well as justify its place of importance in the world of science.

As mentioned earlier, we have the two slightly competing goals of introducing sampling and experimental design to beginning AP Statistics teachers and extending veterans' understanding of it. Because this is a single presentation for both audiences, we expect some readers will experience difficulty wading through the material. Beginning AP Statistics teachers, too, will undoubtedly find it a bit daunting in spots, especially if they have not yet completely grasped the fundamentals of random variables. On the other hand, veterans may need to stifle a yawn now and then.

For teachers at all points on the spectrum of understanding study design, we offer the following strategy. On your first reading, hit the high points and survey the panorama. On your second reading, pick out those aspects of the writing that are of greatest interest or present your greatest challenge, and skip over the remainder. As you teach AP Statistics and your comfort level rises, take this presentation out and reread it. You will see that what you find interesting and challenging will change with your own growth as an AP Statistics teacher. The words will remain the same, but what you read and understand will be different with successive readings.

How This Publication Is Organized

Our general strategy is to present a global view of the research process, outlining the scope of research questions and methods. We first present modern experimental design as an evolution of thought about how objective knowledge of the external world is "created" and why its process is necessarily a bit complicated. Then we focus on the topics of sampling and experimental design. Our exposition extends somewhat beyond the content of the AP Statistics syllabus. We do this not to make professional statisticians out of AP Statistics teachers but rather to increase their comfort level with respect to these topics. Finally, we provide some tips and materials to help implement these ideas in the classroom. It is our hope that as AP Statistics teachers expand their knowledge and facility, these materials will provide them with a template for developing their own teaching materials. We fully expect to see teachers moving beyond the particular contexts we have chosen, writing and sharing exciting activities with their students, and thus providing AP Statistics students with a wealth of experiences as they learn to view the world through the critical lens of statistical thinking.

The Historical and Philosophical Background of Scientific Studies

It is in the nature of cats, deer, and humans to be curious about the events around them. Even toddlers behave like little scientists as they attempt to make sense of their surroundings, translating the chaos of their early experiences into a coherent structure that eventually makes "logical" sense. Each youngster conducts informal experiments, testing causal theories and explanations, and, as language skills develop, is quite happy to share the fruits of those experiments with other "budding scientists." Eventually, the world seems to make sense, more so as formal education begins. Children learn what causes a light to go on, what makes an interesting sound, and—of course—what it takes to "cause" their parents to give them what they want!

The natural development of children's "scientific" behavior parallels the development of scientific knowledge and methodology for the past two millennia. It is helpful to divide the development of scientific thought and experience into before and after periods, with the pivotal point being the beginning of the Scientific Revolution of the seventeenth century.

From the time of medieval scholastics and the rediscovery of Aristotle's writings, observations were used not to generate new knowledge but to support what was already assumed to be true—by virtue of the authority of religious or philosophical authority. As the story goes, the learned men of Galileo's time had no need to look through his telescope—they already *knew* that the universe was geocentric.

Of course, since the beginning humans have been making observations and advancing knowledge. The domestication of plants and animals, surely by chance processes of observation and trial and error, testifies to the early use of the powers of human observation. However, observations unaccompanied by systematic methods guaranteed a very slow accretion of knowledge.

Francis Bacon (1561–1626) is generally credited with the invention of modern science, arguing that casual observation simply is not sufficient to establish knowledge of the objects in our world and of the relationships between those objects:

> It cannot be that axioms established by argumentation [that is, arguments using appeal to authoritative writings] should avail for the discovery of new works; since the subtlety of nature is greater many times over than the subtlety of argument. But axioms duly and orderly formed from particulars easily discover the way to new particulars, and thus render sciences active.
>
> —*Novum Organum* (The New Logic), Aphorism XXIV

Hacking (1983) writes of Bacon, "He taught that not only must we observe nature in the raw, but that we must also 'twist the lion's tail,' that is, manipulate our world in order to learn its secrets." It was Bacon's view that while passive observation can tell us much about our world, active manipulation gives us much cleaner knowledge and is much more efficient. After Bacon's time, the word "experiment" came to signify a deliberate action followed by careful observation of subsequent events. This usage is still common in high school science classrooms, but in the larger scientific community a finer distinction is used, based on the particular scientific methodology—i.e., how the observations are made. The Baconian distinction between mere passive observation and twisting the lion's tail is preserved today in the separation of scientific studies into two major groups: observational studies and experiments.

Observational Studies Versus Experiments

It is sometimes said that the difference between observational studies and experiments is that experiments can establish causation and observational studies cannot. This view is an echo of Bacon's distinction between observing and manipulating the environment, but it is to some extent incomplete. In search of causal relationships, *both* observational and experimental studies are used, and used effectively (Holland 1986). In most medical and psychological studies, experimentation presents serious ethical problems. For example, we would not be able to study the effects of severe head trauma by creating two groups and randomly assigning abusive treatments to some! We could only place ourselves in an emergency room, wait for the head-trauma patients to arrive, and then make our observations.

The difference between conclusions drawn from observational and experimental studies is not whether causal relationships can be found, but how efficiently they can be found and how effectively they can be supported. Causal relationships are much easier to support with an experimental study. Establishing causal relationships requires a combination of observations from many studies under different conditions—experimental and/or observational—as well as thoughtful, logical argument and consideration of the existing scientific knowledge base. Much of the logical argument focuses on the implausibility of alternative explanations for X causing Y. We shall see that experimental studies are far superior to observational studies.

In order to better distinguish between observational studies and experiments, consider this example. Suppose a man has a high fever. His physician, up on all the latest medicine, suggests taking some vitamin C. The patient does so, and his temperature falls to an acceptable level. An investigator might ask, "What would have happened had the vitamin C route not been taken?" And the doctor would reply, while writing out his bill, "Why, had that not been done, the temperature would still be high!" Why does the doctor

believe that taking vitamin C causes a drop in temperature? Most likely, this belief is derived from his or her accumulated experience and observation of patients. The physician, over the years, has associated these two events. Association, however, is not the same thing as causation! Suppose that we (or the physician) wish to verify that the drop in temperature occurs *because of* vitamin C. A natural strategy might be to gather together lots of patients with high temperatures and see if taking vitamin C consistently reduces their fever, with their fever remaining (or going higher) if they go without.

We want to examine what *did* happen after the vitamin C—compared with what *would have* happened without vitamin C. Specifically, we want to know whether taking vitamin C causes elevated temperatures to subside, where such temperatures would not have otherwise subsided. It is impossible to make this observation on a single person since one cannot simultaneously take the vitamin C and *not* take the vitamin C. We can, however, calculate the average change in temperature for a group taking vitamin C and compare it with the average change in temperature for a group not taking vitamin C. To do this, we might call our local medical association and ask the doctors to cooperate in a survey by providing initial temperature readings for the next 200 people who come in with high temperatures, instructing all 200 to get plenty of rest, giving just half of them vitamin C tablets, and having all 200 call back in the morning with their new temperature reading. The doctors, if they agree, will then record the temperature change for both those who took the tablets and those who did not. To measure the effect of the vitamin C treatment, we would calculate the difference between the mean temperature change of the people who took vitamin C and the mean of those who did not.

One crucial feature will determine if a study is an observational one or an experiment: Is there active intervention? That is, does someone *decide* which patients will get the vitamin C? If the doctor or investigator performing the study determines who gets which treatment (vitamin C or no vitamin C), then it is an experiment. If treatments are not assigned and existing groups are merely observed, there is no intervention, and so the study is purely observational.

The term **experiment** can carry qualifying adjectives. For example, since in this example we are comparing two groups of people, we can call it a **comparative** experiment. And if we choose who gets which treatment by flipping a coin or using some other random device, we can call it a **randomized** experiment. (Some authorities require that the assignment to treatments be randomized before a study legitimately can be termed an "experiment.") Further, by attempting to lessen the potentially disrupting effects of other variables by having both groups of patients get plenty of rest, and by reducing variability across individual patients by considering only those with initially high temperatures, we

say that we are attempting to **control** other variables that might affect patients' recovery from an elevated temperature.

Random Selection Versus Random Assignment

There is another distinction to make about the role of randomness in experiments. R. A. Fisher's work in the mid-1920s is largely responsible for the imposition of chance in the design of a scientific study—now considered a crucial element in the process of making statistically based inferences. Randomness is associated with the two common inferences that we make in AP Statistics: (1) inferring characteristics of a population from a sample and (2) inferring a cause-effect relationship between treatment and response variables.

In a scientific study whose goal is to generalize from a sample to a larger population, **random selection** from a well-defined population is essential. This is typically referred to as **random sampling**. Random sampling tends to produce samples that represent the population in all its diversity. It is certainly possible to get a "bad" sample, but luckily we know in advance the probability of getting such a nonrepresentative sample.

In a scientific study with a goal of inferring a cause-effect relationship, **random assignment** to treatments provides the underlying mechanism for causal inferences, as we shall see in the paragraphs to follow. Random assignment tends to produce treatment groups with the same mix of values for variables extraneous to the study, so that the different treatments have a "fair chance" to demonstrate their efficacy. When treatment groups have the same (or at least very similar) mixes of values for extraneous variables, no treatment has a systematic advantage in an experiment. As with random selection, it is certainly possible to get a "bad" assignment of treatments, but again, random assignment allows us to gauge the probability of getting dissimilar treatment groups.

Here is a summary of the inferences based on considerations of random selection and random assignment:

- If neither random selection nor random assignment to treatments is performed, there is virtually no statistical inference that can be made from the study. Only information about the sample can be described.
- If random selection from a known population is performed, one may infer that the characteristics of the sample tend to mirror corresponding characteristics of the population.
- If random assignment of treatments is performed, cause-effect inferences may be drawn about the results within the sample(s) at hand.

- If both random selection and random assignment to treatments are performed, one may draw cause-effect inferences about the sample results, as well as generalize to the larger population from which the sample was drawn.

Ramsey and Schafer (2002) recapitulate these considerations wonderfully in Table 1:

Table 1: Selection and Assignment

		Assignment of Units to Groups		
		By Randomization	**Not by Randomization**	
Selection of units	**At random**	A random sample is selected from one population; units are then randomly assigned to different treatment groups.	Random samples are selected from existing distinct populations.	**Inferences to the populations can be drawn.**
	Not at random	A group of study units is found; units are then randomly assigned to treatment groups.	Collections of available units from distinct groups are examined.	
		Causal inferences can be drawn.		

It might seem, therefore, that observational studies involving neither random sampling nor random treatment assignments are useless. In reality, most discovery is exploratory by nature, and purely observational studies are quite common—indeed, much can be learned from them. They certainly can suggest causal relationships and stimulate the formation of hypotheses about features of a population. But inferring *beyond* the sample is subjective, based on the belief (i.e., not statistically supported) that the sample is "representative" of a larger population, or that the treatment groups are "similar to one another" in nontreatment aspects. Such inferences about a causal relationship have no statistical support! Even "reasonable-sounding" inferences from such observational studies must be made and viewed very critically. Randomized experiments are the most effective way to examine well-formulated questions or hypotheses based on observation. And this type of confirmatory investigation may be developed only after extensive exploratory investigation using observational studies.

Observation and Experimentation: Understanding the What and the Why

Sampling and the assignment of treatments are at the heart of planning and conducting a study. In a scientific study, deciding how to handle the problems of sampling and assignment to treatments dictates the sort of conclusions that may legitimately follow from a study. We will now consider some representative studies in light of the above discussion.

Both observational studies and surveys are considered descriptive studies, while experiments are generally designed to be explanatory studies. Descriptive studies are sometimes used simply to describe a sample or a population but may also explore relationships for further study. A descriptive study, such as a survey, is one in which characteristics of a sample derived from a population are observed and measured. While some measures can be taken unobtrusively (i.e., from hiding!), a common methodology for exploratory studies is direct observation in a laboratory or, in the case of human subjects, asking them to respond to survey questions. For example, if an investigator is interested in the occurrence and kinds of bullying at a local school, he or she might walk around the halls, classrooms, and lunchrooms, observing student behavior. Alternatively, the investigator might write a set of questions and use them to solicit information from local students and teachers. The presentation of such a study's results might describe the frequency and types of bullying behaviors or the times and places they are most prevalent. Associations between the variables might be discovered. Perhaps bullying among boys tends to occur in the afternoon and is more physical, whereas bullying among girls is more of a morning event and tends to be verbal.

The purpose of a descriptive study is to observe and gather data. In the example above, no particular theory of bullying behavior need be hypothesized in advance; information is being gathered at a particular point in time, and the resulting data are analyzed with the usual univariate and bivariate descriptive statistics. For this purpose, the nature of the sampling is very important. To be able to reliably generalize from observed features of the sample to the larger, mostly unobserved population, the investigator needs the sample data to mirror that of the population so that the descriptions of the sample behaviors are dependably close to the "true" state of the population of behaviors.

Some descriptive studies focus on relationships between variables, without concern for establishing causal relationships. While we may not be able to manipulate one variable in a potential causal chain—or even have a clear understanding of the relationship between variables—we still may be able to capitalize on a known and stable association between variables by identifying the potential existence, form, and strength of relationships between and among variables. For example, does family size appear to be related to family

income? Is success in college related to scores on standardized tests such as the SAT and ACT? These questions address associations between variables.

If we can establish the direction and strength of a relationship between two variables, we may "predict" the value of one variable after measuring the other variable. In some cases, this "prediction" is simply the substitution of an easily measured variable for one that is harder to measure. In other cases, we may actually be attempting to "predict the future."

Imagine, for example, that we want to sample salmon returning to spawn in order to determine their total biomass as it relates to the following year's fishing regulation. Weighing the salmon presents the problem of catching them—and the ones you manage to catch may be the slower or less slippery ones, which may mean they are systematically different from the typical salmon. Also, the salmon might not sit idly by on a scale, but rather flop around and generally make measurement attempts next to impossible. Luckily, individual salmon have very similar shapes, and their lengths and masses can be modeled approximately with a simple mathematical function. An ideal measurement strategy would be to video salmon passing through a glass chute on their way upstream—a glass chute with a scale drawn on it. From individual length measures, individual mass measures could be "predicted," using the mathematical model of the relationship between these two variables.

The ability to predict the future within a small margin of error can also be very useful in decision-making situations such as college admissions. If an applicant's SAT score can effectively predict his or her first-semester GPA, a college may use the SAT (as well as other information) in its admissions decisions. If crime rates can be reasonably predicted using measurable economic factors such as average wages and unemployment figures, a government might decide to allocate additional resources to police departments in cities with low wages and high unemployment, without caring "why" the variables are related. No particular "causal mechanism" is needed for such a predictive study—only associations between and among the variables. For useful prediction, all that is needed is that the relationship be strong and stable enough over time.

Descriptive studies of relationships may—or may not—suggest possible causal relationships for future study, but the examples above are still observational, and the issues of sample selection are of paramount concern. Regardless of any causal relationship, the descriptive information is only as useful as the sample is a near mirror image of the population—and a valid sampling plan is what establishes the credibility of the mirror image and provides a basis for quantifying uncertainty.

Causation: Establishing the Why

Through appropriate intervention—a well-designed experiment—we address the problem of demonstrating causation. When an investigator conducts an experiment, he or she has one or both of the following goals: (1) to demonstrate and possibly explain the causes of an observed outcome and (2) to assess the effect of intervention by manipulating a causal variable with the intent of producing a particular effect. Explaining or identifying the causes of an observed outcome may serve just "pure science," or it may be part of a larger scientific effort leading to an important application of the results. When conducting an experiment, an investigator proceeds from data to causal inference, which then leads to a further accumulation of knowledge about the objects and forces under investigation.

As illustrated earlier in Table 1, explaining the why requires, *at the very least*, random assignment to treatments and in most cases suggests the use of additional experimental methodology. In some experimental studies, an investigator also will wish to establish that the causal mechanism that is demonstrated experimentally will generalize to some larger population. To justify that additional claim, selection in the form of random sampling is required once again.

Even assuming for the moment that X causes Y, establishing that relationship is not guaranteed by mere random assignment to treatments. Random assignment to treatments *plus* sound experimental methodology, topped off by a healthy dose of logic—together this allows an inference of a causal relationship between X and Y. To give a "big picture" perspective of the process of establishing causation, we will review some of the history and philosophy of science relevant to the question, "Just how is it that empirical methods—observation and experimentation—permit the claim of a causal relationship to be justified?"

Our big-picture tour begins with mathematician Rene Descartes. Descartes, a rationalist, argued that to acquire knowledge of the real world, observations were unnecessary; all knowledge could be reasonably constructed from first principles—i.e., statements "known" to be true. For the rationalists, a demonstration that A causes B was only a matter of logic. In contrast, seventeenth-century empiricists such as John Locke argued that the human mind at birth was a blank slate, and that all knowledge of the outside world came from sensory experience. For the empiricists, causation could be demonstrated only by showing that two events were "associated," and this could be done only by gathering data. This difference of opinion was settled for all intents and purposes by David Hume, an eighteenth-century Scot, who showed that neither logic nor observation alone can demonstrate causation. For Hume, past observations of an association did not constitute a demonstration of causation. This assertion still survives in

the form of one of the most prevalent and recognizable dictums in all of statistical science: **correlation does not imply causation**.

In the nineteenth century, philosopher John Stuart Mill addressed the problem of how a scientist might, as a practical matter, demonstrate the existence of causal relationships using observation and logic—i.e., what sort of evidence would be required to demonstrate causation? In Mill's view, a cause-effect relationship exists between events C ("cause") and E ("effect") if:

1. C precedes E;
2. C is related to, or associated with, E; and
3. No plausible alternative explanation for E, other than C, can be found.

These three requirements are closely mirrored in the accepted methodology of the modern scientific experiment:

1. The investigator manipulates the presumed cause, and the outcome of the experiment is observed *after* the manipulation.
2. The investigator observes whether or not the variation in the presumed cause is associated with a variation in the observed effect.
3. Through various techniques of experimental "design," the plausibility of alternative explanations for the variation in the effect is reduced to a predetermined, acceptable level.

The "gold standard" for demonstrating causal relationships is, quite aptly, the well-designed experiment. Observational studies, by contrast, can provide only weak evidence for causation, and such claims of causation must be supported by very complex and context-specific arguments. The possible inferences that can be made based on a study—whether an investigator is planning a study or reading the results of someone else's—will be determined by the answers to these questions:

- How were the observed units selected for the study?
- How were the observed units allocated to groups for study?
- Was an association observed between variables?
- Was there an identifiable time sequence, i.e., C preceding E?
- Are there plausible alternatives to concluding that C "caused" E?

Below are a few examples of studies that will help illuminate some of these five questions. The studies were performed by investigators in a variety of fields and are gleaned from the scientific literature. As you read and study them a bit, begin to ask yourself each of the above questions. With a little practice, this kind of analysis will become a natural part of how you understand and plan studies.

Examples of Studies

Example 1: Establishing an Association/Difference—a Survey

Our first example concerns the study of the ecology of black-tailed jackrabbits (*Lepus californicus*) in northern Utah. These animals are unpopular with local farmers because they tend to graze in fields and reduce crop values. To measure the extent of the jackrabbit problem, ecologists need good methods of estimating the size of the jackrabbit population, which in turn depends on good methods for estimating the density of the jackrabbits. Estimating population size is frequently done by first estimating the population density and then multiplying. Population density is estimated by sampling in well-defined areas called transects. Then the jackrabbit (JR) population size is estimated by solving the following equation:

$$\frac{\text{JR in transect}}{\text{Area of transect}} = \frac{\widehat{\text{JR}}}{\text{Population area}}$$

Two common transect methods are to (1) walk around in a square and count the number of jackrabbits seen or (2) ride in a straight-line transect on horseback and count the number of jackrabbits seen. In both methods, the jackrabbits are counted as they are flushed from cover. This study was conducted to see if using these two different transect methods produced different results.

The investigators, Wywialowski and Stoddart (1988), used 78 square transects drawn from a previous study and also randomly located 64 straight transects in the original study area. The two transect methods were compared over a two-year period, with the results shown in Table 2.

Table 2: Transect Data

Year	Method	*N*	Estimated Density (Jackrabbits/sq. km)	Std. Error	95% CI
1978	Walk/square	49	25.8	4.8	16.4–35.2
1978	Ride/linear	108	42.6	4.8	32.2–51.9
1979	Walk/square	138	70.6	8.0	54.9–86.3
1979	Ride/linear	218	124.4	10.3	104.3–144.5

On the basis of these results, the investigators concluded that densities estimated from straight horseback transects do differ from densities estimated from square walked transects. They also felt it reasonable to conclude from their statistics that the density estimates using the two methods also differ. The individual estimates for each walk and

ride were assumed to be independent, since jackrabbits move widely over a large area; in addition, the transect lines and areas had been randomly chosen.

Notice that in both years (1978 and 1979) the straight horseback transects produced higher estimates of jackrabbit density than did the square walked transect. The investigators speculated that an observer up on horseback is probably better able to see the jackrabbits flush from cover. He or she is also able to pay greater attention to counting (since the horse, rather than the rider, looks down to avoid obstacles). However, this speculation is not the same as logical proof of a causal relationship between transect method and results. Recall that the transect methods were not randomly assigned to the areas observed: the square transects were previously existing—only the line transects were new to the study. Indeed, the square transects were located so that the observer "could return to [their] vehicle upon completion of each transect." It is entirely possible that the square transects, being accessible by gas-fed vehicle, are somehow significantly different from the line transects that are accessible by oat-fed vehicle. This hypothesized difference is confounded with the transect methods, possibly producing a biased result. Because the transects themselves were randomly chosen (from all possible areas accessible by vehicle), it *does* appear that the results could be generalized to larger populations of transects.

In summary, investigators in this study found a difference between the two methods. And they were convinced that the difference in samples reflected a real difference in the populations represented by the two methods. However, they could not conclude that the difference was due to the linear transect versus the square transect. They found instead that other factors that differed for the two types of transects—such as vehicle accessibility—may have caused the difference in results.

Example 2: Establishing an Association and a Time Sequence I—The Prospective Study

One study strategy that can establish an association and also has an unambiguous time sequence is variously known in scientific literature as **prospective**, **cohort**, or **follow-up**. In a prospective study, individuals are selected at a single point in time, data are gathered, and then follow-up data are gathered in the future. This type of study is popular in epidemiology, where health risks cannot, ethically, be randomly assigned. A prospective study generally aims to identify a group at a point in time and then measure a variable that might affect future events. For example, young adults might be asked if they currently smoke. Then, at some future time, they might be asked about significant lung or heart events. The idea is to investigate the possibility that smoking may "cause" an increased risk of such health problems.

The chief explanatory advantage of a prospective study lies in its time sequence; there is no doubt which of the associated variables came first! With a prospective study not only can we establish an association, but we also know the time sequence of the variables.

Consider the following example, a prospective study concerning asthma. It appears that many factors play a role in the onset of childhood asthma, some of which are thought to be environmental exposures, especially during infancy. One theory is that early exposure to allergens may activate a child's genetic predisposition to allergic disease. Cats and dogs are two very common environmental culprits. Might it be possible that exposure to these animals in infancy increases the risk of subsequent allergy troubles? Or, as some studies suggest, might it be that early exposure actually *decreases* the risk of subsequent allergy?

In a recent prospective study, Ownby, et al. (2002) investigated just such a relationship between early exposure to cats and dogs and the risk of children's future allergies to them. Investigators interviewed HMO-enrolled pregnant women living in suburban Detroit regarding their level of education, presence of allergies (specifically hay fever and asthma), and smoking habits. When their infants were 1 year old, the mothers were contacted again and asked about the presence and number of pets in the home—including cats and dogs—during the first year of the child's life. Then, when the children were between six and seven, they were tested for allergic sensitization and asthma.

The study results appear to be in favor of cats and dogs! The proportions of children with allergen-specific sensitivity as determined by two different procedures are given in Table 3.

Table 3: Percentage of Children with Positive Reactions to Pet Allergens at 6 to 7 Years of Age

Test for Reaction	No Exposure to Cat or Dog	Exposure to One Cat or Dog	Exposure to Two or More Cats or Dogs
Skin prick test	33.6	34.3	15.4
Blood test	38.5	41.2	17.9

From these data, what can be inferred about the presence of cats and dogs "causing" a resistance to allergies? For both of the allergen-sensitivity tests, the proportions of children who were exposed to two or more cats or dogs are significantly greater than the proportions for no exposure and exposure to one cat or dog. The results would seem to be in favor of having two or more of these pets.

Thus there does appear to be an association between the proportions of youngsters who test positive and their prior exposure to two or more cats or dogs. Furthermore, it seems that a case could be made that the extensive prior exposure *caused* the protection. The selection of women in the study controlled for such variables as education, smoking habits, and their own history of allergies, which should result in a homogeneous group. The time sequence is certainly correct; the potential exposure to the animals would occur before the testing for sensitivity to allergens, so it would seem the allergens could not "cause" the exposure to allergens. It might seem a stretch to view the association between exposure to animals and a positive test as the result of being allergic, but we might construct at least a plausible chain of reasoning as follows. Suppose, for example, that allergies to ragweed or pollen appear earlier in life. Parents might keep their youngsters inside more often, getting them pets to compensate for being cooped up. The allergies to cats and dogs subsequently develop not because of the cats and dogs in the house but because of the children's prior disposition to allergies in general.

One key difficulty (from a causal standpoint) of the Ownby study design is that the investigator could not randomly assign treatments. Clearly it would not be feasible for investigators to force parents and children to live with some specific number and combination of pets for six to seven years! From a study-design standpoint, this presents the possibility of uncontrolled confounding variables. In addition, families with pets could differ in some other way that is related to allergens. For example, perhaps they travel more and thus expose their children to a wider variety of allergens, increasing the likelihood of triggering allergic reactions in general. Or it could be that some of the households had fewer pets because the father was allergic to them. If allergies have a genetic component, then the children in those households would be more likely to

develop an allergic reaction. Thus there could be a third confounding variable—the father's disposition to allergies—that causes both lower pet ownership and higher susceptibility to allergies in the children.

While the association certainly seems enticing, we cannot proceed to make a statistical claim in support of causality. It is also unclear what population we would be able to generalize to. The greater the attempt to enroll a homogenous group of pregnant women in the study, the less potential there is for generalization. It is certainly possible that pregnant women in the Detroit suburbs, and/or women in HMOs, are not representative of pregnant women across the country.

Example 3: Establishing an Association and a Time Sequence II—the Retrospective Study

A second strategy that can nail down John Stuart Mill's requirement for association and the correct time sequence is what is known as a **retrospective**, **case-control**, or **case-history** study. In a retrospective study, an investigator notices the potential effect first, and then looks backward in time to locate an association with a variable that might be a potential cause of an effect.

The case-control design is used primarily in the biomedical field, and it is probably epidemiologists' greatest contribution to research methodology. It is not hard to see why case-control methodology would appeal to epidemiologists, whose major research questions revolve around understanding what has caused their observed effects—such as an outbreak of some newly discovered exotic disease or a spate of food poisoning among pedestrians. The case-control design has many practical advantages over, say, a prospective study. First, diseases or other health events that occur very infrequently can be efficiently investigated, whereas the corresponding cohort study would require a massive number of people for an initial sample. Second, a disease that takes a long time to develop can be studied in a shorter amount of time with less expense than might be necessary with a prospective study.

The case-control design also has some disadvantages. First, the possible causes of the event must be identified *after* the event appears. If the event occurs with a significant period of latency, as frequently happens with such health problems as heart disease and cancer, the list of potential causes can be very long. The second major problem with case-control studies is that the sampling is taken from two different populations. Recall that the "effect" in a cause-effect relationship is theoretically the difference between what happened after the appearance of the alleged cause and what would have happened to the *same* population absent the alleged cause. In a prospective study, the initial sample and its characteristics are established by sampling from a single population. But in a retrospective study, the investigator must artificially create an appropriate second population equivalent to the population that has experienced the event of interest, and there is no technically valid way to do this; it is always a matter of judgment.

Our example of a retrospective study involves observations to ascertain a possible cause of a serious health risk—if you are a gazelle. In the Serengeti National Park in Tanzania, cheetahs hunt Thomson's gazelles (*Gazella thomsoni*), stalking their prey and generally moving to within 30 meters of the selected victim before chasing it for a distance shorter than 300 meters. Once the cheetah chooses its victim and starts the chase, it seems not to be dissuaded from its choice by another, slower-moving gazelle.

Fitzgibbon (1989) conducted a retrospective study to look into the question of cheetah choice. What is it that "causes" them to choose one gazelle over another? Her research question was, "Would a stalking cheetah be more likely to pick a gazelle that was not looking for danger?" Grazing gazelles are generally engaging in one of two activities: munching on grass or looking for predators. During the stalk, cheetahs can assess the behavior of a gazelle and may increase the likelihood of a successful kill by picking out a gazelle that is grazing more and gazing less.

In her study, Fitzgibbon filmed 16 stalks and analyzed the choice behavior of the cheetah. After each stalk, Fitzgibbon matched the gazelle selected by the cheetah (the "case") with the nearest actively feeding, same-sex adult within a five-meter radius and at the edge of a group of gazelles (the "control"). If gazing versus grazing time is a factor in prey selection, on average the gazing percentage should be less for the selected victim than for the nearest neighbor who, after all, just as easily could have been chased. Fitzgibbon tested the hypothesis that the mean gazing percentage of selected gazelles and nonselected gazelles is equal—in other words, that the cheetah does *not* tend to select the least-vigilant prey. Note that each selected prey is compared to its nearest neighbors, the selected and ignored gazelles are not independently chosen in the sampling process, and hence paired-*t* procedures were used in the analysis. The hypothesis was rejected ($t = 3.62$, $p < 0.005$, $df = 15$) in favor of the alternate hypothesis, that the selected gazelles *are* gazing less than the unselected ones.

What can be said with respect to inference in this particular study? It seems reasonable to suggest that the stalks and kills filmed by Fitzgibbon are representative, though technically they are not a random sample. On the other hand, as with the jackrabbits study, these kills were filmed from the safety of an SUV—in plain view, a slow-moving, nonvigilant SUV!—suggesting that cheetah kills in general could unfold differently without an SUV present. Is the prey selection "caused" by a gazelle's grazing rather than by its gazing behavior? The data are certainly consistent with that view, but remember: The pairs of gazelles were not assigned to the gazing-versus-grazing treatments randomly. There may be other characteristics associated with grazing that are relevant to the cheetah choices. Perhaps, for example, younger, older, and/or weaker gazelles have a greater need for food and are less finicky about their surroundings, leading them to graze more. Cheetahs could be picking up on these characteristics, representing other possible causes for the animal's choice of meal. If so, parent gazelles teaching their young to gaze more and graze less may not be conferring any life-saving lessons.

Example 4: Establishing an Association, a Time Sequence, and the Elimination of Plausible Alternative Explanations—the Randomized Experiment

The principal reason for assigning experimental units at random to treatment groups is that the influence of unknown, not measurable, and hence unwanted extraneous factors would tend to be similar among the treatment groups. The random assignment of experimental units ("subjects") to treatments will bring the effects of the extraneous factors under the laws of probability. We cannot say that their effects are neutralized every time, but we can say that *on average* the effects are zero, and we will be able to quantify the chance that observed differences in groups' responses are due merely to "unlucky" treatment assignment.

Let us revisit Mill's requirements, paired with how the randomized experiment satisfies them:

- The investigator manipulates the presumed cause, and the outcome of the experiment is observed after the manipulation. (There is no ambiguity of time.)
- The investigator observes whether or not the variation in the presumed cause is associated with a variation in the observed effect. (This is shown by either a statistically significant correlation or a statistically significant difference between the means being compared.)
- Through various techniques of experimental "design," the plausibility of alternative explanations for the variation in the effect is reduced to a predetermined, acceptable level. (In addition to other techniques, randomization reduces or eliminates association of the treatment group and variables representing alternative explanations for the treatment effect on the response variable.)

A study by Ratnaswamy, et al. (1997) on attempts to help sea turtles serves as a good example of a randomized experiment. Sea turtles are protected under the Endangered Species Act of 1973 and the Marine Turtle Conservation Act of 2004, but unfortunately the laws' authority does not extend to raccoons. Raccoon predation's effect on turtle eggs can be severe; on some beaches, raccoons may depredate 20 to 80 percent of sea turtle nests. Raccoons are an essential part of the ecosystem; simply trapping and removing them might have unintended consequences for the surrounding area's ecology. At the Canaveral National Seashore near the Kennedy Space Center in Florida, Ratnaswamy and her colleagues evaluated three experimental treatments designed to keep raccoons away from the turtle eggs: removal from the area, conditioned taste aversion (CTA), and placement of screens around the turtle nests. The removal treatment involved trapping, anesthetizing, and removing the raccoons from the area. The CTA treatment consisted of injecting chicken eggs with oral estrogen, apparently not the most tasty additive, and

placing them in artificial nests. The screening method consisted of encircling the nests with a screen with holes large enough to allow the hatchling turtles to escape to the sea but too small for the raccoons to gain entry.

The experimental sites exhibited a great deal of variability—some were near paved roads, parking lots, and boardwalks; others were not easily accessible to the public. Because of this variability, the experimental treatments were "blocked by location," and the investigators used what is known as a randomized complete block (RCB) design. (We will discuss the RCB design later.) In the raccoon study, each "block" consisted of a set of four nests at the same location. Within each block, the four nests were randomly assigned to the three treatments and a control group.

The Canaveral National Seashore is a long, thin stretch of beach, making it easy for the National Park Service personnel to locate all the sea turtle nests. A random sample of the nests was used for purposes of this experiment. Analysis of the data revealed that, when compared to control nests, only nest screening showed a statistically significant, reduced level of turtle-nest depredation.

Can we conclude from this experiment that the screening "caused" the decrease in sea turtle depredation? It would appear that we have a strong statistical case. This experiment used sound methodology throughout: random sampling, control of extraneous environmental variables through the method of blocking, and random assignment to treatments within blocks. The time sequence is clear, too: at the time of the initiation of the treatments, the turtle eggs were in fine shape! The only cloud on the inferential horizon might be the amount of generalization that can be done. It could very well be that other beaches differ from the Canaveral National Seashore, making generalization beyond Canaveral, strictly speaking, unjustified.

A Postscript

In this introduction, we have discussed the history and philosophy of scientific, statistically based inference and provided an overview of the ideas that make up the AP Statistics topic of planning and conducting a study. We have attempted to outline the topic's "big picture." In the pages that follow, we will flesh out sampling and experimental design with greater detail, but we pause here to note the importance of both a big-picture understanding and a familiarity with the details of planning and conducting studies. The big picture provides an understanding of the "why" behind decisions made in the planning and execution of studies; the methodological details provide the "how." It is all too easy to get lost or confused in the details of a study, especially as you learn more terminology and study new strategies. When this happens, take a deep breath and go back to the big picture. The questions we address here about the highlighted studies are the very same questions that should be asked when planning a study. The forest remains the same, though the trees may differ.

References

Fitzgibbon, C. D. 1989. "A Cost to Individuals with Reduced Vigilance in Groups of Thomson's Gazelles Hunted by Cheetahs." *Animal Behavior* 37 (3): 508–510.

Hacking, I. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge, England: Oxford University Press.

Holland, P. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:396, 945–960.

Ownby, D. R., C. C. Johnson, and E. L. Peterson. 2002. "Exposure to Dogs and Cats in the First Year of Life and Risk of Allergic Sensitization at Six to Seven Years of Age." *JAMA* 288, no. 8 (August 28).

Ramsey, F. L., and D. W. Schafer. 2002. *The Statistical Sleuth: A Course in Methods of Data Analysis*. 2nd ed. Pacific Grove, California: Duxbury.

Ratnaswamy, M. J., et al. 1997. "Comparisons of Lethal and Nonlethal Techniques to Reduce Raccoon Depredation of Sea Turtle Nests." *J. Wildl. Manage* 61 (2): 368–376.

Wywialowski, A. P., and L. C. Stoddart. 1988. "Estimation of Jack Rabbit Density: Methodology Makes a Difference." *J. Wildl. Manage* 52 (1): 57–59.

Design of Experiments

Roxy Peck
California Polytechnic State University
San Luis Obispo, California

Statistics is the study of variation—how to quantify it, how to control it, and how to draw conclusions in the face of it. As we consider designing experiments, we can expand this definition to include trying to identify "causes," or sources, of variation.

Experiments are usually conducted to collect data that will allow us to answer questions like "What happens when . . . ?" or "What is the effect of . . . ?" For example, the directions for a particular brand of microwave popcorn say to cook the corn on high for 3 to 4 minutes. How does the number of kernels that remain unpopped vary according to cook time—when it is 3 minutes as compared with 3½ or 4 minutes? An experiment could be designed to investigate this question.

It would be nice if we could just take three bags of popcorn, cook one for 3 minutes, one for 3½ minutes, and one for 4 minutes, and then compare the number of unpopped kernels. However, we know that there will be variability in the number of unpopped kernels even for bags cooked for the same length of time. If we take two bags of microwave popcorn and cook each one by setting the microwave for 3 minutes, we will most likely find a different number of unpopped kernels in the two bags. There are many reasons for this: small variations in environmental conditions, differing number of kernels placed in the bags during the filling process, slightly different orientations of the bag in the microwave, and so on. This creates chancelike variability in the number of unpopped kernels from bag to bag. If we want to be able to compare different cooking times, we need to be able to distinguish the variability in the number of unpopped kernels that is caused by differences in the cook time from the chancelike variability. A well-designed experiment produces data that allow us to do this.

In an experiment, the value of a response variable (e.g., the number of unpopped kernels) is measured under different sets of circumstances (e.g., cook times) created for the experiment and assigned by the investigator. These sets of circumstances, determined by the researcher after consideration of his or her research hypothesis, are called **treatments**.

An **experimental unit** is the smallest unit to which a treatment is applied at random and a response is obtained. In the popping experiment above, the individual *bag* is an experimental unit. To further clarify this distinction, imagine an experiment with 10 mice in a cage and 10 cages. If a treatment is randomly applied to the cage, then it is the *cages* and not the individual mice that are the experimental units. When humans are the

experimental units, they are usually referred to as "subjects." The **design** of an experiment is the overall plan for conducting the experiment. A good design makes it possible to obtain information that will give unambiguous answers to the questions the experiment was designed to answer. It does this by allowing us to separate response variability due to differing treatments from other sources of variability in the responses.

The design for an experiment can accomplish this by employing a variety of strategies, including:

1. Eliminating some sources of variability
2. Isolating some sources of variability so that we can separate out their effect on the variability in the response
3. Ensuring that remaining sources of variability (those not eliminated or isolated) produce chancelike variability

For example, in the popcorn experiment we might eliminate variability due to differences between microwave ovens by choosing to use only a single microwave to pop all the bags of popcorn. If we plan to pop six bags of popcorn at each cook time and the popcorn comes in boxes of six bags each, we might try to isolate any box-to-box variability that might occur due to the freshness of the popcorn or changes in the filling process at the popcorn factory. If we plan the experiment carefully, we can separate out the box-to-box variability so that we are better able to compare variability due to cook time against this single microwave's chance variability.

But using only one oven presents a problem: Most investigators would not be interested in an experiment that provided information about just one microwave! Most investigators would wish to generalize to more than that microwave. This aspect of an experiment is known as its **scope of inference**. If only my microwave is used, then the scope of inference—those conditions to which I may generalize my results—is limited to my microwave. An experiment would be more useful if a random sample of Brand X microwaves of the same wattage were used; then the scope of inference would be all Brand X microwaves of that wattage. Or, going further, perhaps a random sample of each brand of all known microwaves could be chosen. In that case, the scope of inference would be all known microwaves. As a rule, the sampling procedure will determine the scope of inference for an experiment. Part of the planning of an experiment involves an understanding of the natural tension that exists between lessening or eliminating sources of variability and compromising the scope of inference.

We also can get into trouble if the design of the experiment allows some systematic (as opposed to chancelike) source of variation that we can't isolate. For example, suppose that we use three different microwave ovens in our experiment and that one is used for all bags cooked for 3 minutes, the second oven is used for all bags cooked 3½ minutes, and the third for all bags cooked 4 minutes. If there are differences among the ovens' microwave activity when the oven is set at high power, this will produce variability in the response (the number of unpopped kernels) indistinguishable from variability in the response due to the treatments in which we are interested (the cook times).

When we cannot distinguish between the effects of two variables on the response, the effects of the two variables are said to be **confounded**. In the situation just described where three different microwaves and three different cooking times are used, the oven used is called a **confounding variable**, and we also would say that "microwave used" and "cook time" are confounded. If we observe a difference in the number of unpopped kernels for the three cook times, we will not be able to attribute it to cook time, since we can't tell if the difference is due to cook time, the oven used, or some combination of both. A well-designed experiment will protect against such potential confounding variables.

A common conceptual error is to think of a confounding variable as any variable that is related to the response variable. To be confounded with the treatments, a confounding variable must also be associated with the experimental groups. For example, we described a situation in the context of the popcorn experiment where microwave used would be confounded with cook times (the treatments) because all of the bags cooked for 3 minutes were done in one microwave, all the bags cooked for 3½ minutes were done in a different microwave oven, and so on. In this case, oven used is associated with experimental groups because specifying which oven is used also identifies the cook time. But consider an alternate approach in which three microwaves are used, but all three cook times are tried the same number of times in each oven. In that case, even though the microwave used might be related to the response, it is no longer associated with the experimental treatments (the 3-minute bags, the 3½-minute bags, and the 4-minute bags). Knowing which oven was used to pop a particular bag provides no information about which cook time was used. Here, oven used would not be a confounding variable.

Design Strategies

The goal of an experiment is to allow us to determine the effect of the treatments on the response variable of interest. As outlined above, to do this we must consider other potential sources of variability in the response and take care to ensure that our experimental design eliminates them, isolates them, or ensures that they produce chancelike (as opposed to systematic) variability in the response.

Eliminating Sources of Variability Through Direct Control

An experiment can be designed to eliminate some sources of variability through **direct control**. Direct control means holding a potential source of variability constant at some fixed level, which removes any variability in response due to this source. For example, in the popcorn experiment we might think that the microwave oven used and the orientation of the bag of popcorn in the oven are possible sources of variability in the number of unpopped kernels. We could eliminate these sources of variability through direct control by using just one microwave oven and by having a fixed orientation that would be used for all bags. Again, recall that the first half of this elimination of a source of variability will compromise the scope of inference. Whether or not to eliminate that variability will depend on the purposes of the study.

Blocking/Grouping to Reduce Variability of Treatment Means

The effects of some potential sources of variability can be partially isolated (separated out from variability due to differing treatments and from the chancelike variability) by blocking or grouping. Although blocking and grouping are slightly different strategies, often at the introductory level they are both called **blocking**. In our discussion we are specifically, though implicitly, considering blocking and blocks with reference to the **randomized complete** block design—the oldest, simplest, and most pervasive of blocking designs. Both strategies, blocking and grouping, create groups of experimental units that are as similar as possible with respect to one or more variables thought to be potentially large sources of variability in the experiment's response of interest.

In blocking, the experimental units are divided into **blocks**, which are sets of experimental units that are similar in some feature for which we would like to control variability. In the simplest case, the size of each block is equal to the number of treatments in the experiment. Each treatment is then applied to one of the experimental units in each block, so that all treatments are tried in each block. For example, consider an experiment to assess the effect practicing has on the time it takes to assemble a puzzle. Two treatments are to be compared. Subjects in the first experimental group will be allowed to assemble the puzzle once as a practice trial, and then they will be timed as they assemble the puzzle a second time. Subjects in the second experimental group will not have a practice trial, and they will be timed as they assemble the puzzle for the first time. To control for the possibility that the subject's age might play an independent role in determining the response (time to assemble the puzzle), researchers may use random assignment of subjects to the two treatment groups, resulting in different age distributions in the two treatment groups. If age does influence response times, left unchecked it would add undesirable variability to the observed mean response times for the groups; part of the variation in the mean response times would be due to the random

difference in age distributions resulting from the random assignment of subjects to groups.

To lessen variability in mean response times for the two treatment groups, the researchers could block on the age of the subjects. Since there are two treatments, they would create blocks, each consisting of two subjects of similar age. This would be done by first determining the subjects' ages and then placing the two oldest subjects in a block, the next two oldest in a second block, and so on. Subjects would be randomly assigned to treatments within each block, and the difference in response times for the two treatments would be observed within each block. In this way it would be possible to separate out variability in the response times (time to assemble the puzzle) that is due to treatments from variability due to block-to-block differences (differences in ages).

Grouping is similar to blocking (and as previously mentioned is sometimes also called blocking at the introductory level). The difference between grouping and blocking is that while the goal of grouping is still to create groups that are as similar as possible with respect to some variable that is thought to influence the response, the group size need not be equal to the number of treatments. Once groups are created, all treatments are tried in each group. For example, consider an experiment to assess the effect of room temperature on the attention span of 8-year-old children, and suppose the researchers plan to compare two room temperatures—say 70 and 80 degrees. If the researchers believe that boys and girls might tend to have different attention spans at any given room temperature, they might choose to group the subjects by gender, creating two groups. They would then make sure that both room temperatures were used with subjects from both groups. This strategy, like blocking, makes it possible to separate out and study variability in the response (attention span) that is attributable to group-to-group differences (gender).

Two special cases of blocking are worth mentioning. In some experiments that compare two treatments, the same subjects are used in both treatment groups, with each subject receiving both treatments. Randomization is incorporated into this design by determining the order in which each subject receives the two treatments at random. As long as it is possible to randomize the order of the treatments for each subject, this design can be thought of as a randomized block design, with each subject constituting a block.

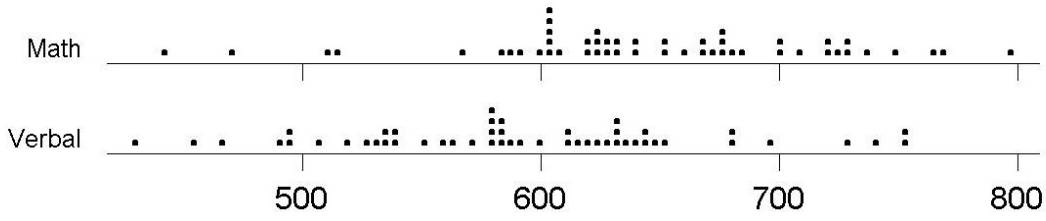
Another special case of blocking uses matched pairs. In a **matched-pairs design**, the subjects available for the experiment are paired based on the value of some variable thought to be related to the response variable. The experiment that assesses the effect of practice on the time required to assemble a puzzle matches subjects on the basis of age, and that is an example of a matched-pairs design.

Ensuring That Remaining Sources of Variability Produce Chancelike Variability: Randomization

We can eliminate some sources of variability through direct control and isolate others through blocking or grouping of experimental units. But what about other sources of variability, such as the number of kernels or the amount of oil placed in each bag during the manufacturing process? These sources of variability are beyond direct control or blocking, and they are best handled by the use of random assignment to experimental groups—a process called **randomization**. Randomizing the assignment to treatments ensures that our experiment does not systematically favor one experimental condition over any other. Random assignment is an essential part of good experimental design.

To get a better sense of how random assignment tends to create similar groups, suppose 50 first-year college students are available to participate as subjects in an experiment to investigate whether completing an online review of course material prior to taking an exam improves exam performance. The 50 subjects vary quite a bit with respect to achievement, which is reflected in their math and verbal SAT scores, as shown in Figure 1.

Figure 1: Dotplots of Math SAT and Verbal SAT Scores for 50 First-Year Students



If these 50 students are to be assigned to the two experimental groups (one that will complete the online review and one that will not), we want to make sure that the assignment of students to groups does not favor one group over the other by tending to assign the higher-achieving students to one group and the lower-achieving students to the other.

It would be difficult to try to create groups of students with similar achievement levels in a way that considered both verbal and math SAT scores simultaneously, so we rely on random assignment. One possible method of random assignment is to randomly pick two students and flip a coin to see which treatments are assigned to which student. Another way is to use a computer to generate a list of random numbers, one for each of the 50 students. Then, we put the 25 students with the smallest random numbers into the first treatment group and other students into the second treatment group. Figure 2A shows boxplots of the math and verbal SAT scores of the students assigned to each of the two

experimental groups for a random assignment of students to groups. Figures 2B and 2C show the boxplots for two other random assignments. Notice that each of the three random assignments produced groups that are quite similar with respect to *both* verbal and math SAT scores. If any of these three random assignments were used and the two groups differed on exam performance, we could rule out differences in math or verbal SAT scores as possible competing explanations for the difference. Randomization also tends to create a similar amount of variability within each experimental group, if the only source of variability is the differences among the experimental units.

Not only will random assignment eliminate any systematic bias in treatment comparisons that could arise from differences in the students' verbal and math SAT scores, but we also can count on it to eliminate systematic bias with respect to other extraneous variables, including those that could not be measured or even identified at the start of the study. (Randomization *can* produce extreme assignments that lead to incorrect conclusions, but the act of randomization makes this a random event with a probability that can be determined as part of the statistical-inference procedure.) As long as the number of subjects is not too small, we can rely on random assignment to regularly produce comparable experimental groups. It is for this reason that random assignment to treatments is an integral part of all well-designed experiments.

Figure 2: Boxplots for Three Different Random Assignments to Two Groups

Figure 2A

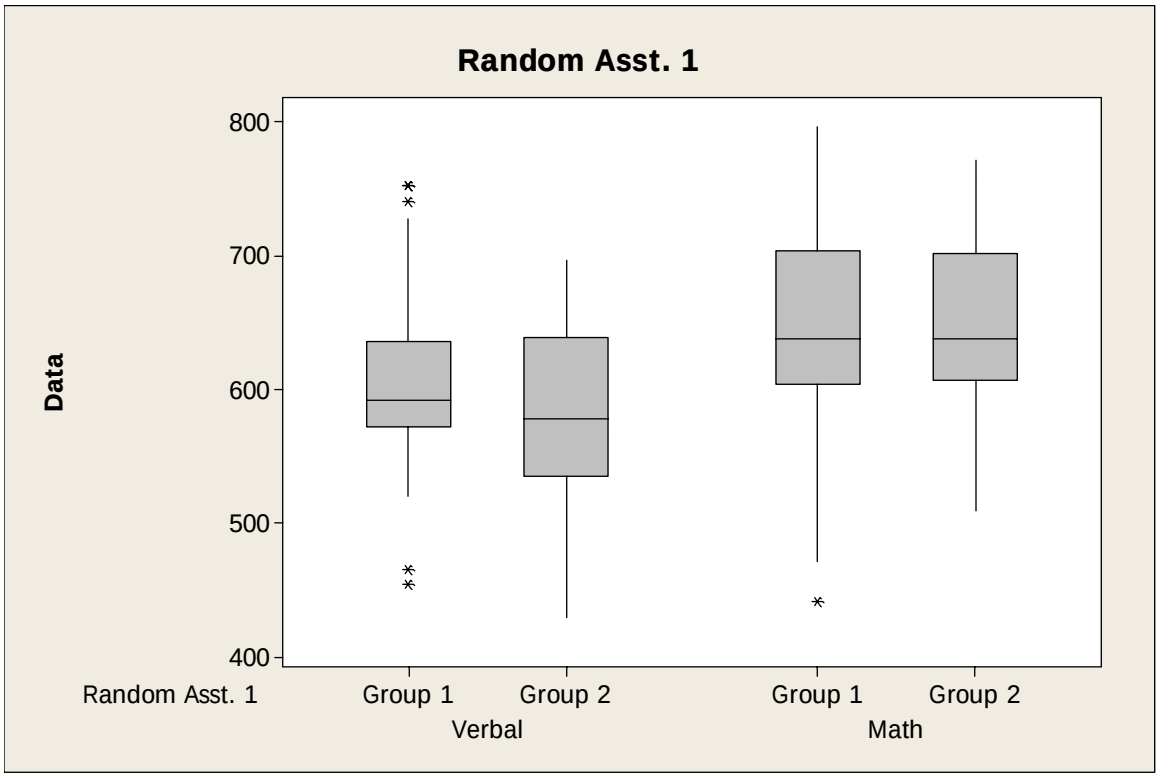


Figure 2B

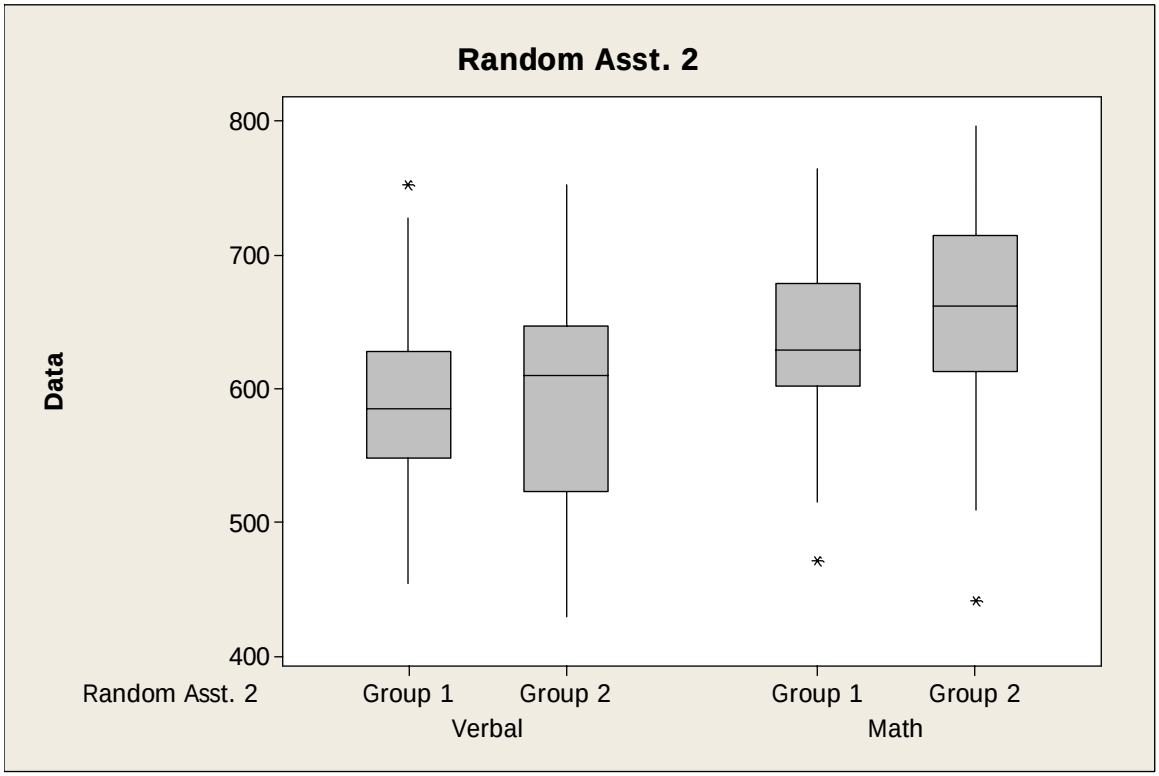
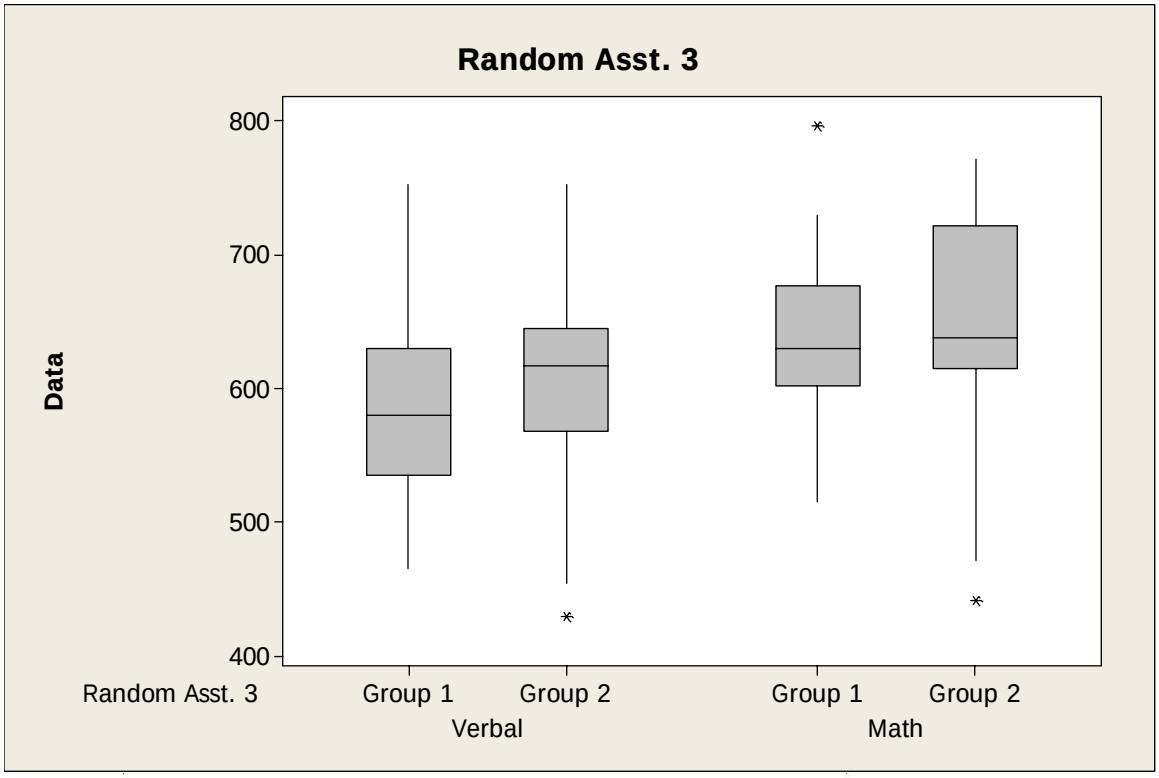


Figure 2C



Not all experiments require the use of human subjects as the experimental units. For example, a researcher might be interested in comparing three different gasoline additives' impact on automobile performance as measured by gas mileage. The experiment might involve using a single car (or more cars, if a larger scope of inference is desired) with an empty tank. One gallon of gas containing one of the additives is put in the tank, and the car is driven along a standard route at a constant speed until it runs out of gas. The total distance traveled on the gallon of gas is then recorded. This is repeated a number of times, 10 for example, with each additive.

The experiment just described can be viewed as consisting of a sequence of trials or **runs**. Because there are a number of extraneous factors that might have an effect on gas mileage (such as variations in environmental conditions like wind speed or humidity and small variations in the condition of the car), it would not be a good idea to use additive 1 for the first 10 trials, additive 2 for the next 10, and additive 3 for the last 10. An approach that would not unintentionally favor any one of the additives would be to randomly assign additive 1 to 10 of the 30 planned trials, and then randomly assign additive 2 to 10 of the remaining 20 trials. The resulting plan for carrying out the experiment might look the assignments shown in Table 1.

Table 1: Random Assignments

Trial	1	2	3	4	5	6	7	. . .	30
Randomly Assigned Additive	2	2	3	3	2	1	2	. . .	1

Random assignment can be effective in evening out the effects of extraneous variables only if the number of subjects or observations in each treatment or experimental condition is large enough for each experimental group to reliably reflect variability in the population. For example, if there were only eight students (rather than 50) participating in the exam-performance experiment, it is less likely that we would get similar groups for comparison, even with random assignment to the sections. The randomization is still necessary, but it may not have "room to work." **Replication** is a design strategy that makes multiple observations for each experimental condition. Together, replication and randomization allow the researcher to be reasonably confident of comparable experimental groups.

When blocking or grouping is part of the experimental design, such randomization should occur separately within each block or group. That is, within each group or block, there should be random assignment of subjects to treatments or of treatments to trials.

In sum, the goal of an experimental design is to provide a method of data collection that accomplishes both of the following:

1. Minimizes or isolates extraneous sources of variability in the response so that any differences in response for various treatments can be more easily assessed
2. Creates experimental groups that are similar with respect to extraneous variables that cannot be controlled either directly or through blocking

Additional Considerations

We now examine some additional considerations that you may need to think about when planning an experiment.

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 if the treatment has had an effect.

It is interesting to note that although we usually think of a control group as one that receives no treatment, in experiments designed to compare a new treatment to an existing standard treatment, the term *control group* is sometimes used to describe the group that receives the current standard treatment.

Not all experiments require the use of a control group. For example, many experiments are designed for the purpose of comparing two or more conditions, as in an experiment to compare density for three different formulations of bar soap or an experiment to determine how oven temperature affects the cooking time of a particular type of cake. And, in the popcorn experiment, there is no "control time" for popping. Nevertheless, you will sometimes see a control group included even when the ultimate goal is to compare two or more different treatments. An experiment with two treatments and no control group may allow us to determine whether or not there is a difference between the two treatments, and even to assess the magnitude of the difference if one exists, but it does not allow us to assess the individual effect of either treatment. For example, without a control group, we might be able to say that there is no difference in the increase in mileage for two different gasoline additives, but we wouldn't be able to tell if this was because both increased gas mileage by a similar amount or because neither had any effect on gas mileage.

Use of a Placebo

In experiments that use human subjects, use of a control group may not be enough to determine if a treatment really does have an effect. People sometimes respond merely to the power of suggestion! For example, consider a study designed to determine if a particular herbal supplement is effective in promoting weight loss. Suppose the study is designed with one experimental group that takes the herbal supplement and a control group that takes nothing. It is possible that those taking the herbal supplement believe that they are taking something that will help them lose weight and therefore may be more motivated to change their eating behavior or activity level, resulting in weight loss. The belief itself may be the actual agent of change.

Although there is debate about the degree to which people respond, many studies have shown that people sometimes respond to treatments with no active ingredients, such as sugar pills or solutions that are nothing more than colored water, reporting that such "treatments" relieve pain or reduce symptoms such as nausea or dizziness—a phenomenon called the **placebo effect**. The message here is that if an experiment is to enable researchers to determine if a treatment has an effect on the subjects, comparing a treatment group to a control group may not be enough.

To address this problem, many experiments use a placebo. A **placebo** is something that is identical (in appearance, taste, feel, and so on) to the treatment received by the treatment group, except that it contains no active ingredients.

In the herbal supplement example, rather than using a control group that receives *no* treatment, the researchers might want to include a placebo group. Individuals in the placebo group would take a pill that looks just like the herbal supplement, but which does not contain the herb or any other active ingredient. As long as the subjects do not know whether they are taking the herb or the placebo, the placebo group will provide a better basis for comparison and allow the researchers to determine if the herbal supplement has any real effect beyond the placebo effect.

Single-Blind and Double-Blind Experiments

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

An experiment in which subjects do not know what treatment they receive 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 impossible to hide from participants whether they are in the swimming group or the jogging group. However, when possible, is it generally a good strategy to "blind" the experiment's subjects.

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 he or she records the response, it is also a good idea to make sure that this person is blind to which treatment an individual subject received. For example, as part of a medical experiment to determine if a new vaccine reduces the risk of getting the flu, doctors must decide whether or not a particular individual who is not feeling well actually has the flu or some other unrelated illness. If the doctor knows whether a participant with flulike symptoms has been vaccinated with the new flu vaccine, he or she may be less likely to determine that the participant has the flu and more likely to interpret the symptoms as being the result of some other illness—supporting a finding that the new shot works.

There are two ways an experiment can use blinding. One involves blinding the participants, while the other involves blinding the individuals who measure the response. And if neither the participant nor the one measuring the response knows which treatment the participant received, the experiment is described as **double-blind**.

Experimental Units and Replication

An **experimental unit** is the smallest unit to which a treatment is randomly applied and a response obtained. In the language of experimental design, treatments are assigned at random to experimental units, and **replication** means that each treatment is applied to more than one experimental unit.

Replication is important for two reasons. We already have discussed the fact that sufficient replication is an effective way of creating similar experimental groups. It also is used in order to get a sense of the variability in the response values for individuals who receive the same treatment. This information is important because it enables us to use statistical methods to decide whether differences in the responses in different treatment groups can be attributed to the treatment received, or if they can be readily explained by chance variation (the natural variability seen in the responses to a single treatment).

Be careful when designing an experiment to ensure that there is replication. For example, suppose that children in two third-grade classes are available to participate in an

experiment that compares two different methods for teaching arithmetic. At first it might seem reasonable to select one class at random for one method and then assign the other method to the remaining class. But what are the experimental units here? Treatments are randomly assigned to classes, and so classes are the experimental units. Since there are only two classes, with one assigned to each treatment, it is necessarily an experiment with no replication, even though there are many children in each class. We would *not* be able to determine if there was a difference between the two methods based on data from this experiment, since we would have only one observation per treatment.

Replication is achieved in completely randomized designs by assigning more than one experimental unit to each treatment. Generally, the best strategy is to have the same number of experimental units assigned to each treatment. Replication in a randomized block design can be achieved by using more than one block. Better estimates of treatment means and better estimates of their standard errors are achieved by using more experimental units. In a randomized block design, this generally corresponds to using more blocks.

One final note on replication: don't confuse replication in an experimental design with "replicating" an experiment. When investigators talk about replicating an experiment, they mean conducting a new experiment using the same experimental design as a previous experiment. Replicating an experiment is a way of confirming conclusions based on a previous experiment, but it does not eliminate the need for *replication* in each of the individual experiments.

Using Volunteers as Subjects in an Experiment

Although it is preferable to randomly sample the experimental units from the population of interest, it is often impractical. Hence the common practice of using volunteers as subjects. While the use of volunteers in a study is not a preferred practice because it limits the researchers' ability to generalize inferences to a larger population, random assignment of the volunteers to treatments allows inferences about treatment effects to be made for the particular group of volunteers used in the experiment. The strength of arguments for generalizing those inferences to a larger population will depend on how well the researchers can show that the volunteers used in the study are representative of the larger population. Such arguments are generally open to considerable debate.

Comparing Two Means: The Algebra of the Completely Randomized (CR) and Randomized Complete Block (RCB) Designs

The two experimental designs in the AP Statistics content outline are the completely randomized design and the randomized complete block design. We will now discuss these two designs at some length, highlighting their characteristics and advantages. To simplify our discussion, we will assume that only two groups or treatments are being compared. When only two groups or treatments are studied, the randomized complete block is frequently referred to as a matched-pairs experiment, another term we will use as we proceed.

Completely randomized designs are characterized by the random assignment of each subject to a treatment. In comparing two treatments, this usually means that $2n$ individuals are randomly assigned to the two treatments, n to each treatment. Their sample means are compared, and an inference is made about the difference in population means or treatment effects. For example, suppose we wish to compare two hybrids of Iowa corn, H1 and H2, to see on average which hybrid results in taller corn. (This is necessary to support the state's "tall corn" reputation.) Iowa has 99 counties, and we will suppose that the Iowa Department of Agriculture and Land Stewardship has designated two acres per county for just such experimentation. (Another statistical benefit to using Iowa is that the 99 counties work perfectly with a random-number table.) We will list the counties alphabetically and assign numeric values to them. Thus in the random-number table we could associate Adair with 01, Allamakee with 02, and so on. Suppose we have selected 10 counties' plots for our experiment. With the completely randomized design, we randomly assign the H1 and H2 treatments. If the completely randomized design is **balanced**, we will assign five plots to H1 and five to H2. A balanced design is characterized by having the same number of observations for each treatment, and balancing is generally preferred. (For simplicity, we will assume that our experimental design is balanced.) The hybrids might, for example, be assigned to the test acres in these counties shown in Table 2.

Table 2: Random Assignment of Hybrids

Hybrid 1	Hybrid 2
26. Davis	72. Osceola
25. Dallas	10. Buchanan
73. Page	22. Clayton
59. Lucas	82. Scott
93. Wayne	49. Jackson

As shown on the map in Figure 3, these randomly assigned counties are scattered across Iowa. Nevertheless, there seems to be a disturbing result. Even though these counties were randomly chosen and assigned treatments, Hybrid 1 seems to be concentrated in the southwest half of the state, while Hybrid 2 appears to lie in the northeast half. One can imagine soil characteristics and perhaps mean rainfall differing across Iowa. And if the variation happens to divide across a southwest-northeast line, our process of randomization will not provide groups of counties that on average consist of approximately equally good soil and moisture for tall-corn growth. We might say that this experiment is a victim of "bad luck," but the randomization must be used to preserve the integrity of the experimental procedure. (So yes, bad luck—but experimentally sound!)

Figure 3: Our Experimental Material—Iowa Counties

The above randomization exercise illustrates one disadvantage of the completely randomized design: it is possible for an "unfair" random assignment of treatments to occur in the sense that one treatment is assigned to most of the "better" experimental units. In many experimental situations, unfortunately, we may not always be able to discern the unfairness quite so easily. To reduce this variability due to unfair assignment—whether we can see it or not—we may elect to use the matched-pairs (RCB) design. The RCB design is characterized by the recognition that while the experimental material may vary over the population, it is likely to be made up of local pockets of homogeneity (e.g., where the soil quality is uniform) for practical experimental purposes.

These local pockets of relatively homogeneous material make up the "blocks" in the randomized block design. In the randomized block design, each experimental treatment is applied within the block. (Ideally, there would be no variability in corn height resulting from soil differences, but in the real world there always will be some amount of natural variability.) The advantage of the randomized block design is that this variability due to soil differences—rather than treatment differences—will be reduced from the amount of soil variability over the whole population to the soil variation within the individual blocks. And we have chosen the blocks to achieve relatively small variability among experimental units within each block with respect to soil types and other growing conditions that affect corn heights.

In the case of the matched-pairs design, where a block would support two hybrids, both would be planted in a randomly assigned half of the block. For our hybrid experiment, we consider that available experimental two-acre rectangles within an Iowa county as a block, and we select five counties for our experiment. We might, for example, choose the blocks shown in Table 3 using our random number table.

Table 3: Randomly Selected Blocks

Blocks for H1 and H2
10. Buchanan
15. Cass
64. Marshall
17. Cerro Gordo
81. Sac

Having chosen our blocks, we would randomly assign Hybrid 1 to one half of the block in each county and Hybrid 2 to the other half of the block. The usual claim is that using these blocks will "reduce variability," but it is not particularly obvious how this reduction in variability is accomplished. Therefore, we would like to give some indication in mathematical terms of what is actually meant by the reduction of variability, and why the randomized blocks (matched-pairs) design is generally superior to the completely randomized design because of this reduction in variability.

We first remind the reader that for both the completely randomized and matched-pairs designs for our tall-corn experiment, we are interested in the difference in means between two treatments. (For more than two treatments, the mathematics as well as the statistical analysis gets a bit more complicated.) We are interested in the difference in mean height of corn that can be attributed to the differing effectiveness of the two hybrids; a common inferential strategy would be to test the hypothesis of equal-mean corn heights for the two

hybrid treatments. In general, we will compare the difference in sample means with what we expect the difference to be if our null hypothesis is true. Then we compare this observed difference to the amount of variability we would expect to see if the experiment were replicated many times. In mathematical terms, we will calculate a test statistic:

$$\frac{\left(\text{Difference in sample means}\right)-\left(\text{Expected difference in sample means}\right)}{\text{Standard deviation of difference in sample means}}$$

The difference in sample means is a statistic and therefore a random variable. We can reformulate this mathematical representation algebraically using the natural notation,

$$\frac{\left(\bar{X}_1-\bar{X}_2\right)-\mu_{\left(\bar{X}_1-\bar{X}_2\right)}}{\sigma_{\left(\bar{X}_1-\bar{X}_2\right)}}$$

The expected value, or mean, of the difference in sample means is equal to the difference in the population means, according to the familiar algebra of random variables:

$$\begin{aligned}\mu_{\left(\bar{X}_1-\bar{X}_2\right)} &= \mu_{\bar{X}_1}-\mu_{\bar{X}_2}\\ &= \mu_1-\mu_2.\end{aligned}$$

In the matched-pairs design, the reduction in variability that comes from blocking refers to a reduction in the value of $\sigma_{\left(\bar{X}_1-\bar{X}_2\right)}$. To see how this occurs, we will appeal once again to the algebra of random variables. When we consider a random variable *X*, we usually think of the values clustered around the mean of *X*, μ_X. When we consider the variability of *X*, it is natural (or at least consistent!) to measure it as we would the variability of raw data—as an average squared difference between the value of *X* and μ_X. Thus, the variance of a random variable *X* is defined as follows:

$$\sigma_X^2 = E\left[\left(X-\mu_X\right)^2\right],$$

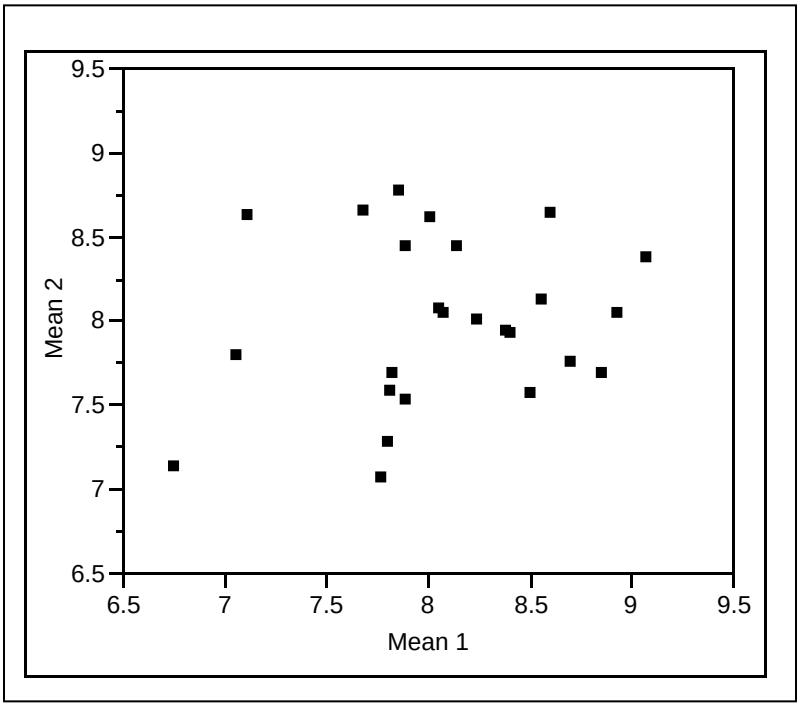
where σ_X^2 is the variance of random variable *X*, and μ_X is the mean of the random variable *X*. When convenient, for easier reading we will also use the notation $V(X)$ and $E\left(X\right)$, respectively for σ_X^2 and μ_X.

In our two-treatment blocking experiment, the random variable of interest is the difference between the two sample means, $\bar{X}_1 - \bar{X}_2$. Using our notation above,

$$\begin{aligned} V\left(\bar{X}_1 - \bar{X}_2\right) &= E\left[\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_{\bar{X}_1} - \mu_{\bar{X}_2}\right)\right]^2 \\ &= E\left[\left(\bar{X}_1 - \mu_{\bar{X}_1}\right) - \left(\bar{X}_2 - \mu_{\bar{X}_2}\right)\right]^2 \\ &= E\left[\left(\bar{X}_1 - \mu_{\bar{X}_1}\right)^2 - 2\left(\bar{X}_1 - \mu_{\bar{X}_1}\right)\left(\bar{X}_2 - \mu_{\bar{X}_2}\right) + \left(\bar{X}_2 - \mu_{\bar{X}_2}\right)^2\right]. \end{aligned}$$

Before we take the final steps, let us review our tall-corn experiment. What we are attempting to do through blocking is capitalize on the homogeneous pockets by sampling blocks from the population of possible blocks. Let us suppose that in general the mean height of corn in Iowa is eight feet, with a standard deviation of about 0.5 feet. If we took *random pairs* of simple random samples, one from a plot with Hybrid 1 and one from a plot with Hybrid 2, and plotted the mean sample heights for Hybrids 1 and 2 in a standard scatterplot, we might get something like the plot in Figure 4.

Figure 4: Scatterplot of Pairs of Means (Completely Randomized Design)

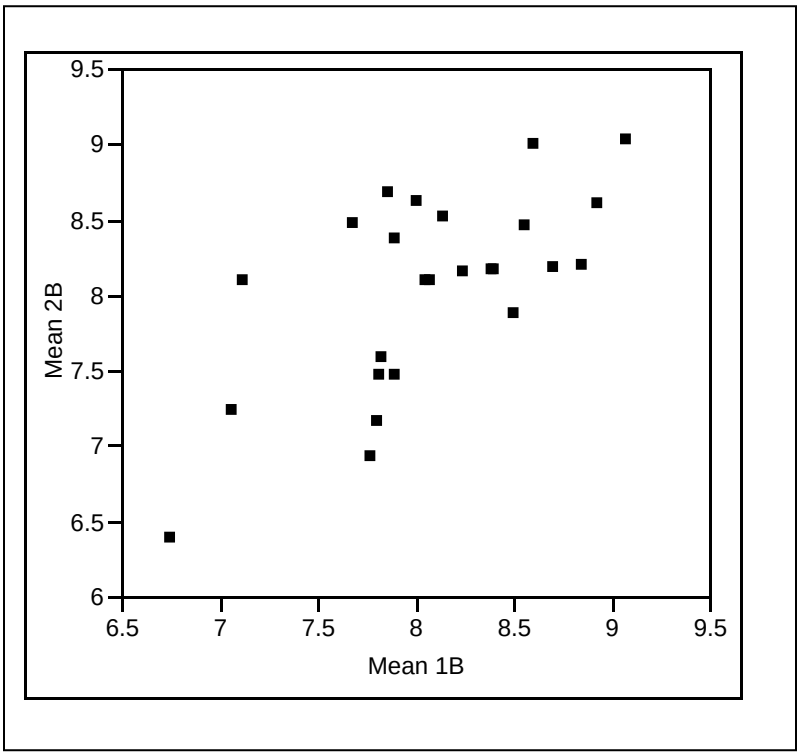


This scatterplot is the sort of plot we would expect to see from means if we randomly assigned our treatments; the observed heights exhibit a certain variability, most likely due to the variability of characteristics of the soil, amount of local rainfall, insect manifestations, fungal diseases, and other aspects of local growing conditions. There appears to be, at most, a very weak relation between the heights for Hybrid 1 and Hybrid 2.

Now suppose that we randomly assigned the Hybrid 1 and 2 treatments within randomly selected blocks in our population. We would ordinarily expect less variation due to soil characteristics and rainfall and other growing conditions within these blocks. In general, the heights of cornstalks for the two hybrids within each block should be closer to each other than heights for the two hybrids planted in two different randomly selected counties in Iowa. Consequently, it should not be surprising that the pairs of mean heights from randomly selected blocks should be more aligned along the line $y = x$.

In other words, pairs of mean cornstalk heights for the two hybrids from randomly selected *blocks* should be correlated! Thus a scatterplot generated from randomly selected blocks should look something like the plot in Figure 5.

Figure 5: Scatterplot of Pairs of Means (Randomized Block Design)



We will now pick up our algebraic thread. Taking a closer look at the variance, we find:

$$\begin{aligned} V\left(\bar{X}_1-\bar{X}_2\right) &= E\left[\left(\bar{X}_1-\mu_{\bar{X}_1}\right)^2-2\left(\bar{X}_1-\mu_{\bar{X}_1}\right)\left(\bar{X}_2-\mu_{\bar{X}_2}\right)+\left(\bar{X}_2-\mu_{\bar{X}_2}\right)^2\right] \\ &= E\left[\left(\bar{X}_1-\mu_{\bar{X}_1}\right)^2-\frac{2\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}\left(\bar{X}_2-\mu_{\bar{X}_2}\right)\left(\bar{X}_1-\mu_{\bar{X}_1}\right)}{\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}}+\left(\bar{X}_2-\mu_{\bar{X}_2}\right)^2\right] \\ &= E\left[\left(\bar{X}_1-\mu_{\bar{X}_1}\right)^2-2\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}\frac{\left(\bar{X}_1-\mu_{\bar{X}_1}\right)}{\sigma_{\bar{X}_1}}\frac{\left(\bar{X}_2-\mu_{\bar{X}_2}\right)}{\sigma_{\bar{X}_2}}+\left(\bar{X}_2-\mu_{\bar{X}_2}\right)^2\right]. \end{aligned}$$

Part of the term in the middle of this expression may look familiar to teachers of AP Statistics, as well it should; it is a product of Z-scores for random variables, as well as the random variable analog of the population correlation coefficient, ρ. Since the expected value of this product is the population correlation coefficient, ρ, we will substitute ρ into the expression and perform some reasonable (and more importantly, correct) algebra of random variables (Ross 2002, section 7.3), giving:

$$\begin{aligned} V\left(\bar{X}_1-\bar{X}_2\right) &= E\left[\left(\bar{X}_1-\mu_{\bar{X}_1}\right)^2-2\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}\rho+\left(\bar{X}_2-\mu_{\bar{X}_2}\right)^2\right] \\ &= E\left[\left(\bar{X}_1-\mu_{\bar{X}_1}\right)^2+\left(\bar{X}_2-\mu_{\bar{X}_2}\right)^2\right]-2\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}\rho \\ &= E\left[\left(\bar{X}_1-\mu_{\bar{X}_1}\right)^2\right]+E\left[\left(\bar{X}_2-\mu_{\bar{X}_2}\right)^2\right]-2\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}\rho \\ &= V\left(\bar{X}_1\right)+V\left(\bar{X}_2\right)-2\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}\rho \\ &= \sigma^2_{\bar{X}_1}+\sigma^2_{\bar{X}_2}-2\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}\rho. \end{aligned}$$

We are now in a position to compare the variability of the sampling distributions of the difference between two means, $\bar{X}_1-\bar{X}_2$, under both the completely randomized and the randomized complete block design, and we will see what variability is reduced and by how much when blocking is used. If the assignment of hybrids to test acres is completely random, the sample means are independent of each other, and $\rho=0$, giving

$$V\left(\bar{X}_1-\bar{X}_2\right)=\sigma^2_{\bar{X}_1}+\sigma^2_{\bar{X}_2}.$$

If, on the other hand, assignments are made to blocks, then

$$V\left(\bar{X}_1-\bar{X}_2\right)=\sigma^2_{\bar{X}_1}+\sigma^2_{\bar{X}_2}-2\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}\rho,$$

where ρ is the within block correlation—positive if the blocking is effective. The more homogeneous the blocks, the greater the value of ρ and the greater the reduction of variability of the estimated difference in sample means from one experiment to another. One should note that a blocking strategy may not always be the best strategy to employ, since the statistical analysis for the matched-pairs experiment analysis will be performed with only about half the number of degrees of freedom. If blocking results in very little similarity of growing conditions within blocks, the decreased variability in the estimated difference in treatment means due to blocking may not compensate for the reduction in degrees of freedom. The lesson here is that when working with homogenous pockets of experimental material in a heterogeneous population, blocking should be used wisely, with knowledge of the context of the variables that make up the experimental design problem. The risk of adverse effects by blocking is usually quite small, and it is therefore generally wise to use blocking if you are confident that you can create blocks of relatively homogeneous experimental units.

Postscript to Experimental Design

In this section we have provided some in-depth discussion of the two experimental designs in the AP Statistics syllabus. As you reread and study this section, note that 90 percent of this material is intended as background for you, the instructor. We hope that this helps you become much more familiar with the terminology of experimental design. It also should help you more fully grasp the differences between the completely randomized design and the randomized complete block design and help you understand what is meant by the assertion that the RCB design "reduces variability." With these concepts now firmly in hand, designing an experiment should be seen less as magic and more as art!

Reference

Ross, S. 2002. *A First Course in Probability*. 6th ed. Upper Saddle River, New Jersey: Prentice-Hall.

Planning Experiments in a Classroom Setting

Peter Flanagan-Hyde
Phoenix Country Day School
Paradise Valley, Arizona

In the preceding section, we discussed some important aspects of designing an experiment. Of course, planning an experiment in theory and executing it in real life are complementary but different tasks. Professional statisticians and scientists learn much about the practical aspects of their craft by wisely keeping track of what goes wrong with their experiments and then consulting their notes when planning the next one. Every good experimenter builds into his or her procedures a "pilot" study where the practicalities of the experiment are tested on a small scale to tease out possible errors in advance. This has two implications for classroom instruction. First, it is essential that students plan experiments "from scratch" and not just analyze experiments that have already been performed. Second, it is important that they actually carry out the procedures of their planned experiments and confront the practical problems of experimentation. Knowing up front that things don't always—or even usually—go according to plan will help inoculate your students against their fear of failure.

In the paragraphs to follow, we will outline a sequence of steps to help teachers guide students through the process of planning and executing an experiment. Strictly speaking, these steps are not sequential in the usual sense. Experiments are frequently planned by considering some aspects simultaneously, so the steps to follow should be thought of as a list that makes some sequential sense rather than as a "magic" sequence to follow blindly.

In the classroom, two intertwined processes will run simultaneously: (1) students designing experiments and (2) teachers monitoring students designing experiments. The monitoring process not only involves guiding the planning process but also anticipating some of the practical difficulties. (The teacher is the wise and enlightened person in the room with the long list of reasons why classroom-designed experiments have failed in the past.) Mindful of how these two parallel processes unfold in the classroom, we will present the steps for experimental design as a sequence of teacher decisions that demand the design and possible execution of an experiment by students. The first step in designing the classroom activity is the most important: framing the assignment.

What Kind of Assignment/Activity Will This Be?

We suggest that there are three types of experiments that call for student design, all of which are reasonable for enhancing students' knowledge of experimental design. Let us

call them (1) "thought" experiments, (2) "practical" experiments, and (3) "real" experiments.

We consider a **thought experiment** one in which the actual practicalities of sampling and measurement are, in effect, a done deal. In a thought experiment, the task is to plan the experiment, but students are not responsible for its practical execution. Students are deemed to have solved the sampling and measurement problems after settling on acceptable and appropriate methods. The methods may be very time consuming, expensive, or possibly even impossible, but that is okay because the primary goal of the assignment is that the experiment work in theory. A thought experiment generally involves a student's written product, possibly supplemented by a visual and/or oral presentation to the class. Each student's task is to consider the usual issues of sampling and experimental design, assisted by the teacher according to usual classroom customs.

We understand a **practical experiment** to be one that students actually carry out. They must perform all the steps in a thought experiment but with some restrictions: in a practical experiment, the design must be executable in practice, the experimental units must be "accessible" to the student, and measurement must be safe and effective. If the experimental units are people, both student and teacher must adhere to ethical principles. While it is not a topic in the AP Statistics syllabus, a discussion of ethical issues should precede *any* and *every* interaction between students and their "subjects." Since not every classroom teacher is an expert on ethics, standard practice should be that teachers follow school district guidelines and seek approval from appropriate district authorities before the execution of any practical student experiments. For a detailed discussion of ethical problems, teachers may refer to Crook and Dean (1999). Gould (2002) is also a good general reference for the ethics of experimentation.

In practical experiments, as opposed to thought experiments, students should attend more closely to the problems of sampling and execution. Students, aided as needed by their teacher, should develop a *practical* sampling plan and then use it to randomly sample from the available experimental units. Students in an AP Statistics classroom should not be expected to get a sample of some gargantuan population such as "all high school students"—even Ph.D. students are limited to sampling from an available population. However, students *should* be expected to carefully define a sampling frame, justify any generalization beyond the available population, and discuss the limitations imposed by the available source of experimental units. Students need to justify anticipated measurements, carefully define the measurement protocol, and then carry out that protocol on a few subjects, as would be done in any pilot study. Each student work product should include a section on any difficulties that arose during the encounter with experimental units or subjects and offer perfecting amendments to the procedure. An

added role for teachers during the planning process is to harness students' natural desire to do a "perfect" experiment, while assuring them that even professional researchers are usually unable to sample as they would prefer, and that it is okay to limit the scope of their sample to something practicable. Students should also be reminded that missteps and mistakes are a normal part of the process, even for professionals. One very good idea is for the teacher to act as the students' first experimental subject to help them uncover measurement problems at the outset.

A **real experiment**, as we use the term, is one that has all the characteristics of a practical experiment but is deemed to be of greater consequence to the student. Experiments in this category are often designed for competition submission, such as to the Siemens Westinghouse Competition or the Young Epidemiology Scholars Competition, or perhaps to a state or regional science competition. Generally, designing an experiment such as this would not be a classroom assignment but rather a student-initiated project that could double as an AP Statistics assignment. In this case, the student may work in concert with people outside the classroom, such as a mentor in a scientific discipline and/or a parent working in a related professional area. In such a situation, the AP Statistics teacher also should act as mentor and communicate with the science or professional person. It should *not* be supposed that these outside mentors have adequate knowledge of the statistical aspects of the problem the student has chosen—in real life, most professionals consult statisticians for advice about experimentation. Because of the increased importance of the experiment to the student and the fact that these competitions may be judged at least partially on statistical merit, the teacher should be more vigilant than usual of the design, sampling, and measurement issues at hand. In some cases the student may wish to perform an experiment that requires knowledge of methodology outside the AP Statistics curriculum. This may present more of a challenge, and sometimes the best an AP Statistics teacher can do is ask a professional statistician to consult on the project. The American Statistical Association Web site, www.amstat.org, can be useful in locating nearby statisticians.

Step 1: Clearly state the problem to be addressed.

It is very easy for students to consider choosing a problem that sounds impressive but is still vague enough that they can "think about it later." It will not occur to most students that the detailed planning for an experiment must take place well in advance—an experiment cannot be "crammed" at the last minute. While it is generally true in our modern computer era that data analysis need not be time consuming, every other part of a well-designed and well-executed experiment is.

The most crucial aspect of stating the problem is for the student to identify the explanatory and response variables, and the student must have a clear idea of how these

variables will be measured. Variables involving lengths, times, and weights should present little difficulty under normal circumstances, but all other variables must be considered in some detail right from the start.

Step 2: Identify an appropriate randomization method.

This step is a relatively simple one, usually involving a choice between a simple random sample or a stratified sample. For the most part, students must make a judgment about the homogeneity versus heterogeneity of the population with respect to extraneous variables that might affect the experimental unit's response. If the case can be made for heterogeneous experimental material, students should opt for a stratified sampling method. Depending on the variables being investigated, typical heterogeneous populations are males versus females and first-year students versus sophomores versus juniors versus seniors. It is probably wise to consider just two strata so that student resources are not spread over a number of strata, each of which only delivers a small number of experimental units. In other words, for data analysis purposes, it is probably better to sample 40 first-years and 40 sophomores rather than 20 from each of four classes.

The assignment of experimental units to treatments, usually a boilerplate description involving calculator generation of random numbers or possibly a random number table, should still be described in some detail—not just with a throwaway statement such as "Assign the subjects using a random number table" or "Assign subjects using a coin flip."

Step 3: Identify needs and methods for controlling extraneous variables.

A very important part of experimental design is the identification and control of potential confounding variables. In a thought experiment, potential confounding variables should not only be identified but also justified. That is, students should explain why they believe that a particular variable is a potential confounder and offer some method of control. A rote listing of the "usual suspect" potential confounding variables is not good experimental design, and controlling for variables that will not affect the response is a waste of valuable research resources. This part of the design process relies most heavily on knowledge of the experiment's subject matter. Identification of potential confounding variables is, for the most part, a science problem, not a statistical one. (Students should be encouraged to seek help from science teachers or other faculty in identifying potential confounders. Such teachers and faculty may need a quick lesson in confounding; for example, many science teachers' idea of control in an experiment is to "only vary one variable at a time," which is a great deal easier to do in physics and chemistry labs.)

Students, having done their homework, may consider any reasonable variable to be of potential concern and should offer possible ways to control that variable. However, as a

practical matter, the only strategy available to students for their statistical analysis will be to use a matched pair (randomized complete block design with one blocking variable); the answer will be to use randomization to control for all but one potential confounding variable. Despite this being the only reasonable practical strategy, students should be encouraged to consider direct control (e.g., "constant temperature and pressure") and blocking as potential (but more costly) strategies. If direct control is considered, the value for the variables should be specified and justified. The possible limitation on the scope of inference also should be considered.

If the experiment is a real experiment, a professional statistician should be consulted if at all possible.

Step 4: Run the experiment in "pilot" mode.

Students should run a small number of trials, executing the sampling plan and performing all measurements. Students should be encouraged to start their trials early enough so that problems can be discussed with the teacher and suggestions made for improvement while they are still in the pilot mode. Students can learn from their errors during pilot mode, and that is a major reason for assigning a practical experiment, not just a thought experiment!

Step 5: Run the experiment in "real" mode if the resources for doing so are available.

In "real" mode, tinkering with the process ends! Any errors or anomalies that occur are no longer fixable; their impact must be considered in the statistical analysis and reported as part of the student's work product. Of course, if a measurement in a trial is clearly anomalous for an identifiable reason unrelated to the relationship between explanatory and response variables, it is okay to delete it for that trial, but in general this should be discouraged.

Step 6: Report the results.

The reporting of the results will, in general, vary in its demands on students. While it would be exemplary for students to follow the professional reporting guidelines for the discipline of their scientific topic, that is very difficult in practice. Different journals have different suggested styles, and different disciplines have different requirements. Perhaps the best solution is to standardize the style as much as possible, according to, say, the *Chicago Manual of Style*, but great leeway should be allowed according to the level of the students. Students will not have been taught this sort of technical writing in their English classes, and they may have encountered conflicting style demands (or no demands) from the various science classes they have taken.

In the case of a real experiment, recent copies of relevant scientific journals also should be consulted!

A Closing Note

Practical experimentation in the classroom is a time-consuming activity, but nothing is as effective in nailing down an understanding of the concepts and the practice of experimental design.

References

Crook, L. S., and M. Dean. 1999. "'Lost in a Shopping Mall'—a Breach of Professional Ethics." *Ethics and Behavior* 9 (1): 39–50.

Gould, J. E. 2002. *Concise Handbook of Experimental Methods for the Behavioral and Biological Sciences*. Boca Raton, Florida: CRC Press.

Examples of Experimental Design Problems

Peter Flanagan-Hyde
Phoenix Country Day School
Paradise Valley, Arizona

The following experimental design problems are intended to provide some examples that might be used in classroom situations. Depending on the individual classroom, some of these may be used as practical (in the sense we have used the term) experiments. It is not necessary that experimental design problems be incredibly important or definitive in any sense; the time for a single grand experiment that settles issues all by itself has passed, if it ever really existed. The experiment, like statistics itself, is usable by ordinary people in ordinary circumstances.

Some of the following ideas for experiments are taken from scientific literature, and others are "made up" from whole cloth, ideas borne of simple observation and reflection about our ordinary world. Both kinds of experiments will help students understand the roles of experimentation, which are: (1) as a tool for exploration of scientific theories and (2) as a practical "action-research" tool for everyday people.

Experimental design problems that seem to work best with students have these common characteristics:

- The problems are authentic.
- The scenario of the problem is easily understood and well within the realm of student experience.
- The explanatory and response variables are easily identified.
- Potential confounding variables are identifiable and perhaps controllable through a blocking strategy.
- The measurements are not terribly difficult.
- There is flexibility in how one might approach the design of the experiment.

Please use these examples as "seed" problems for student assignments or possibly as starting points for students to use while casting about for their own ideas. Both teachers and students benefit from discussing new and different experimental design problems in the classroom.

Example 1: The Rites of Spring

When spring arrives, the sun begins to warm, birds begin to chirp, the sports-minded turn their thoughts to baseball, and—oh, yes—dandelions start to grow. The common dandelion (*Taraxacum officinale*) is also known as lion's tooth, puffball, and monk's head. The genus *Taraxacum* consists of about 40 species worldwide. It originated in Europe and has been used as a potherb and medicinal plant since the Roman times. It has a high vitamin and mineral content, and leaves from the mature plant are often dried and used to make a mild tea. Its roots are often used to make a stronger tea or are dried and used for medicinal purposes.

Lawn fanatics generally regard dandelions as pests, using chemical pesticides to control their growth. But not every chemical pesticide works well on every lawn; the effects of a pesticide may depend on the type of soil, nitrogen level, and kind of fertilization that has been attempted in the past. Suppose that you have two pesticides to choose from, and you wish to perform an experiment to see which appears to be more effective on your lawn.

Your assignment is to design an experiment to decide which pesticide to use on *your* lawn. (If you don't have a lawn, ask your instructor for a drawing to use.)

1. Draw a scale model of your lawn, including streets, driveways, buildings, and any other aspects that might be important for understanding the dandelion growth.
2. Decide on a measure of the effectiveness of the two pesticides.
3. Design an experiment to help identify which pesticide is more effective.

Example 2: Those Guys Are Just Unbearable!

Due to the availability of food in public areas of national parks, black bears are becoming a serious nuisance. These animals, wanderers with a high level of curiosity, often find their way into areas where humans—and their food—are. Once the bears find food, they may go on to damage property and even threaten or injure park visitors. A destructive behavior pattern, reinforced by the promise of food, is likely to continue.

A possible nonlethal solution is to transport problem bears to remote areas of parks, far away from areas where they can find food. It is not well understood how bears find their way back to old haunts, but it is believed that they will travel only limited distances before simply establishing a new home range. Effective planning for translocation may take into account more than the straight-line distance from release point to old haunt. For example, topographic features may be important, such as the number of ridges (drainage divides) that must be navigated or the total elevation gain that the bear would encounter as it attempted to return. Or it may be that bears, finding evidence of human habitation such as roads or trails, will follow them in their search for food.

It is interesting to note that some female bears, for reasons unknown, are better able to find their way back to their haunts. (No, they *don't* ask for directions!)

Your assignment is to:

1. Locate a topographic map of Glacier National Park or Yellowstone National Park.
2. Pick out an area that is frequented by humans.
3. Define a distance, elevation, or other reasonable measure of difficulty-of-returning.
4. Design an experiment to determine how far away a bear should be moved for effective translocation, in terms of your measure in part (3).

Example 3: Quiet Skies *Are* Friendly Skies . . .

The flying public understands that air travel can be annoying and frustrating—especially when there are infants on board. Changes in elevation during the flight can make infants cry with ear pain, which results in subsequent pain for the adult passengers.

One airline is considering a novel solution: offering "official" pacifiers to infants' parents for use during the flight. The idea is that if the child starts to cry, the pacifier will not only comfort the child but also hopefully play a similar role to gum-chewing in alleviating pain due to unequal air pressure on the eardrum. Of course, it is possible that other aspects of flying may contribute to infant crying. For example, the jet noise may be louder or engine vibration may be more intense in some parts of the plane.

Your assignment is to design an experiment that will help determine whether or not the pacifier effectively reduces an infant's discomfort—and therefore its crying.

Example 4: Deep in the Heart of . . . Rhode Island?

Gustav Fechner (1801–1887) was one of the first psychologists to apply techniques of experimentation to psychology. One of the topics he studied was the relationship between the physical world and the world of the senses. For example, recent experiments have suggested that the relative size of objects as they are perceived may be different in memory. That is, if people look at two circles, they may judge their relative areas to be different than if they recall them from memory. This certainly is not surprising; what is interesting is that the judgment is biased: memory seems to "compress" the relative size of objects so that the ratio of the larger-to-smaller object is smaller than the ratio as judged when perceived.

To expand on this memory-compression research, one might examine whether the phenomenon extends to familiar "objects," not just abstract symbols. For example, subjects might be asked to judge the ratio of the areas of two U.S. states presented to them drawn to scale, while others might be asked to judge the ratio after "imagining" and comparing the two states from memory.

Your assignment is to design an experiment to investigate any difference between judgments based on the ratios of areas of U.S. states as they are remembered versus ratios of the areas as they are perceived. Several issues need to be considered:

- To account for some students' knowledge of some states, each subject should see a set of 20 randomly selected pairs of states. (There are a possible $50 \times 49 = 2450$ unique pairs.) What would a "representative" sample of these ratios look like?
- It might be possible that the "memory-compression effect" only works if the ratios are "large" (whatever that might mean). Does that mean only certain pairs should be considered? If so, which ones?
- It might be possible to reduce variability by using the same students for the perception and the memory groups. Should they see the same pairs of states in both conditions?
- It might also be possible that the positions of the two states (e.g., larger on the left, smaller on the right) might affect the students' judgments. How should this factor be neutralized?

Example 5: Les Caprices de la Fortune (Fickle Fortune's Favor)

In the eighteenth-century painting *Lady Luck* by Jean Antoine Pierron, a young woman floats above the countryside, apparently dispensing good fortune where she pleases. In like manner, everyday conversation uses words about chance that may be just as fickle. Here are some words and expressions that are commonly used to express probability:

Certain	Always	Virtually always	High probability	Consistent with
Probable	Best bet	Likely	Significant chance	Moderate probability
Not unreasonable	Cannot be excluded	Moderate risk	Possible	Sometimes
Not certain	Unlikely	Doubtful	Low probability	Never

But what do the recipients of these phrases understand them to mean? That is, what sort of real probability do they associate with these phrases?

Your assignment is to design an experiment to determine whether there is a difference between the probabilities attributed to these words and phrases by individuals who have not studied probability and by those who have. You should construct a list of these words and phrases, with the following directions:

> Shown below are expressions you might encounter when you ask your teacher what the likelihood is of getting a B grade or better at midterm. Please read each expression carefully and estimate the numerical probability (0 to 100 percent) that you associate with each expression.

It is possible that these numbers might fluctuate over time, so you should construct *two* lists of these words and phrases, listed in random order, and present them to your subjects approximately two weeks apart. (How might you measure the consistency of the subjects' probability assignments?)

The Design and Analysis of Sample Surveys

Dick Scheaffer
Professor Emeritus
University of Florida
Gainesville, Florida

What Is a Sample Survey?

Sample surveys are designed and conducted to obtain information from a well-defined population in such a way that we can produce estimates of parameters for that population, such as a population mean or a population proportion. These statistical estimates of a population parameter must be accompanied by a measure of variation that tells us how close to the population value we can reasonably expect our estimate to be. This measure of variability is historically and technically referred to as **error**. The use of the word *error* is unfortunate since it seems to imply a mistake on the part of the researcher, which is *not* the case!

Information from sample surveys affects almost every facet of our daily lives. For example, such information determines government policies on the control of the economy and the promotion of social programs. Opinion polls are frequently cited by the various news media, and the ratings of television shows determine future lineups.

A **census** can be thought of as a survey of the whole population. One often thinks of a country's census as simply a count of its residents, but a census usually measures other characteristics of the residents, too. (In fact, the term *statistics* has as its root the word *state* because it originally referred to collections of data about the conditions of the state. The word's first use has been traced to the writings of a German statistician, Gottfried Achenwall, in 1749.) One of the first censuses ever attempted was taken in Scotland by John Sinclair in the latter part of the eighteenth century. He called the facts he collected "statisticks" and defined the process as "an inquiry into the state of a country, for the purpose of ascertaining the quantum of happiness enjoyed by its inhabitants, and the means of its future improvement." Webster's first dictionary, printed in 1806, defined statistics as "a statement or view of the civil condition of a people," a similar but less-romantic definition.

The first census of U.S. residents was taken in 1790, and since then a census has been taken every 10 years. The U.S. Census Bureau does, indeed, attempt to contact every household in the country in order to count the population. But each decennial census does far more than simply enumerate people. In the 2000 census, the short-form questionnaire that went to all households had questions covering tenure (whether a

housing unit is owned or rented), name, sex, age, relationship to householder, if of Hispanic origin, and race. The long-form questionnaire, which goes to a sample of one in six households, has the short-form questions plus additional ones (40 or so) on such topics as the social characteristics of the respondents, marital status, place of birth, citizenship, educational attainment, ancestry, language spoken at home, veteran status, occupation, income, and housing conditions. The resulting information is used by the following:

- The federal government in determining allocation of funds to states and cities
- Businesses to forecast sales, manage personnel, and establish future site locations
- Urban and regional planners to plan land use, transportation networks, and energy consumption
- Social scientists to study economic conditions, racial balance, and other quality-of-life issues

Other large government surveys are conducted by the U.S. Bureau of Labor Statistics (BLS). The BLS routinely conducts over 20 surveys, with some of the best known and most widely used being the surveys that establish the consumer price index (CPI). The CPI is a measure of price change for a fixed-market basket of goods and services over time. It is used as a measure of inflation and serves as an economic indicator for government policies. Businesses tie wage rates and pension plans to the CPI. Federal health and welfare programs, as well as many state and local programs, link their eligibility requirements to the CPI. Rate-increase clauses in leases and mortgages are based on the CPI. Clearly this one index, determined on the basis of sample surveys, plays a fundamental role in our society.

One of the most noticeable of the BLS data collection efforts is the Current Population Survey (CPS), a monthly survey of households that provides a comprehensive body of data on the labor force, employment, unemployment, and persons not in the labor force. The CPS collects information on the labor-force status of the civilian, noninstitutional population 15 years of age and older—although labor-force estimates are reported only for those 16 and older—using a probability sample of approximately 60,000 households. Respondents are assured that all information obtained is confidential and used only for the purpose of statistical analysis.

Numerous research centers at universities are known for their expertise in sampling, among them the National Opinion Research Center (NORC) at the University of Chicago and the Survey Research Center (SRC) at the University of Michigan. NORC conducts a variety of studies for government agencies, educational institutions, foundations, and private corporations (including a study of the Florida voting controversy of 2000), but it

is probably best known for its General Social Survey (GSS). The GSS assesses social changes in contemporary America through a standard core of demographic and attitudinal variables, in addition to topics of special interest that are rotated in and out. The SRC specializes in interdisciplinary, social-science research involving the collection and analysis of data taken from scientific sample surveys, with a solid mix of basic research; applied, survey-based research; and the propagation of the scientific method of survey research through teaching and training.

Opinion polls are constantly in the news, making names like Gallup and Harris well known to most people. These polls, or sample surveys, reflect citizens' attitudes and opinions on everything from politics and religion to sports and entertainment. Gallup specializes in tracking the public's attitudes on virtually every political, social, and economic issue of the day, including highly sensitive or controversial subjects. The organization takes pride in the fact it carries out its polls independently and objectively, without taking money from special-interest groups. Best known for The Harris Poll®, Harris Interactive is a worldwide market-research and consulting firm that has pioneered the use of market research on the Web.

The Nielsen Company, another polling group, uses sampling in a variety of interesting and important ways. A. C. Nielsen provides market research, information, and analysis to the consumer-products and service industries. Nielsen Media Research, the famous TV ratings company, provides television-audience measurement and related media-research services. Nielsen/NetRatings provides Internet-audience measurement and analysis, an increasingly important index in the modern age.

Variation and Bias

It is well known that a sample will not always produce an exact copy of the features of the population being studied. In fact, any two samples of the same size from the same population are likely to produce slightly different results. Samples are subject to variation because measurements are obtained from only a randomly selected subset of the population units, and the measurement procedure may not be completely accurate. Statistics can be thought of as the study of variation—how to quantify it, how to control it, how to draw conclusions in the face of it—and sample survey design and analysis requires careful consideration of all of these aspects of statistics.

Survey errors can be divided into two major groups: **errors of nonobservation**, where the sampled elements comprise only part of the target population, and **errors of observation**, where recorded data deviate from the truth. Errors of nonobservation can be attributed to sampling, coverage, or nonresponse. Errors of observation can be attributed to the interviewer (data collector), respondent, instrument, or method of data collection. All

except the first of these (sampling error) can be major contributors to bias in the reported results of a sample survey, as well as to the variation. Let's take a closer look at these two types of errors.

Errors of Nonobservation

Generally, the data observed in a sample will not precisely mirror the data in the population from which that sample was selected, even if the sampled units are measured with extreme care and accuracy. This deviation between an estimate obtained from sample data and the true population value is the **sampling error** that occurs for the simple reason that a sample, not a census, is being taken. Generally, samples include only a small fraction of the population units. Since each possible sample results in a value for sampling error, one can imagine a distribution of all possible such errors. Formulas for mean and variance of such distributions of the errors can be derived theoretically and estimated from the sample data for samples selected according to an appropriate sampling design (that is, a plan for selecting the units to be in the sample). It is important to note that sampling error can be reduced both by good survey designs and appropriate choice of sample size. Thus, the investigator has some control over this component of error; there are books on sample survey methods that cover management methods. In addition, sampling error does not lead to bias in the results so long as appropriate random sampling is built into survey design.

In almost all surveys, the sample is selected from a list of units called a sampling frame. Most often, this sampling frame does not match up perfectly with the target population, leading to errors of **coverage**. For telephone surveys, telephone directories are considered inadequate because of unlisted numbers. For mail surveys of property owners, for example, the most recent list of addresses available at the county court house will be out of date because some nonresident owners will have recently moved or sold their property. For surveys of hunters or anglers, lists of license purchases are inadequate because children are not required to purchase a license. This lack of coverage introduces an error in the sampling process, an error that often is not easily measured or corrected. Known or suspected problems of coverage should be explicitly discussed in any results of a sample survey so that those using the results can see clearly how the sampled population differs from the target population.

Probably the most serious of all the nonobservational errors is **nonresponse**. This is a particularly difficult and important problem in surveys that attempt to collect information directly from people using some form of interview. Nonresponse rates are easily obtained because both the sample size and the number of responses to the survey are known. Sometimes nonresponse rates are mistakenly used to judge the quality of a survey. A survey with a small nonresponse rate might still miss an important part of the

population—say, anyone over age 70. On the other hand, data from a survey with a high nonresponse rate could still be informative if the distribution of nonrespondents mirrors the distribution of the respondents in the entire population with respect to all important characteristics. The important consideration here is not the response rate, but rather the nature of the nonrespondents. A good survey plan includes a strategy to follow up on some nonrespondents in order to measure how far from or close to the respondent group they may be.

Nonresponse arises in one of three ways: the inability to contact the sampled element (a person or household, for example), the inability of the person responding to come up with an answer to the question of interest, or the refusal to answer. Some might think it reasonable to simply substitute a "similar person" for the nonresponder, but data must be collected from precisely those elements that were selected by the randomization scheme used in the survey design. An interviewer must not substitute a next-door neighbor, who just happens to be home at 3 p.m., for the person actually selected for the sample but who isn't answering the door. This type of substitution might lead to a survey that is biased because too many families with children or too many retired persons or too many people who work nights are being interviewed. (Callbacks at different times of the day can lower nonresponse appreciably.) In addition to these obvious biases, haphazard substitutions also can alter the probabilistic structure of the design and may make it impossible to estimate the sampling error. For example, cell phones are now a big problem for surveys that randomly select telephone numbers from directories of listed numbers, because a higher proportion of younger adults may own only cell phones with unlisted numbers, skewing the population.

The inability of the interviewed person to answer the question of interest is a serious problem, particularly in questions dealing with fact. A question on opinion can have a "don't know" option, and the survey design can account for a certain percentage being in this category. A survey on businesses' economic impact on a community, however, can be seriously biased if a few of the larger businesses do not know how much they spend on transportation, for example. Still, with more-thorough checking, questions of fact are often the type of questions for which an answer can be found.

The most serious aspect of the nonresponse problem today is refusal to answer. Perhaps because of the proliferation of surveys, perhaps because of fear related to increases in crime, and, undoubtedly, for a variety of other reasons, people are increasingly refusing to answer survey questions. Many surveys report that their response rates are as good as ever and have not decreased in recent years. On closer scrutiny, however, these "stable" response rates are often due to an increased effort to replace the refusals with others who *will* respond, the shortcomings of which we already have discussed.

What do survey designers and analysts know about those who tend to refuse to answer surveys? The highest refusal rates occur among the elderly and the poorly educated—although this is not uniformly true—and the pattern seems to exist across ethnic and salary groups. Single-person households are more likely to refuse an interview than are multiple-person households, but household size is confounded with age because many of the elderly belong to single-person households. For the poorly educated and the elderly, surveys often suggest that someone else (perhaps the government) is attempting to gain more power over them. Thus, by refusing to participate in the survey, they refuse to give those in "authority" any more ammunition. Of course, the proliferation of surveys is causing a widespread and tremendous intrusion on privacy, especially since most people group sales calls (which may begin with a comment about conducting a survey) together with serious surveys. If a survey produces a high refusal rate, it behooves the investigator to find some information on those refusing to answer in order to reduce a potentially sizable bias.

Careful planning can lower refusal rates. For example, alerting respondents in advance, with a letter or telephone call, that they have been selected for a survey may help improve the response rate. This is especially true if the letter is from a "prestigious" organization (in the eyes of the potential respondents), if the letter explains that the survey can be beneficial to them and others, and if the letter explains why it is important that the person actually selected must respond in order to make the survey valid. In general, a potential respondent may not initially comprehend why his or her next-door neighbor cannot be substituted. (After all, he is home all the time and loves to talk.) Explaining the nature of random sampling in nontechnical language sometimes helps. Long introductions about the technical merits of the survey and its outcomes, however, are not considered effective.

Groves et al. (2002) give a comprehensive assessment of what is known about nonresponse and point out effective ways to lessen its effect. Here are some of their main points. Surveys are governed by the principles of social exchange. Small gestures (personalized letters, reminder notes, tokens of appreciation) can help reap big response rates, furthering a major goal, which is to build trust between the interviewer and the respondent. Interestingly, using authority to increase response rate is not all it's cracked up to be; one study showed a 26 percent compliance rate when "university" and "scientific research" were invoked, as compared to a 54 percent compliance rate with a nonauthoritative, personal appeal ("I would like your help.") Topic saliency improves the response rate, too, as respondents may want to give their opinions on important matters, especially if they belong to a group that can be potentially advantaged (or disadvantaged) by the results of a survey. Interviewer effects can be huge (though they carry a risk of bias), and experienced interviewers can work to bring saliency to a topic and thereby improve response rates. They can "tailor" the nature of the interview to information

provided by the respondent. The social skills of the interviewer appear to be more important than attributes such as age, race, and sex. Length of interview, especially in telephone interviews, is a critical determinant of response rate. In one study, a mention of the fact that the interview would be about 15 minutes got a 36 percent compliance rate, while a mention of a 10-minute interview got a 43 percent compliance rate, and no mention of time at all got a 66 percent compliance rate.

Errors of Observation

Once a person (or other sampling unit) is in place and ready to be "measured," there are still more errors that can creep into the survey. These are errors of observation, and they can be attributed to the interviewer, the respondent, the measurement instrument, or the method of data collection. As mentioned above, all of these can contribute to measurement bias.

Interviewers have a direct and dramatic effect on the way a person responds to a question; for example, reading a question with inappropriate emphasis or intonation can nudge a person to respond in a particular way. Most people who agree to an interview do not want to appear disagreeable and will tend to side with the view apparently favored by the interviewer—especially when the respondent does not have a strong opinion. Friendly interviewers have more success, of course, than the overtly forceful ones. How gender affects interviews is not clear. Male interviewers get a higher rate of cooperation from male respondents than do female interviewers. In general, interviewers of the same gender, racial, and ethnic groups as those being interviewed are slightly more successful.

Respondents differ greatly in their motivation to answer "correctly" and in their ability to do so. Each respondent must understand the entire question and be clear about the answer options. In personal interviews, flashcards showing the question in written form can help this process. That means that questions must be clearly phrased and the questionnaire should not be too long because people will quickly tire of the survey. Obtaining an honest response to sensitive questions, on business or sexual practices for example, is particularly difficult and may require special techniques. It appears that most response errors are due to the following:

- Recall bias (the respondent simply does not remember correctly)
- Prestige bias (the respondent exaggerates a little on hunting success or income)
- Intentional deception (the respondent will not admit breaking a law or has a particular gripe against an agency)
- Incorrect measurement (the respondent did not understand the units and reported feet instead of inches or did not fully understand the definition of children, reporting grandchildren as well)

Incorrect measurement refers to the **measurement instrument** as a source of error. In any measurement, the unit of measurement must be clearly defined, whether it be inches on a tape measure, pounds on a scale, or glasses of water (where a "glass" could be any standard size, such as 12 ounces). Inaccurate responses are often caused by errors of definition in survey questions. Some examples are: (1) As alluded to above, the word *children* must be clearly defined. (2) What does the term *unemployed* mean? Should the unemployed include those who have given up looking for work, teenagers who cannot find summer jobs, and those who lost part-time jobs? (3) Does *education* include formal schooling and technical training, on-the-job classes, and summer institutes? Measurements to be taken as part of a survey must be precisely and unambiguously defined.

The interviewer, the respondent, and the instrument are brought together in various ways, depending on the **method of data collection**. The most commonly used methods of data collection in sample surveys are in-person interviews and telephone interviews. These methods, with appropriately trained interviewers and carefully planned callbacks, commonly achieve response rates of 60 to 75 percent and sometimes even higher. A questionnaire mailed to a specific group can achieve good results, but response rates for this type of data collection are generally so low that the reported results are suspect. Frequently, more-objective information can be found through direct observation rather than from a phone interview or mailed questionnaire.

In today's technological age, Web surveys are very popular and are improving in quality due to standardized software, user-friendly interfaces, high-speed transmission, and low cost. But nonresponse and incorrect response problems are even more serious for Web surveys than for other modes of sampling. Emailed invitations to participate in a survey and follow-up memos can be easily ignored, and there are plenty of technical glitches that can cause problems along the way. The responses that *are* completed tend to be completed quickly, so the follow-up time frame has to be shorter than what it would be for a mailed questionnaire. On the other hand, since first-responders tend to be young and more technically astute, enough time must be allowed for others to respond so as to not seriously bias the results.

Random Selection

The key to being able to quantify **sampling error** variability and make probability statements about potential sampling errors is the random selection of units in the sample. We will use some examples to illustrate the essential properties of random samples.

Let us assume that a population consists of a very large number of integers (0, l, 2, . . . , 9) in equal proportions. We may think of these integers as stored in a table (like a random-

number table) or as generated by a random-number generator in a calculator. Since all integers occur in equal proportions, the relative-frequency histogram, which shows the distribution of the population measurements, is as shown in Figure 1. These relative frequencies can be thought of in probabilistic terms. If a number is selected **at random**, then the **probability** that the selected number will be a 4 is 1/10.

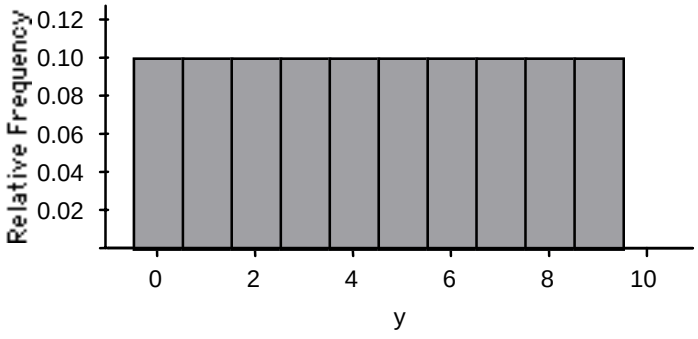
Suppose a number is to be selected at random from this population of digits, with its value denoted by y. Then the possible values for y are 0, 1, 2, . . . , 9, in this case, and a probability of 0.10 is associated with each of these 10 possible values. This constitutes the **probability distribution** for the **random variable *Y***. (Note that capital Y is used to represent the random variable, and lowercase y is used to represent the particular values Y can be.) The probability associated with a particular outcome y is denoted by $p(y)$. For instance, in this example, $p(0) = 0.1$, as do $p(1)$ through $p(9)$.

One of the numerical measures used to summarize the characteristics of a population is the **expected value** of Y or functions of y. The expected value of Y, denoted by $E(Y)$, is by definition

$$E(Y) = \sum_{y} yp(y),$$

where the summation is over all values of y for which $p(y) > 0$.

Figure 1: Distribution of Population Containing Integers 0 Through 9 with Equal Frequency



For the population and random variable Y under study,

$$\begin{aligned} E(Y) &= \sum_y yp(y) \\ &= 0p(0)+1p(1)+2p(2)+3p(3)+4p(4)+5p(5) \\ &\qquad +6p(6)+7p(7)+8p(8)+9p(9) \\ &= \frac{1}{10}(45) \\ &= 4.5. \end{aligned}$$

One can see that $E(Y)$ is equal to the average value, or mean value, of all the measurements in our population. In general, a population mean will be denoted by μ, and it follows that

$$\mu = E(Y),$$

where the random variable Y represents the value of a single measurement chosen at random from the population.

The variability of population measurements can be quantified by the **variance**, which is defined as the expected value, or average value, of the square of the deviation between a randomly selected measurement Y and its mean value, $\mu = E(Y)$. Thus the variance of Y, $V(Y)$, is given by

$$V(Y) = E(Y-\mu)^2 = \sum_y (y-\mu)^2 p(y).$$

For the population of random digits under study,

$$\begin{aligned} V(Y) = E(Y-\mu)^2 &= \sum_y (y-\mu)^2 p(y) \\ &= (0-4.5)^2\left(\frac{1}{10}\right)+(1-4.5)^2\left(\frac{1}{10}\right)+\ldots+(9-4.5)^2\left(\frac{1}{10}\right) \\ &= \frac{1}{10}\left[(0-4.5)^2+(1-4.5)^2+\ldots+(9-4.5)^2\right] \\ &= \frac{1}{10}[82.5] \\ &= 8.25. \end{aligned}$$

The variance $V(Y)$ is commonly denoted by σ^2.

The **standard deviation** (SD) of Y is defined to be the square root of its variance, and it is denoted by $\sigma = \sqrt{V(Y)} = \sqrt{\sigma^2}$. For the specific population under discussion,

$$\sigma = \sqrt{8.25} = 2.9 .$$

Suppose we now have a random sample of n measurements taken from the population of random digits, with the sample measurements denoted by $y_1, y_2, \ldots, y_n$. **Random sampling** in this infinite-population case means that each sampled value has the same probability distribution, e.g., the one value we considered above, and all sampled values are mutually independent of one another. Roughly speaking, mutual independence means that the outcome for any one unit in the sample is not influenced by the outcomes for any of the other units in the sample. The mean, variance, and standard deviation of a sample are given, respectively, by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i ,$$

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1},$$

$$\mathrm{s} = \sqrt{s^2} .$$

The sample mean is a statistic derived from the results of a chance experiment and is thus a random variable, usually denoted as $\bar{Y}$ (uppercase) to distinguish it from a particular sample mean of data (lowercase). As the sample mean $\left(\bar{Y}\right)$ is one of the most widely used statistics, it is essential to know two of its properties, namely its mean and variance. For randomly selected samples from infinite populations, mathematical properties of expected value can be used to derive the facts that

$$E\left(\bar{Y}\right) = \mu \quad \text{and} \quad V\left(\bar{Y}\right) = \frac{\sigma^2}{n}.$$

The expectation of the sample mean is the population mean, and the variance of the sample mean becomes smaller when the sample size is increased. It can also be shown that the variance of the sample mean can be estimated from sample data without bias using

$$\hat{V}\left(\bar{Y}\right) = \frac{s^2}{n}.$$

Detailed derivations of all results presented here and in the sections to follow can be found in the references at the end of this article.

In lieu of mathematical derivations, we will present a few simulation studies to demonstrate properties of the sample mean in sampling from both infinite and finite populations. Simulations are quite appropriate for studying the behavior of statistical techniques and are widely used to study complex methodologies that have no closed-form mathematical solutions.

A Simulation Study

Figure 2 shows a histogram of the observed sample means for 200 random samples, each of size 10, taken from the random-digit population. Figure 3 shows a corresponding histogram for samples of size 40. In each case the histogram suggests that the distribution of possible sample means is mound shaped and nearly symmetrical (approximately normal). Table 1 summarizes the means and standard deviations both for the simulated distributions and from the theory outlined above. Recall that the population mean for the random digits is 4.5, and the population standard deviation is 2.9.

Table 1: Sampling from an Infinite Population

	Simulation	Theory
Mean, $n = 10$	4.51	$\mu = 4.5$
Mean, $n = 40$	4.48	$\mu = 4.5$
SD, $n = 10$	0.945	$\sigma/\sqrt{n} = 0.917$
SD, $n = 40$	0.471	$\sigma/\sqrt{n} = 0.458$

Observe that the means and standard deviations from the simulations are quite close to what the theory says they should be.

Figure 2: Means from Samples of Size 10 from an Infinite Population

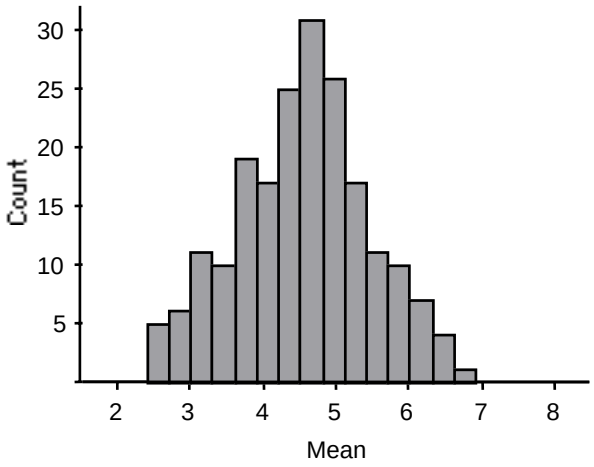
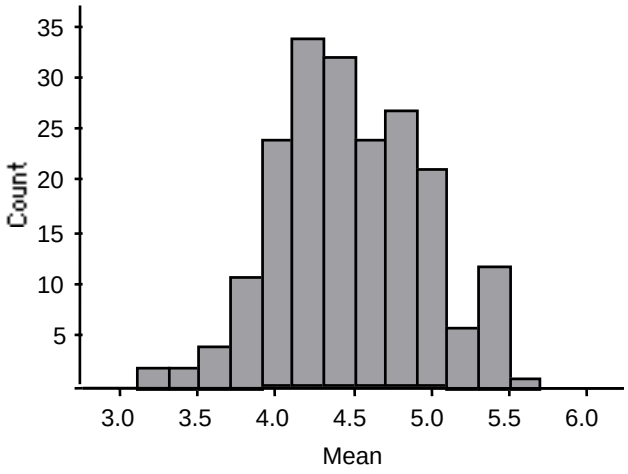


Figure 3: Samples of 40 from an Infinite Population



Now the population will be changed to a finite population consisting of 100 digits, 10 of each of the integers 0 through 9. **Random sampling** from such a finite population (without replacement) means that **each possible sample of size *n* has the same chance of being selected**. In practice, this is like mixing the numbered chips in a box and then pulling out *n* chips, using sampling with replacement. Sampling with replacement is achieved by putting a selected chip back into the box and mixing the chips before the next chip is selected. In this way, the box contains the original population of chips for each selection.

Figure 4 shows a simulated sampling distribution for means of random samples of size 10 from this population; Figure 5 shows the distribution for means of random samples of size 40. Note that both distributions still appear to be approximately normal in shape. Table 2 shows the summary statistics for the sample means computed from the simulated samples and makes comparisons to the theoretical expectations and standard deviations for the sample means.

Figure 4: Samples of Size 10 from a Finite Population

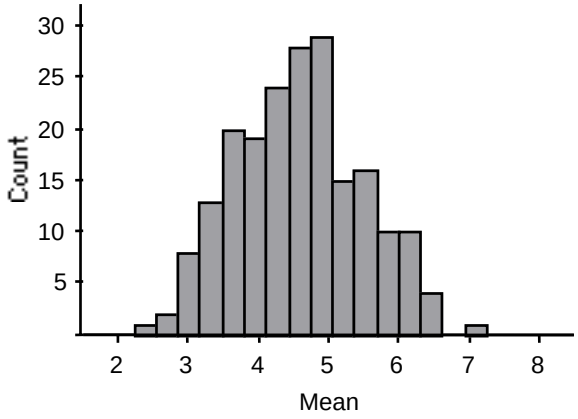


Figure 5: Samples of 40 from a Finite Population

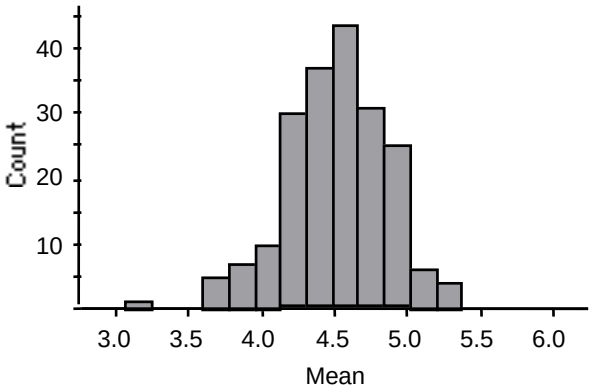


Table 2: Sampling from a Finite Population

	Simulation	Theory
Mean, $n = 10$	4.57	$\mu = 4.5$
Mean, $n = 40$	4.50	$\mu = 4.5$
SD, $n = 10$	0.904	$\sigma/\sqrt{n} = 0.917$
SD, $n = 40$	0.349	$\sigma/\sqrt{n} = 0.458$

Table 2 shows that the means of the simulated distributions are still very close to what the theory suggests; the same theory as used in the infinite population situation appears to work for finite populations. But something is wrong with the standard deviations. For sampling 10 items from 100, the infinite-population theoretical value appears to hold up

reasonably well, but for sampling 40 out of 100, that theoretical value appears to be too large. On giving this a little thought, it should seem reasonable that sampling from a finite population should produce means with less variability than when sampling from an infinite population. After all, if the sample size equaled the population size, then all samples would be identical to the population, and the sample means would equal the population mean, with no variability present. It follows, then, that the theoretical variance of sample means used in sampling from infinite populations must be scaled down a bit when sampling from finite populations. The "scaling down" factor is presented next.

Properties of the Sample Mean in Simple Random Sampling

Simple random sampling without replacement is achieved by randomly selecting units from a population without putting units that were previously selected back into the population. After each selection, the remaining population from which the next unit is selected changes slightly. If the number of units in the population (the population size) is very large, and the sample size is small, random sampling without replacement is for all practical purposes the same as random sampling with replacement. There can be noticeable differences, however, for small populations. For a random sample of size n selected without replacement from a finite population of size N (a simple random sample without replacement), probability theory provides the following results for the sample mean and its variance. (We are listing the highlights here; a more complete discussion can be found in the references at the end of this article.)

As in the infinite population case,

$$E\left(\overline{Y}\right) = \mu,$$

but the variance of the sample mean gets slightly more complicated:

$$V\left(\overline{Y}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

It is smaller than the variance of the sample mean for a sample of size n selected from the same population using simple random sampling with replacement. Considering the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

it can be shown that

$$E\left(S^2\right)=\left(\frac{N}{N-1}\right)\sigma^2.$$

Putting these formulas together and substituting the sample information, we can see that $V\left(\bar{Y}\right)$ can be estimated from the sample without bias by

$$\widehat{V\left(\bar{Y}\right)}=\frac{s^2\left(\frac{N-1}{N}\right)}{n}\left(\frac{N-n}{N-1}\right)$$
$$=\frac{s^2}{n}\left(\frac{N-n}{N}\right).$$

The variance of the estimator $\bar{Y}$ is the same as that given in an introductory course except that it is multiplied by a correction factor to adjust for sampling from a finite population. The correction factor takes into account the fact that an estimate based on a sample $n = 10$ from a population of $N = 20$ contains more information about the population mean than a sample of $n = 10$ from a population of $N = 20{,}000$. Back in Table 2, if the theoretical standard deviation for samples of size 40 (0.458) is adjusted by the finite population correction with $N = 100$ and $n = 40$, the resulting value is 0.353. This is much closer to the value of the standard deviation of the sample means obtained in the simulation, which was 0.349.

To illustrate how these results hold, consider all possible random samples of size $n = 2$ selected from the population {1, 2, 3, 4}. Table 3 shows the six possible samples of size 2 and the related sample statistics.

Table 3: Simple Random Sampling of a Finite Population, n = 2

Sample	Probability of Sample	$\bar{Y}$	S^2	$\hat{V}\left(\bar{Y}\right)$
{1, 2}	1/6	1.5	0.5	0.125
{1, 3}	1/6	2.0	2.0	0.500
{1, 4}	1/6	2.5	4.5	1.125
{2, 3}	1/6	2.5	0.5	0.125
{2, 4}	1/6	3.0	2.0	0.500
{3, 4}	1/6	3.5	0.5	0.125

If a single observation Y is selected at random from this population, then Y can take on any of the four possible values, each with probability $\frac{1}{4}$. Thus,

$$\mu = E(Y) = \sum yp(y) = 1\left(\frac{1}{4}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{4}\right) + 4\left(\frac{1}{4}\right)$$
$$= \left(\frac{1}{4}\right)(1+2+3+4) = \frac{1}{4}(10) = 2.50,$$

and

$$\sigma^2 = V(Y) = E(Y-\mu)^2 = \sum (y-\mu)^2 p(y)$$
$$= (1-2.5)^2\left(\frac{1}{4}\right) + (2-2.5)^2\left(\frac{1}{4}\right) + (3-2.5)^2 + (4-2.5)^2\left(\frac{1}{4}\right)$$
$$= \frac{5}{4}.$$

Since each of these sample means can occur with probability $\frac{1}{6}$, we can exactly compute $E(\bar{Y})$ and $V(\bar{Y})$. From our definition of expected value,

$$E(\bar{Y}) = \sum \bar{y}p(\bar{y}) \qquad \text{(summed over all values of } \bar{y}\text{)}$$
$$= (1.5)\left(\frac{1}{6}\right) + (2.0)\left(\frac{1}{6}\right) + (2.5)\left(\frac{1}{6}\right) + (2.5)\left(\frac{1}{6}\right)$$
$$+ (3.0)\left(\frac{1}{6}\right) + (3.5)\left(\frac{1}{6}\right)$$
$$= 2.50 = \mu,$$

and

$$V(\bar{Y}) = E(\bar{Y}-\mu)^2 = \sum (\bar{y}-\mu)^2 p(\bar{y})$$
$$= (1.5-2.5)^2\left(\frac{1}{6}\right) + (2.0-2.5)^2\left(\frac{1}{6}\right) + (2.5-2.5)^2\left(\frac{1}{6}\right)$$
$$+ (2.5-2.5)^2\left(\frac{1}{6}\right) + (3.0-2.5)^2\left(\frac{1}{6}\right) + (3.5-2.5)^2\left(\frac{1}{6}\right)$$
$$= (2.5)\left(\frac{1}{6}\right) = \frac{5}{12}.$$

Recall that for this example $\sigma^2 = \frac{5}{4}$, $N = 4$, and $n = 2$, and we have

$$\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{(5/4)}{2}\left(\frac{4-2}{4-1}\right)$$
$$= \frac{5}{8}\left(\frac{2}{3}\right) = \frac{5}{12}$$

Considering the sample variances, we have

$$E\left(S^2\right) = \frac{(0.5+2.0+4.5+0.5+2.0+0.6)}{6} = \frac{5}{3}$$
$$= \left(\frac{4}{4-1}\right)\frac{5}{4} = \frac{5}{3}.$$

Thus we see that $E\left(S^2\right) = \left(\frac{N}{N-1}\right)\sigma^2$.

Also, we can demonstrate that our estimator of $V\left(\bar{Y}\right)$ is unbiased. Using our values of $\hat{V}\left(\bar{Y}\right)$ from Table 3,

$$E\left(\hat{V}\left(\bar{Y}\right)\right) = \sum\left[\left(\hat{V}\left(\bar{Y}\right)\right)P\left(\hat{V}\left(\bar{Y}\right)\right)\right] \text{ [summed over all values of } \hat{V}\left(\bar{Y}\right)\text{]}$$
$$= (0.125)\left(\frac{1}{6}\right) + (0.5)\left(\frac{1}{6}\right) + (1.125)\left(\frac{1}{6}\right)$$
$$+ (0.125)\left(\frac{1}{6}\right) + (0.5)\left(\frac{1}{6}\right) + (0.125)\left(\frac{1}{6}\right)$$
$$= \frac{5}{12}$$
$$= V\left(\bar{Y}\right).$$

At this point we have demonstrated the following:

1. $E\left(\bar{Y}\right) = \mu$
2. $V\left(\bar{Y}\right) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$
3. $E\left(\hat{V}\left(\bar{Y}\right)\right) = V\left(\bar{Y}\right)$

Note that this demonstration is not a proof or derivation of a general result; such a proof would require much deeper probabilistic arguments than those used in deriving results for sampling from infinite populations.

Stratified Random Sampling

Often it is convenient and appropriate to divide a population into nonoverlapping groups and then sample from within each group. A population of high school students, for example, could be divided by grade level, with a sample taken from each. National surveys, like the Gallup Poll, divide the United States into geographical regions and then take a sample from each region. Such subdivisions of a population are called **strata**. A stratified random sample consists of dividing a population into strata and then selecting a simple random sample from each stratum.

Stratified sampling designs are often used for convenience (it is nearly impossible to select a random sample from the totality of the United States) or because estimates are desired for each of the strata (you may want to know how seniors differ from sophomores on opinion-poll questions). One of the main statistical purposes for stratification, however, is that a properly stratified sample survey can produce estimates with smaller variance than those from a simple random sample of the same size. Thus, stratification can provide more information per dollar than a simple random sample. This aspect of stratification will be illustrated below.

In developing the stratified random sampling estimator (a function of the sample data and other known constants) of a population mean, it helps to begin by considering the estimation of a population total. You do not see estimates of totals in introductory statistics books because a total does not make sense for an infinite population. But often a survey's objective is a total, such as estimating the total crop yield for a county or the total value of all properties in a neighborhood.

Suppose a population is divided into L strata. Let $\bar{Y}_i$ denote the sample mean for a simple random sample of n_i units selected from stratum i. Let N_i denote the population size for stratum i, μ_i the population mean for stratum i, and τ_i the population total for stratum i.

Then the population total is equal to $\tau_1 + \tau_2 + ... + \tau_L$. In a simple random sample within each stratum, $\bar{Y}_i$ is an unbiased estimator of μ_i, and $N_i\bar{Y}_i$ is an unbiased estimator of the stratum total. It would seem reasonable to form an estimator of the sum of the τ_i's by summing the estimators of the τ_i's. Since the population mean μ equals the population total divided by $N = N_1 + N_2 + ... + N_L$, an unbiased estimator of μ is obtained by summing the estimators of the τ_i's over all strata and then dividing by *N*. This estimator is denoted by $\bar{Y}_{st}$, where the subscript *st* indicates that stratified random sampling is used:

$$\bar{Y}_{st} = \frac{1}{N}\left[N_1\bar{Y}_1 + N_2\bar{Y}_2 + ... + N_L\bar{Y}_L \right] = \frac{1}{N}\sum_{i=1}^{L} N_i\bar{Y}_i .$$

An estimated variance is produced by summing the estimated variances of the components of the stratified sampling mean. Such a summation of variances is only valid if the samples are independent of one another, which is a general condition for stratified random sampling. The variance of the estimated mean is

$$V\left(\bar{Y}_{st}\right) = \frac{1}{N^2}\left[{N_1}^2 V(\bar{Y}_1) + {N_2}^2 V\left(\bar{Y}_2\right) + ... + {N_L}^2 V\left(\bar{Y}_L\right) \right].$$

The unbiased estimate of this variance is found by substitution of the individual estimates of the variances for the strata:

$$\widehat{V\left(\bar{Y}_{st}\right)} = \frac{1}{N^2}\left[{N_1}^2\left(\frac{N_1 - n_1}{N_1}\right)\left(\frac{{s_1}^2}{n_1}\right) + {N_2}^2\left(\frac{N_2 - n_1}{N_2}\right)\left(\frac{{s_2}^2}{n_2}\right) + ... + {N_L}^2\left(\frac{N_L - n_L}{N_L}\right)\left(\frac{{s_L}^2}{n_L}\right)\right]$$
$$= \frac{1}{N^2}\sum_{i=1}^{L} {N_i}^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{{s_i}^2}{n_i}\right).$$

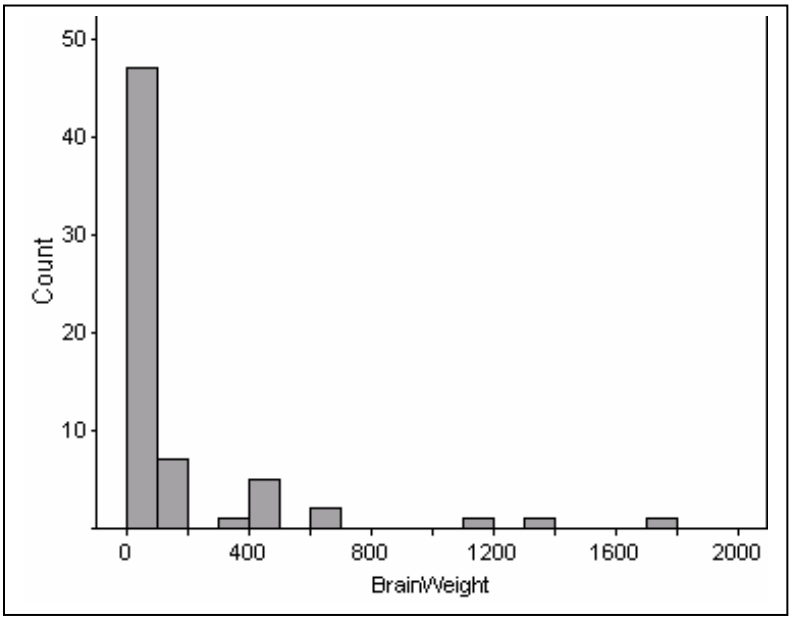
A Simulation Study

The data shown in Table 4 are the brain weights for a selection of 65 animals that will be our population for a simulation study that compares simple random sampling to stratified random sampling. While data such as these are of interest to ecologists and biologists, our use of them will be blatantly pedagogical. We simply wish to use a more interesting population than mathematically derived random numbers; the techniques we will discuss apply to all sorts of populations, even ones that are not distributed normally or uniformly on some variable. Animals are stratified into two groups, large and small, based on their weights. The mean of this population is 149 grams, and the standard deviation is 323 grams; the distribution is skewed toward the larger values, as can be seen in the histogram of Figure 6.

Table 4: Brain Weight Data

Species	Brain Weight (grams)	Size 1=Large	Species	Brain Weight (grams)	Size 1=Large
African giant pouched rat	6.6	2	Little brown bat	0.25	2
Arctic fox	44.5	2	Loon	6.12	2
Arctic ground squirrel	5.7	2	Mackerel	0.64	2
Baboon	179.5	1	Man	1320	1
Barracuda	3.83	2	Mole rat	3	2
Big brown bat	0.3	2	Musk shrew	0.33	2
Brown trout	0.57	2	Nine-banded armadillo	10.8	2
Canary	0.85	2	North American opossum	6.3	2
Cat	25.6	2	Northern trout	1.23	2
Catfish	1.84	2	Ostrich	42.11	2
Chimpanzee	440	1	Owl monkey	15.5	2
Chinchilla	6.4	2	Pheasant	3.29	2
Cow	423	1	Pig	180	1
Crow	9.3	2	Pigeon	2.69	2
Desert hedgehog	2.4	2	Porpoise	1735	1
Donkey	419	1	Rabbit	12.1	2
Eastern American mole	1.2	2	Raccoon	39.2	2
European hedgehog	3.5	2	Rat	1.9	2
Giant armadillo	81	1	Red fox	50.4	1
Giraffe	680	1	Rhesus monkey	179	1
Goat	115	1	Roe deer	98.2	1
Golden hamster	1	2	Salmon	1.26	2
Gorilla	406	1	Seal	442	1
Gray seal	325	1	Sheep	175	1
Gray wolf	119.5	1	Stork	16.24	2
Ground squirrel	4	2	Tree shrew	2.5	2
Guinea pig	5.5	2	Tuna	3.09	2
Flamingo	8.05	2	Vulture	19.6	2
Horse	655	1	Walrus	1126	1
Jaguar	157	1	Water opossum	3.9	2
Kangaroo	56	1	Yellow-bellied marmot	17	2
Lesser short-tailed shrew	0.14	2			

Figure 6: Distribution of Population of Brain Weights



First, 200 simple random samples of size 10 were selected from this population to produce a simulated approximation to the sampling distribution shown in Figure 7 (bottom). Notice that the sampling distribution is slightly skewed because of the extreme skewness in the population and the small sample size. Second, the population was divided somewhat arbitrarily into "large" and "small" animals, and stratified random samples of size 5 from each of the two strata were selected 200 times. This simulated distribution, shown in Figure 7 (top), still has some skewness, but its spread is much smaller than that for simple random sampling. Table 5 provides a comparison of the summary statistics. The means of these sampling distributions are both close to the population mean of 149 grams, but the standard deviation of the sample mean is cut almost in half by the stratified design, even though the same overall sample size of 10 is used.

Figure 7: Simple Random Sampling Mean Versus Stratified Random Sampling Mean

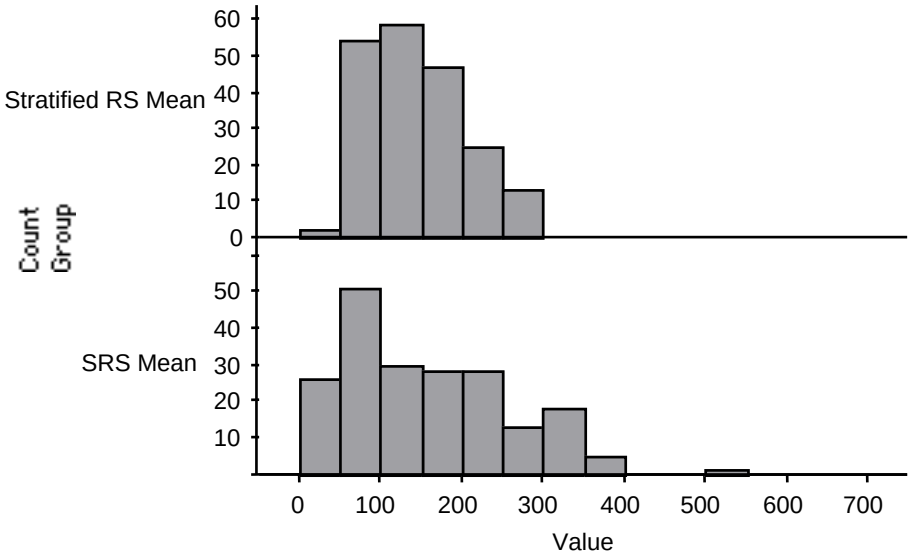


Table 5: Summary Statistics, Simple Random Versus Stratified Random Sampling

	Simple Random Sample $n = 10$ $N = 65$	**Stratified Random Sample** $n_1 = 5,\ n_2 = 5$ $N_1 = 22,\ N_2 = 43$
Mean	156.4	144.0
Standard Deviations	100.6	59.1

Stratifying the population into large and small animals also stratified the brain weights into large and small. The large animals have a mean brain weight of 425 grams and the small ones 8 grams. That difference in group means is the key factor in how the stratification produces a sample estimate of the population mean with a smaller standard deviation than a simple random sample of the same size would produce. Essentially, the deviations (and hence the variance) calculated around the separate sample means will be smaller than the deviations (and hence the variance) calculated around the "composite" mean. Even greater reduction in variation is achieved if the resulting strata standard deviations are smaller than the population standard deviations. The rules for stratification, then, are to choose strata such that:

- The stratum means differ as much as possible.
- The data values within a stratum vary as little as possible.

For simplicity, consider the special case where the strata sample sizes are proportional to the strata population sizes, i.e., $n_i = n\left(\frac{N_i}{N}\right)$.

It can be shown mathematically that the difference between the variance of the sample mean for simple random sampling and the variance of the stratified sampling mean is

$$V(SRS) - V(StrRS) = \frac{1}{nN}\left(\frac{N-n}{N-1}\right)\left[\sum N_i\left(\mu_i - \mu\right)^2 - \frac{1}{N}\sum\left(N - N_i\right)V\left(\overline{Y}_i\right)\right],$$

where μ_i represents the mean for stratum i and $\mu = \frac{N_i\mu_i}{N}$ represents the overall population mean. From this formula we can see that stratified random sampling pays dividends in terms of smaller variance if strata are constructed to have very different population means. Wise use of stratification will produce an estimator of the population mean with smaller variance than the estimator obtained from simple random sampling with the same total sample size. It is mathematically possible for $V(SRS) < V(StrRS)$; that is,

$$\sum N_i\left(\mu_i - \mu\right)^2 < \frac{1}{N}\sum\left(N - N_i\right)V\left(\overline{Y}_i\right).$$

This occurs when the stratum means are nearly the same, and there is large variability within some strata resulting in relatively large values of $V\left(\overline{Y}_i\right)$. This cannot be exactly verified, of course, in the planning stage of most surveys because precise information on stratum means and within stratum variance will not be available, but stratified sampling should be considered if previous experience suggests that strata can be constructed such that within each stratum variability is relatively small, and there are substantial differences among stratum means.

Cluster Sampling

It might be quite difficult to actually locate and invite student participation in a random-sample survey about school policies. (Think of some possible reasons why this may be so.) It would be easier and quicker to select a random sample of classrooms and then ask every student in those classes to complete a questionnaire. This method is called **cluster sampling**. Cluster sampling often is used in sampling situations where it is difficult or impossible to develop a list of the elements of the population you would like to sample. In a survey of households in a city, for example, it is difficult to develop an accurate list of currently occupied housing units. It is much easier to list the city blocks, select a random sample of blocks, and then go interview a person in each household within each of the sampled blocks. This also may reduce the cost of conducting the survey by producing substantial reductions in travel time and expense relative to what would be required to travel to households selected in a simple random sample of city households.

Cluster sampling is simple random sampling within sampling units (clusters), with each sampling unit containing a number of elements, each of which is included in the sample. Hence, the estimator of the population mean μ is equal to that for simple random sampling.

The following notation is used in this section:

N = Number of clusters in the population
n = Number of clusters selected in a simple random sample
m_i = Number of elements in cluster i, $i = 1, \ldots, N$
$\overline{m} = \frac{1}{n}\sum_{i=1}^{n} m_i =$ Average cluster size for the sample
$M = \sum_{i=1}^{N} m_i$ = Number of elements in the population
$\overline{M} = \frac{M}{N}$ = Average cluster size for the population
$\overline{y}_i$ = Average of all observations in the i^{th} cluster

The estimator of the population mean μ is the sample mean per element $\overline{Y}$, which is given by

$$\overline{Y} = \frac{\sum_{i=1}^{n} m_i \overline{y}_i}{\sum_{i=1}^{n} m_i} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}.$$

Notice that hidden inside the algebra is the simple notion that we are *still* just adding the values for the observations and then dividing by the number of observations to get $\bar{Y}$. We multiply the sample means by the sample sizes to get sample totals and then divide by the sum of the sample sizes.

This estimator also can be viewed as a weighted average of the cluster means with the cluster sizes serving as the weights. Consequently, larger clusters have greater influence on the estimator than smaller clusters. The estimated variance of $\bar{Y}$ (which is a little too complicated to derive here) is given by

$$\hat{V}(\bar{Y}) = \left(\frac{N-n}{Nn\bar{m}^2}\right) s_r^2,$$

where

$$s_r^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y}m_i)^2}{n-1}.$$

This estimated variance is biased and a good estimator of the true $V\left(\bar{Y}\right)$ only if n is large—say, $n \geq 20$. The bias disappears if the cluster sizes $m_1, m_2, ..., m_N$ are equal. The following example will illustrate the calculations for estimation based on a cluster sample.

For estimating a city's per-capita annual income, a random sample of 25 blocks is selected from the city's total of 415 blocks. The data on incomes are presented in Table 6. We want to use the data to estimate the per-capita income in the city and find a margin of error for the estimate.

Table 6: Per-Capita Income

Cluster	Number of Residents, m_i	Total Income per Cluster, y_i	$y_i - \bar{y}m_i$
1	8	96,000	25589.404
2	12	121,000	15384.106
3	4	42,000	6794.702
4	5	65,000	20993.377
5	6	52,000	–807.947
6	6	40,000	–12807.947
7	7	75,000	13390.728
8	5	65,000	20993.377
9	8	45,000	–25410.596
10	3	50,000	2359.026
11	2	85,000	67397.351
12	6	43,000	–9807.947
13	5	54,000	9993.377
14	10	49,000	–39013.245
15	9	53,000	–26211.921
16	3	50,000	23596.026
17	6	32,000	–20807.947
18	5	22,000	22006.623
19	5	45,000	993.377
20	4	37,000	1794.702
21	6	51,000	–18079.470
22	8	30,000	–40410.596
23	7	39,000	–22609.272
24	3	47,000	20596.026
25	8	41,000	–29410.596

$$\sum_{i=1}^{25} m_i = 151 \qquad \sum_{i=1}^{25} y_i = 1,329,000$$

A summary of the basic statistics for these data on a cluster basis is as follows.

	N	Mean	Total	SD
Number of Residents	25	6.040	151	2.371
Income/Cluster	25	53,160	1,329,000	21,784
$y_i - \bar{y}m_i$	25	0	0	$s_r = 25,189.31$

The best estimate of the population mean is

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} = \frac{\$1,329,000}{151} = \frac{\$53,160}{6.04} = \$8,801.$$

The $\bar{m}$ needed for the variance calculation is

$$\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n} = \frac{151}{25} = 6.04.$$

Then, putting all this together,

$$\widehat{V(\bar{Y})} = \left(\frac{N-n}{Nn\bar{m}^2}\right) s_r^2$$
$$= \left[\frac{415-25}{(415)(25)(6.04)^2}\right](25,189)^2 = 653,785$$

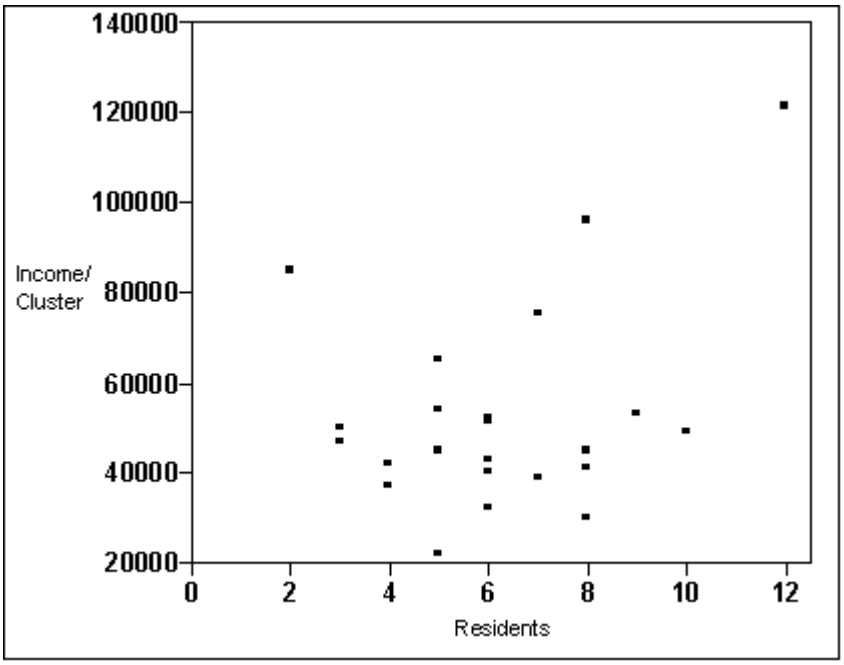
Thus, the estimate of μ, the population mean, with an appropriate margin of error is given by

$$\bar{y} \pm 2\sqrt{\widehat{V(\bar{Y})}} = 8801 \pm 2\sqrt{653,785} = 8801 \pm 1617.$$

The best estimate of the average per-capita income is \$8,801, and the error of estimation should be less than \$1,617, with probability close to 0.95. This margin of error is rather large and could be reduced by sampling more clusters.

Cluster sampling produces estimators with small variance when the cluster totals are directly proportional to the cluster sizes. In such cases, the s_r^2 component of the variance is small, and the plot of cluster totals versus cluster sizes has a linear pattern with strong positive association. The plot for the data of Table 6, seen in Figure 8, shows a relatively weak positive association, which is why the margin of error is so large for this example.

Figure 8: Plot of Data from Table 6



If cluster totals in a cluster sample were directly proportional to their cluster sizes, with all clusters having the same constant of proportionality, then the cluster means would be equal to the population mean. Thus cluster sampling will produce an estimate of the population mean with a relatively small variance when:

- The cluster means are nearly alike.
- The data values within a cluster exhibit a great deal of variability.

Notice that this is just the opposite of the optimal way of constructing strata, which called for differing strata means and homogeneity among data values within strata. The basic reason for this difference is that in cluster sampling the only random selection is at the cluster level. If a cluster design samples five classrooms at random, and each contains 20 students to interview, then it is desirable that each classroom have a variety of students in them rather than a whole set of 20 students who think alike. The investigator would like to get 100 opinions on the issues at hand, not five opinions replicated 20 times. On the other hand, if students at each grade level tend to think alike on the issues being investigated, then it would pay to stratify on grade level and randomly select, say, 25 students from each of the four grades. If there were no apparent differences of opinion across grade levels, then it would be best to select a simple random sample of 100 students from the school. Cluster sampling, usually done for convenience or to save on costs,

generally results in greater variation of estimates (with a fixed sample size) than the other two designs discussed above. That is because cluster sampling involves less randomization than the other two designs. A comparison of these three designs is provided in the simulation that follows.

Estimating Mean GPA by Sampling

Consider a population of GPAs taken from a class of 60 undergraduate students taking an introductory statistics class. These data are shown in Table 7.

Table 7: GPA Data

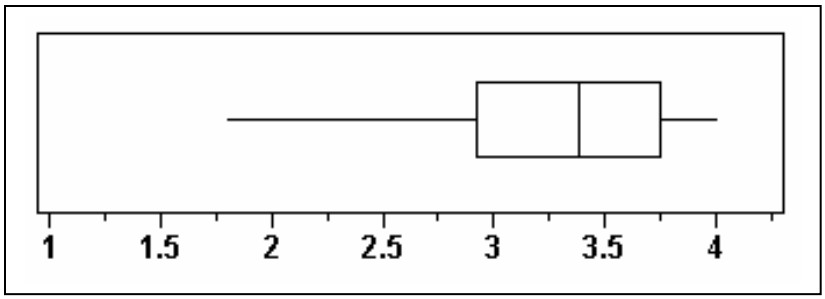
GPA	Gender 1 = F, 2 = M	Ordered	Random	Cluster Groups
2.75	1	1.80	2.75	1
2.80	1	2.00	3.49	1
2.80	1	2.27	2.80	1
2.87	1	2.30	3.50	1
2.90	1	2.30	2.87	1
3.00	1	2.32	3.80	2
3.00	1	2.47	3.80	2
3.03	1	2.50	4.00	2
3.10	1	2.60	3.81	2
3.15	1	2.75	1.80	2
3.20	1	2.80	3.80	3
3.25	1	2.80	4.00	3
3.25	1	2.87	3.16	3
3.30	1	2.90	3.40	3
3.30	1	2.90	4.00	3
3.30	1	3.00	3.80	4
3.37	1	3.00	3.76	4
3.40	1	3.00	3.70	4
3.40	1	3.00	3.50	4
3.49	1	3.03	4.00	4
3.50	1	3.10	3.60	5
3.50	1	3.15	4.00	5
3.50	1	3.16	2.30	5
3.50	1	3.20	3.57	5
3.53	1	3.25	2.32	5

Table 7: GPA Data (continued)

GPA	Gender 1 = F, 2 = M	Ordered	Random	Cluster Groups
3.60	1	3.25	3.50	6
3.66	1	3.30	3.73	6
3.73	1	3.30	3.20	6
3.76	1	3.37	2.90	6
3.78	1	3.40	2.50	7
3.78	1	3.40	3.00	7
3.78	1	3.49	3.78	7
3.80	1	3.50	2.47	7
3.80	1	3.50	2.80	7
3.80	1	3.50	3.30	8
3.81	1	3.50	3.25	8
3.87	1	3.53	3.30	8
4.00	1	3.57	2.00	8
4.00	1	3.60	3.25	8
4.00	1	3.66	3.87	9
4.00	1	3.66	3.00	9
2.60	1	3.70	3.37	9
2.00	1	3.70	3.10	9
3.00	1	3.73	2.60	9
3.16	2	3.76	3.78	10
2.47	2	3.78	3.40	10
2.30	2	3.78	3.03	10
1.80	2	3.78	3.30	10
2.30	2	3.80	3.66	10
3.70	2	3.80	2.90	11
3.00	2	3.80	3.15	11
3.70	2	3.80	3.70	11
3.80	2	3.81	3.50	11
2.27	2	3.87	2.30	11
2.90	2	4.00	2.27	12
3.57	2	4.00	3.78	12
4.00	2	4.00	3.53	12
2.32	2	4.00	3.00	12
2.50	2	4.00	3.66	12

Figure 9 shows a boxplot of these data.

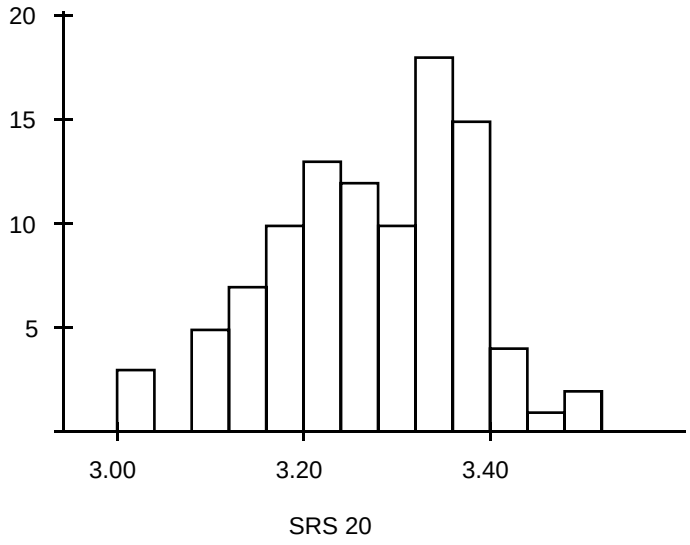
Figure 9: Distribution of Class GPAs



Summary of GPA data:		
$N = 60$	Mean 3.27	Standard deviation 0.55

Simple random samples of size $n = 20$ each selected from this population resulted in the distribution of sample means shown in Figure 10.

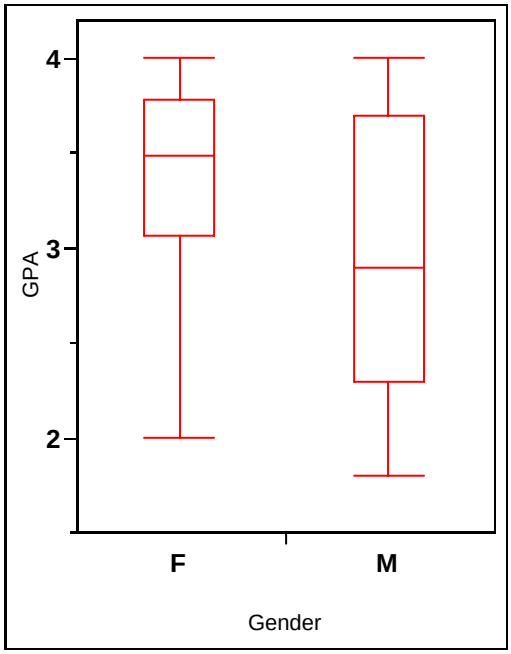
Figure 10: Distribution of Sample Means from Simple Random Sampling



Summary of means from simple random sampling:	
Mean 3.27	Standard deviation 0.105

Since the female students taking this course tend to do a little better than male students, perhaps stratification on gender would produce sample estimates of the mean with less variability. The boxplots of the GPAs by gender shown in Figure 11 (1 = female) show that females do, indeed, have the higher mean, and so stratifying on gender may be a good thing to do.

Figure 11: GPAs by Gender

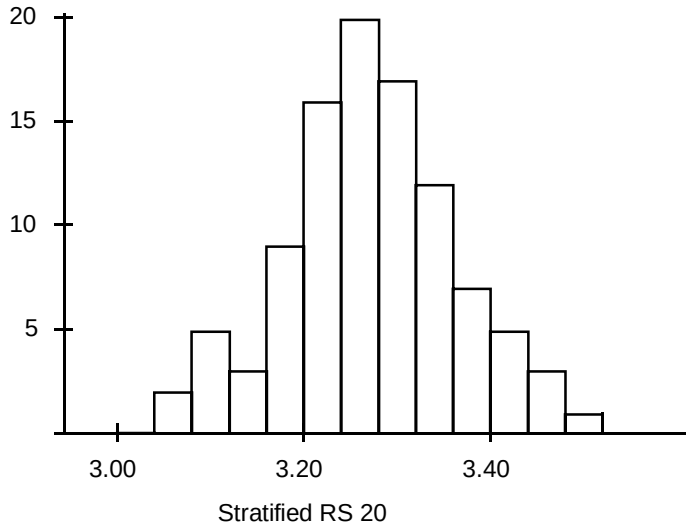


Summary of GPA data by gender			
	Count	Mean	Standard deviation
F	45	3.39	0.436
M	15	2.92	0.698

Now the object is to select stratified random samples of size $n = 20$ each to compare the results with simple random sampling, also using samples of size $n = 20$. A good way to choose sample sizes for the strata is to make them proportional to the sizes of the strata in the population. Since females comprise 75 percent of the population, they may as well comprise 75 percent of the sample. Thus the sample sizes were set at 15 and 5 for females and males, respectively.

The simulation results for the stratified sample means are provided in Figure 12. Notice that stratification reduced the standard deviation of the distribution of mean estimates from 0.105 to 0.091, a small decrease. The improvement due to stratification would have been more dramatic if the means for the genders were farther apart and if the standard deviations within the two groups were much smaller than the overall population standard deviation. In this case, the standard deviation for male GPAs is nearly as large as the overall-population standard deviation.

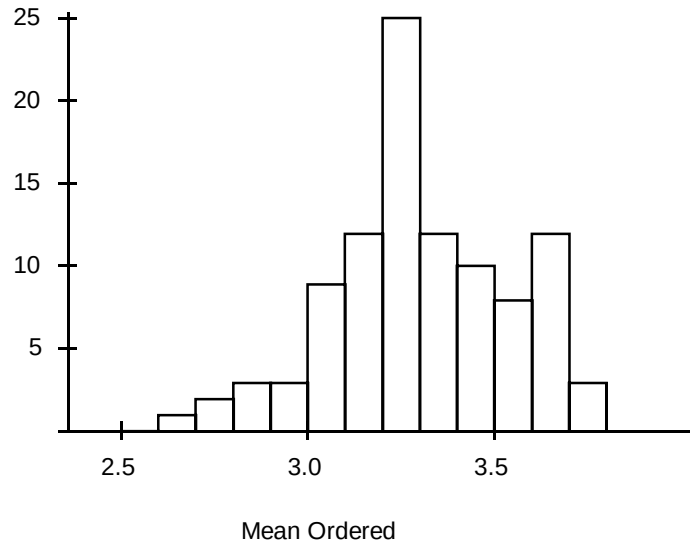
Figure 12: Distribution of Sample Means from Stratified Random Sampling



Summary of means from stratified random sampling		
Mean 3.27	Standard deviation 0.091	
$n = 20$	$n_1 = 15$	$n_2 = 5$

Going on to cluster sampling, the first consideration is how to form clusters. For illustrative purposes, this will be done in two ways. First, the population data values are ordered from smallest to largest and then grouped into clusters of size $m = 5$. Randomly selecting four such clusters results in a sample size of 20, as was used above. Notice this method will produce clusters with different means but with fairly homogeneous data values within each cluster. The resulting distribution of sample means from repeating this sampling design many times is shown in Figure 13. The standard deviation of sample means from this method of clustering is more than *double* that of either simple random sampling or stratified random sampling.

Figure 13: Distribution of Sample Means from Cluster Sampling, Ordered Clusters

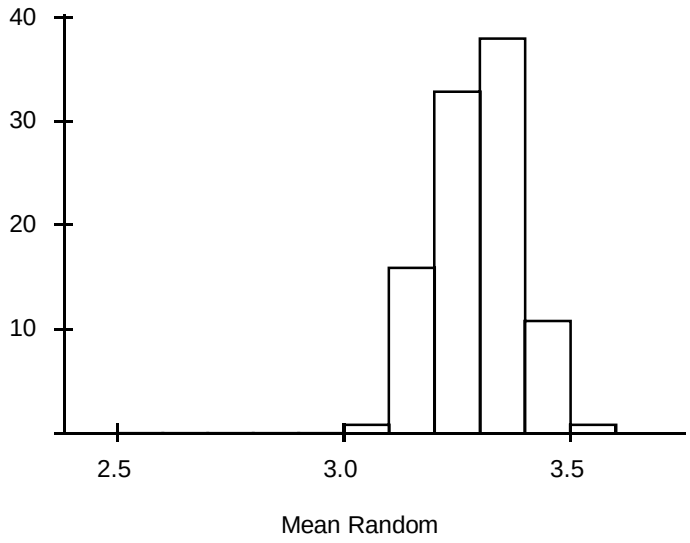


Summary of means (ordered)	
Mean 3.30	Standard deviation 0.232

But the clustering can be accomplished in other ways. In the second method of clustering, the population data values are arranged in random order, and clusters are formed by sequentially grouping the randomly ordered data.

This produces clusters that tend to have similar means and tend to have values within the clusters that vary a great deal. Again, repeated selection of four clusters each of size 5 produced the distribution of sample means shown in Figure 14. The standard deviation of these sample means dropped to about the level of the stratified random sampling design. In general, this method of random grouping to form clusters should give results that are approximately equivalent to what would be acquired from simple random sampling.

Figure 14: Distribution of Sample Means from Cluster Sampling, Random Clusters



Summary of means (random)	
Mean 3.30	Standard deviation 0.089

Sample Surveys Versus Experiments

A reiteration of some of the issues discussed in the introduction may be helpful here, and we offer a closing note on the comparison between sample surveys and experiments. These two types of statistical investigations have some common elements, as each requires randomization both for purposes of reducing bias and building a foundation for statistical inference. Each provides data needed to apply commonly used inference mechanisms of confidence interval estimation and hypothesis testing. But these two types of investigations have very different objectives and requirements. Sample surveys are used to estimate or make decisions about parameters of populations, and they require a well-defined, fixed population as their main ingredient. Experiments are used to estimate or compare the effects of treatments and require well-defined treatments and experimental units on which to study those treatments. Randomization comes into play in sample surveys because a random sample is required in order to generalize the results. Randomization comes into play in experiments because random assignment to treatments facilitates the search for cause-and-effect conclusions. These are very different uses of the concept of randomization, and they have different consequences.

Estimating the proportion of city residents that would support an increase in taxes for education requires a sample survey. If the random selection of residents is done in an appropriate manner, then the results from the sample can be expanded to represent the

population from which the sample is selected. A measure of sampling error can be calculated to ascertain how far the estimate is likely to be from the true value.

Testing to see if a new medication to improve breathing for asthma patients produces greater lung capacity than a standard medication requires an experiment in which a group of patients who have consented to participate in the study are randomly assigned to either the new or the standard medication. With this type of randomized comparative design, an investigator can determine, with a measured degree of uncertainty, whether or not the new medication caused an improvement in lung capacity. Generalization extends only to the types of units used in the experiment, however, as the experimental units are not usually sampled randomly from a larger population. Arguments used to extrapolate conclusions to a larger population must be based on information, expertise, or opinions developed outside of the experiment. To confidently generalize to a larger class of experimental units, more experiments would have to be conducted. That is one reason why replication of studies is a hallmark of good science.

References

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Groves, Robert M., Donald Dillman, John Eltinge, and Roderick Little, eds. 2002. *Survey Nonresponse*. New York: Wiley.

Lohr, S. 1999. *Sampling: Design and Analysis*. Pacific Grove, California: Brooks Cole.

Scheaffer, Richard, William Mendenhall, and R. Lyman Ott. 1996. *Elementary Survey Sampling*. 5th ed. Belmont, California: Duxbury Press.

Using Sampling in a Classroom Setting

Peter Flanagan-Hyde
Phoenix Country Day School
Paradise Valley, Arizona

In the preceding section, we discussed details about the theory and motivation of sampling; now it is time to put these ideas into practice. As a classroom teacher, you have a class of students, some of whom are eager to learn and others who are less intrinsically motivated. Conducting a survey that probes areas of student concern is an excellent way to bring home to them the ideas of sampling and surveys. This is best done with students just as it occurs in "real life"—with a researchable problem about a population that the students are engaged with, often one in which they, indeed, are members. Examples of these populations might be your school community or teenagers of the same (or opposite) gender. Your students will be very receptive if you can provide them with the tools necessary to answer their research questions.

Conducting a survey will also present them with situations in which they can experience some of the frustrations of sampling—problems with nonresponse and sampling instruments as well as issues of the cost (in time and energy) of gathering data. This, however, is not necessarily a bad thing, since it will influence many of the choices made in the planning of a study.

Of course, the gulf between a research problem and its survey solution is wide and challenging; while it is best to leap over the Grand Canyon in one step, planning a survey is quite another thing. In the paragraphs to follow, we will discuss the planning of surveys, step by step.

Part 1: Performing the Initial Planning

The first steps in planning a survey are to identify a population of interest and the research question. What is the population that you want to know something about? This is an important first question that must be clear before you begin. This population must be one to which you have reasonably easy access. Planning a survey on recording artists' opinions about trading songs on the Web might be interesting but not practical. Examples of populations to which you would have reasonable access are:

- Seniors at my school who are applying to college
- Teachers at my school

Given your research question, what would you like to know about this population? Which aspects or characteristics are related to the problem you are interested in? Formulating the general purpose of the survey is an important step in the process. Then you will clarify exactly what questions you will ask or what measurements you will obtain from members of the population.

Part 2: Designing Your Sampling Instrument

When setting out to design a sample survey, it is important to begin with a clear picture of the population characteristic(s) that you hope to measure. You will need to decide on appropriate units as well, since for some characteristics there may be more than one way to measure what you are asking about. For example, let us say you want to survey your school's students to determine their average commute to school. Without a clearly worded question, one student might respond, "20 minutes," while her older brother might say, "five miles." Either answer might be an appropriate measure, but you should decide in advance which units you will use to measure the characteristic of interest. Also, think about whether you want to measure a proportion of the population that has a particular characteristic or whether you want to take a numerical measure. With a question like "Do you plan to take calculus?" you will be estimating a proportion. A numerical measure, on the other hand, will follow in response to the question, "How many math classes do you plan on taking in high school?"

When framing questions, think about how to achieve consistent responses. In casual conversation, you might ask someone, "Do you go to Major League Baseball games?" But, as part of a survey about entertainment options, this would generate a wide array of responses, many of which would be difficult to compare. A better question might be "How many Major League Baseball games did you attend last season?" This is unambiguous and will generate responses on a consistent scale.

Be on the lookout for confusing questions; often your intent may not be clear to all the survey respondents. If you are interested in students' religious practices, you might ask the question, "Do you regularly attend church?" This question may be clear in your mind, but would a Muslim or Buddhist student who regularly attends religious events say yes? Perhaps the question rephrased as "Do you regularly attend church or other religious services?" would better elicit the responses that you want.

Questions should not be too long and should be phrased in neutral language that doesn't hint at an outcome. And if the question is asked in person, you will want the language of the question to be especially simple and straightforward. If several interviewers will be used, you will want to provide some training and practice run-throughs to ensure that the surveys are conducted in a uniform manner. (For example, what should the interviewer do if a respondent is reluctant or evasive in answering a question?)

Part 3: Choosing a Sampling Method

There are several sampling methods that might be employed to complete a survey. The most simple sampling strategies result in **simple random samples**, **stratified random samples**, and **cluster samples**. The following set of questions is designed to guide you to an appropriate choice of sampling method for your population.

Question 1: Is the population relatively homogeneous?

If yes, you can consider a **simple random sample** of your population. To do this, assign a number to each member of the population on the list. Using a random number process (a random number table, calculator, or computer), you can select a simple random sample of your population of the desired sample size.

> Example: For the population of teachers at your large public high school, you'd like to estimate the average length of time that they have been employed there. There are approximately 200 teachers, and a list of their names, by department, is available on your school's Web site. Choose a simple random sample of 20 teachers and ask them about their length of service.

If no, then a simple random sample is probably not your best choice. If different subsets of the population have different characteristics that are important to your survey, consider a **stratified random sample** of your population. To perform a stratified random sample, group the members of the population that are *most alike* with respect to this characteristic into separate **strata**. If the strata all have the same number of members, then sample an equal number from each strata, using the procedure for making a simple random sample. If the strata are different in size, you can choose a proportional number from each strata.

> Example: Your school is contemplating a policy change regarding off-campus lunch privileges, which are currently limited to seniors. As a member of the student council, you suggest conducting a survey to gauge student opinion. Since you think that there may be differences between how seniors would view this change as compared to juniors or sophomores, you could stratify by grade level. This means you separately collect the names of seniors, juniors, and sophomores—these are your three strata. Choosing a random sample from each of these strata will guarantee that an equal number or proportion of each of the three classes has a chance to express their views.
>
> Because of this guarantee of equal representation from each stratum, a stratified random sample is more likely to produce a result that is closer to the true student opinions than would a simple random sample, which might happen to have more seniors than juniors in it.

Question 2: Would it be difficult to gather information from randomly chosen members of the population due to time or geographic constraints?

If yes, then a simple or stratified random sample may be difficult to carry out. If the population has convenient groups that are easy to gather data from *and* that you think are reasonably representative of the general population, then you may consider a **cluster sample**. From among the convenient groups (or clusters), randomly select a group (or groups) to gather information from.

> Example: You are writing an article for the school newspaper about your community's music scene, and you'd like to know what proportion of the student population has attended a rock concert within the last six months. It would be cumbersome to interview a random sample of the more than 1,800 students in your school. However, some mornings the students meet in homeroom for attendance and announcements. You know that these homerooms are more or less randomly assigned to students, so each homeroom is probably reasonably representative of the larger student body. You randomly select four of the homerooms and ask the teachers to distribute your survey.
>
> This is a quick and easy way to get a reasonably good sample of the student population, and it relies on the fact that the homerooms are similar in makeup to the overall population. This wouldn't necessarily be the case with academic classes at the school, which would not be good clusters to use. For example, an AP Statistics class or a first-year English class may differ in makeup from the general student body and thus would not provide a good representation.

If no, then a cluster sample would not be a good choice.

Part 4: Conducting Your Sample Survey

Having decided on a sampling method, you will now choose which members of your population will participate in the survey. Deciding on a sample size is usually a balancing act between better estimates (larger sample) and ease of completion and cost (smaller sample). Using a random number process, select your sample.

If you are using a stratified random sample, first make sure you have correctly grouped your population into distinct strata that share the specific characteristics you think are related to the response. For the sample size you have determined to be appropriate, you'll need to calculate a proportional representation from each of the strata. If the strata are equal in number, then each of the subsamples will be as well. If the strata are not equal in number, you will need to use sample sizes that are proportional to the representation of the strata in the population. For example, if you are stratifying on grade level, and your

school has 560 first-year students, 500 sophomores, 480 juniors, and 460 seniors, a sample of 100 students should include 28 first-years, 25 sophomores, 24 juniors, and 23 seniors.

For a cluster sampling approach, compile a list of the clusters. Then decide, based on the size of the clusters and the desired overall sample size, how many you need to choose. (You will need to consult a textbook about sampling to locate the applicable formulas.)

Having chosen the sample, it is very important that those chosen actually participate! A common error in doing surveys is to think that you can substitute different members of the population for persons who might be unavailable. This is especially tempting in a stratified design, where, if a given student is absent, you might think you can just substitute another member of his or her stratum. This, however, destroys the randomization and can possibly introduce bias into the sampling procedure. You may need to work very hard to make sure that the last few members of the sample are contacted and measured, despite difficulties that may arise, but that hard work will be rewarded with the confidence you will have in the results! One very good idea is to include more students in your sample than needed to allow for anticipated nonresponse.

As you carry out the sampling process, make sure that the questions are asked consistently and clearly. Record your data, making sure that you don't skip some questions or enter the same data twice. As you gather the data, the questions and answers will probably begin to seem boring to you, and you may lapse into inattention. Don't let that happen to you! First, you don't want to be the source of errors and inaccuracy. Second, the respondents to the survey may respond differently if you appear to have a "bad attitude" about the survey.

Part 5: Analyzing and Reporting Your Results

Analyzing and reporting what you find will be the most satisfying part of the process, especially if the results are interesting or surprising. Think about how to use graphical displays to present your findings. If you are reporting your results using a margin of error or as a confidence interval, double-check your calculations. Be careful with the language of your conclusion so that you don't overstate your findings. Your role as a statistician is to be objective and clear about the limitations of your results.

Sampling Activities

Peter Flanagan-Hyde
Phoenix Country Day School
Paradise Valley, Arizona

To perform these activities, you will need technology available to generate random numbers and display statistical results. Your calculator may have commands slightly different from those given in these instructions, and if you are using computer software, the procedures will be very different indeed. Consult your calculator guidebook or the software's help facilities if you are in doubt.

Activity 1: The Effect of Different Sample Sizes When Sampling from an Infinite Population

This activity will demonstrate the effect of different sample sizes on the typical accuracy of a sample. To do this, we will sample with replacement from a very well-defined population: the set of digits from 0 to 9. This is equivalent to sampling from an infinite population in which each of the 10 digits is replicated an infinite number of times; you can continue to select digits as long as you like, and at each select you have the same probability (0.1) of selecting any particular value. Adding up the possible digits and dividing by 10 produces the mean of the population, 4.5. The standard deviation of the population can also be calculated and is $\sigma = \sqrt{8.25} \approx 2.872$.

In this activity, you will choose a sample of 10 random digits and calculate the mean of that sample. Repeating this exercise a number of times will allow you to see the typical sample results for this sampling situation. Repeating the process again with samples of 40 will show you the difference that changing the size of the sample will make.

Your calculator can choose a sample of random digits from this population using the command `randInt(0,9,10)`. Begin by pressing `MATH` then choosing the menu `PRB`. To get the mean of this sample, press `2nd LIST`, choose menu `MATH`, then choose `3:mean(`. Together with the preceding, this will give you the following command:

```
mean(randInt(0,9,10)).
```

Entering this command should produce a single number, the mean of the 10 randomly selected digits.

To easily repeat this 200 times and store the results for analysis, the sequence command is used. Press 2nd LIST, choose menu OPS, then choose 5:seq(. This can be stored for later analysis in L1. Here is the final command:

```
seq(mean(randInt(0,9,10)),X,1,200)->L1.
```

This can then be repeated, changing the sample size to 40 and storing the list results in L2.

Use Figure 1 when you are answering questions 1 and 2 below.

Figure 1: Simulation—Different Sample Sizes

200 samples of size 10, stored in L1: `seq(mean(randInt(0,9,`**`10`**`)),X,1,200)->`**`L1`**	200 samples of size 40, stored in L2: `seq(mean(randInt(0,9,`**`40`**`)),X,1,200)->`**`L2`**
To see a numerical summary of your results, press: STAT…CALC…1:1-VarStats L1. Mean of your 200 samples, $\bar{x}$ = _____ Standard deviation of your 200 samples, S_x = _____	To see a numerical summary of your results, press: STAT…CALC…1:1-VarStats L2. Mean of your 200 samples, $\bar{x}$ = _____ Standard deviation of your 200 samples, S_x = _____
Make a histogram of your results, with the following window parameters: Xmin=1.8, Xmax=7.2, Xscl=0.3 Ymin=-10, Ymax=60	
Histogram for 200 samples:	Histogram for 200 samples:

1. Consider the means of the distributions of means calculated in Figure 1.

    a) How does the mean of your sample of means for a sample size of 10 differ from the population mean? Why do you suppose it differs as much or as little as it does?

    b) How does the mean of your sample of means for a sample size of 40 differ from the population mean? Why do you suppose it differs as much or as little as it does?

2. Consider the standard deviation of the distributions of means calculated in Figure 1.

    a) How does the standard deviation of your sample of means for a sample size of 10 differ from the population standard deviation? Why do you suppose it differs as much or as little as it does?

    b) How does the standard deviation of your sample of means for a sample size of 40 differ from the population standard deviation? Why do you suppose it differs as much or as little as it does?

Activity 2: Sampling *with* Replacement Versus Sampling *without* Replacement

In the preceding example, sampling was done from a theoretically infinite population. This is often called "sampling with replacement," since it's as if you've thrown back each digit after you choose it. This guarantees that the probabilities of choosing a particular digit don't change as you proceed. Most sampling is not actually done this way, since we don't have access to infinite populations! When we sample from real populations, sampling is usually performed *without* replacement from a finite population. As each member of the population is chosen, it is set aside so that it won't be chosen again. This activity will help you see how the results for these two methods differ.

You already have results from the previous activity, where sampling was done with replacement. The procedure for sampling without replacement is a little more complicated; we'll need a calculator program to help us. The finite population will consist of a set of 100 digits, 0 through 9, with exactly 10 of each digit. Notice that this population has the same mean and standard deviation as the infinite population,

$$\mu = 4.5 \text{ and } \sigma = \sqrt{8.25} \approx 2.872 .$$

In the appendix, there is a complete program, `FINITE`, that will perform this simulation. Enter it into your calculator. The program begins by asking the sample size. Enter 10 for the first run, and then run the program again with samples of 40. When the program is finished running, `L5` will hold the means of each of the 200 different samples. This program takes a few minutes to run, so please be patient!

Transfer your results from the simulation into Figure 2 below.

Figure 2: Simulation—with and without Replacement

200 samples of size 10, without replacement:	200 samples of size 40, without replacement:
Mean of your 200 samples, $\bar{x}$ = _____ Standard deviation of your 200 samples, S_x = _____ **Please save your results!** *L5->L6*	Mean of your 200 samples, $\bar{x}$ = _____ Standard deviation of your 200 samples, S_x = _____
Make a histogram of your results, with the following window parameters: Xmin=1.8, Xmax=7.2, Xscl=0.3 Ymin=-10, Ymax=60	
Histogram for 200 samples:	Histogram for 200 samples:

Write a few sentences about the differences that you see between the results in this activity (sampling without replacement) and the first (sampling with replacement). Consider the location, spread, and shape of the distribution of your data.

Activity 3: Stratified Random Sampling

If you can anticipate that a population has identifiable characteristics that differ from subgroup to subgroup, it is often a good idea to do stratified random sampling. In this activity, you will see the effect of stratification on sample variability: it makes it smaller. Imagine you are doing research in biology and wish to estimate the average brain weight among a varied group of 65 animals. Of course, the best solution would be to find the brain weights of all 65, but measuring each one would require careful dissection—not an easy task. By using a stratified sample, you typically get a more accurate result than by using a simple random sample.

First, as a point of comparison, let's look at the results from 200 simple random samples from this population, with $n = 10$. We would expect about the same mean as the population mean, which, in this demonstration, is 149 g. (As the experiment's biological researcher, of course, you wouldn't yet know this figure.) Now we'll see what happens when we stratify by the size of the animal, dividing the animals into categories of "large" and "small." (It is reasonable to assume that body size correlates to brain size.)

In the appendix are two calculator programs that will perform the sampling, either with a simple random sample (`SRS`) or stratified (`STRAT`). Here is how we organize the data on our calculator:

1. Data list `LBRAIN`, the brain sizes of the 65 animals (presumably this is unknown)
2. Data list `LSIZE`, which identifies the strata, either 1 for large or 2 for small
3. Program `SRS`
4. Program `STRAT`

Once you have linked, run each of the programs. In both cases, enter 200 for *K* (the number of different samples) and 10 for *N* (the sample size). Then complete the table in Figure 3 for each type of sample. Table 1 provides the brain weight data.

Figure 3: Simulation—SRS Versus STRAT

200 samples of size 10, SRS:	200 samples of size 10, STRAT:
Mean of your 200 samples, $\bar{x}$ = _____ Standard deviation of your 200 samples, s_x = _____	Mean of your 200 samples, $\bar{x}$ = _____ Standard deviation of your 200 samples, s_x = _____
Histogram for 200 samples:	Histogram for 200 samples:
You can use a boxplot to help you compare the results of the two types of sampling. Create this display, then write a few sentences that compare the two methods.	Boxplots to compare SRS to STRAT:

Table 1: Brain Weight Data

Species	Brain Weight (grams)	Size 1=Large	Species	Brain Weight (grams)	Size 1=Large
African giant pouched rat	6.6	2	Little brown bat	0.25	2
Arctic fox	44.5	2	Loon	6.12	2
Arctic ground squirrel	5.7	2	Mackerel	0.64	2
Baboon	179.5	1	Man	1320	1
Barracuda	3.83	2	Mole rat	3	2
Big brown bat	0.3	2	Musk shrew	0.33	2
Brown trout	0.57	2	Nine-banded armadillo	10.8	2
Canary	0.85	2	North American opossum	6.3	2
Cat	25.6	2	Northern trout	1.23	2
Catfish	1.84	2	Ostrich	42.11	2
Chimpanzee	440	1	Owl monkey	15.5	2
Chinchilla	6.4	2	Pheasant	3.29	2
Cow	423	1	Pig	180	1
Crow	9.3	2	Pigeon	2.69	2
Desert hedgehog	2.4	2	Porpoise	1735	1
Donkey	419	1	Rabbit	12.1	2
Eastern American mole	1.2	2	Raccoon	39.2	2
European hedgehog	3.5	2	Rat	1.9	2
Giant armadillo	81	1	Red fox	50.4	1
Giraffe	680	1	Rhesus monkey	179	1
Goat	115	1	Roe deer	98.2	1
Golden hamster	1	2	Salmon	1.26	2
Gorilla	406	1	Seal	442	1
Gray seal	325	1	Sheep	175	1
Gray wolf	119.5	1	Stork	16.24	2
Ground squirrel	4	2	Tree shrew	2.5	2
Guinea pig	5.5	2	Tuna	3.09	2
Flamingo	8.05	2	Vulture	19.6	2
Horse	655	1	Walrus	1126	1
Jaguar	157	1	Water opossum	3.9	2
Kangaroo	56	1	Yellow-bellied marmot	17	2
Lesser short-tailed shrew	0.14	2			

Activity 4: Cluster Sampling

Cluster sampling relies on being able to easily find groups among the population that are deemed reasonably representative of the larger population and that are relatively inexpensive to sample. For example, a scenario in which cluster sampling might be appropriate would be if you need to find the average household income for your city. A city's population is usually quite mobile, with people moving to different geographic areas within the city. The streets and roads of the city are generally parallel and divide the city into (say) 2,000 city blocks of approximately the same area. With a cluster sample you pick a random sample of blocks, say 25 of them. If you can reasonably assume that each of the 25 blocks contains a representative group of households—little "mirror images" of the population—then cluster sampling may work well. Your sample will consist of randomly chosen *blocks*, rather than randomly chosen *households*. For each block chosen, you will interview *each* household about their income. Cluster sampling will not necessarily produce more accurate results than random sampling, as is the case with stratified sampling, but it will make it easier to complete the sampling process.

In the appendix is a calculator program that will perform cluster sampling. You must link several items to your calculator:

1. Data list `LCLUSM`, the number of households in a given block (presumably this is unknown)
2. Data list `LCLUSY`, the total income for the households in a given block (also unknown)
3. Program `CLUSTER`

Once you have linked, run the program. In both cases, enter 200 for *K* (the number of different samples) and 25 for *N* (the number of clusters to sample). Then complete the table in Figure 4.

Figure 4: Simulation—Cluster Sampling

25 clusters sampled
Mean of your 200 samples, $\bar{x}$ = _____ Standard deviation of your 200 samples, S_x = _____
Histogram of your results:

You may notice that the results don't seem particularly accurate—that the standard deviation of the results is quite high. The best results for cluster sampling will occur when the individuals in each of the clusters are a good representation of the population as a whole. You would hope that each cluster has a wide range of different individuals and that the means of each of the clusters are very close together. To the degree that these ideals are not met, cluster sampling will be limited in its accuracy. In this example, the mean household income is about $8,910, which should be somewhat near the center of your distribution of outcomes.

Appendix: Code for Calculator Programs

```
Program FINITE                          Samples without replacement
:Prompt N                               N = sample size
:ClrLst L5
:seq(int(X/10),X,1,100)->L3             Create population
:For(I,1,200)                           Loop . . .
:rand(100)->L4                          Assign a random number
:sortA(L4,L3)                           Sort by random number
:mean(seq(L3(X),X,1,N)->L5(I)           Choose sample
:End
```

```
Program SRS                             Simple random sample of L1
:Prompt K                               Number of samples
:Prompt N                               Sample size
:ClrList L5
:dim(L1)->D                             Find length of data list
:For(I,1,K)                             Loop . . .
:rand(D)->L3                            Assign random numbers
:SortA(L3,L1)                           Sort by random number
:mean(seq(L1(X),X,1,N))->L5(I)          Mean of sample, store result
:End
```

```
Program STRAT                          Stratified random sample of
                                       data in L1
:Prompt K                              Number of samples
:Prompt N                              Sample size
:SortD(L2,L1)                          Sort by stratified code (L2)
:dim(L1)->D                            Overall length of data list
:round(D(mean(L2)-1),0)->S             Length of first stratum
:D-S->T                                Length of second stratum
:seq(L1(X),X,1,S)->ʟSTR1               Separate strata
:seq(L1(X),X,S+1,D)->ʟSTR2
:ClrList L4
:For(I,1,K)                            Loop . . .
:rand(S)->L5                           Assign random number
:SortA(L5, ʟSTR1)                      Sort by random number
:rand(T)->L5                           Assign random number
:SortA(L5,ʟSTR2)                       Sort by random number
:mean(seq(ʟSTR1(X),X,1,N/2))->U        Mean of first stratum
:mean(seq(ʟSTR2(X),X,1,N/2))->V        Mean of second stratum
:(SU+TV)/D->L4(I)                      Weighted average of strata
:End
```

```
Program CLUSTER                          Cluster sample
:Input K                                 Number of samples
:Input N                                 Clusters to sample
:ClrList L4,L5
:LCLUSM->L1                              CLUSM has number in each
                                         cluster
:LCLUSY->L2                              CLUSY has total in each
                                         cluster
:For(I,1,K)                              Loop . . .
:rand (200)->L3                          Assign random number
:SortA(L3,L1,L2)                         Sort by random number
:sum(seq(L1(X),X,1,N)->L4(I)             Total members
:sum(seq(L2(X),X,1,N)->L5(I)             Total amount
:End
:L5/L4->L6                               Estimated mean
:1-Var Stats L6

Program CLUSINT                          Create an interval estimate
:Prompt N
:LCLUSM->L1
:LCLUSY->L2
:dim(L1)->D
:rand (200)->L3
:SortA(L3,L1,L2)
:seq(L1(X),X,1,N)->LMS
:seq(L2(X),X,1,N)->LYS
:sum(LMS)->M
:sum(LYS)->Y
:Y/M->E
:M/N->F
:sum((LYS-E*LMS)^2/(N-1))->S
:((D-N)/D)(1/(NF^2))S->V
:Disp E
:Disp E-2√(V)
:Disp " TO"
:Disp E+2√(V)
```

Contributors

Peter Flanagan-Hyde teaches mathematics and statistics at Phoenix Country Day School in Paradise Valley, Arizona. He has taught AP Statistics since its beginnings in the1996–1997 school year. He is an AP Statistics Exam Reader (currently a Table Leader) and has conducted numerous College Board workshops and summer institutes. He is on the editorial board of the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE), and he writes reviews of Web-based material for the Multimedia Educational Resource for Learning and Online Teaching (MERLOT) project. He has written a variety of articles and reviews for AP Central and contributes regularly to the AP Statistics Electronic Discussion Group. He is on the editorial board of *Stats* magazine, where he writes a regular column that focuses on AP Statistics.

Chris Olsen has taught statistics at George Washington High School in Cedar Rapids, Iowa, for over 25 years and has taught AP Statistics since its inception. He is a past member of the AP Statistics Development Committee. He is currently on "temporary detached assignment," as they say in the military, serving as the math/assessment facilitator for Cedar Rapids Community Schools. He has been involved nationally in statistics-related workshops and conferences for over 20 years. He has written a variety of articles and reviews for AP Central and currently serves as the AP Statistics content adviser for AP Central. He is a regular poster to the AP Statistics Electronic Discussion Group. He has reviewed materials for the *Mathematics Teacher*, *American Statistician*, and the *Journal of the American Statistical Association*, and he currently writes a column for *Stats* magazine. He is coauthor of a leading textbook on introductory statistics.

Roxy Peck has been a professor of statistics at California Polytechnic State University: San Luis Obispo since 1979, serving for six years as chair of the Statistics Department, and is currently in her eighth year as associate dean of the College of Science and Mathematics. Nationally known in the area of statistics education, she was made a fellow of the American Statistical Association (ASA) in 1998, and in 2003 she received ASA's Founders Award in recognition of her contributions to K–12 and undergraduate statistics education. In addition to coauthoring two leading textbooks on introductory statistics, she is also editor of a collection of case studies and a collection of expository papers that showcase applications of statistical methods. She is currently chair of the joint ASA/NCTM Committee on Curriculum in Statistics and Probability for Grades K–12, and she served from 1999 to 2003 as the Chief Reader for the AP Statistics Exam.

Richard L. Scheaffer is professor emeritus of statistics at the University of Florida, where he served as department chair for 12 years. His research interests are in the areas of sampling and applied probability, especially with regard to their application to industrial processes. He has published numerous papers in statistical journals and is coauthor of five college-level textbooks covering introductory statistics and aspects of sample survey design, probability, and mathematical statistics. In recent years, much of his effort has been directed toward statistics education in high school and college curricula. He was one of the developers of the Quantitative Literacy Project in the United States, which formed the basis of the data-analysis emphasis in the mathematics curriculum standards recommended by the National Council of Teachers of Mathematics (NCTM). He also directed the task force that developed the AP Statistics program, for which he served as the first Chief Reader. He continues to work on educational projects at the elementary, secondary, and college levels. He is an ASA Fellow, a past president, and recipient of its Founders Award. He earned his Ph.D. in statistics from Florida State University.