Descriptive Statistics

**Instructors Introduction**

This lesson and the next lesson for this course are the beginning of the [Data Foundations Nanodegree Program](https://www.udacity.com/course/data-foundations-nanodegree--nd100). Though this is considered prerequisite knowledge, these lessons are provided as a refresher for some of the ideas you may need to review.

Additionally, you might hear/see information about projects or other items that are not true about the **Data Analyst Nanodegree Program**. Please ignore this information and follow the information that is true to your programs projects and schedule. Cheers!

**What Is Coming Up?**

By first project, I mean **next** project.

**What Is Data?**

**Data Types**

**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.).

**QUESTION 1 OF 2**

Can you identify the data types below as either **quantitative** or **categorical**?

Variable - Datatypes

Zip Code - Categorical

Age - Quantitative

Income - Quantitative

Marital Status (Single, Married, Divorced, etc.) - Categorical

Height - Quantitative

**Thanks for completing that!**

Nice! You know your variables! The zip code is tricky. Even though this is a number, it isn't a number with which we can perform mathematical operations (add, subtract, etc.) and get another value that makes sense. Therefore, we consider it a categorical variable, not quantitative.

**QUESTION 2 OF 2**

Can you identify the data types below as either **quantitative** or **categorical**?

Variable - Datatypes

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

Travel Distance to Work - Quantitative

Ratings on a Survey (Poor, Ok, Great) - Categorical

Temperature - Quantitative

Average Speed - Quantitative

Nice! You got them all!

**Categorical Ordinal & Nominal Data**

**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).

**Data Types Summary**

**Recap**

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 below variables are categorical. It is your job to **check** all of the below which are **nominal**. Do not check the ordinal categorical variables.

Nominal –

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

That's right! Remember ordinal variables have a rank ordering associated with each. Nominal variables could be placed in any order.

**C**

**QUIZ QUESTION**

This quiz will assure you have a clear understanding of the differences between quantitative continuous vs. discrete variables. All of the below variables are quantitative. It is your job to **check** all of the below which are **continuous**. Do not check the discrete variables.

* Travel Distance from Home to Work
* Amount of Rain in a Year
* Time to Run a Mile
* Amount of Water Consumed in a Day

Nice! You know your data types! Continuous data types are those that can take on decimal values, where discrete data types are those that are countable.

**Introduction To Summary Statistics**

**Calculating The Mean**

**Analyzing Quantitative Data**

**Four Aspects for Quantitative Data**

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

* Measures of Center
* Measures of Spread
* The Shape of the data.
* 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.

**QUESTION 1 OF 2**

Which of the below are measures of center (Check all that apply)?

ANSWER:

Mean, Median, Mode

That's right! There are 3 M's to our measures of center - means, medians, and modes.

**QUESTION 2 OF 2**

If we have the data:

5, 8, 15, 7, 10, 22, 3, 1, 15

What is the mean?

Ans: 9.56

Nice! That is the mean!

**The Median**

**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!

**QUESTION 1 OF 2**

If we have the data:

5, 8, 15, 7, 10, 22, 3, 1, 15

What is the median?

Ans:

1, 3, 5, 7, 8, 10, 15, 15, 22

Hence, 8

Nice! That's right, the median is the middle number when our data are ordered.

**QUESTION 2 OF 2**

If we have the data:

5, 8, 15, 7, 10, 22, 3, 1, 15, 2

What is the median?

Ans:

1,2,3,5, 7, 8, 10, 15, 15, 22

Hence, 7.5

Nice! That's right, when we have an even number of values, we need to average the two values in the middle together.

**Measures Of Center - The Mode**

**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.

**QUESTION 1 OF 4**

Check all of the below that are true with regards to our measures of center.

* The mode is the middle number in the dataset when the numbers are rank ordered.
* The median is the middle number in the dataset when the numbers are rank ordered.    ANSWER
* The mean is always the best measure of center for any dataset.
* The mean is always less than the median.
* The median is always the best measure of center for any dataset.
* The mode is always the best measure of center for any dataset.

Nice! These are tricky, and you got them all!

**QUESTION 2 OF 4**

If we have the data:

5, 8, 15, 7, 10, 22, 3, 1, 15

What is the mode?

Ans:

1, 3, 5, 7, 8, 10, 15, 15, 22

Hence, 15

Nice! You know your measures of center! The mode is the most frequent value in our dataset.

**QUESTION 3 OF 4**

For the dataset below match the correct measure to the value:

8, 12, 32, 10, 3, 4, 4, 4, 4, 5, 12, 20

Ans:

3, 4, 4, 4, 4, 5, 8, 10, 12, 12, 20, 32

Measure – Value

Mean – 9.83

Median – 6.5

Mode – 4

None of the Above – 8

Nice! You know your measures of center!

**CONTINUE**

**QUESTION 4 OF 4**

If we have the data:

5, 8, 15, 7, 10, 22, 3, 1, 15, 10

Mark all statements that are true.

Ans:

1,3, 5, 7,8, 10, 10, 15, 15, 22

Mean = 9.6

Median = 9

Mode = 10,15

Hence,

* The mode is 15.       True
* The mean is 15.       False
* The mode is 10.       True
* None of the above are true.     False

Nice! That's right there are two modes in this dataset. If all the values appear the same number of times, we usually say there is no mode. However, if more than one value appears the most frequent number of times, we count all of these values as modes.

**What Is Notation?**

**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:

* **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.
* **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.
* **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.
* **Notation For Random Variables**
* **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.

**Example Dataset**

An example of the data we might have collected in the previous video is shown here:

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

**QUESTION 1 OF 2**

What type of variable is the random variable X in the video in the previous concept?

Quantitative – Continuous

You got it! Still keeping those data types straight!

**QUESTION 2 OF 2**

What type of variable is the random variable Y in the video in the previous concept?

Categorical – Nominal

That's right! Whether or not an individual 'buys' is a category without order involved.

**Random & Observed Values**

**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.

**x**

**1**

= 10,       **x**

**2**

= 20       **x**

**3**

= 45       **x**

**4**

= 12       **x**

**5**

= 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%**.

Consider we have the following table:

|  |  |  |
| --- | --- | --- |
| **Years Experience** | **Department** | **Part/Full Time** |
| 5 | IT | Part Time |
| 10 | Finance | Full Time |
| 8 | HR | Full Time |
| 1 | Finance | Part Time |

Consider we have the following labels:

**X**= years of experience

**Y**= Department

**Z**= Part/Full Time

Match the following notation to their corresponding:

A. **x**

**1**

B. **y**

**2**

C. **z**

**3**

D. **n**

**NOTATION - VALUE**

A. - 5

B. - Finance

C. - Full Time

D. - 4

Nice! Looks like you know your notation!

**There Must Be A Better Way**

**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:

**x**

**1**

+**x**

**2**

+**x**

**3**

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

**x**

**1**

+**x**

**2**

+**x**

**3**

+**x**

**4**

+**x**

**5**

+**x**

**6**

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

**Aggregations**

**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 Σ.

**Example 1**

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

**x1**= 10,… **x9**= 5

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

**x1**+**x2**+**x3**

In our new notation, we can write:

*i*=1∑3*xi*.

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

So all of the following are equal to one another:

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

**Other Aggregations**

The Σ 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 Π, 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 ∫ 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!

**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.

*n*

1

*i*=1

∑

*n*

*x*

*i*

Instead of writing out all of the above, we commonly write

*x*

¯

to represent the mean of a dataset. Although, similar to the first video, we could use any variable. Therefore, we might also write

*y*

¯

, or any other letter.

We also could index using any other letter, not just *i*. We could just as easily use *j*, *k*, or *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

*i*=1∑5*xi*=*x*1+*x*2+*x*3+*x*4+*x*5.

The *xi*is missing here in front of the summation.

**Match The Notation**

For this quiz, you will be matching the notation attached the letters below to the corresponding numeric value to make sure you understand exactly what is being done with each part of the notation.

Imagine, we have the following table of values:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **x**  **1** | **x**  **2** | **x**  **3** | **x**  **4** | **x**  **5** | **x**  **6** | **x**  **7** |
| 5 | 15 | 3 | 3 | 8 | 10 | 12 |

A. **n**

B.

**i**=**1**∑**nxi**

C.

**j**=**2**∑**7xj**+**6**

D. **x5**

E.

**n**−**1i**=**3**∑**6xi**

**QUIZ QUESTION**

Use the letters, numbers, and notation as defined above to match each letter to the appropriate value.

ANSWER:

7, 56, 57, 8, 4 (🡪 3.5 rounded off to 4)

Wow! Spectacular! These were tough, and you got them all!

**Notation for Quizzes**

For the below quiz, let the following letters denote the corresponding notation:

A. **X**

B. **Y**

C. **x1**

D. **n**

E.

**i**=**1**

∑**xi**

**n**

**QUESTION 1 OF 2**

Use the letter next to the notation above to match the notation to the description of what the notation represents.

**DESCRIPTION**

A- The notation for a random variable.

B- The notation for a random variable.

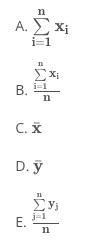
C- The notation for the first observed value of a random variable.

D- The notation for the number of rows in our dataset.

E- The notation for the sum of all the values in our dataset.

**Notation for Quizzes**

For the below quiz, let the following letters denote the corresponding notation:



Answers:

B,C,D,E

That's right! The letter we use to denote the dataset isn't so important - we could use either 'x' or 'y'. Therefore, we might choose any of the last four options as the notation for the mean of our dataset.

**Notation Recap**

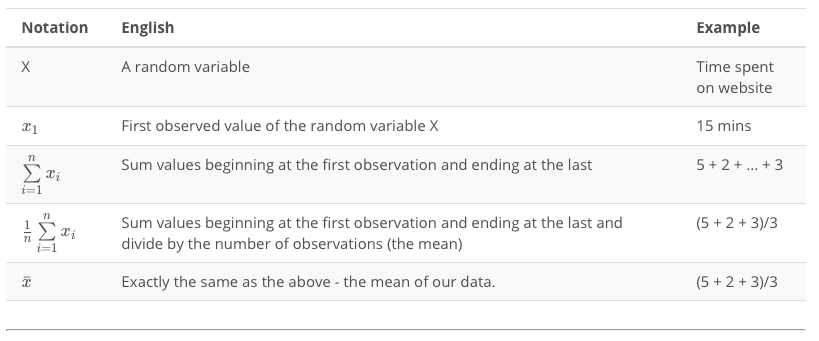
Notation is an essential tool for communicating mathematical ideas. We have introduced the fundamentals of notation in this lesson that will allow you to read, write, and communicate with others using your new skills!

**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 **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 for the Mean**

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

**i**=**1**

(1/n)∑**xi**

**n**

In the next lesson, 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!**

**NEXT**