Data and its Importance

21 December 2021 07:56

Data

Data is defined as distinct pieces of info. and it can come in many forms. From numbers in spreadsheet, text to video and Databases to images and audio recordings, utilizing data in its different forms is of new way of the world.

Data is used to understand and improve nearly every facet of our lives. So, no matter what field you are in, you can utilize data to make better decisions and accomplish your goals.

We will start this lesson with an overview of data types and the most common statistics used when analyzing data.

We'll discuss :

- Measures of center and spread.
- Common shapes that data takes on and how to handle outliers
- How to use spreadsheets to handle these calculations
- How to build visuals to communicate calculations

Data Types (Quantitative vs Categorical)

23 December 2021 11:18

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).

sum, subtn, multi, divisn

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

Categorical - Ordinal and Nominal Data

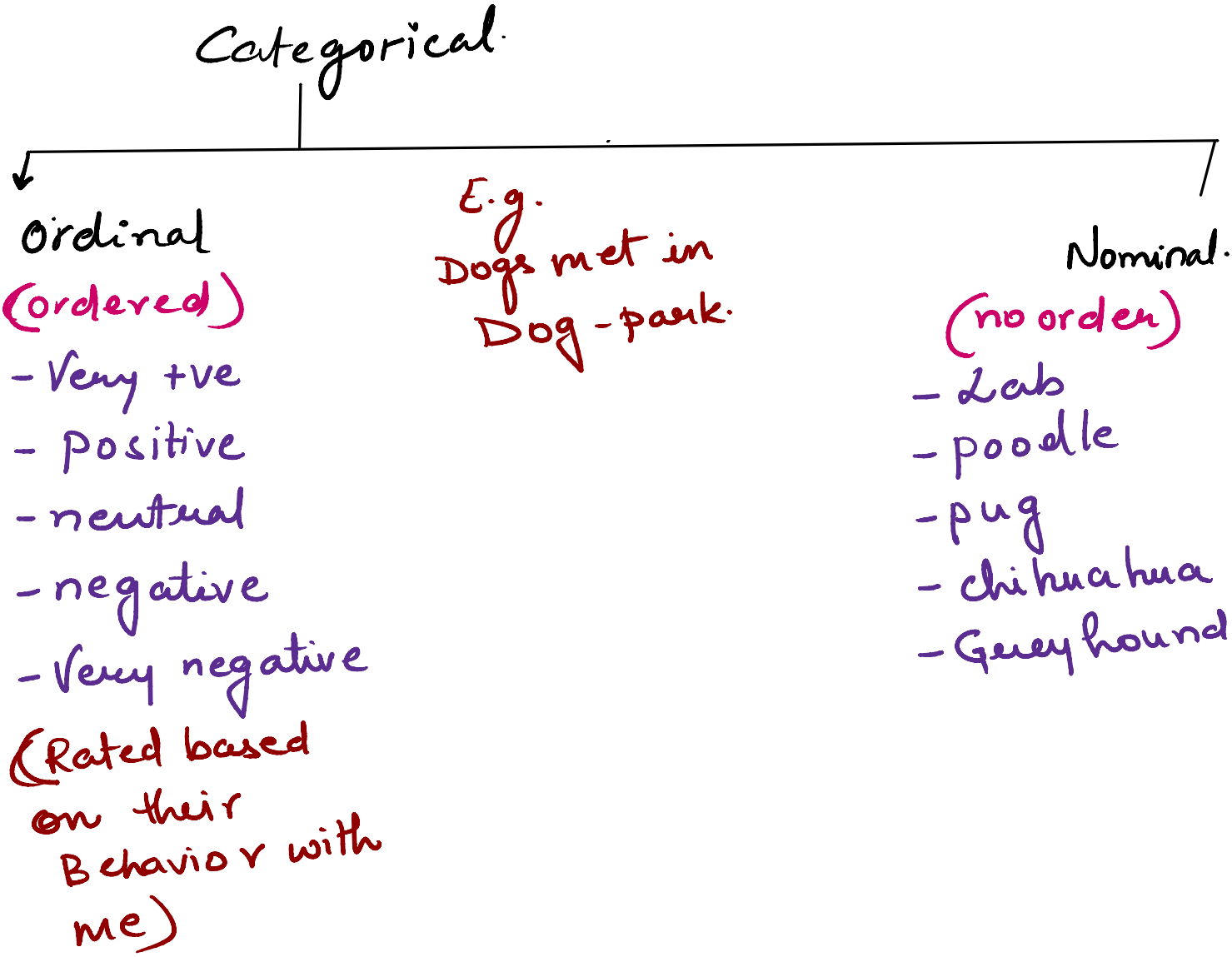
23 December 2021 11:34

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).



Quantitative - Discrete and Continuous

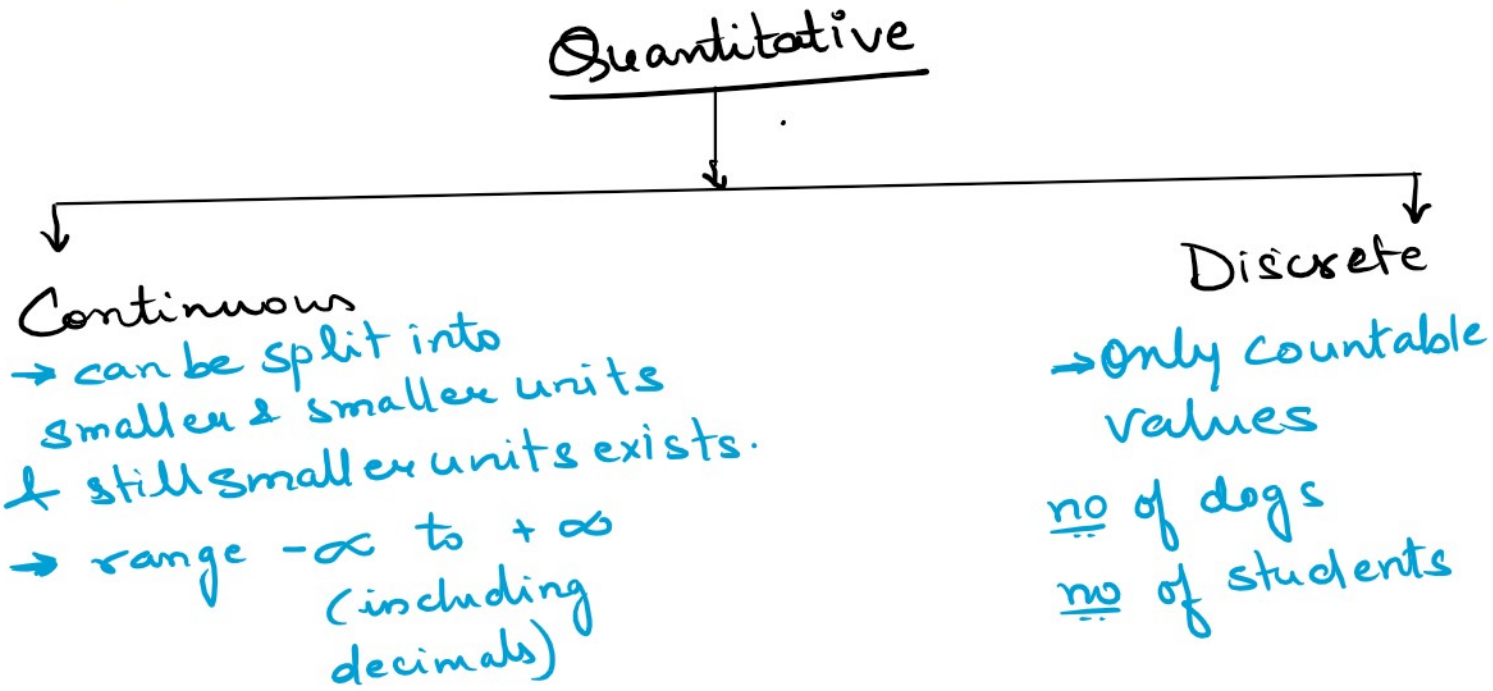
23 December 2021 11:40

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.



QUIZ QUESTION

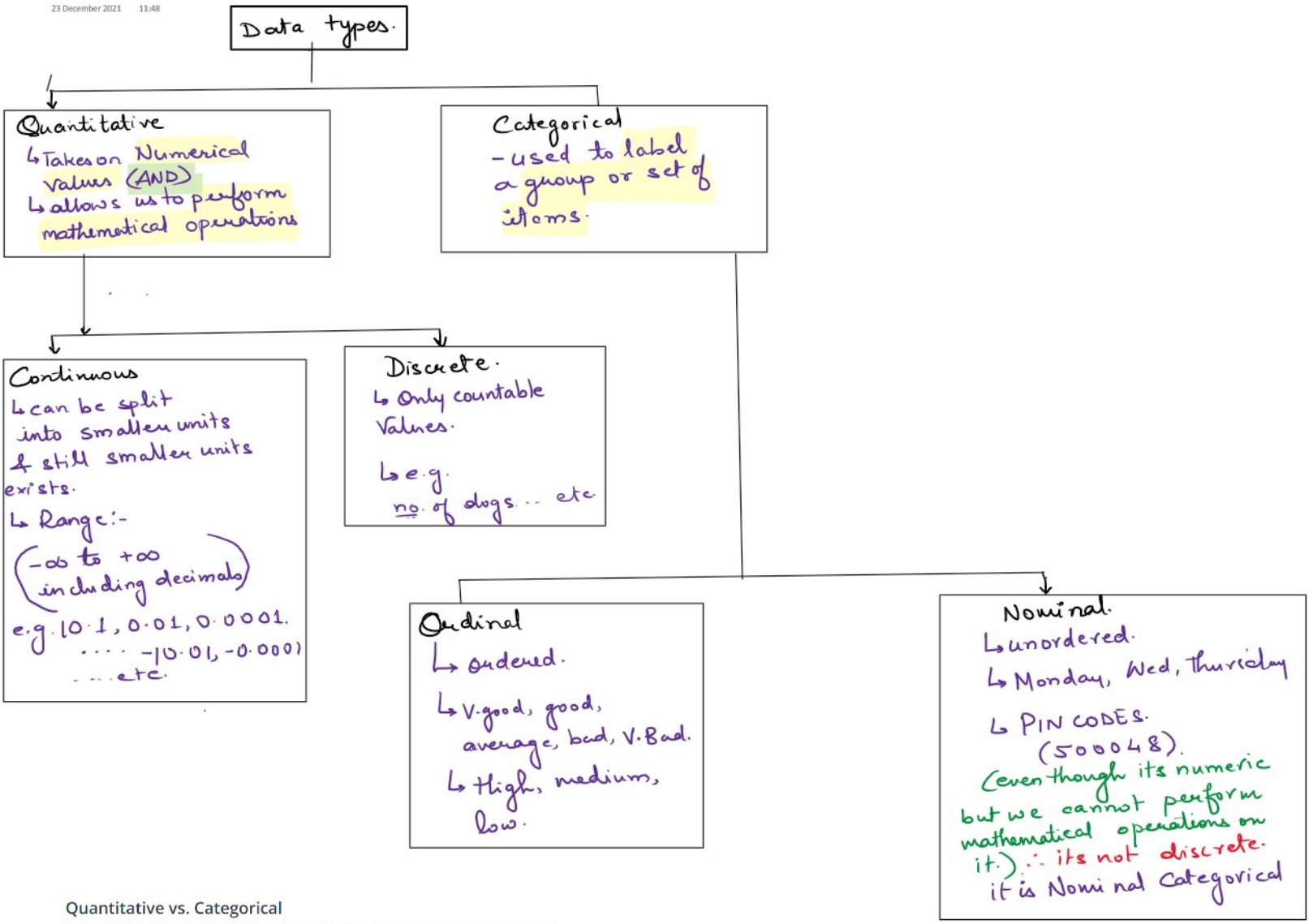
This quiz will ensure you have a clear understanding of the differences between quantitative continuous vs. discrete variables. All of the variables below are quantitative. Your task is to check the box next to each variable that is **continuous**; do not check the discrete variables.

- [ ] Travel Distance from Home to Work
- [ ] Number of Pages in a Book
- [ ] Amount of Rain in a Year
- [ ] Time to Run a Mile
- [ ] Number of Movies Watched in a Week
- [ ] Amount of Water Consumed in a Day
- [ ] Number of Phones per Household

- [x] Travel Distance from Home to Work
- [ ] Number of Pages in a Book
- [x] Amount of Rain in a Year
- [x] Time to Run a Mile
- [ ] Number of Movies Watched in a Week
- [x] Amount of Water Consumed in a Day
- [ ] Number of Phones per Household

Data Types Summary

23 December 2021 11:48



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!

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 check the 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
- [ ] Education (HS, Associates, Bachelors, Masters, PhD, etc.)

- [ ] Letter Grades (A, B+, B, B-, etc.)
- [x] Types of Fruit (Apple, Banana, etc.)
- [ ] Ratings on a Survey (Poor, Ok, Great)
- [x] Types of Dog Breeds (German Shepherd, Collie, etc.)
- [x] Genres of Movies (Horror, Comedy, etc.)
- [x] Gender
- [x] Nationality
- [ ] Education (HS, Associates, Bachelors, Masters, PhD, etc.)

Notation

27 December 2021 08:00

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.**

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.

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.
1. **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. If you really want to understand how this algorithm works, you need to be able to read and understand notation.
1. **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, the notation can convey an idea better than words. Usually, those ideas are related to mathematics, but I am not here to stifle your creativity.

Supporting Materials
Wikipedia on Gradient boosting.

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 multiple 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

27 December 2021 08:08

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 spend 10 minutes, the second spends 20 minutes, the third spend 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.

\bold{x_1}**x1** = 10, \bold{x_2}**x2** = 20 \bold{x_3}**x3** = 45 \bold{x_4}**x4** = 12 \bold{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 \geq≥ 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%**.

From <https://classroom.udacity.com/nanodegrees/nd098-sc-1/parts/7a7defcf-d048-488a-b7d5-3b5c4e62921a/modules/90d0cfd8-ffa9-42ab-b989-7e4a95eb4278/lessons/23759628-47eb-49d0-a7e7-914f1a1f1e01/concepts/143b9451-4e2f-4c44-8e99-77c3860862d2>

Summation

27 December 2021 08:09

Aggregations

An **aggregation** is a way to turn multiple numbers into fewer numbers (commonly one number).

Summation is a common aggregation. The notation used to sum our values is a greek symbol called sigma \SigmaΣ.

Example 1

Imagine we are looking at the amount of time individuals spend on our website. We collect data from nine individuals:

\bold{x_1}**x1** = 10, \bold{x_2}**x2** = 20 \bold{x_3}**x3** = 45 \bold{x_4}**x4** = 12 \bold{x_5}**x5** = 8 \bold{x_6}**x6** = 12, \bold{x_7}**x7** = 3 \bold{x_8}**x8** = 68 \bold{x_9}**x9** = 5

If we want to sum the **first three values** together in our previous notation, we write:

\bold{x_1} + \bold{x_2} + \bold{x_3}**x1**+**x2**+**x3**

In our new notation, we can write:

\sum\limits_{i = 1}^3 x_i*i*=1∑3*xi*.

Notice, our notation starts at the first observation (i=1*i*=1) and ends at 3 (the number at the top of our summation).

So all of the following are equal to one another:

\sum\limits_{i = 1}^3 x_i*i*=1∑3*xi* **=** \bold{x_1} + \bold{x_2} + \bold{x_3}**x1**+**x2**+**x3 = 10 + 20 + 45 = 75**

Example 2

Now, imagine we want to sum the **last three values** together.

\bold{x_7} + \bold{x_8} + \bold{x_9}**x7**+**x8**+**x9**

In our new notation, we can write:

∑{i = 7}^9 x_ii=7∑9$x$$i$.

Notice, our notation starts at the seventh observation (i=7i=7) and ends at 9 (the number at the top of our summation).

Other Aggregations

The \SigmaΣ sign is used for aggregating using summation, but we might choose to aggregate in other ways. Summing is one of the most common ways to need to aggregate. However, we might need to aggregate in alternative ways. If we wanted to multiply all of our values together we would use a product sign \PiΠ , capital Greek letter pi. The way we aggregate continuous values is with something known as integration (a common technique in calculus), which uses the following symbol \int∫ which is just a long s. We will not be using integrals or products for quizzes in this class, but you may see them in the future!

From <https://classroom.udacity.com/nanodegrees/nd098-sc-1/parts/7a7defcf-d048-488a-b7d5-3b5c4e62921a/modules/90d0cfd8-ffa9-42ab-b989-7e4a95eb4278/lessons/23759628-47eb-49d0-a7e7-914f1a1f1e01/concepts/bc8078ce-25aa-4c0f-ab68-e9a77a6bba35>

Notation for mean

27 December 2021 08:11

Final Steps for Calculating the Mean

To finalize our calculation of the mean, we introduce **n** as the total number of values in our dataset. We can use this notation both at the top of our summation, as well as for the value that we divide by when calculating the mean.

\frac{1}{n}\sum\limits_{i=1}^nx_in1*i*=1∑*nxi*

Instead of writing out all of the above, we commonly write \bar{x}*x*¯ to represent the mean of a dataset. Although similar to the first video, we could use any variable. Therefore, we might also write \bar{y}*y*¯, or any other letter.

We also could index using any other letter, not just i*i*. We could just as easily use j*j*, k*k*, or m*m* to index each of our data values. The quizzes on the next concept will help reinforce this idea.

Notice

At second 0:12, this should say \sum\limits_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5*i*=1∑5*xi*=*x*1+*x*2+*x*3+*x*4+*x*5. The x_i*xi* is missing here in front of the summation.

From <https://classroom.udacity.com/nanodegrees/nd098-sc-1/parts/7a7defcf-d048-488a-b7d5-3b5c4e62921a/modules/90d0cfd8-ffa9-42ab-b989-7e4a95eb4278/lessons/23759628-47eb-49d0-a7e7-914f1a1f1e01/concepts/e475fea8-92a4-4c81-a824-90ccf174f3c6>

Summary

27 December 2021 08:12

Notation Recap

Notation is an essential tool for communicating mathematical ideas. We have introduced the fundamentals of notation in this lesson that will allow you to read, write, and communicate with others using your new skills!

Notation and Random Variables

As a quick recap, **capital letters** signify **random variables**. When we look at **individual instances** of a particular random variable, we identify these as **lowercase letters** with subscripts attach themselves to each specific observation.

For example, we might have **X** be the amount of time an individual spends on our website. Our first visitor arrives and spends 10 minutes on our website, and we would say \bold{x_1}**x1** is 10 minutes.

We might imagine the random variables as columns in our dataset, while a particular value would be notated with the lower case letters.

Notation	**English**	**Example**
X	A random variable	Time spent on website
x_1*x*1	First observed value of the random variable X	15 mins
\sum\limits_{i=1}^nx_i*i*=1∑*nxi*	Sum values beginning at the first observation and ending at the last	5 + 2 + ... + 3
\frac{1}{n}\sum\limits_{i=1}^nx_i*n*1*i*=1∑*nxi*	Sum values beginning at the first observation and ending at the last and divide by the number of observations (the mean)	(5 + 2 + 3)/3
\bar{x}*x*¯	Exactly the same as the above - the mean of our data.	(5 + 2 + 3)/3

Notation for the Mean

We took our notation even further by introducing the notation for summation \sum∑. Using this we were able to calculate the mean as:

\bold{\frac{1}{n}\sum\limits_{i=1}^nx_i}**n1i=1∑nxi**

In the next section, you will see this notation used to assist in your understanding of calculating various measures of spread. Notation can take time to fully grasp. Understanding notation not only helps in conveying mathematical ideas but also in writing computer programs - if you decide you want to learn that too! Soon you will analyze data using spreadsheets. When that happens, many of these operations will be hidden by the functions you will be using. But until we get to spreadsheets, it is important to understand how mathematical ideas are commonly communicated. **This isn't easy, but you can do it!**

From <https://classroom.udacity.com/nanodegrees/nd098-sc-1/parts/7a7defcf-d048-488a-b7d5-3b5c4e62921a/modules/90d0cfd8-ffa9-42ab-b989-7e4a95eb4278/lessons/23759628-47eb-49d0-a7e7-914f1a1f1e01/concepts/cbdfe9ac-5695-4fff-9ade-206eeb161165>