

## 1.3: Collecting Data and Sampling Techniques

### Qualitative vs. Quantitative

Variables can be either quantitative or qualitative. Quantitative variables are numeric values that count or measure an individual. Qualitative variables are words or categories used to describe a quality of an individual. Qualitative variables are also called categorical variables and can sometimes have numeric responses that represent a category or word.

**Qualitative or categorical variable** – answer is a word or name that describes a quality of the individual.

**Quantitative or numerical variable** – answer is a number (quantity), something that can be counted or measured from the individual.

Each type of variable has different graphs, parameters and statistics that you find. Quantitative variables usually have a number line associated with graphical displays. Qualitative variables usually have a category name associated with graphical displays.

Examples of quantitative variables are number of people per household, age, height, weight, time (usually things we can count or measure). Examples of qualitative variables are eye color, gender, sports team, yes/no (usually things that we can name).

When setting up survey questions it is important to know what statistical questions you would like the data to answer. For example, a company is trying to target the best age group to market a new game. They put out a survey with the ordinal age groupings: baby, toddler, adolescent, teenager, adult, and elderly. We could narrow down a range of ages for, say, teenagers to 13-19, although many 19-year-olds may record their response as an adult. The company wants to run an ad for the new game on television and they realize that 13-year-olds do not watch the same shows nor in the same time slots as 19-year-olds. To narrow down the age range the survey question could have just asked the person's age. Then the company could look at a graph or average to decide more specifically that 17-year-olds would be the best target audience.

### Types of Measurement Scales

There are four types of data measurement scales: nominal, ordinal, interval and ratio.

**Nominal data** is categorical data that has no order or rank, for example the color of your car, ethnicity, race, or gender.

**Ordinal data** is categorical data that has a natural order to it, for example, year in school (freshman, sophomore, junior, senior), a letter grade (A, B, C, D, F), the size of a soft drink (small, medium, large) or Likert scales. A Likert scale is a numeric scale that indicates the extent to which they agree or disagree with a series of statements. Interval data is numeric where there is a known difference between values, but zero does not mean “nothing.”

**Interval data** is ordinal, but you can now subtract one value from another and that subtraction makes sense. You can do arithmetic on this data. For example, Fahrenheit temperature, 0° is cold but it does not mean that no temperature exists. Time, dates and IQ scores are other examples.

**Ratio data** is numeric data that has a true zero, meaning when the variable is zero nothing is there. Most measurement data are ratio data. Some examples are height, weight, age, distance, or time running a race.

Here are some ways to help you decide if the data are nominal, ordinal, interval, or ratio. First, if the variable is words instead of numbers then it is either nominal or ordinal data. Now ask yourself if you can put the data in a particular order. If you can order the names then this is ordinal data. Otherwise, it is nominal data. If the variable is numbers (not including words coded as numbers like Yes = 1 and No = 0), then it is either interval or ratio data. For ratio data, a value of zero means there is no measurement. This is known as the absolute zero. If there is an absolute zero in the data, then it means it is ratio. If there is no absolute zero, then the data are interval. An example of an absolute zero is if you have \$0 in your bank account, then you are without money. The amount of money in your bank account is ratio data. Word of caution, sometimes ordinal data is displayed using numbers, such as 5 being strongly agree, and 1 being strongly disagree. These numbers are not really numbers. Instead, they are used to assign numerical values to ordinal data. In reality, you should not perform any computations on this data, though many people do. If there are numbers, make sure the numbers are inherent numbers, and not numbers that were randomly assigned.

Likert scales are frequently misused as quantitative data. A Likert scale is a numeric scale that indicates the extent to which they agree or disagree with a series of statements. For example, if we look at the following 5-point Likert Scale:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Nominal and ordinal data are qualitative, while interval and ratio data are quantitative.

Likert scales are ordinal in that one can easily see that the larger number corresponds to a higher level of agreeableness. Some people argue that since there is a one-unit difference between the numeric values Likert scales should be interval data. However, the number 1 is just a placeholder for someone that strongly disagrees. There is no way to quantify a one-unit difference between two different subjects that answered 1 or 2 on the scale. For example, one person's response for strongly disagree could stem from the exact same reasoning behind another person's response of disagree. People view subjects at different intensities that is not quantifiable.

## Discrete vs. Continuous

Quantitative variables are discrete or continuous. This difference will be important later on when we are working with probability. Discrete variables have gaps between points that are countable, usually integers like the number of cars in a parking garage or how many people per household. A continuous variable can take on any value and is measurable, like height, time running a race, distance between two buildings. Usually, just asking yourself if you can count the variable then it is discrete and if you can measure the variable then it is continuous. If you can actually count the number of outcomes (even if you are counting to infinity), then the variable is discrete.

**Discrete** variables can only take on particular values like integers.

Discrete variables have outcomes you can count.

**Continuous** variables can take on any value.

Continuous variables have outcomes you can measure.

For example, think of someone's age. They may report in a survey an integer value like 28 years-old. The person is not exactly 28 years-old though. From the time of their birth to the point in time that the survey respondent recorded, their age is a measurable number in some unit of time. A person's true age has a decimal place that can keep going as far as the best clock can measure time. It is more convenient to round our age to an integer rather than 28 years 5 months, 8 days, 14 hours, 12 minutes, 27 seconds, 5 milliseconds or as a decimal 28.440206335775. Therefore, age is continuous.

However, a continuous variable like age could be broken into discrete bins, for example, instead of the question asking for a numeric response for a person's age they could have had discrete age ranges where the survey respondent just checks a box.

1. Under 18
2. 18-24
3. 25-35
4. 36-45
5. 46-62
6. Over 62

When a survey question takes a continuous variable and chunks it into discrete categories, especially categories with different widths, you limit what type of statistics you can do on that data.

Figure 1-2 is a breakdown of the different variable and data types.

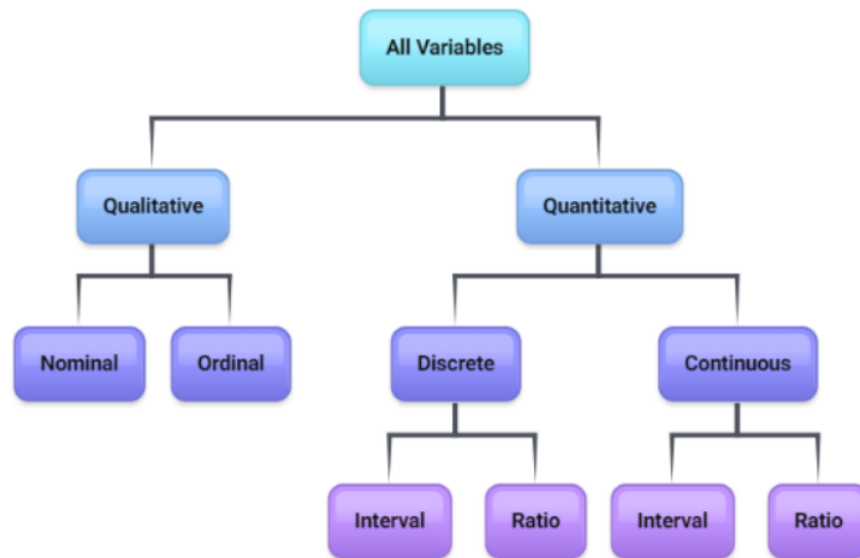


Figure 1-2

## Types of Sampling

If you want to know something about a population, it is often impossible or impractical to examine the entire population. It might be too expensive in terms of time or money to survey the population. It might be impractical: you cannot test all batteries for their length of lifetime because there would not be any batteries left to sell.

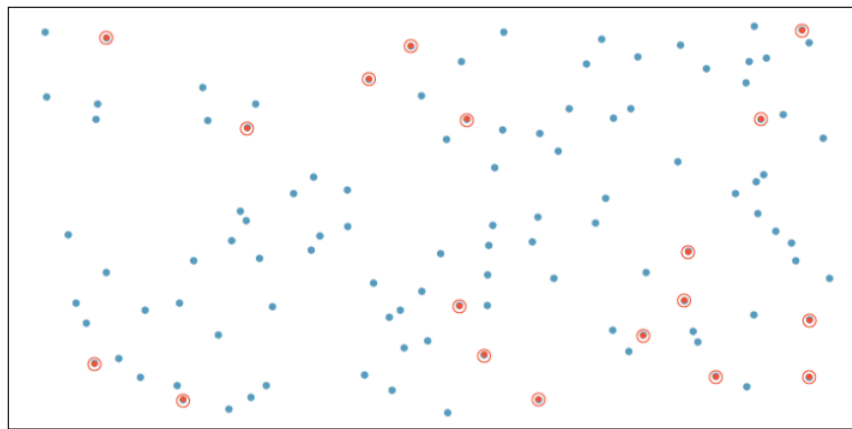
When you choose a sample, you want it to be as similar to the population as possible. If you want to test a new painkiller for adults, you would want the sample to include people of different weights, age, etc. so that the sample would represent all the demographics of the population that would potentially take the painkiller. The more similar the sample is to the population, the better our statistical estimates will be in predicting the population parameters.

There are many ways to collect a sample. No sampling technique is perfect, and there is no guarantee that you will collect a representative sample. That is unfortunately the limitation of sampling. However, several techniques can result in samples that give you a semi-accurate picture of the population. Just remember to be aware that the sample may not be representative of the whole population. As an example, you can take a random sample of a group of people that are equally distributed across all income groups, yet by chance, everyone you choose is only in the high-income group. If this happens, it may be a good idea to collect a new sample if you have the time and money.

When setting up a study there are different ways to sample the population of interest. The five main sampling techniques are:

1. Simple Random Sample
2. Systematic Sample
3. Stratified Sample
4. Cluster Sample
5. Convenience Sample

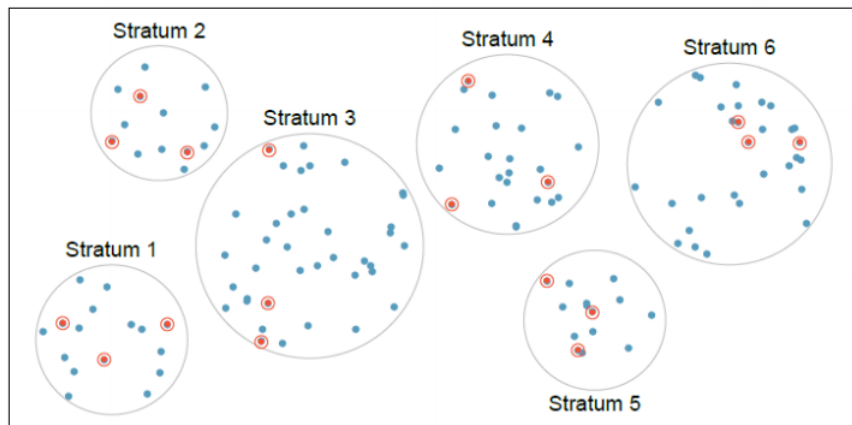
A **simple random sample** (SRS) means selecting a sample size of  $n$  objects from the population so that every sample of the same size  $n$  has equal probability of being selected as every other possible sample of the same size from that population. For example, we have a database of all PSU student data and we use a random number generator to randomly select students to receive a questionnaire on the type of transportation they use to get to school. See Figure 1-3. Simple random sampling was used to randomly select the 18 cases.



Retrieved from [OpenIntroStatistics](https://openintrostatistics.org/).

Figure 1-3

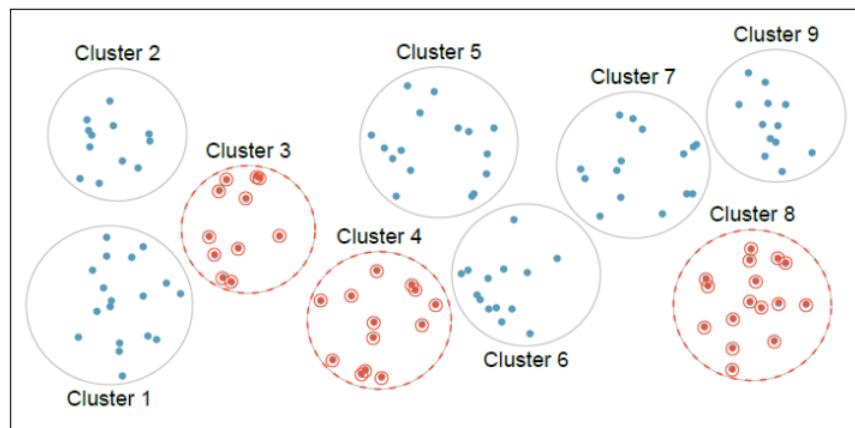
A **stratified sample** is where the population is split into groups called strata, then a random sample is taken from each stratum. For instance, we divide Portland by ZIP code and then randomly select  $n$  registered voters out of each ZIP code. See Figure 1-4. Cases were grouped into strata, then simple random sampling was employed within each stratum.



Retrieved from [OpenIntroStatistics](https://openintrostatistics.org/).

Figure 1-4

A **cluster sample** is where the population is split up into groups called clusters, then one or more clusters are randomly selected and all individuals in the chosen clusters are sampled. Similar to the previous example, we split Portland up by ZIP code, randomly pick 5 ZIP codes and then sample every registered voter in those 5 ZIP codes. See Figure 1-5. Data were binned into nine clusters, three of these clusters were sampled, and all observations within these three clusters were included in the sample.



Retrieved from [OpenIntroStatistics](https://openintrostatistics.org/).

Figure 1-5

A **systematic sample** is where we list the entire population, then randomly pick a starting point at the  $n^{\text{th}}$  object, and then take every  $n^{\text{th}}$  value until the sample size is reached. For example, we alphabetize every PSU student, randomly choose the number 7. We would sample the 7<sup>th</sup>, 14<sup>th</sup>, 21<sup>st</sup>, 28<sup>th</sup>, 35<sup>th</sup>, etc. student.

A **convenience sample** is picking a sample that is conveniently at hand. For example, asking other students in your statistics course or using social media to take your survey. Most convenience samples will give biased views and are not encouraged.

There are many more types of sampling, snowball, multistage, voluntary, purposive, and quota sampling to name some of the ways to sample from a population. We can also combine the different sampling methods. For example, we could stratify by rural, suburban and urban school districts, then take 3rd grade classrooms as clusters.

### Guidelines for planning a statistical study

1. Identify the individuals that you are interested in studying. Realize that you can only make conclusions for these individuals. As an example, if you use a fertilizer on a certain genus of plant, you cannot say how the fertilizer will work on any other types of plants. However, if you diversify too much, then you may not be able to tell if there really is an improvement since you have too many factors to consider.
2. Specify the variable. You want to make sure the variable is something that you can measure, and make sure that you control for all other factors too. For example, if you are trying to determine if a fertilizer works by measuring the height of the plants on a particular day, you need to make sure you can control how much fertilizer you put on the plants (which is what we call a treatment), and make sure that all the plants receive the same amount of sunlight, water, and temperature.
3. Specify the population. This is important in order for you to know for whom and what conclusions you can make.
4. Specify the method for taking measurements or making observations.
5. Determine if you are taking a census or sample. If taking a sample, decide on the sampling method.
6. Collect the data.
7. Use appropriate descriptive statistics methods and make decisions using appropriate inferential statistics methods.
8. Note any concerns you might have about your data collection methods and list any recommendations for future.

### Observational vs. Experimental

The section is an introduction to experimental design. This is a brief introduction on how to design an experiment or a survey so that they are statistically sound. Experimental design is a very involved process, so this is just a small overview.

There are two types of studies:

1. An **observational study** is when the investigator collects data by observing, measuring, counting, watching or asking questions. The investigator does not change anything.
2. An **experiment** is when the investigator changes a variable or imposes a treatment to determine its effect.

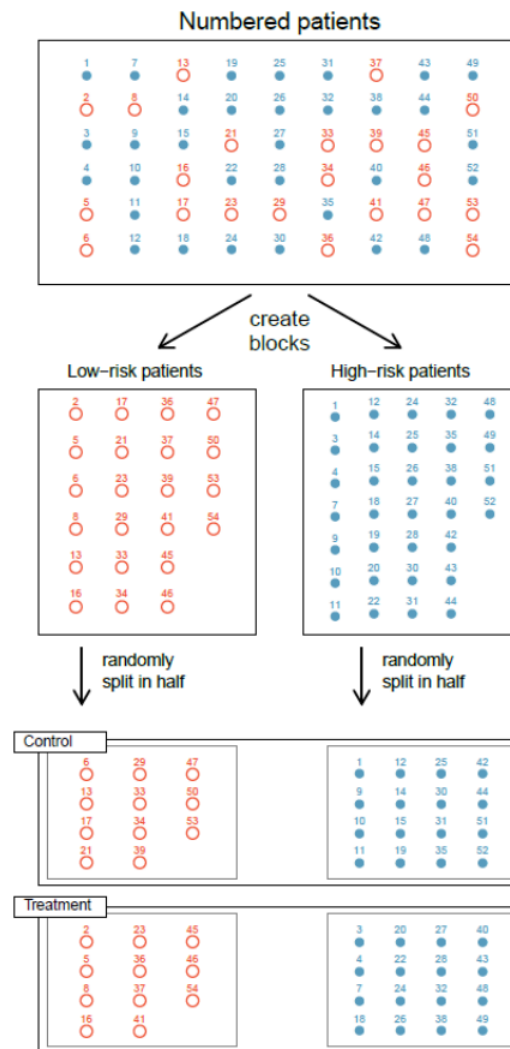
For instance, if you were to poll students to see if they favor increasing tuition, this would be an observational study since you are asking a question and getting data. Give a patient a medication that lowers their blood pressure. This is an experiment since you are giving the treatment and then getting the data.

Many observational studies involve surveys. A **survey** uses questions to collect the data and needs to be written so that there is no bias.

Bias is the tendency of a statistic to incorrectly estimate a parameter. There are many ways bias can seep into statistics. Sometimes we don't ask the correct question, give enough options for answers, survey the wrong people, misinterpret data, sampling or measurement errors, or unrepresentative samples.

In an experiment, there are different options to assign treatments.

1. **Completely Randomized Experiment:** In this experiment, the individuals are randomly placed into two or more groups. One group gets either no treatment or a placebo (a fake treatment); this group is called the control group. The groups getting the treatment are called the treatment groups. The idea of the placebo is that a person thinks they are receiving a treatment, but in reality, they are receiving a sugar pill or fake treatment. Doing this helps to account for the placebo effect, which is where a person's mind makes their body respond to a treatment because they think they are taking the treatment when they are not really taking the treatment. Note, not every experiment needs a placebo, such as when using animals or plants. In addition, you cannot always use a placebo or no treatment. For example, if you are testing a new Ebola vaccination you cannot give a person with the disease a placebo or no treatment because of ethical reasons.
2. **Matched Pairs Design:** This is a subset of the randomized block design where the treatments are given to two groups that can be matched up with each other in some way. One example would be to measure the effectiveness of a muscle relaxer cream on the right arm and the left arm of individuals, and then for each individual you can match up their right arm measurement with their left arm. Another example of this would be before and after experiments, such as weight of a person before and weight after a diet.
3. **Randomized Block Design:** A block is a group of subjects that are considered similar or the same subject measured multiple times, but the blocks differ from each other. Then randomly assign treatments to subjects inside each block. For instance, a company has several new stitching methods for a soccer ball and would like to pick the ball that travels the fastest. We would expect variation in different soccer player's abilities which we do not want affect our results. We randomly choose players to kick each of the new types of balls where the order of the ball design is also randomized. Figure 1-6 shows blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.
4. **Factorial Design:** This design has two or more independent categorical variables called factors. Each factor has two or more different treatment levels. The factorial design allows the researcher to test the effect of the different factors simultaneously on the dependent variable. For example, an educator believes that both the time of day (morning, afternoon, evening) and the way an exam is delivered (multiple-choice paper, short answer paper, multiplechoice electronic, short answer electronic) affects a student's grade on their exam.



Retrieved from [OpenIntroStatistics](https://openintrostatistics.org/).

Figure 1-6

No matter which experiment type you conduct, you should also consider the following:

**Replication:** repetition of an experiment on more than one subject so you can make sure that the sample is large enough to distinguish true effects from random effects. It is also the ability for someone else to duplicate the results of the experiment.

**Blind study** is where the individual does not know which treatment they are getting or if they are getting the treatment or a placebo.

**Double-blind study** is where neither the individual nor the researcher knows who is getting the treatment and who is getting the placebo. This is important so that there can be no bias in the results created by either the individual or the researcher.

One last consideration is the time-period that you are collecting the data. There are different time-periods that you can consider.

**Cross-sectional study:** observational data collected at a single point in time.

**Retrospective study:** observational data collected from the past using records, interviews, and other similar artifacts.

**Prospective (or longitudinal or cohort) study:** Subjects are measured from a starting point over time for the occurrence of the condition of interest.

This page titled [1.3: Collecting Data and Sampling Techniques](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Rachel Webb](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.