4.4: Sampling Techniques

When doing research, it is critical to obtain a sample that is representative of the population. Non- representative or biased samples will produce invalid inferences, regardless of the sample size. For example, it is far better to have a representative sample of 500 observations, than a biased sample of 50,000 observations. In this section we will explore methods of sampling that have the highest chance of producing a representative sample.

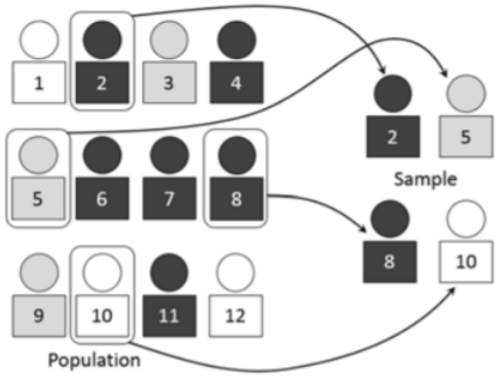
A word of caution: even if you carefully attempt to create a representative sample, there is always a chance you will select non-representative outlier sample. However, if you use one of these appropriate methods of sampling, you have a small probability of selecting an outlier sample.

The best methods of sampling are those in which the probability of getting a representative sample can be calculated. The methods are called **probability sampling methods**. Other **non-probability sampling methods** have immeasurable bias and need to be avoided when conducting research

Probability Sampling Methods

These methods will usually produce a sample that is representative of the population. These methods are also called scientific sampling.

Simple Random Sampling[46]



A **simple random sample** is a subset of a population in which all members of the population have the same chance of being chosen and are mutually independent of each other. Think of random sampling as a raffle or lottery in which all names are put in a bowl and then some names are randomly selected.

Random samples in practice are almost impossible to obtain as it is difficult to list every member of the population.

Advantages of Simple Random Sampling:

- no possibility of bias in the sampling method
- no knowledge of population demographics needed
- easy to measure precision

Disadvantages of Simple Random Sampling:

- often impossible to conduct due to difficulty of cataloguing population
- high expense
- often less precise than a stratified sample

Example: Custom control searching

Before leaving customs at several international airports, all passengers must push a button. If the button is red, you will be required to go through an intensive search. If the button is green, you will not be searched.[47] The button is totally random and has a 20% chance of being red. Passengers who are subject to the intensive search are a true simple random sample of the entire population of arriving passengers.

Systematic Sampling[48]

A **systematic sample** is a subset of the population in which the first member of the sample is selected at random and all subsequent members are chosen by a fixed periodic interval. An example would be having a list of the entire population and then taking every 3rd person on the list.

Advantages of Systematic Sampling:

- easy to design and explain
- more economical than random sampling
- avoids random clustering (several adjacent values)

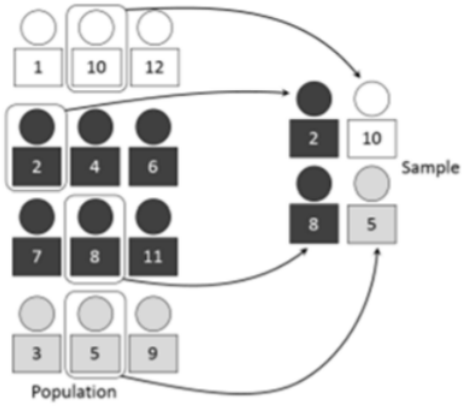
Disadvantages of Systematic Sampling:

- may be biased if population is patterned or has a periodic trait
- easier for researcher to wrongly influence data
- population size needs to be known in advance

Example: Random drug testing of employees

A shipping company has approximately 20,000 employees. The company decided to administer a random drug test to 5% of the employees, a sample size of 1000. The company has a list of all employees sorted by social security number. A random number is selected between 1 and 20. Starting with that person, every subsequent 20th person is also sampled. For example, if the selected number is 16, then the company would select persons 16, 36, 56, 76, ... , 19996 for drug testing.

Stratified Sampling[49]



A **stratified sample** is designed by breaking the population into subgroups called strata, and then sampling so the proportion of each subgroup in the sample matches the proportion of each subgroup in the population. For example, if a population is known to be 60% female and 40% male, then a sample of 1000 people would have 600 women and 400 men.

Advantages of Stratified Sampling:

- minimizes selection bias as all strata are fairly represented
- each subgroup receives proper representation
- high precision (low standard deviation) compared to other methods

Disadvantages of Stratified Sampling:

- high knowledge of population demographics needed
- not all populations are easily stratified
- time consuming and expensive

Example: Social media conversations about race

In 2016, Pew Research Center conducted a study to examine how people use social media such as Twitter or Facebook.[50] The study focused on the content and hash tags used on people's comments about events involving racially motivated attacks by the police and differences in opinions about groups such as Black Lives Matter.

Since the study involved people's opinions about race, it was important that Pew used stratified sampling by race. Particular care was taken to make sure that there was appropriate representation in the sample from traditionally undersampled African American and Latino groups.

Cluster Sampling[51]

A **cluster sample** is created by first breaking the population into groups called clusters, and then taking a sample of clusters. An example of cluster sampling is randomly selected several classes at a college and then sampling all the students in those selected classes.

Advantages of Cluster Sampling:

- most economical form of sampling because only clusters need to be randomized
- study can be completed in less time
- suitable for surveying populations that are broken into natural clusters

Disadvantages of Cluster Sampling:

- sample may not be as diverse as population
- clusters may have a similar bias, causing sample to be biased
- less precision (higher standard deviation)

Example: Police attitudes

In 2017, Pew Research Center conducted a survey of 8000 police officers called Behind the Badge. 52 The goal was to draw on the attitudes and experiences of police officers especially in light of highly publicized and controversial killings of Black Americans by the police.

To conduct this survey, the researchers had to select police departments throughout the country that they felt were representative of the population of departments. Then they surveyed police officers in those departments. One potential problem reported by the researchers was that only police departments with at least 100 officers were sampled. This is a example of potential similarity bias that sometimes arises in cluster sampling.

Example: Student homelessness[53]

The Bill Wilson Center of Santa Clara County provides social services for children, teens and adults. In 2017, the center conducted a study documenting homeless youth populations, surveying both high school students and community college students.[54]

For community college students, the researchers chose two community colleges from the eight in Santa Clara County and surveyed students from Winter 2017 to Spring 2017. One finding was that a staggering 44% of community college students surveyed at these two colleges reported that they were homeless. (Homeless in this study means living on the street, living in cars, or couch surfing).

This study is an example of cluster sampling. Out of the eight Santa Clara County community colleges, the researchers chose 2. Although not reported in the study, it would be important that the demographics of the two chosen colleges match the average of all community college students in the county.

Non-probability Sampling Methods

There are non-scientific methods of sampling being conducted today that have immeasurable biases and should not be used in scientific research. The only advantage of these methods is that they are inexpensive and can generate very large samples. However, these samples will often fail to create a representative sample and therefore have no value in research. Worse yet, these biased samples may be presented as more accurate or better than scientific studies because of the large sample size. However, a biased sample of any size has little or no value -- a big pile of garbage is still garbage.

Convenience Sampling

A **convenience sample** is simply a sample of people who are easy to reach.

Example: Marijuana usage

A 21 year old student wants to conduct a survey on marijuana usage. He asks his friends on Facebook to fill out a survey. The results of his survey show that 65% of respondents frequently use marijuana.

The student's Facebook friends were easy to sample but are not representative of the population. For example, if the student frequently uses marijuana, it is more likely that his Facebook friends would also use marijuana.

Self-selected Sampling

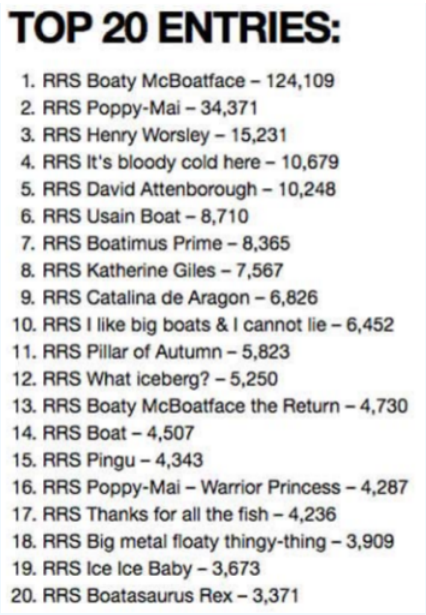
A **self-selected sample** is one in which the participants volunteer to be sampled. This would include Internet polls and studies that advertise for volunteers.

Do not confuse self-selected sampling with scientific studies that ask for volunteers from an initial representative sample. Researchers take care to avoid bias making sure the demographics of the volunteers match the demographics of the representative sample.

Example: Boaty McBoatface

The Natural Environment Research Council (NERC), an agency of the British government, decided to let the Internet suggest a name for a $287 million polar research ship. A public relations professional and former BBC employee started a social media frenzy by suggesting people vote for the name "Boaty McBoatface."[55]

The final result of this self-selected poll showed that Boaty McBoatface was the overwhelming winner. You can see that the top 20 entries included many other humorous choices, along with some more traditional names.[56]



The NERC eventually chose a more serious name, the RSS Sir David Attenborough, but as a consolation to the voters, the agency named a remotely operated underwater research vessel Boaty McBoatface . [57]

The results of the poll do not reflect what the public wanted. What happened instead was many people, through social media, were inspired to vote for Boaty McBoatface as a joke.

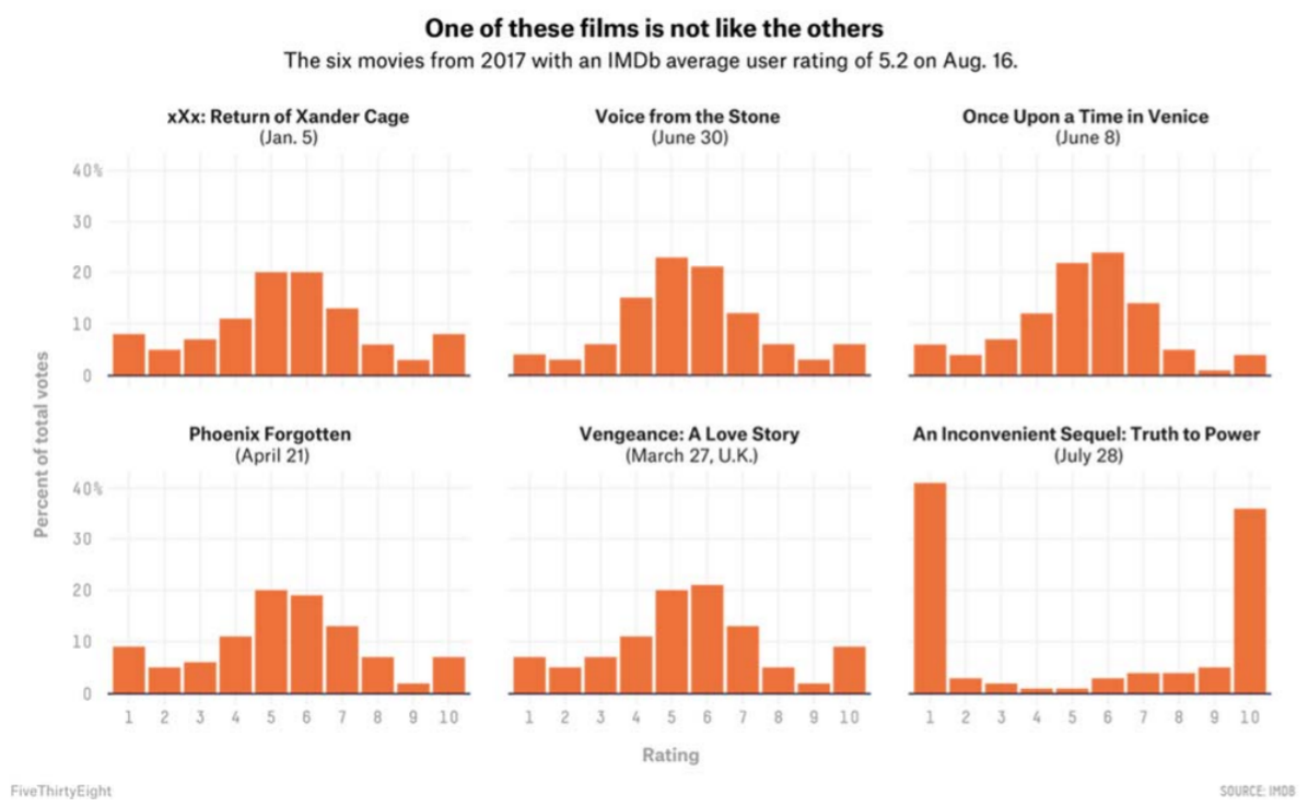
Example: Online Movie Ratings

Many people use online rating services, such as Google, Yelp, Rotten Tomatoes, IMDb and Rate My Professor to make decisions about restaurants, products, services, movies or what college class to take.

All of these ratings systems are examples of self-selected sampling as users volunteer to write reviews. This can lead to ratings that may be extremely inaccurate.

The Internet Movie Database (IMDb) maintains movie reviews and ratings by users. Movies are rated on a scale of 1 (the worst) to 10 (the best). On July 28, 2017, Al Gore's "An Inconvenient Sequel: Truth to Power" was released as a follow-up to his original documentary about climate change, "An Inconvenient Truth". The IMDb overall rating for the movie was 5.2, which is the average of all ratings by users.

The website fivethirtyeight.com conducted an analysis of this overall rating by comparing "An Inconvenient Sequel" to other movies with similar ratings. [58]



It is clear that the graph "An Inconvenient Sequel" was far different from the other five movies with that also had an average rating of 5.2; in this case, most people voted either 1 or 10. The fivethirtyeight.com study also found that many of the reviews were written before the movie release date. Also, traditional critics rated the movie much higher. The IMDb rating in this case was not a true movie rating but an attempt to discredit or to support climate change.

> *The conclusion by fivethirtyeight.com was a warning about these popular online rating systems: "Say what you will, but in addition to being controversial, "An Inconvenient Sequel" was ambitious: Few films involve Arctic expeditions, inside access to the Paris Climate Conference, interviews with the sitting secretary of state and a globe-trotting look at catastrophic weather conditions. If ambitious-yet-controversial films are boiled down to a single number that makes them look identical to mediocre films, what incentive does Hollywood have to continue investing in movies that challenge the audience? "The democratization of film reviews has been one of the most substantial structural changes in the movie business in some time, but there are dangerous side effects. The people who make movies are terrified. IMDb scores represent a few thousand mostly male reviewers who might have seen the film but maybe didn't, and they're influencing the scoring system of one of the most popular entertainment sites on the planet."*

We will all continue to use online rating services, but we must keep in mind the reviews could be fake, manipulated or extremely biased.