# Data Types

In this video, two data types are introduced: **Quantitative** and **Categorical**.

**Quantitative** data takes on numeric values that allow us to perform mathematical operations (like the number of dogs).

**Categorical** are used to label a group or set of items (like dog breeds - Collies, Labs, Poodles, etc.).

# Categorical Ordinal vs. Categorical Nominal

We can divide categorical data further into two types: **Ordinal** and **Nominal**.

**Categorical Ordinal** data take on a ranked ordering (like a ranked interaction on a scale from Very Poor to Very Good with the dogs).

**Categorical Nominal** data do not have an order or ranking (like the breeds of the dog).

# Continuous vs. Discrete

We can think of quantitative data as being either **continuous** or **discrete**.

**Continuous** data can be split into smaller and smaller units, and still a smaller unit exists. An example of this is the age of the dog - we can measure the units of the age in years, months, days, hours, seconds, but there are still smaller units that could be associated with the age.

**Discrete** data only takes on countable values. The number of dogs we interact with is an example of a discrete data type.

# Recap of Previous Video

The table below summarizes our data types. To expand on the information in the table, you can look through the text that follows.

| **Data Types** |  |  |
| --- | --- | --- |
| **Quantitative:** | **Continuous** | **Discrete** |
|  | Height, Age, Income | Pages in a Book, Trees in Yard, Dogs at a Coffee Shop |
|  |  |  |
| **Categorical:** | **Ordinal** | **Nominal** |
|  | Letter Grade, Survey Rating | Gender, Marital Status, Breakfast Items |

Below is a little more detail of the information shared in the above table.

**Another Look**

To break down our data types, there are two main blocks:

**Quantitative** and **Categorical**

**Quantitative** can be further divided into Continuous or Discrete.

**Categorical** data can be divided into Ordinal or Nominal.

You should have now mastered what types of data in the world around us falls into each of these four buckets: Discrete, Continuous, Nominal, and Ordinal. In the next sections, we will work through the numeric summaries that relate specifically to quantitative variables.

**Quantitative vs. Categorical**

Some of these can be a bit tricky - notice even though zip codes are a number, they aren’t really a quantitative variable. If we add two zip codes together, we do not obtain any useful information from this new value. Therefore, this is a categorical variable.

**Height**, **Age**, the **Number of Pages in a Book** and **Annual Income** all take on values that we can add, subtract and perform other operations with to gain useful insight. Hence, these are quantitative.

**Gender**, **Letter Grade**, **Breakfast Type**, **Marital Status**, and **Zip Code** can be thought of as labels for a group of items or individuals. Hence, these are categorical.

**Continuous vs. Discrete**

To consider if we have continuous or discrete data, we should see if we can split our data into smaller and smaller units. Consider time - we could measure an event in years, months, days, hours, minutes, or seconds, and even at seconds we know there are smaller units we could measure time in. Therefore, we know this data type is continuous. **Height**, **age**, and **income** are all examples of continuous data. Alternatively, the **number of pages in a book**, **dogs I count outside a coffee shop**, or **trees in a yard** are discrete data. We would not want to split our dogs in half.

**Ordinal vs. Nominal**

In looking at categorical variables, we found **Gender**, **Marital Status**, **Zip Code** and your **Breakfast items** are nominal variables where there is no order ranking associated with this type of data. Whether you ate cereal, toast, eggs, or only coffee for breakfast; there is no rank ordering associated with your breakfast.

Alternatively, the **Letter Grade** or **Survey Ratings** have a rank ordering associated with it, as ordinal data. If you receive an A, this is higher than an A-. An A- is ranked higher than a B+, and so on... Ordinal variables frequently occur on rating scales from very poor to very good. In many cases we turn these ordinal variables into numbers, as we can more easily analyze them, but more on this later!

**Final Words**

In this section, we looked at the different data types we might work with in the world around us. When we work with data in the real world, it might not be very clean - sometimes there are typos or missing values. When this is the case, simply having some expertise regarding the data and knowing the data type can assist in our ability to ‘clean’ this data. Understanding data types can also assist in our ability to build visuals to best explain the data. But more on this very soon!

**QUIZ QUESTION**

This quiz will assure you have a clear understanding of the differences between categorical nominal vs. categorical ordinal variables. All of the variables below are categorical. Your task is to select the **check** box next to each variable that is **nominal**; do not check the ordinal categorical variables.

* 

Letter Grades (A, B+, B, B-, etc.)

* Types of Fruit (Apple, Banana, etc.)
* 

Ratings on a Survey (Poor, Ok, Great)

* Types of Dog Breeds (German Shepherd, Collie, etc.)
* Genres of Movies (Horror, Comedy, etc.)
* Gender
* Nationality

This quiz will ensure you have a clear understanding of the differences between quantitative continuous vs. discrete variables. All of the variables below are quantitative. Your task is select the check box next to each variable that is continuous; do not check the discrete variable:

Travel Distance from Home to Work

Amount of Rain in a Year

Time to Run a Mile

Amount of Water Consumed in a Day

# Analyzing Quantitative Data

**Four Aspects for Quantitative Data**

There are four main aspects to analyzing **Quantitative** data.

1. Measures of Center
2. Measures of Spread
3. The Shape of the data.
4. Outliers

**Analyzing Categorical Data**

Though not discussed in the video, analyzing categorical data has fewer parts to consider. **Categorical** data is analyzed usually be looking at the counts or proportion of individuals that fall into each group. For example if we were looking at the breeds of the dogs, we would care about how many dogs are of each breed, or what proportion of dogs are of each breed type.

**Measures of Center**

There are three measures of center:

1. Mean
2. Median
3. Mode

**The Mean**

In this video, we focused on the calculation of the mean. The mean is often called the average or the **expected value** in mathematics. We calculate the mean by adding all of our values together, and dividing by the number of values in our dataset.

The remaining measures of the median and mode will be discussed in detail in the upcoming quizzes and videos.

# The Median

The **median** splits our data so that 50% of our values are lower and 50% are higher. We found in this video that how we calculate the median depends on if we have an even number of observations or an odd number of observations.

**Median for Odd Values**

If we have an **odd** number of observations, the **median** is simply the number in the **direct middle**. For example, if we have 7 observations, the median is the fourth value when our numbers are ordered from smallest to largest. If we have 9 observations, the median is the fifth value.

**Median for Even Values**

If we have an **even** number of observations, the **median** is the **average of the two values in the middle**. For example, if we have 8 observations, we average the fourth and fifth values together when our numbers are ordered from smallest to largest.

In order to compute the median we MUST sort our values first.

Whether we use the mean or median to describe a dataset is largely dependent on the **shape** of our dataset and if there are any **outliers**. We will talk about this in just a bit!

# The Mode

The **mode** is the most frequently observed value in our dataset.

There might be multiple modes for a particular dataset, or no mode at all.

**No Mode**

If all observations in our dataset are observed with the same frequency, there is no mode. If we have the dataset:

1, 1, 2, 2, 3, 3, 4, 4

There is no mode, because all observations occur the same number of times.

**Many Modes**

If two (or more) numbers share the maximum value, then there is more than one mode. If we have the dataset:

1, 2, 3, 3, 3, 4, 5, 6, 6, 6, 7, 8, 9

There are two modes 3 and 6, because these values share the maximum frequencies at 3 times, while all other values only appear once.

# Notation

Notation is a common language used to communicate mathematical ideas. **Think of notation as a universal language used by academic and industry professionals to convey mathematical ideas.** In the next videos, you might see things that seem confusing. Use the quizzes to assist with your understanding of the concepts.

You likely already know some notation. Plus, minus, multiply, division, and equal signs all have mathematical symbols that you are likely familiar with. Each of these symbols replaces an idea for how numbers interact with one another. In the coming concepts, you will be introduced to some additional ideas related to notation. Though you will not need to use notation to complete the project, it does have the following properties:

1. **Understanding how to correctly use notation makes you seem really smart.** Knowing how to read and write in notation is like learning a new language. A language that is used to convey ideas associated with mathematics.
2. **It allows you to read documentation, and implement an idea to your own problem.** Notation is used to convey how problems are solved all the time. One really popular mathematical algorithm that is used to solve some of the world's most difficult problems is known as Gradient Boosting. The way that it solves problems is explained here: <https://en.wikipedia.org/wiki/Gradient_boosting>. If you really want to understand how this algorithm works, you need to be able to read and understand notation.
3. **It makes ideas that are hard to say in words easier to convey.** Sometimes we just don't have the right words to say. For those situations, I prefer to use notation to convey the message. Similar to the way an emoji or meme might convey a feeling better than words, notation can convey an idea better than words. Usually those ideas are related to mathematics, but I am not here to stifle your creativity.

# Example to Introduce Notation

There is a lot going on in this video - here is a recap of the big ideas.

**Rows and Columns**

If you aren't familiar with spreadsheets, this will be covered in detail in future lessons. Spreadsheets are a common way to hold data. They are composed of rows and columns. Rows run horizontally, while columns run vertically. Each column in a spreadsheet commonly holds a specific **variable**, while each row is commonly called an **instance** or **individual**.

The example used in the video is shown below.

| **Date** | **Day of Week** | **Time Spent On Site (X)** | **Buy (Y)** |
| --- | --- | --- | --- |
| June 15 | Thursday | 5 | No |
| June 15 | Thursday | 10 | Yes |
| June 16 | Friday | 20 | Yes |

This is a **row**:

| **Date** | **Day of Week** | **Time Spent On Site (X)** | **Buy (Y)** |
| --- | --- | --- | --- |
| June 15 | Thursday | 5 | No |

This is a **column**:

| **Time Spent On Site (X)** |
| --- |
| 5 |
| 10 |
| 20 |

**Before Collecting Data**

**Before collecting data, we usually start with a question, or many questions, that we would like to answer. The purpose of data is to help us in answering these questions.**

**Random Variables**

A **random variable** is a placeholder for the possible values of some process (mostly... the term 'some process' is a bit ambiguous). As was stated before, notation is useful in that it helps us take complex ideas and simplify (often to a single letter or single symbol). We see random variables represented by capital letters (**X**, **Y**, or **Z** are common ways to represent a random variable).

We might have the random variable **X**, which is a holder for the possible values of the amount of time someone spends on our site. Or the random variable **Y**, which is a holder for the possible values of whether or not an individual purchases a product.

**X** is 'a holder' of the values that could possibly occur for the amount of time spent on our website. Any number from 0 to infinity really.

# Capital vs. Lower Case Letters

**Random variables** are represented by capital letters. Once we observe an outcome of these random variables, we notate it as a lower case of the same letter.

**Example 1**

For example, the **amount of time someone spends on our site** is a **random variable** (we are not sure what the outcome will be for any particular visitor), and we would notate this with **X**. Then when the first person visits the website, if they spend 5 minutes, we have now observed this outcome of our random variable. We would notate any outcome as a lowercase letter with a subscript associated with the order that we observed the outcome.

If 5 individuals visit our website, the first spends 10 minutes, the second spends 20 minutes, the third spends 45 mins, the fourth spends 12 minutes, and the fifth spends 8 minutes; we can notate this problem in the following way:

**X** is the amount of time an individual spends on the website.

**x1**​ = 10,       {x\_2}**x2**​ = 20       {x\_3}**x3**​ = 45       {x\_4}**x4**​ = 12       {x\_5}**x5**​ = 8.

The capital **X** is associated with this idea of a **random variable**, while the observations of the random variable take on lowercase **x** values.

**Example 2**

Taking this one step further, we could ask:

**What is the probability someone spends more than 20 minutes in our website?**

In notation, we would write:

**P(X > 20)?**

Here **P** stands for **probability**, while the parentheses encompass the statement for which we would like to find the probability. Since **X** represents the amount of time spent on the website, this notation represents the probability the amount of time on the website is greater than 20.

We could find this in the above example by noticing that only one of the 5 observations exceeds 20. So, we would say there is a **1** (the 45) **in 5 or 20%** chance that an individual spends more than 20 minutes on our website (based on this dataset).

**Example 3**

If we asked: **What is the probability of an individual spending 20 or more minutes on our website?** We could notate this as:

**P(X** ≥**20)?**

We could then find this by noticing there are two out of the five individuals that spent 20 or more minutes on the website. So this probability is **2 out of 5 or 40%**.

# Notation for Calculating the Mean

We know that the mean is calculated as the sum of all our values divided by the number of values in our dataset.

In our current notation, adding all of our values together can be extremely tedious. If we want to add 3 values of some random variable together, we would use the notation:

**x1**​+**x2**​+**x3**​

If we want to add 6 values together, we would use the notation:

**x1**​+**x2**​+**x3**​+**x4**​+**x5**​+**x6**​

To extend this to add one hundred, one thousand, or one million values would be ridiculous! How can we make this easier to communicate?!

# Measures of Spread

**Measures of Spread** are used to provide us an idea of how spread out our data are from one another. Common measures of spread include:

1. **Range**
2. **Interquartile Range (IQR)**
3. **Standard Deviation**
4. **Variance**

Throughout this lesson you will learn how to calculate these, as well as why we would use one measure of spread over another.

# Histograms

Histograms are super useful to understanding the different aspects of quantitative data. In the upcoming concepts, you will see histograms used all the time to help you understand the four aspects we outlined earlier regarding a quantitative variable:

* center
* spread
* shape
* outliers

# Calculating the 5 Number Summary

The five number summary consist of 5 values:

1. **Minimum:** The smallest number in the dataset.
2. **Q1**​: The value such that 25% of the data fall below.
3. **Q2**​: The value such that 50% of the data fall below.
4. **Q3**​: The value such that 75% of the data fall below.
5. **Maximum:** The largest value in the dataset.

Range

The **range** is then calculated as the difference between the **maximum** and the **minimum**.

The **interquartile range** is calculated as the difference between **Q3**​ and **Q1**​.

# Standard Deviation and Variance

The **standard deviation** is one of the most common measures for talking about the spread of data. It is defined as **the average distance of each observation from the mean**.

In the above video we saw this as how far individuals were from the average distance from work (the example distances shown are examples from the full data set, the mean of just those 4 numbers is 38.5. The mean of 18 shown later in the video is the mean of the full data set which is not shown in the video). In the next video, you will see exactly how this is calculated.

# Important Final Points

The variance is used to compare the spread of two different groups. A set of data with higher variance is more spread out than a dataset with lower variance. Be careful though, there might just be an outlier (or outliers) that is increasing the variance, when most of the data are actually very close.

When comparing the spread between two datasets, the units of each must be the same.

When data are related to money or the economy, higher variance (or standard deviation) is associated with higher risk.

The standard deviation is used more often in practice than the variance, because it shares the units of the original dataset.

**Use in the World**

The standard deviation is associated with risk in finance, assists in determining the significance of drugs in medical studies, and measures the error of our results for predicting anything from the amount of rainfall we can expect tomorrow to your predicted commute time tomorrow.

These applications are beyond the scope of this lesson as they pertain to specific fields, but know that understanding the spread of a particular set of data is extremely important to many areas. In this lesson you mastered the calculation of the most common measures of spread.

# Histograms

We learned how to build a **histogram** in this video, as this is the most popular visual for quantitative data.

**Shape**

From a histogram we can quickly identify the shape of our data, which helps influence all of the measures we learned in the previous concepts. We learned that the distribution of our data is frequently associated with one of the three **shapes**:

**1. Right-skewed**

**2. Left-skewed**

**3. Symmetric** (frequently normally distributed)

**Summary**

| **Shape** | **Mean vs. Median** | **Real World Applications** |
| --- | --- | --- |
| Symmetric (Normal) | Mean equals Median | Height, Weight, Errors, Precipitation |
| Right-skewed | Mean greater than Median | Amount of drug remaining in a blood stream, Time between phone calls at a call center, Time until light bulb dies |
| Left-skewed | Mean less than Median | Grades as a percentage in many universities, Age of death, Asset price changes |

The mode of a distribution is essentially the tallest bar in a histogram. There may be multiple modes depending on the number of peaks in our histogram.

# When outliers are present

we should consider the following points.

**1.** Noting they exist and the impact on summary statistics.

**2.** If typo - remove or fix

**3.** Understanding why they exist, and the impact on questions we are trying to answer about our data.

**4.** Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.

**5.** Be careful in reporting. Know how to ask the right questions.

1. **Population** - our entire group of interest.
2. **Parameter** - numeric summary about a population
3. **Sample** - subset of the population
4. **Statistic** - numeric summary about a sample

# Binomial:

A screenshot of a social media post

Description automatically generated

# Conditional Probability

A screenshot of a social media post

Description automatically generated

# Bayes Rule:

**Step by Step Walkthrough**

The step-by-step breakdown of the solution is pretty quick. Let's recap what's covered in the solution video.

Let's start with what we know:

**Prior Probabilities**

The robot is perfectly ignorant about where it is, so prior probabilities are as follows:

P(at red} ) = 0.5*P*(at red)=0.5

P(at green} ) = 0.5*P*(at green)=0.5

**Conditional Probabilities**

The robot's sensors are not perfect. Just because the robot *sees* red does **not** mean the robot is *at* red.

P(red} | at red} ) = 0.8*P*(see red∣at red)=0.8

P(green} | at green} ) = 0.8*P*(see green∣at green)=0.8

**Posterior Probabilities**

From these prior and posterior probabilities we are asked to calculate the following posterior probabilities after the robot sees red:

1. P(at red} | red} )*P*(at red∣see red)
2. P(at green} | red} )*P*(at green∣see red)

and as a reminder, Bayes' rule looks like this:

P(A|B ) = \frac{P(B|A) \cdot P(A)}{P(B)}*P*(*A*∣*B*)=*P*(*B*)*P*(*B*∣*A*)⋅*P*(*A*)​

or, if we want to use our "versions" of A and B (for posterior #1)...

P(at red}|red} ) = \frac{P(red}|at red}) \cdot P(at red})}{P(red})}*P*(at red∣see red)=*P*(see red)*P*(see red∣at red)⋅*P*(at red)​

Now, we can read two of those terms from what we already know about our prior and conditional probabilities which means we can rewrite this as...

P(at red}|red} ) = \frac{0.8 \cdot 0.5}{P(red})}*P*(at red∣see red)=*P*(see red)0.8⋅0.5​

But we still have one unknown! What was the probability that we would see red? The answer is 0.5 and there are two ways I can convince myself of that. The first is intuitive and the second is mathematical.

**Why is P(see red) 0.5?**

**Argument 1: Intuitive**

Of course it's 0.5! What else could it be? The robot had a 50% belief that it was in red and a 50% belief that it was in green. Sure, its sensors are unreliable but that unreliability is symmetric and **not** biased towards mistakenly seeing either color.

So whatever the probability of seeing red is, that will also be the probability of seeing green. Since these two colors are the only possible colors the probability MUST be 50% for each!

**Argument 2: Mathematical (Law of Total Probability)**

There are exactly two situations where the robot would see red.

1. When the robot is in a red square and its sensors work correctly.
2. When the robot is in a green square and its sensors make a mistake.

I just need to add up these two probabilities to get the total probability of seeing red.

P(red} ) = P(at red}) \cdot P(red} | at red}) + P(at green}) \cdot P(red} | at green})*P*(see red)=*P*(at red)⋅*P*(see red∣at red)+*P*(at green)⋅*P*(see red∣at green)

I can read these quantities from above!

P(red} ) = 0.5 \cdot 0.8 + 0.5 \cdot 0.2*P*(see red)=0.5⋅0.8+0.5⋅0.2

P(red} ) = 0.4 + 0.1*P*(see red)=0.4+0.1

P(red} ) = 0.5*P*(see red)=0.5

# Sampling Distributions Notes

We have already learned some really valuable ideas about sampling distributions:

First, we have defined **sampling distributions** as **the distribution of a statistic**.

This is fundamental - I cannot stress the importance of this idea. We simulated the creation of sampling distributions in the previous ipython notebook for samples of size 5 and size 20, which is something you will do more than once in the upcoming concepts and lessons.

Second, we found out some interesting ideas about sampling distributions that will be iterated later in this lesson as well. We found that for proportions (and also means, as proportions are just the mean of 1 and 0 values), the following characteristics hold.

1. The sampling distribution is centered on the original parameter value.
2. The sampling distribution decreases its variance depending on the sample size used. Specifically, the variance of the sampling distribution is equal to the variance of the original data divided by the sample size used. This is always true for the variance of a sample mean!

In notation, we say if we have a random variable, \bold{X}**X**, with variance of \bold{\sigma^2}*σ***2**, then the distribution of \bold{\bar{X}}**X**¯ (the sampling distribution of the sample mean) has a variance of \bold{\frac{\sigma^2}{n}}**n***σ***2**​

# Common hypothesis tests include:

1. Testing a population mean [**(One sample t-test)**](http://sites.utexas.edu/sos/guided/inferential/numeric/claim/one-sample-t/).
2. Testing the difference in means [**(Two sample t-test)**](https://www.isixsigma.com/tools-templates/hypothesis-testing/making-sense-two-sample-t-test/)
3. Testing the difference before and after some treatment on the same individual [**(Paired t-test)**](http://www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf)
4. Testing a population proportion [**(One sample z-test)**](http://stattrek.com/statistics/dictionary.aspx?definition=one-sample%20z-test)
5. Testing the difference between population proportions [**(Two sample z-test)**](https://onlinecourses.science.psu.edu/stat414/node/268)

You can use one of these sites to provide a t-table or z-table to support one of the above approaches:

* [**t-table**](https://s3.amazonaws.com/udacity-hosted-downloads/t-table.jpg)
* [**t-table or z-table**](http://www.z-table.com/t-value-table.html)

**There are literally hundreds of different hypothesis tests!** However, instead of memorizing how to perform all of these tests, you can find the statistic(s) that best estimates the parameter(s) you want to estimate, you can bootstrap to simulate the sampling distribution. Then you can use your sampling distribution to assist in choosing the appropriate hypothesis.

# Difficulties in A/B Testing

As you saw in the scenarios above, there are many factors to consider when designing an A/B test and drawing conclusions based on its results. To conclude, here are some common ones to consider.

* Novelty effect and change aversion when existing users first experience a change
* Sufficient traffic and conversions to have significant and repeatable results
* Best metric choice for making the ultimate decision (eg. measuring revenue vs. clicks)
* Long enough run time for the experiment to account for changes in behavior based on time of day/week or seasonal events.
* Practical significance of a conversion rate (the cost of launching a new feature vs. the gain from the increase in conversion)
* Consistency among test subjects in the control and experiment group (imbalance in the population represented in each group can lead to situations like [**Simpson's Paradox**](https://en.wikipedia.org/wiki/Simpson%27s_paradox))