# Data types

This lesson is about identifying different types of variables. By the end, you will be able to:

* Distinguish datafrom variables
* Identify and distinguish ordinal categorical variables, nominal categorical variables, continuous numerical variables, and discrete numerical variables

A variableis any measurable characteristic of an observation unit. The idea is that the value of the variablecan *vary* among sampling units (hence the name 'variable').

A variablecontains three pieces of information. The first is what the variablerepresents, the second is the measurement unit, and the third is a description of the observation unit. For example, if we wanted to observe the long jump distance for high school students, then our variablemight be "Long jump distance (cm) for each student". Breaking it down, 'long jump distance' is what the variablerepresents, 'centimetres' is the measurement unit, and 'student' is the observation unit.

## Types of variables

*Numerical* variables are those where the data is numeric, whereas categoricalvariables are those where the data is a qualitative description.

Categoricalvariables have no measurement units because they are qualitative descriptions. *Numerical* variables, however, have measurement units that indicates the scale (e.g., centimeters, seconds, kilograms). There are several subtypes within both categorical and numerical; we will look at four here.

* **Continuous numerical variable** is a variable that can take on continuous numbers. Continuous numbers are those that can take on any value including fractional numbers.
* **Discrete numerical variable** is a variable that can take only take on whole numbers (integers)
* **Ordinal categorical variable** is a variable that can take on qualitative values but where values are from a ranked scale.
* **Nominal categorical variable** is a variable that can take on qualitative values but where values do not have any particular order.

## Descriptive stats

This lesson is about calculating descriptive statistics. By the end, you will be able to:

* Calculate the central tendency for numerical and categorical variables including mean, median, counts, and proportions
* Calculate dispersion for numerical variables including variance, standard deviation, interquartile range, and range

Central tendency describes the typical value in your sample (e.g., mean), and dispersion describes the spread of the values (e.g., variance).

Counts and proportions indicate the central tendency of categorical data. On the other hand, range is used to indicate dispersion, which describes the variation in the response variable.

## Numerical variables

The first approach to calculating descriptive statistics for numerical variables is based on the meanvalue of your sample. The meancharacterizes the central tendency of a numeric variable and is calculated by

1. Summing all of the values in your sample
2. Dividing by the number of data points in your sample.

The variance is a measure of the amount of variation in your sample. It is calculated as the average squared distance of each data point from the sample mean. A related quantity is the standard deviation, which is the square root of variance. We will use the convention that the varianceis given the notation *σ2* and the standard deviationthe notation *σ*. The varianceis calculated as:

1. Calculate the mean for a sample
2. Calculate the difference between each data point and the mean, then square that value
3. Sum the squares of the differences and divide by the number of observations/data points

Quartilesare specific values of the variable that divide your data into ranked groups. For example, if we sort our data from smallest to largest and split it into four roughly equal groups, then each group contains about 25% of the data. What’s cool is that we now know the range of values that are in the lower and upper parts of the dataset. This will allow us to calculate both *central tendency* and *dispersion*.

To begin, we need to calculate the quartiles. The steps are:

1. Sort the data in your sample from lowest to highest value
2. Find the 2nd quartileby splitting the data in half according to whether
   1. The sample has an odd number of observations, in which case the middle value of the dataset is the second quartile.
   2. The sample has an even number of observations, in which case the average of the two values closest to the middle is the second quartile.
3. Find the 1st quartileby creating a subset of the data that is the lower-valued half of the observations, then use the rules in step 2 to find the middle value. The lower-valued subset is created according to whether
   1. The sample has an odd number of observations, in which case the lower-valued subset is all values less than or equal to the second quartile. The subset includes the second quartile*.*
   2. The sample has an even number of observations, in which case the lower-valued subset is all values less than the second quartile. The subset does not include the second quartile.
4. Find the 3rd quartileby repeating step 3 but for the upper-valued half of the observations.
5. To calculate the interquartile range, subtract the 1st quartilefrom the 3rd quartile.

### Quartiles

**Pros** The medianand *interquartile* range are relatively robust to extreme values. Consider the two samples of wait times for Starbucks coffee in the figure below. In the first sample, the seven observations are similar to each other around a value of 3 minutes. In the second sample, there is one unusually long wait time of over 12 minutes. Despite the unusually large wait time in the second sample, the medianand *interquartile* range are similar for both.

**Cons** The medianand *interquartile* range become quite variable for samples with a small number of observations. Imagine that you had just three observations in your sample. The medianand *interquartile* range will be sensitive to the value of the middle observation.

### Means

**Pros** The meanand standard deviationare more robust when there is a small number of observations in the sample.

**Cons** The downside to the meanand standard deviationis that they are sensitive to extreme values. Consider the same two samples of coffee wait times used above, but now looking at the meanand standard deviation, shown in the figure below. There you can see how much of an impact the single extreme value has on both the meanand standard deviationcompared to the quartiles.

# Effect sizes

This lesson is about effect size. By the end, you will be able to:

* Define effect size, and be able to calculate it for numerical and categorical variables
* Distinguish between using a simple differenceversus a ratioto quantify *absolute* effect size

The effect sizethat we are calculating is called the *absolute* effect size, which is the simple change in mean value between groups. Unfortunately, "effect size*"* also refers to a separate topic in inferential statistics. To avoid confusion, we are calculating the *absolute* effect size *for descriptive statistics*. While it is cumbersome, it helps separate the different topics. *Effect* *size* can be calculated either as a *difference* or a *ratio*.

**Difference** calculations are the differences in mean values among groups. For example, a sample of customers at Tim Hortons spent an average of $1.79 for a coffee whereas a sample of customers at Starbucks spent $2.10 for a coffee. The effect sizebased on difference is $2.10-$1.79=$0.31. Starbucks coffee is $0.31 more expensive than Tim Hortons. Using the differenceto calculate effect sizehas the advantage of retaining the original scale. In the coffee example, effect size is still in units of dollars.

**Ratio** calculations are the ratio of mean values among groups. In the above coffee example, the effect sizebased on ratios is $2.10/$1.79=1.17. Starbucks coffee is 1.17 times the cost of Tim Hortons, or 17% more expensive. Using the ratioto calculate effect sizehas the advantage of indicating a relative change, but it loses the original scale.