5.1 Designing samples

population sample

census sampling

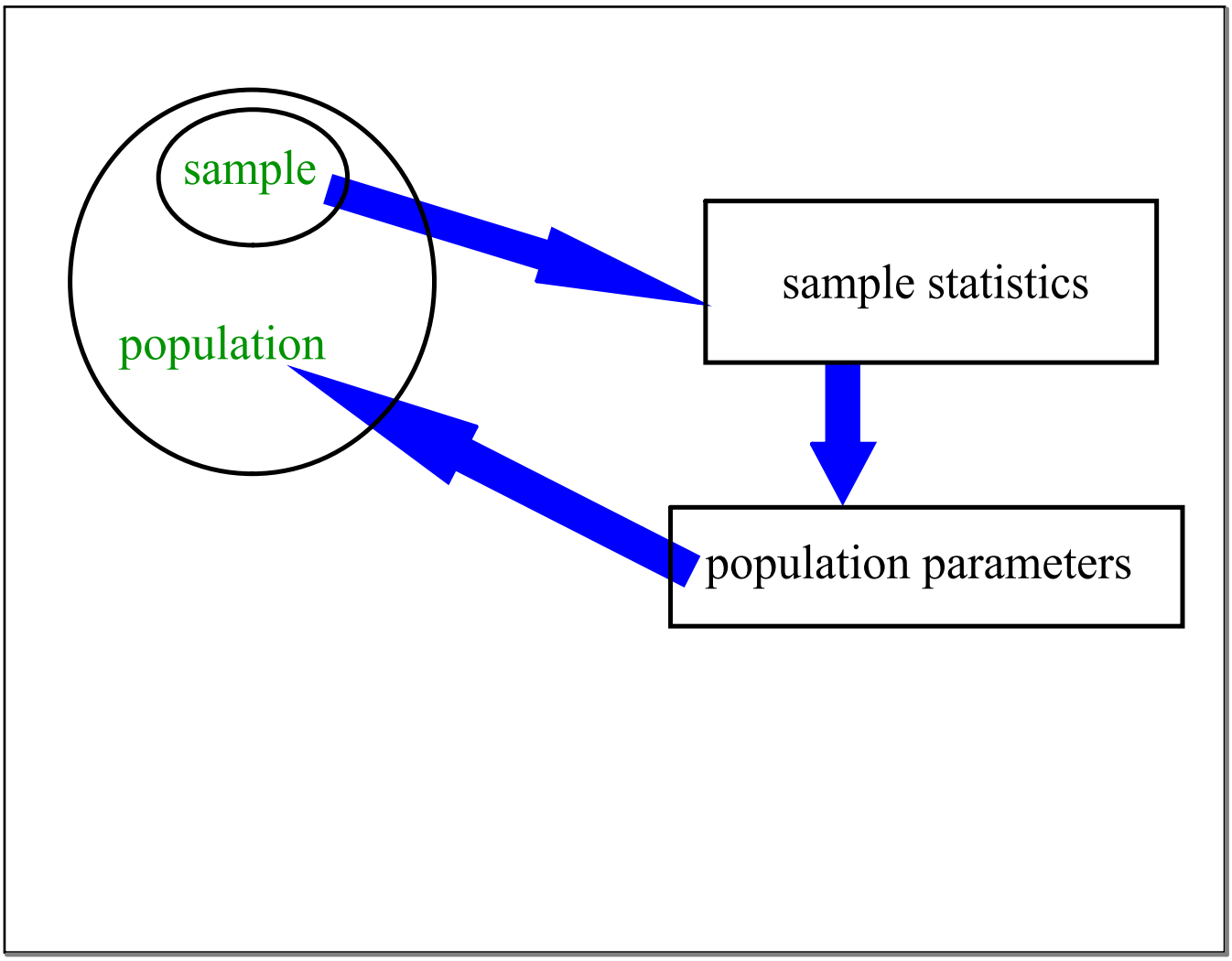
observational study experiment survey

voluntary response sampling convenience sampling

simple random sample probability sample
stratified random sample multistage sample
systematic sample quota sample

bias undercoverage nonresponse
household bias response bias wording effects

random sampling error sampling method error
non-sampling method error


sample
population
sample statistics
population parameters

The entire group of individuals that we want information about is called the **population**. A **sample** is a part of the population that we actually examine in order to gather information.

Sampling involves studying a part in order to gain information about the whole. A **census** attempts to contact every individual in the entire population.

examples
My dad's field of corn is a **population**. He needs to determine the type of insects infesting the field. A **census** of the field would take too long-- he doesn't have time to examine every single plant. By **sampling**, he examines a **sample** of 10 plants from various parts of the field to inspect for insect damage.

The federal government's No Child Left Behind legislation (NCLB) requires that every student be tested in math one time during high school to assess the progress of the nation's students. In Missouri, we use the 10th grade math MAP test to meet NCLB requirements. The **population** of Missouri 10th graders all take the test, so this is a **census** of 10th graders. NCLB won't allow a **sample** of 10th graders. **Sampling** would not measure every student.

ActivStats III.10.1 Sample

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses.

example
Three college students enter a busy dorm restroom, one at a time, and each monitor handwashing of individuals. The students make note of whether the subjects wash their hands. Because each student is in the restroom only a few minutes, none of the observed individuals suspects that their behavior is being watched, so the students do not influence the subjects' behaviors.

An **experiment** deliberately imposes some treatment on individuals in order to observe their responses.

example
Three college students secretly monitor handwashing habits of some acquaintances. The next day, in those same restrooms, the observers post signs that emphasize the importance of washing hands. That day, the college students monitor the handwashing habits of those same acquaintances to determine whether the signs had any effect on behavior.

A **survey** seeks responses from individuals who are knowingly responding to the questions.

example
Three college students stand outside a set of restrooms and ask about handwashing habits of those who exit the restroom.

Sample **design** is the method used to choose a sample from a population and gather data from the sample.

The design of a study is **biased** if it systematically favors certain outcomes.

ActivStats III.10.2-3 Bias

Voluntary response sampling consists of people choosing themselves by responding to a general appeal. It's biased because people with strong opinions, especially negative opinions, are most likely to respond.

Convenience sampling consists of choosing individuals who are easiest to reach. It's biased because it is not likely to represent the entire population.

examples
Radio talk shows and election exit polls are examples of **voluntary response sampling**. Responders choose whether or not they want to participate-- and whether or not they want to tell the truth.

Warning: data reflect the views of only those motivated to respond, often with the most extreme opinions

Remember this photo?
The Chicago Daily Tribune mistakenly ran this due to **convenience sampling**.
Pollsters called prospective voters. Back then, only the rich had phones and the sample consisted mostly of republicans, so the sample did not represent the entire population.

© Associated Press. TITLE: Dewey Defeats Truman
AP PHOTOGRAPHER: BYRON ROLLINS
11/4/1948

Selecting the first 25 shoppers a a store on Tuesday morning would be another example of **convenience sampling**.

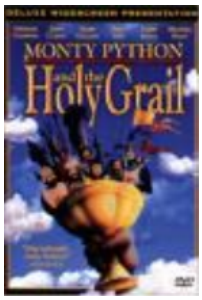
A Cautionary Note: Data reflect the views of only those available at the time taken.

In a voluntary response sample, members of the sample choose whether they participate whereas in a convenience sample, the interviewer chooses who participates. This conscious choosing results in bias.

A **simple random sample** (SRS) of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample selected.

In an SRS, every individual has an equal chance of participating and every sample of size n has an equal chance of being chosen. The participants are chosen randomly. This can be done:
a) by choosing names from a hat
b) by having a computer choose randomly for us
c) by assigning a numerical label to every individual in the population and using a table of random digits to select labels at random

examples
A lottery drawing is a type of SRS. If each student's name is put in a hat and 5 are drawn out to take part in a survey, we have a SRS.



A SRS is the "Holy Grail of selecting" (ideal method sought to select) a sample because there is no selection bias.

A **table of random digits** is a long string of the digits 0 through 9 with these two properties:
1. Each entry in the table is equally likely to be any of the 10 digits 0 through 9.
2. The entries are independent of each other, so knowledge about one part of the table tells you nothing about any other part.

http://bcs.whfreeman.com/yates2e/pages/bcs-main.asp?v=category&s=00020&n=99000&i=99020.01&o=

Simple Random Sample (SRS) examine this in reverse:
Sample: Select a few individuals from a larger population and you have a sample. There are many ways to do this that are not random.

Random: Use some kind of probabilistic selection scheme - draw cards, roll dice, use random numbers, etc - and your sample is random. There are many ways to do this that are not "simple".

Simple: Make all the individuals equally likely to be chosen, placing no restrictions on who might comprise the eventual sample. This is the conceptual equivalent to pulling names out of a hat - any group of people could end up being chosen.

The use of chance to select the sample is the essential principle of statistical sampling.

Other sampling designs

A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.

example
An SRS is a type of probability sample; an SRS gives each member of a population an equal chance of being selected. An SRS of the entire US population is not often practical. Having some computer select 10 individuals from a census roster so that they can participate in a survey would likely have the 10 people so spread out that conducting the survey would be expensive or time consuming.

Valid **probability sampling** methods have two critical characteristics:
1. The interviewers and subjects themselves are not choosing the subject who is interviewed.
2. There is a definite procedure for selecting participants in the sample and that procedure involves the use of probability.

In a **stratified random sample**, we first divide the population into groups of *similar* individuals called strata. We then choose a separate SRS in each stratum and combine these SRS's to form the full sample. Groups are often formed around race, gender, residence, or economic status.

examples

A farmer wishes to work out the farm's average egg yield for each breed of chicken. He has 4 breeds, so he could divide up his chickens into the four sub-groups and take samples from these.

A principal wishes to develop a preferred student parking incentive. She expects that students who are in sports may have different views on what is a desireable parking space from the students who are in the work program or who stay all 7 periods of the day. She assembles lists of these various groups and interviews a couple students from each group.

We select, at random, 5 freshmen, 5 sophomores, 5 juniors, and 5 seniors.

Caution: The groups/strata must be selected so members of any particular group/stratum are homogeneous.

A **Multistage Sample or Multistage Cluster Sample** - is constructed by taking a series of SRS's in stages. In each stage, the SRS is called a cluster. Individuals in each cluster are *heterogeneous*.

examples

A political scientist wants to predict the outcome of an election. First, take a SRS of electoral sub-divisions (clusters) are from a city or state. Second, blocks of houses are selected by SRS from within the electoral sub-divisions. Third, individual houses are selected to be polled by SRS from within the selected blocks of houses.

A survey is to be conducted at school. We us a SRS to randomly select one of the 4 classes. From that class chosen, select a SRS of homerooms. From that list of homerooms, select a SRS of students to survey.

Caution: the methods used to analyze the results of a multi-stage cluster sample differ from the methods used to analyze an SRS.

Don't confuse stratified and multistage cluster sampling.

In stratified sampling we divide up the population based on some factor we believe is important, but in cluster sampling the groups are naturally occurring (picture schools of fish).

In stratified sampling, we randomly select subjects from each strata, but in cluster sampling we randomly select one or more clusters and measure every subject in each selected cluster. (In advanced techniques, samples are taken within the cluster(s)).

In a **Systematic Sample**, we start with a list of all members of the population, then select a systematic way of choosing members.

examples
A farmer wishes to sample the weights of his hogs. All the hogs are already tagged with numbers for identification and record keeping purposes (the USDA requires this, by the way). He decides to weigh one of every 20 hogs. He randomly chooses one hog out of the first 10 and then every 20th hog thereafter. Suppose he selected the 4th one, then the 24th, 44th, 64th, etc. will be weighed.

For a poll, someone randomly selects, from an ordered list, one of the first 100 student numbers in the school and then select every 100th student number after that one.

We must be careful that the methods used to select the sample can't be associated in some way with the way the population is organized. For example, it is not appropriate to choose a day of the week at random and then select every 7th day after that one since the sample will consist only of, for example, Mondays.

In **Quota Sampling** we measure a fixed quota of members of the population, organized around categories like race, gender, residence, income, political party preference, or economic status that are often set to match known or assumed demographic information about the population.

examples
A restaurant marketing research department offers trial-size portions of a new sandwich to a sample that is selected to mirror the U.S. population in terms of the percent consisting of males, females, state of residence, age, and race, believing that those are the characteristics that most influence food choice.

Quota sampling includes surveys with screening questions to select who is or is not included.

The danger with quotas is that the surveyor has the choice of who is interviewed within the constraints of the quotas that have been set.

Sources of bias and cautions about sample surveys

A **sampling frame** is the list of possible subjects who could be selected in a sample (the list of individuals from which a sample is actually selected). If the sampling frame is not equal to the population, the sample will be biased the way the sampling frame is biased.

Most samples of humans suffer from **undercoverage**-- this happens when some groups in the population are left out of the process of choosing the sample.

examples
Surveys of households will not represent the homeless, inmates, and students in dormitories.

Surveys by phone may miss those who only use a cell phone and have no landline phone.

Most sample surveys also suffer from **nonresponse**-- this happens when someone is unavailable for selection or refuses to cooperate. Non-respondents tend to differ from those who are readily available.

examples
Some voters refuse to participate in election exit polls.

Some people sign up for the no-call list or are not at home when a pollster calls.

The surveyors at the mall miss those who do not shop at the mall or who refuse to participate.

Household bias occurs when a sample includes only one member of any given household. This underrepresents the members of large households.

Some sample surveys also suffer from **response bias**-- this happens when someone lies or unintentionally answers falsely.

examples
In an election exit poll, a voter might participate but lie about how he or she voted in hopes that early returns may motivate some voters to get to the polls or to stay home.

A participant may *think* he or she did something recently when it is actually outside the range of time the survey requests.

Some sample surveys also suffer from **wording effects**-- this happens when questions are confusing, leading, or put in a particular order (often the first choice is chosen most often, so the choices should be scrambled on several versions of the survey).

example
Some surveys have leading questions and so, participants may answer in hopes of "getting it right".

Sampling errors

Random Sampling Error occurs due to chance variation.

Sampling Method Error occurs due to the choice of sampling method.

Non-sampling Method Error occurs in responses by members in a sample.

Suppose 24 students are seated in rows of 6 and I want to select a sample of 6. Classify these as SRS, multistage sample, stratified random sample, systematic sample, convenience sample.

I assign numbers 01-24 and choose 6 at random (ignoring repeats)

I pick 3 boys and 3 girls

I put all the names in a hat and draw out 6 names without replacement

I pick the 6 closest to me

I pick a whole row at random

I rolled a 4-sided die and got a 2 so I chose the 2nd, 6th, 10th, and 14th, 18th, and 22nd students.