# The five hierarchical scales of a study

The five hierarchical scales simplify the process of developing a statistical study design. They allow us to spot errors with the design, provide direction for visualizing data, and indicate the most appropriate statistical tests to perform. The following list introduces the five scales from smallest to largest, followed by an explanation of how they are connected.

**Sampling** **unit** is the unit being selected at random. For example, if you randomly selected 100 email addresses to gather data on grocery store preference, then the *sampling unit* is the email address of a person. The *sampling unit* may be the same as the *observation unit*, or it may contain multiple observation units.

**Sample** is the collection of sampling unitsthat you randomly selected. For example, if 72 people replied to your email about their favourite grocery store, then your sample includes the 72 email responses.

**Observation** **unit** is the scale for data collection. You can think of this as the subject of the study. For example, if we were to ask people which grocery store was their favourite, the observation unit would simply be the individual person.

**Statistical population** is the collection of all sampling unitsthat could have been in your *sample*, and represents the true scale in which your statistical conclusions are valid. For example, let's say that you decided to collect your data by sending an email out to a 100 random people from a list of all emails in the City of Toronto. The statistical populationwould then be all people in Toronto with an active email account.

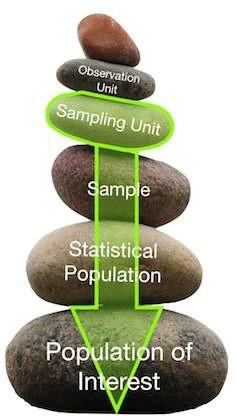
**Population of interest** is the collection of sampling unitsthat you hope to draw a conclusion about. In contrast to the *statistical population*, which is defined by the technical details of your sampling design, the population of interestis defined by the scope of your research question. For example, if your research question is about the proportion of people who shop at large grocery stores as opposed to locally owned corner stores in Toronto, then the population of interestis all the people in Toronto. Ideally the population of interestis the same as your statistical population, but often the population of interest is larger.

These five hierarchical scales provide the base terminology to define all aspects of the sampling process. They are inherently interdependent, and as such, they need to be designed so that they align with each other, and with the research question. To see the interconnectedness of the scales, it helps to restate the relationship among them, starting with the *sampling unit*. As we will see later, the *sampling unit* is also of particular importance because it is the unit of replication for statistical studies.

* **Measurement Unit** The scale of the measurement variable
* **Measurement Variable** The type of data you are collecting
* **Observation Unit** The unit you collect data from
* **Population of Interest** The population that you hope to draw a conclusion about
* **Sample** The collection of all sampling units that were selected
* **Statistical Population** The collection of all sampling units that could be selected to be in your sample
* **Sampling Unit** The unit that is selected at random

# 

## Moving from the sampling unit to larger scales



The *sampling unit* is the scale at which you are selecting things at random. Moving to larger scales, the next two scales are the sampleand the statistical population. It is relatively easy to define the samplesince it is simply the collection of sampling unitsthat you actually collected.

Defining the statistical population, however, can be more challenging because it is based on the details of the sampling process rather than anything thing you observe directly. The statistical populationis the collection of sampling unitsthat *could have been* in your sample based on your study design.

The largest scale is the population of interest, which is defined by your research question rather than the study design. Ideally the statistical populationand the population of interestare the same thing, but that is relatively rare. The key distinction is that the population of interestis defined by your research question, whereas the statistical populationis defined by your study design.

## Moving from the sampling unit to smaller scales

Moving to smaller scales, the *observation unit* is the unit from which you directly collect data. Sometimes the *observation unit* and *sampling unit* are the same thing, but sometimes they are different. For example, if you found prospective voters by randomly sampling addresses from the voter registry, then your *sampling unit* is an address but your *observation unit* is a person. Distinguishing the *sampling unit* from the *observation unit* is important for identifying the study design and for avoiding common statistical errors.

## Measurement variables and units

The last two terms that we need to introduce are the measurement variableand the measurement unit. The measurement variableis what we want to measure about the *observation unit*, such as height, age or voting intent. The measurement unitis the scale of the measurement variable, such as centimetres for height or years for age. If the data are categorical, such as voting intent, then there is no measurement unit.

# Descriptive and Inferential Statistics

Descriptive statisticscharacterize the data in your sample and includes things like averages, tables, and graphs. In contrast, inferential statisticsuses information from your sample to make a probabilistic statement about the statistical population. The tools used to conduct descriptive statisticsand inferential statisticsare be covered in detail throughout later lessons, but here we will talk about the kinds of questions that can be answered with each.

This lesson introduces the distinction between descriptive statisticsand inferential statistics. By the end, you will be able to:

* Describe the steps involved in a statistical study, and the role of descriptive statistics versus inferential statistics
* Identify questions that can be answered using descriptive statistics versus inferential statistics
* Apply the framework to statistical inference for a single population and for comparisons among populations

The overall framework of statistics can be separated into four steps:

**Sampling** is the step of creating your study design and collecting your samples.

**Measuring** is the step of taking measurements from your observation units, which gives you the data with which to work. It may be just a single measurement variable from the observation unit (e.g., weight of a cow) or multiple measurement variables (e.g., weight, age, and health of a cow).

**Calculating descriptive statistics** is the step where you describe the data in your sample. This may include calculating the average value of a measurement variable in your data set, calculating the variation among measurements, or creating graphs. Descriptive statisticsis covered in detail in later lessons.

**Calculating inferential statistics** is the final step where you use the information contained in your data to draw a conclusion about the statistical population. Inferential statisticsis covered in detail in later lessons.

The primary distinction between the last two steps is that descriptive statistics are used to characterize the data you collected, whereas inferential statistics are used to draw conclusions about the statistical population.