Roles on a Data Science Team:

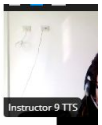
Data Scientist
Data Engineer
Machine LEarning Engineer
Data Architect
Business Analyst/ Domain Experts
Software Engineer
Data Analyst
Data Security Engineer
Data Science Research

Python vs R

Python is a high level programming language:

- Games
- Data analysis
- Machine learning
- Natural language processing
- Deep learning
- Web development
- Data scraping
- Data visualizations

What is Python?



- **Python** is a high level programming language
  - Python is made from C (programming language)
  - C is the original programming language
  - Considered the most popular programming language in the world
  - Considered to be amongst the easiest language when it comes to syntax
  - Python was considered best Data Science platform
  - The issue is when you have to use truly Big Data (then came **Spark**)
    - Big Data = 1TB or more

-R is considered a S Programming Language (Statistical)

-This is considered to be the most used Language in Data Science

-This is because Python is much more Robust, R is a specialty Language

-R is specifically used for Data Science and other program languages are not

-R is heavily used in Finance and the Healthcare industry because of its ability to do what Statisticians once had to do by Hand!

Intro to Data Science

Where Does Data Come From?

- Functional Area Support Systems
- Corporate Databases
- Government Websites
- Commericial Providers
- Academic/Research Institutions
- Myriad Electronic Devices (IoT)
- DIY

TECH TALENT SOUTH

-Digital (websites, apps, instant messages, email, voicemail transcriptions)

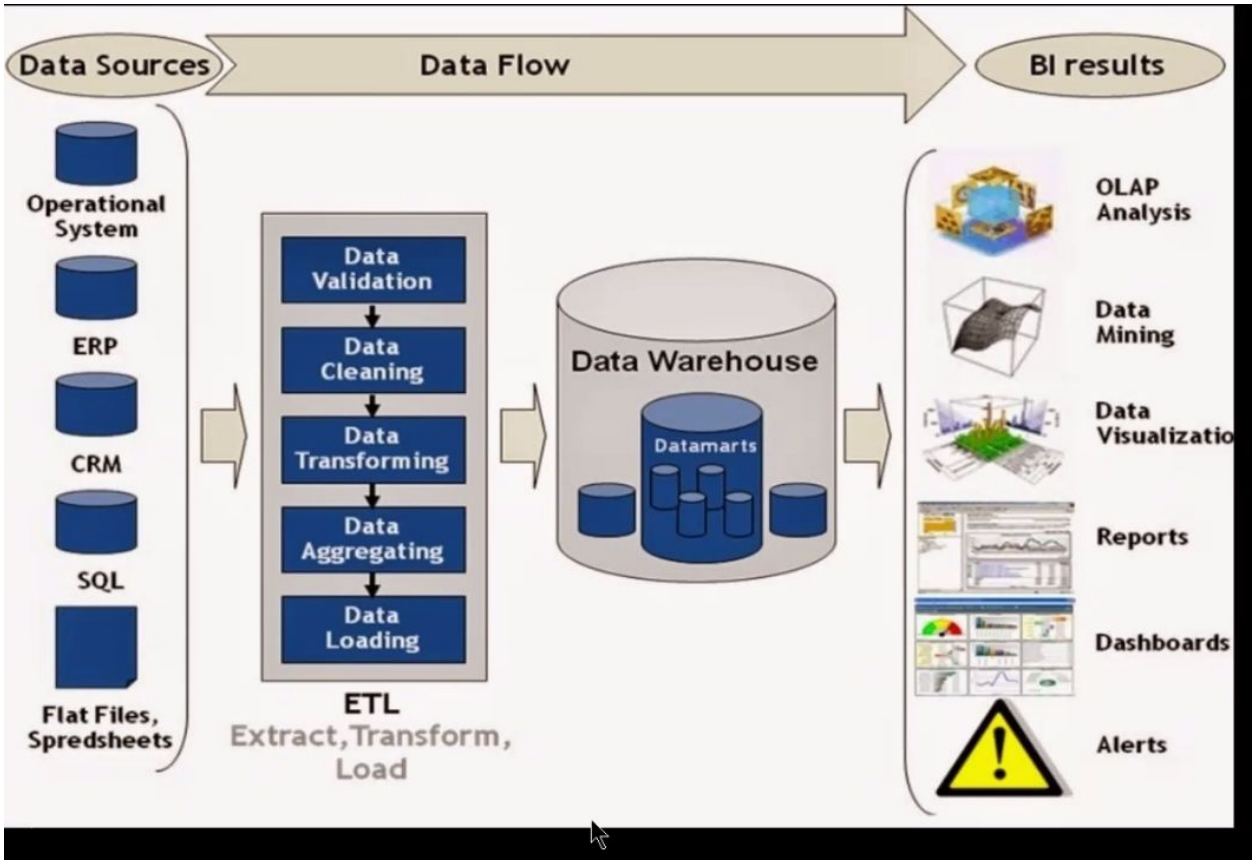
-Physical (GeoLocations, sales transactions, traffic monitors)

What is data science: the expert study of data using statistics.
What is a data scientist an expert in the study of data science

Big Data Conversion Table

1. 1024 MegaBytes = 1 Gigabyte
2. 1024 Gigabytes = 1 Terabyte
3. 1024 Terabytes = 1 Petabyte*
4. 1024 Petabytes = 1 Exabyte*
5. 1024 Exabytes = 1 Zettabyte
6. 1024 Zettabytes = 1 Yottabyte

*Most large scale companies will use Terabytes and Petabytes



Semi-structured Data

```
3,?,alfa-romero,gas,std,two,convertible,rwd,front,88.60,168.80,64.10,48.80,2548,dohc,four,130,mpfi,3.47,2.68,9.00,111,5000,21,27,13495
3,?,alfa-romero,gas,std,two,convertible,rwd,front,88.60,168.80,64.10,48.80,2548,dohc,four,130,mpfi,3.47,2.68,9.00,111,5000,21,27,16500
1,?,alfa-romero,gas,std,two,hatchback,rwd,front,94.50,171.20,65.50,52.40,2823,ohcv,six,152,mpfi,2.68,3.47,9.00,154,5000,19,26,16500
2,164,audi,gas,std,four,sedan,fwd,front,99.80,176.60,66.20,54.30,2337,ohc,four,109,mpfi,3.19,3.40,10.00,102,5500,24,30,13950
2,164,audi,gas,std,four,sedan,4wd,front,99.40,176.60,66.40,54.30,2824,ohc,five,136,mpfi,3.19,3.40,8.00,115,5500,18,22,17450
2,?,audi,gas,std,two,sedan,fwd,front,99.80,177.30,66.30,53.10,2507,ohc,five,136,mpfi,3.19,3.40,8.50,110,5500,19,25,15250
1,158,audi,gas,std,four,sedan,fwd,front,105.80,192.70,71.40,55.70,2844,ohc,five,136,mpfi,3.19,3.40,8.50,110,5500,19,25,17710
1,?,audi,gas,std,four,wagon,fwd,front,105.80,192.70,71.40,55.70,2954,ohc,five,136,mpfi,3.19,3.40,8.50,110,5500,19,25,18920
1,158,audi,gas,turbo,four,sedan,fwd,front,105.80,192.70,71.40,55.90,3086,ohc,five,131,mpfi,3.13,3.40,8.30,140,5500,17,20,23875
0,?,audi,gas,turbo,two,hatchback,4wd,front,99.50,178.20,67.90,52.00,3053,ohc,five,131,mpfi,3.13,3.40,7.00,160,5500,16,22,?
2,192,bmw,gas,std,two,sedan,rwd,front,101.20,176.80,64.80,54.30,2395,ohc,four,108,mpfi,3.50,2.80,8.80,101,5800,23,29,16430
0,192,bmw,gas,std,four,sedan,rwd,front,101.20,176.80,64.80,54.30,2395,ohc,four,108,mpfi,3.50,2.80,8.80,101,5800,23,29,16925
0,188,bmw,gas,std,two,sedan,rwd,front,101.20,176.80,64.80,54.30,2710,ohc,six,164,mpfi,3.31,3.19,9.00,121,4250,21,28,20970
0,188,bmw,gas,std,four,sedan,rwd,front,101.20,176.80,64.80,54.30,2765,ohc,six,164,mpfi,3.31,3.19,9.00,121,4250,21,28,21105
1,?,bmw,gas,std,four,sedan,rwd,front,103.50,189.00,66.90,55.70,3055,ohc,six,164,mpfi,3.31,3.19,9.00,121,4250,20,25,24565
0,?,bmw,gas,std,four,sedan,rwd,front,103.50,189.00,66.90,55.70,3230,ohc,six,209,mpfi,3.62,3.39,8.00,182,5400,16,22,30760
0,?,bmw,gas,std,two,sedan,rwd,front,103.50,193.80,67.90,53.70,3380,ohc,six,209,mpfi,3.62,3.39,8.00,182,5400,16,22,41315
0,?,bmw,gas,std,four,sedan,rwd,front,110.00,197.00,70.90,56.30,3505,ohc,six,209,mpfi,3.62,3.39,8.00,182,5400,15,20,36880
2,121,chevrolet,gas,std,two,hatchback,fwd,front,88.40,141.10,60.30,53.20,1488,l,three,61,2bbl,2.91,3.03,9.50,48,5100,47,53,5151
1,98,chevrolet,gas,std,two,hatchback,fwd,front,94.50,155.90,63.60,52.00,1874,ohc,four,90,2bbl,3.03,3.11,9.60,70,5400,38,43,6295
```

What is wrong with the above: the ?

Structured Data

symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system
2	164.0	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi
2	164.0	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi
1	158.0	audi	gas	std	four	sedan	fwd	front	105.8	...	136	mpfi
1	158.0	audi	gas	turbo	four	sedan	fwd	front	105.8	...	131	mpfi
2	192.0	bmw	gas	std	two	sedan	rwd	front	101.2	...	108	mpfi
0	192.0	bmw	gas	std	four	sedan	rwd	front	101.2	...	108	mpfi
0	188.0	bmw	gas	std	two	sedan	rwd	front	101.2	...	164	mpfi
0	188.0	bmw	gas	std	four	sedan	rwd	front	101.2	...	164	mpfi

Column names are called features in data science.
Rows in data scientist area called attributes.

How is Big Data Stored

1. Database - A database is an organized collection of data, generally stored and accessed electronically from a computer system.

1. Data Lake - Stores all data (structured, semi-structured, and unstructured). With a data lake, you just load in the raw data, as-is, and then when you're ready to use the data, that's when you give it shape and structure. That's called schema-on-read.

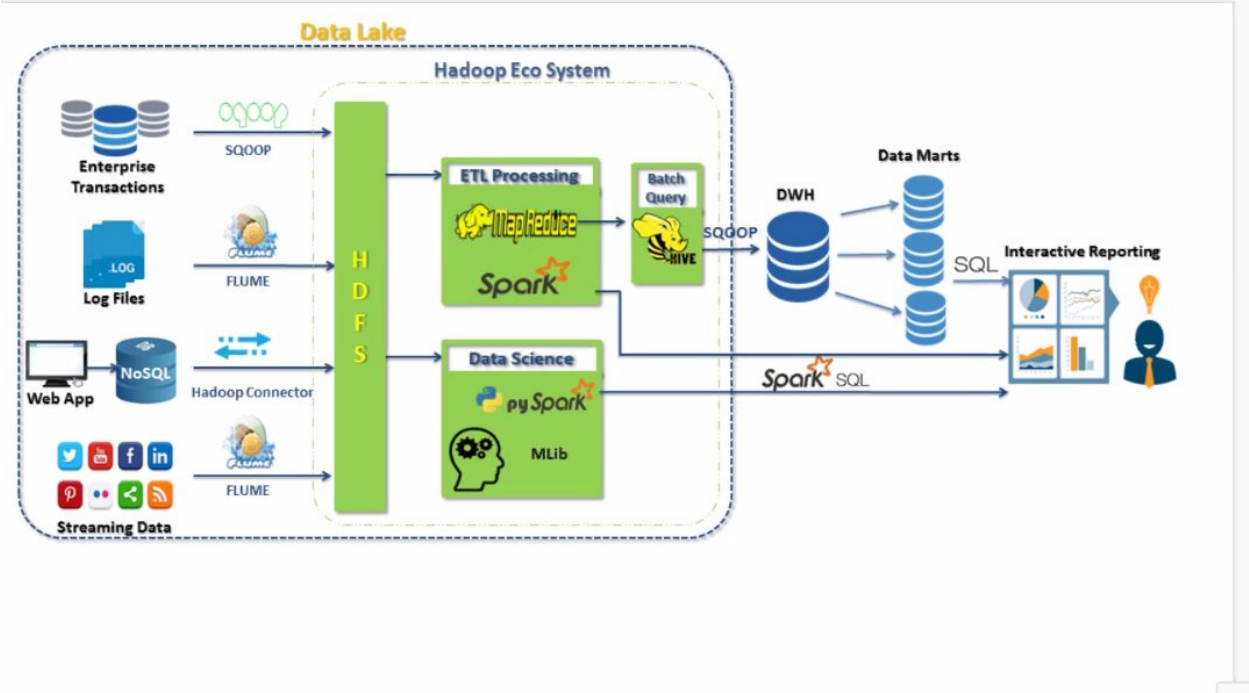
1. Data Warehouse - A data warehouse only stores data that has been modeled/structured (structured data). Before we can load data into a data warehouse, we first need to give it some shape and structure—i.e., we need to model it. That's called schema-on-write.

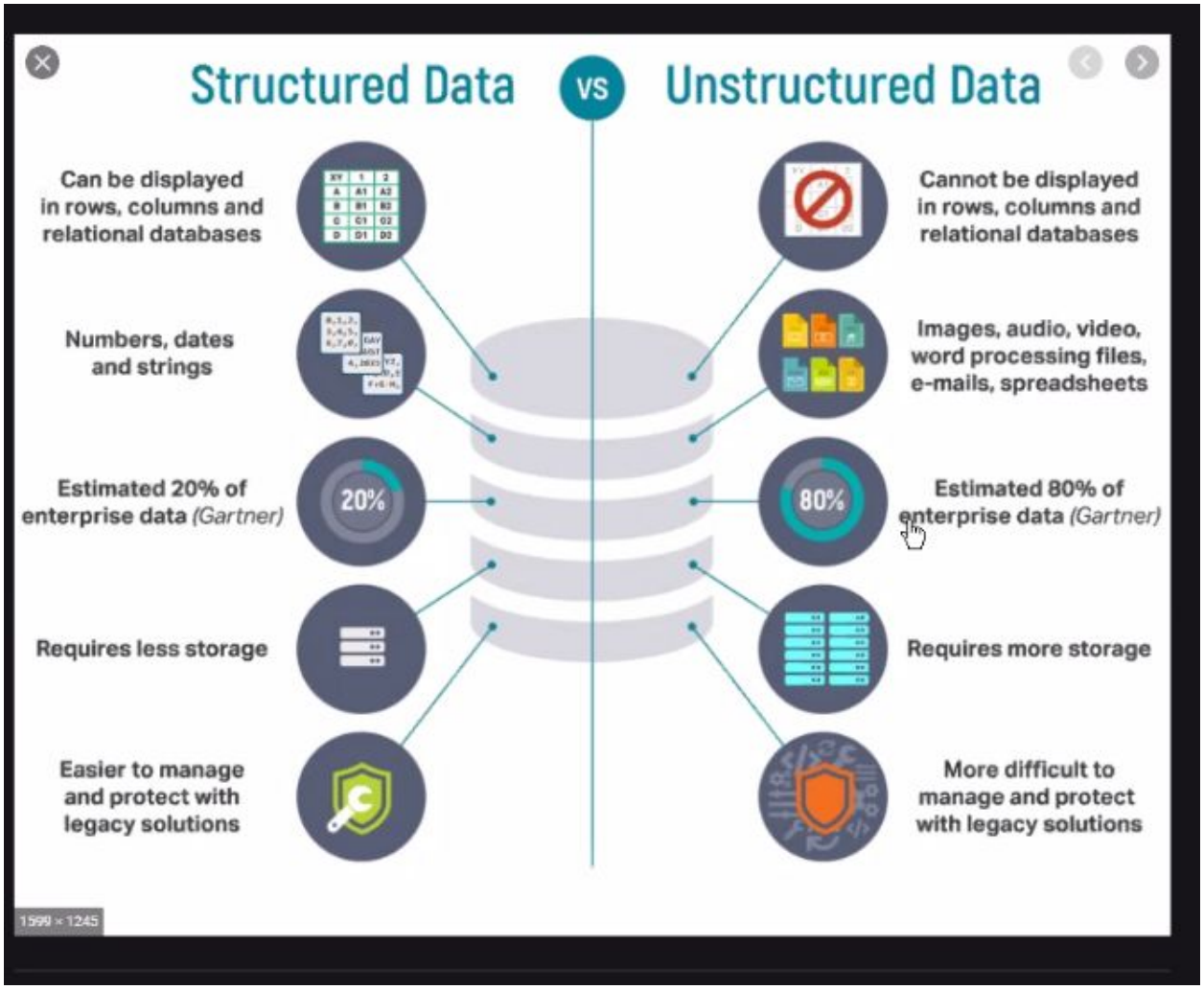
4. Data Mart - A subset of a Data Warehouse

Intro to Data Science

Data Formats

- Flat Files - text-based databases (e.g. csv, tab delimited, JSON, etc.)
- XML Files
- Relational Data
- Unstructured Data




Structured Data
vs
Unstructured Data
Can be displayed in rows, columns and relational databases
Cannot be displayed in rows, columns and relational databases
Numbers, dates and strings
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 20% of enterprise data (Gartner)
20%
80%
Estimated 80% of enterprise data (Gartner)
Requires less storage
Requires more storage
Easier to manage and protect with legacy solutions
More difficult to manage and protect with legacy solutions
1599 × 1245

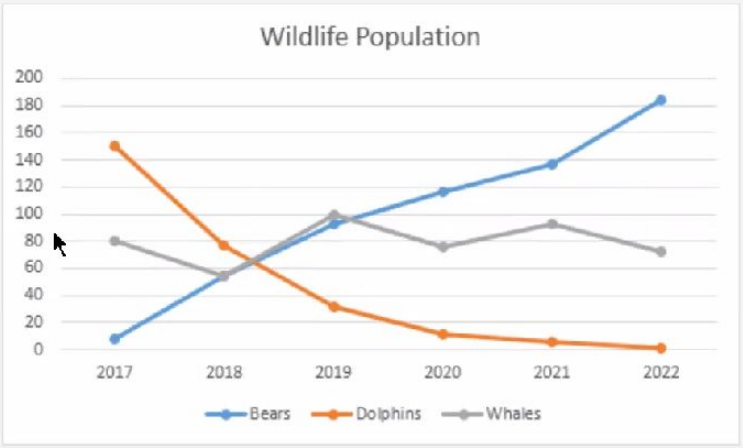
Intro to Data Science

Line Chart

Line charts show continuous data over time on an evenly scaled axis. They are ideal for showing trends in data at equal intervals, such as months, quarters,or years.

In a line chart, category data is distributed evenly along the horizontal axis and value data is distributed evenly along the vertical axis.

Intro to Data Science



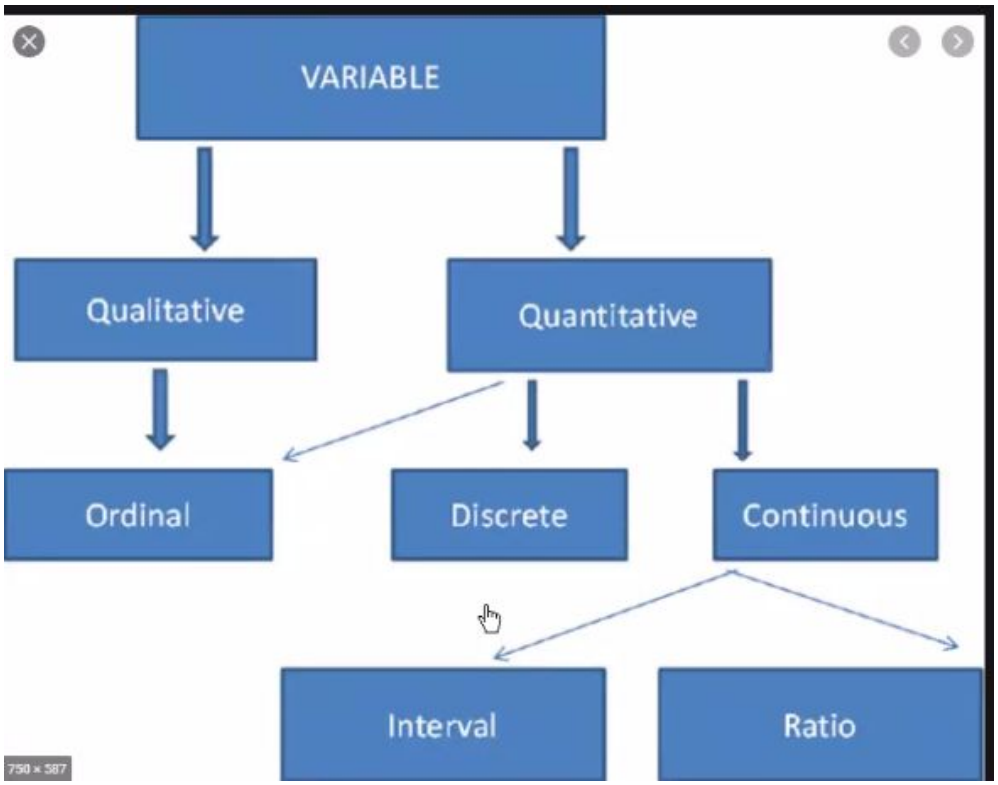
TECH TALENT SOUTH

X variable vs y variable.
The x variable is always horizontal and the y variable is always vertical.
X variable = independent variable
Y variable = dependent variable

The X variable will determine your y variable.

Continuous data is just numerical data.
He said there are two type of data: numerical or categorical
Quantitative data is just numerical

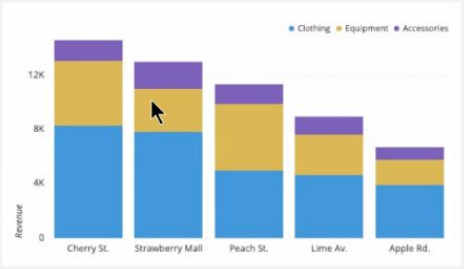
Categorical = Qualitative
Quantitative = Parametric = Numeric
qualitative = non parametric = categorical



Intro to Data Science

Bar Chart

Bar charts illustrate comparisons among individual items. The bar chart has a few sub-types, including clustered bar and stacked bar.



Stacked Bar Chart

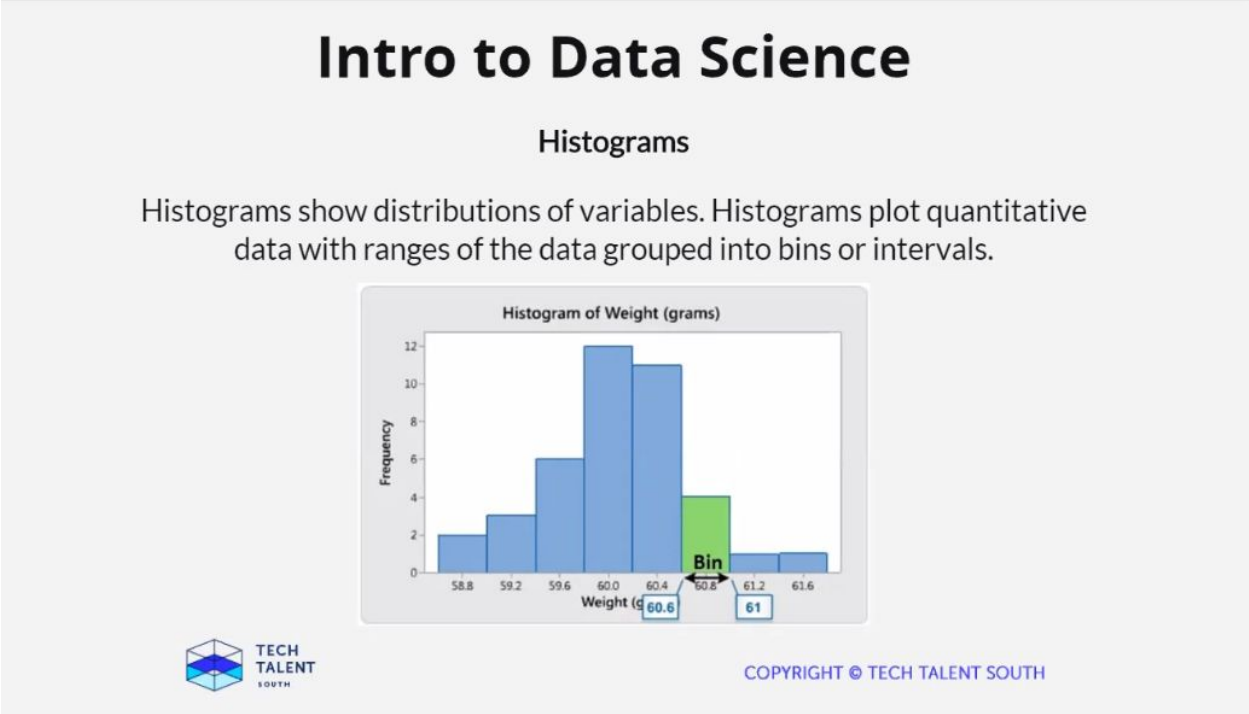
TECH TALENT SOUTH

Intro to Data Science

Area Chart

An area chart is a line chart with the areas below the lines filled in. They can be used to plot change over time and draw attention to the total value across a trend. By showing the sum of the plotted values, an area chart also shows the relationship of parts to a whole.

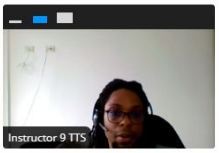
TECH TALENT SOUTH



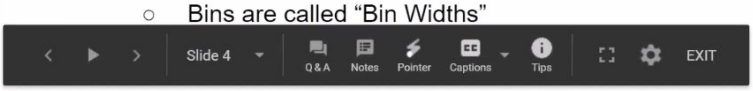
A histogram displays numerical data by grouping data into "bins" of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in that bin.

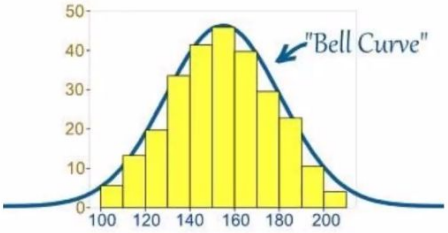
Bins are also sometimes called "intervals", "classes", or "buckets

What is Normal Distribution?

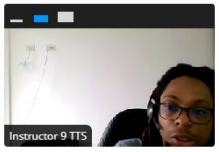


- Data that is Normally Distributed
- This tell us most values are near the mean
- **When data is normally distributed this means the Mean and Median are equal**
  - Median is simply the middle number
    - 1,2,3,4,5 (Median is 3)
    - 1,2,3,4,5,6 (Median is 3+4 divided by 2 = 3.5)
- When data is not normally distributed that means a bias to the left or right exists
  - Represented by a Standard Bell Curve'
  - The yellow marks represent bins
  - Bins are called "Bin Widths"

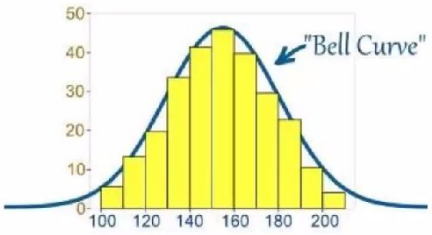




What is Normal Distribution?



- When data is not normally distributed that means a bias to the left or right exists
  - Represented by a Standard Bell Curve'
  - The yellow marks represent bins
  - Bins are called "Bin Widths"
- **Bin Widths represent data represented in groups instead of individual data points**
- In this example it is represented by 20s but you can adjust this when plotting to 10 and etc.
- The total **area under** any **normal curve** is SD of 1
  - Tells us most of the data is within 1 standard deviation



What is Standard Deviation?

- **A number that represents how one group differs from the mean value of entire group or data set**

- **How is Standard Deviation Calculated (when we don't have a set number to calculate a STDeviation)?**
  - What is 68-95-99.7 Rule?
    - **The Percentages or Standard Deviation from the mean**
    - **68% of values are within 1 Standard Deviations**
    - **95% of values are within 2 Standard Deviations**
    - **99.7% of values are within 3 Standard Deviations**
    - **Example (If Mean = 100, Standard Deviation = 15)**
      - **1 Standard Deviation = 85 and 115 = 68%**
      - **2 Standard Deviations = 70 and 130 = 95%**
      - **3 Standard Deviations = 55 and 145 = 99.7%**

The Rules of Standard Deviation?

- A Standard Deviation over 1 is considered high
- A Standard Deviation from 0 to 1 considered normal
- A Standard Deviation of 3 means that 99.7% of people are not Millionaires
- In General the Standard Deviation should not be greater than the mean
- If it is greater this means the data is skewed

EXAMPLE:

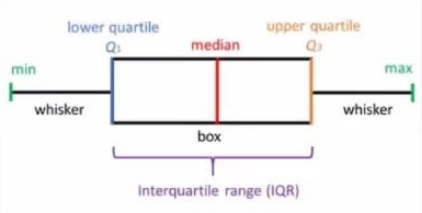
Lets say 5 unemployed women are in a survey about their net worth. One of these people doing the Survey are Michelle Obama.

The mean net worth of the 5 women would be highly skewed and represent a high Standard Deviation!

Intro to Data Science

Box and Whisker

A box and whisker plot displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.



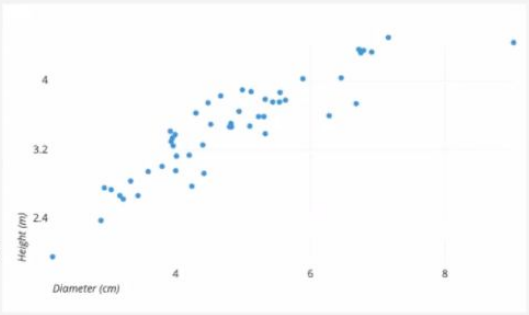
TECH TALENT SOUTH

Good for survey data helps find abnormal data

Intro to Data Science

Scatter Plot

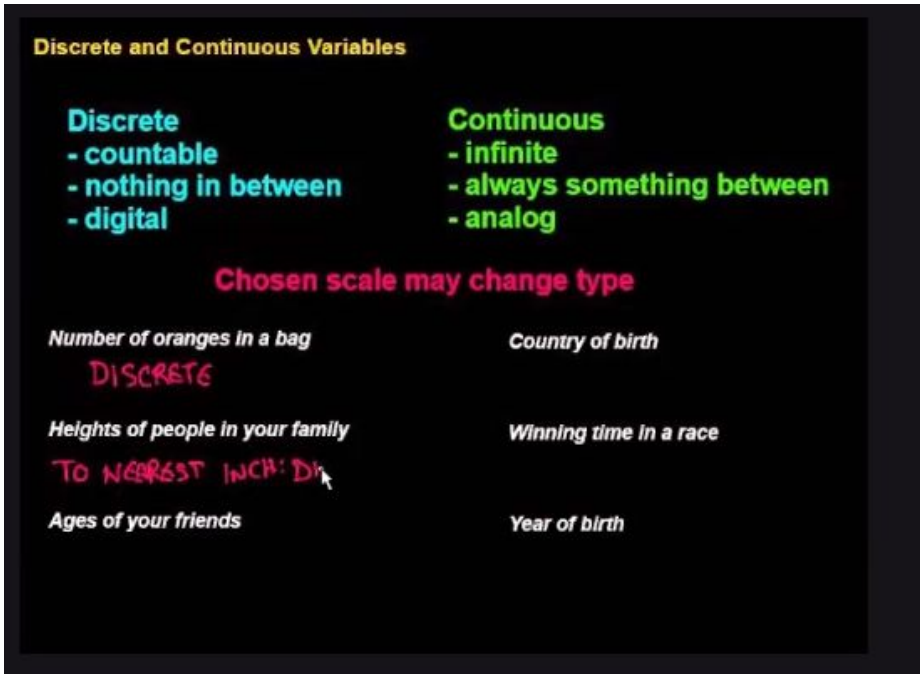
A scatter plot shows scientific XY data. Scatter plots are often used to find out if there is a relationship (correlation) between variable X and Y.



TECH TALENT SOUTH

Table of Contents

Discrete is a number that can only take a certain value. Or a whole number



Intro to Data Science

Discrete Data

Discrete data, can only take in certain values. Usually, these are values that can be counted, like the number of students in a class. In this case, we could have any whole number of students. (We couldn't have half a student!)

Intro to Data Science

Discrete Data

Counts - variables representing frequency of an occurence of an event

- Number of people in a school
- Number of people who voted on a bill

Proportions - also known as bounded counts are the ratios of counts

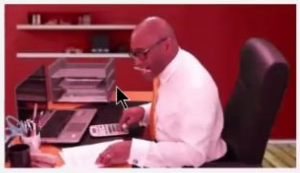
- Number of students in a school divided by the number of teachers in a school
- Number of people who voted "Yes" on a bill

Intro to Data Science

Continuous Data

Continuous data is an unfixed number of possible measurements between two realistic points. The data can be any number is not restricted like discrete data is.

Continous data often contain decimal points and can provide great detail. It is also usually contains numbers within an expected range.



TECH TALENT SOUTH

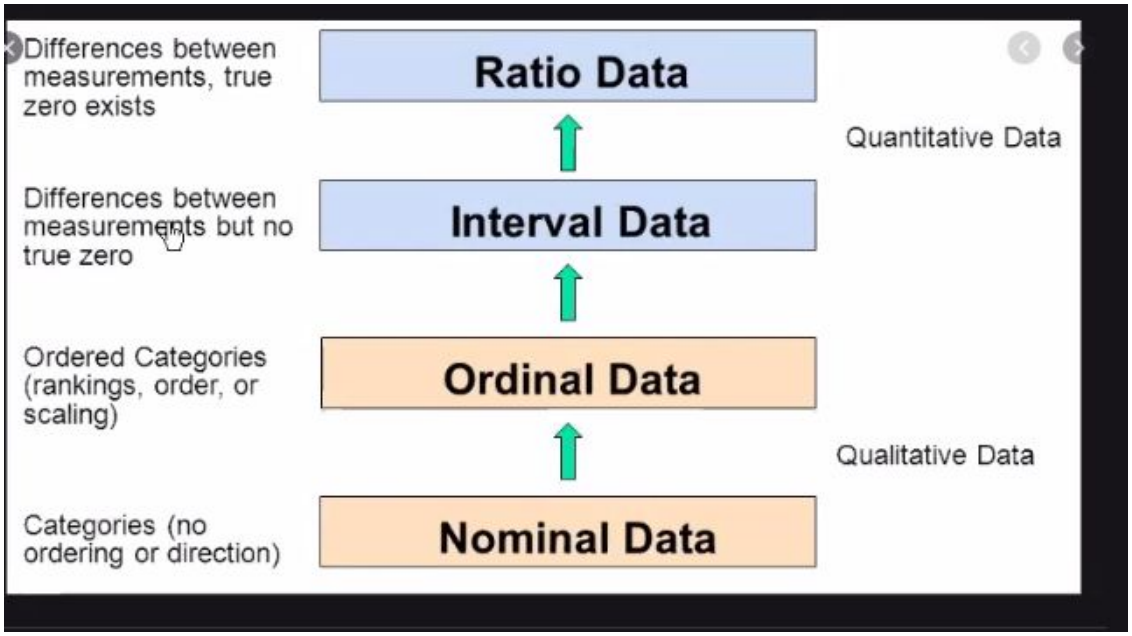
Internal has a stopping point. Temperature, SAT scores, grades etc

Continuous Data

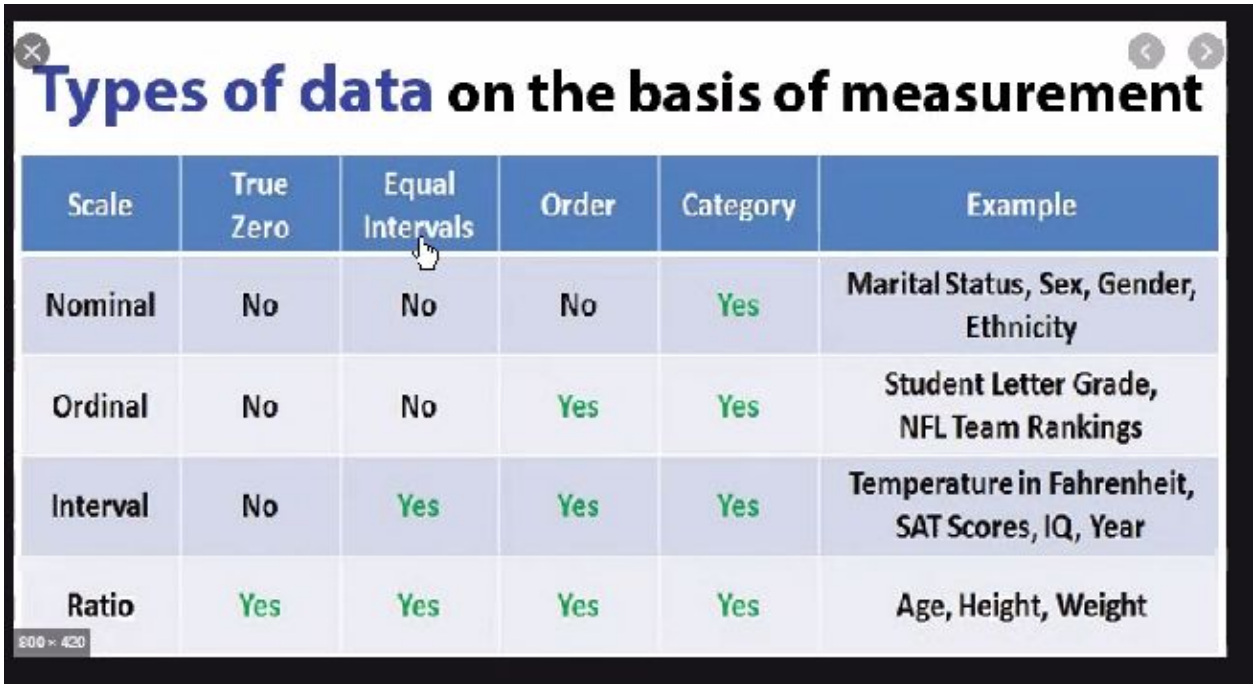
Possible examples include:

- A person's height
- A person's weight
- The temperature
- Inches of rain

Continuous variables have categories within them. Within in continuous variables there are two categories: interval and ratio



Ordinal means rankings



Types of data on the basis of measurement

Scale	True Zero	Equal Intervals	Order	Category	Example
Nominal	No	No	No	Yes	Marital Status, Sex, Gender, Ethnicity
Ordinal	No	No	Yes	Yes	Student Letter Grade, NFL Team Rankings
Interval	No	Yes	Yes	Yes	Temperature in Fahrenheit, SAT Scores, IQ, Year
Ratio	Yes	Yes	Yes	Yes	Age, Height, Weight

Intro to Data Science

Interval Data

Ordered units with the same difference. For example, describing the temperature from a list of options such as:

-10, -5, 0, 5, 10

Interval data does not have a "true zero." (In the above example, there is no option for "no temperature.")

We can add and subtract, but cannot multiply or divide to calculate ratios.

Intro to Data Science

Discrete Data

Discrete data can be measured in different ways: ordered or unordered.

- Nominal Variables (Unordered): gender, location, religion, etc.
- Ordinal (ordered) variables: grade levels, income brackets
- Continuous variables: grouped into a small number of categories (intervals) - income grouped into subsets, blood pressure levels (normal, high-normal etc)

Intro to Data Science

Measurement Hierarchy

nominal < ordinal < interval

Methods applicable to a lower type of variable can be used for a higher one, but not the other way around.

For example, you could use methods designed for nominal data for interval data, but not methods designed for interval data with nominal data.

Intro to Data Science

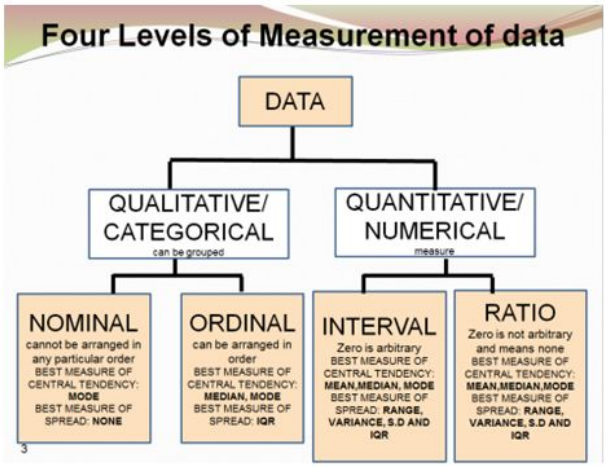
Measurement Hierarchy

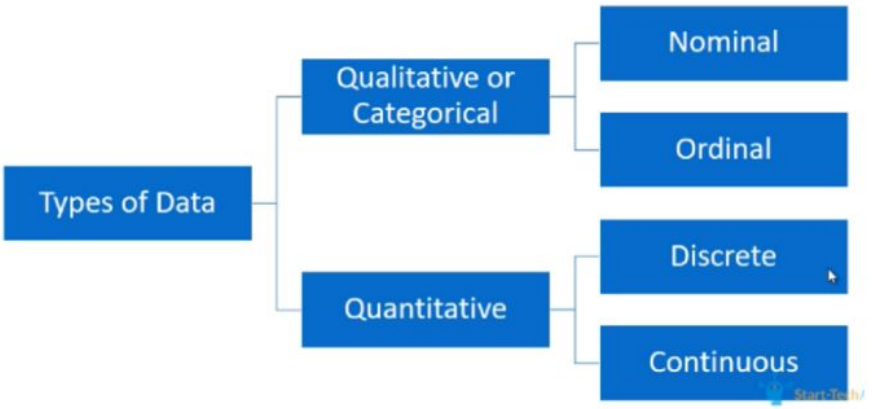
What types of variables could be used to answer the following question?

Have you studied abroad?
What is your interest level in data science: low, medium, or high?
Based on your test score, what letter grade did you get?

1. Categorial
2. Ordinal
3. Ordinal

https://medium.com/@rndayala/data-levels-of-measurement-4af33d9ab51a





Good sources:

ResearchGate - Researchgate.net
Medium
Scikit-learn

Intro to Data Science

Discuss

Why does the measurement hierarchy matter, and how does it affect data analysis?

What are some uses for discrete data that you can think of at this point?

Intro to Data Science

Contingency Tables

Used to summarize discrete data. Contains at least two categories and data to compare the results of the counts.

Preference: Dogs or Cats

	Male	Female	Total
Dog	20	15	35
Cat	12	15	27
Total	32	30	62

Intro to Data Science

Qualitative Data

Qualitative data is used to categorize. It is not numerical in nature.

This includes data from interviews, focus groups, and observational studies.

Even though it is does not provide concrete numerical information, it can still be very useful.

Intro to Data Science

Qualitative Data

Qualitative data is used to categorize. It is not numerical in nature.

This includes data from interviews, focus groups, and observational studies.

Even though it is does not provide concrete numerical information, it can still be very useful.

Intro to Data Science

Qualitative Data

There are many ways to gather qualitative data, including:

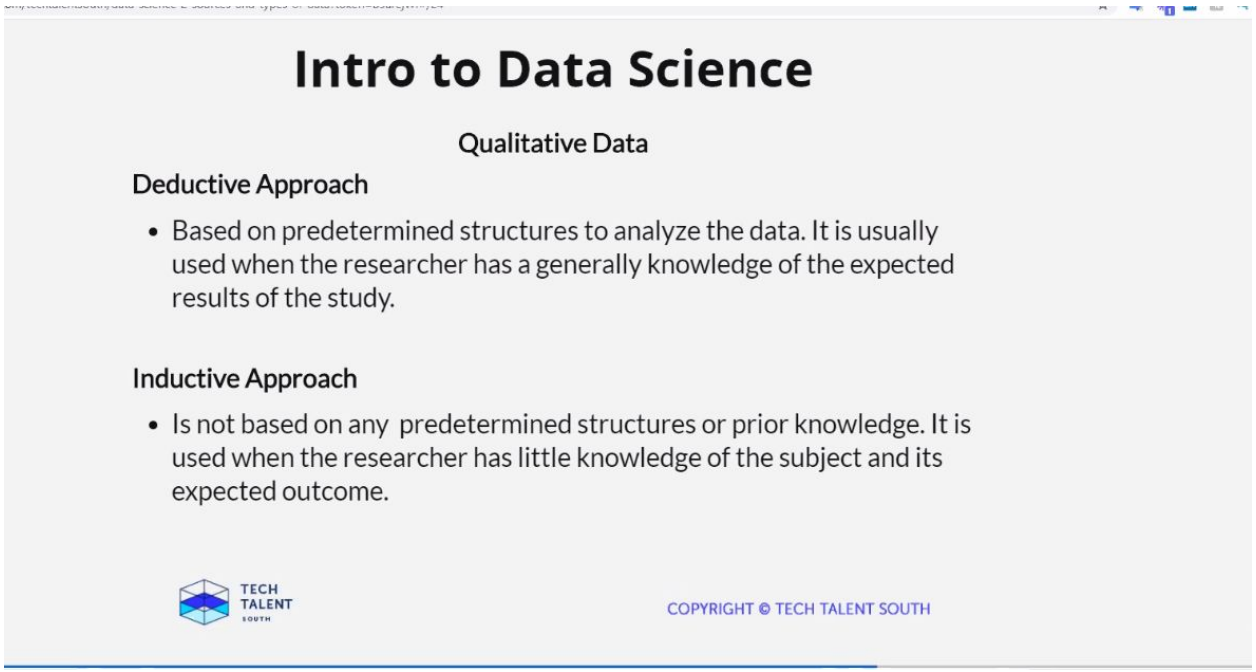
1. Interviews: Researchers ask questions and keep track of the results
2. Focus Groups: Groups are picked out by a researcher, often within a similar demographic, and are asked questions while reactions and feedback are recorded.
3. Observation: Researchers observes settings where respondents are and records relevant information.

Intro to Data Science

Qualitative Data

4. Longitudinal Studies: Data collection from the same source over an extended period of time.

5. Case Studies: An individual occurence or event is studied in depth.



Deductive Approach = Frequentist approach
Inductive Approach = Bayesian Approach

Bayesian vs. frequentist inference

frequentist:

1) Deductive hypothesis testing of Popper--ruling out alternative explanations

Falsification: can prove that a theory is false by finding contradicting evidence but, cannot confirm a theory (may find a better theory or a falsification)

2) Statistical methods of Fisher – everything you have learned this semester

Intro to Data Science

Qualitative Data

Disadvantages

- **Time consuming** - it takes much longer and is more expensive to perform a qualitative test/often a smaller sample size must be used
- **Hard to Generalize** - smaller sample sizes make it harder to draw broad conclusions
- **Skill-Dependen** - quality of gathering data depends on researchers ability to interview and observe

Intro to Data Science

Selection Bias

This occurs when a sample population does not reflect the true population.

For example, say you want to research the effects of a new heart medication. When selecting your candidates, you select those who already have other preexisting conditions.

If your research shows the heart medications causes complications, you will now be unsure whether it was a result of the medication or the preexisting condition.

TECH TALENT SOUTH

Intro to Data Science

Non-Response Bias

If you attempt to poll a large number of people about a topic, many people will opt to not respond. Only those passionate about a topic one way or the other will respond, leading to a loud minority dictating the results.

TECH TALENT SOUTH

Intro to Data Science

Social Desirability Bias

Subjects may be prone to answer what is considered socially acceptable, but not what they truly believe.

Indirect and non-personal questions can help to avoid this. People will be more likely to answer truthfully.

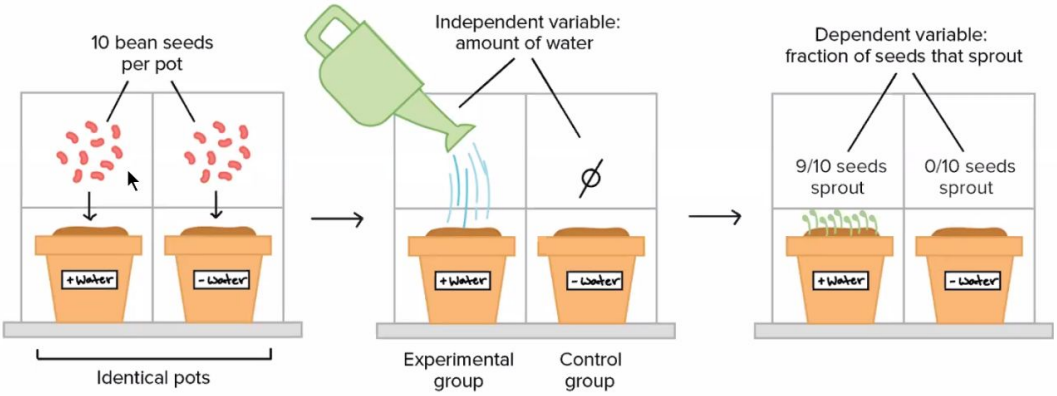
Four steps of hypothesis testing

1. Formulate the **null** and the **alternative** (this includes one- or two-directional test) hypothesis.
2. Select the **significance level** α – a criterion upon which we decide that the claim being tested is true or not.

--- COLLECT DATA ---

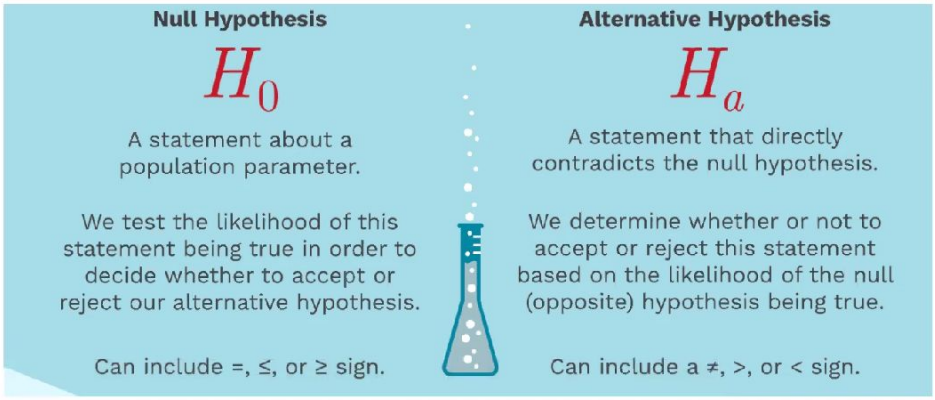
3. Compute the **p-value**. The p-value is the probability that the data would be at least as extreme as those observed, if the null hypothesis were true.
4. Compare the p-value to the α-level. If $p \leq \alpha$, the observed effect is statistically significant, the null is rejected, and the alternative hypothesis is valid.

- **Control Group -** **Otherwise known as the original group without change that we are comparing with or against a experimental group.**

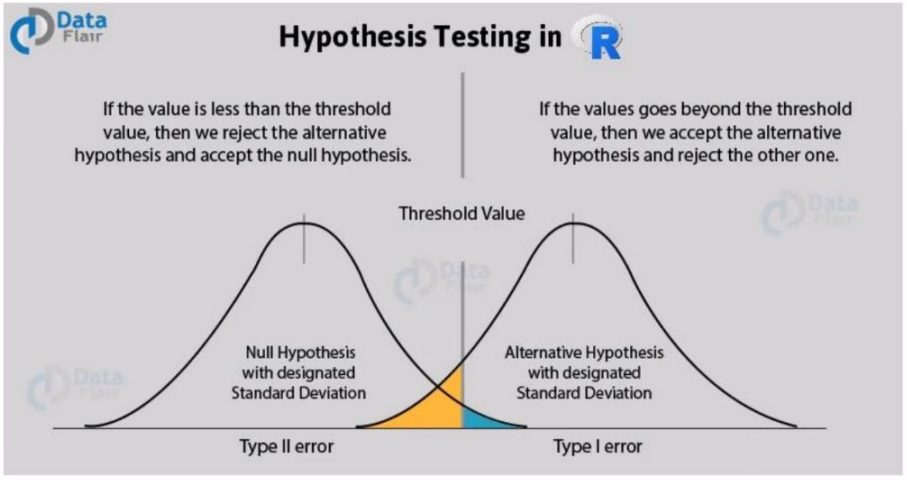


- **A Null Hypothesis -** **Also known as a Valid Claim**
  - **The function of a Null Hypothesis is to make a claim that a relationship between variables, groups, samples that doesn't exist**

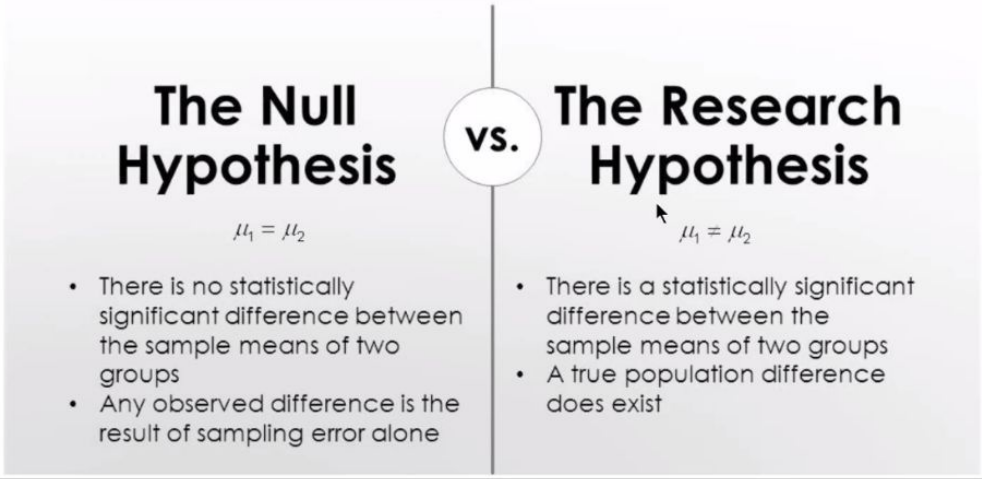
- **Alternative Hypothesis -** A counter to the Null Hypothesis
  - The only function of Alternative Hypothesis is to prove the Null Hypothesis wrong with evidence
  - This is the ultimate goal of every Data Scientist



- **Alternative Hypothesis** (also considered Research Hypothesis)
  - The only function of Alternative Hypothesis is to prove the Null Hypothesis wrong with evidence



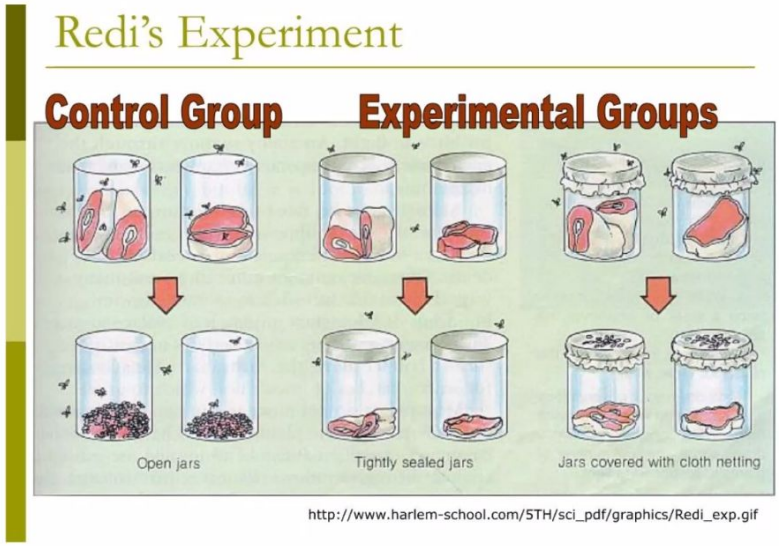
- **Alternative Hypothesis** (also considered Research Hypothesis)
  - The only function of Alternative Hypothesis is to prove the Null Hypothesis wrong with evidence



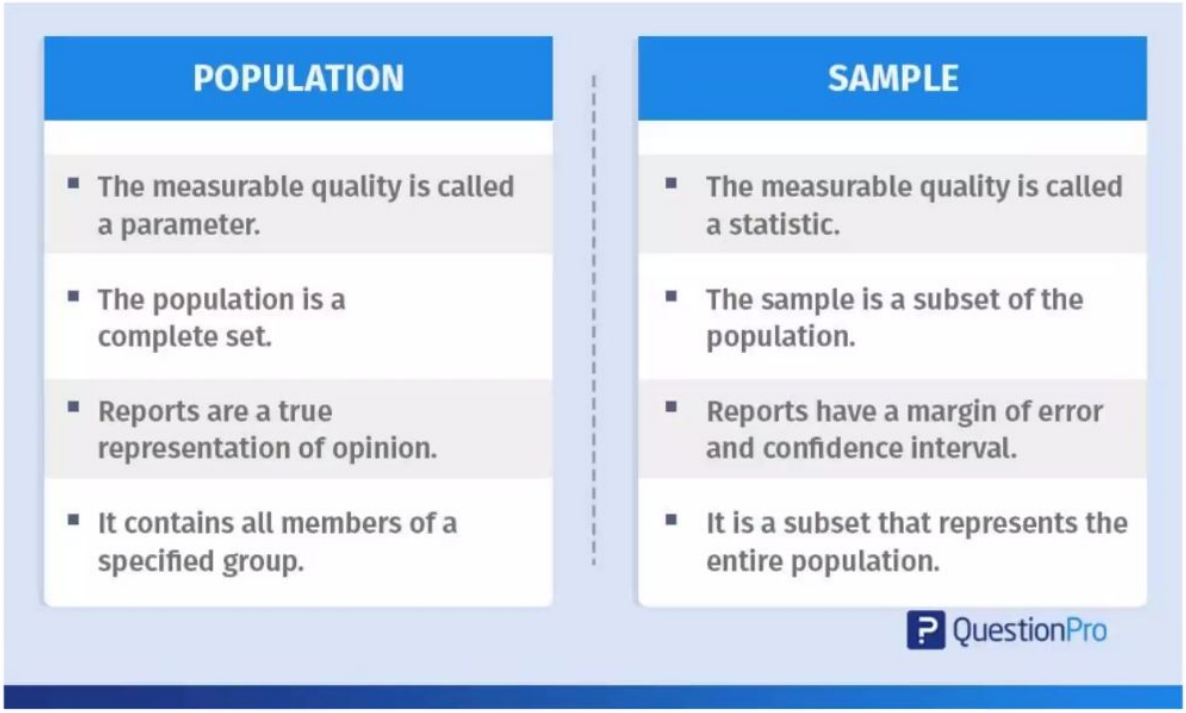
That u in the formula is the mean or population mean.

https://www.rapidtables.com/math/symbols/Statistical_Symbols.html

- **Control Group -** Otherwise known as the original group without change that we are comparing with or against a experimental group.



- **Population -** The total amount of a specific group
- **Sample -** A portion of the population you want to study



- **Parameter - is considered a statistical fact about a population**
  - **80% of the people are Tall in this group**

- **Parameter** – Fact about a population
- **Statistic** – Fact about the sample
- Parameter : Population :: Statistic : Sample
- The theory of sampling is the statistic should be the same or at least close to the parameter

- **Parameter - is considered a statistical fact about a population**
  - **80% of the people are over 6'1 in this group**

Table 1. Comparison of Sample Statistics and Population Parameters

	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Standard deviation	s	sigma
Variance	s^2	sigma^2

Symbols can change depending on the use case.

- **Parameter -** is considered a statistical fact about a population
  - 80% of the people are over 6'1 in this group

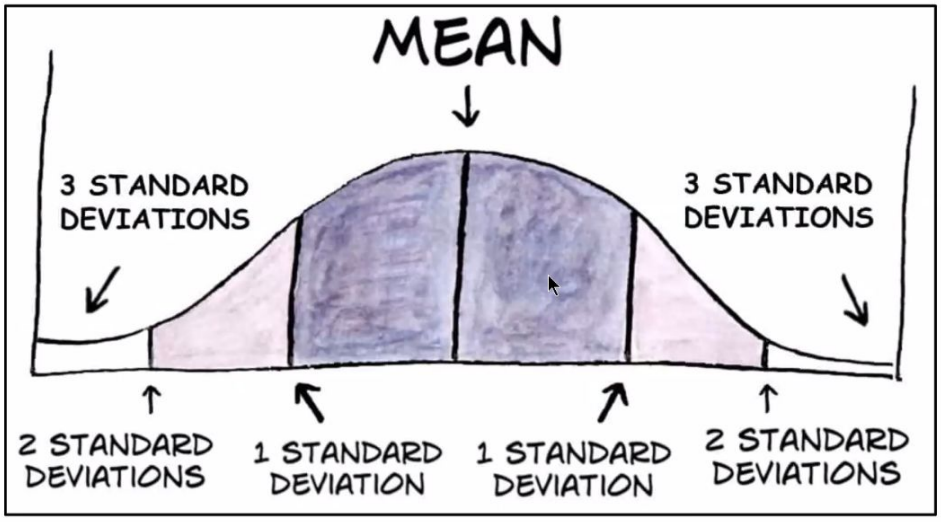
	Sample Statistics	Actual Population Parameters[a]	Estimated Population Parameters	Formula for Estimate	Sampling Distribution Parameters to be Estimated[b]	Formula for Unbiased Estimate of Sampling Distribution Parameter[c]
Mean	$\bar{x}$	μ	$\hat{\mu}$	$\bar{x} = \hat{\mu}$	$\mu_{\bar{x}}$	$\bar{x} = \hat{\mu} = \hat{\mu}_{\bar{x}}$
Standard Deviation	s	σ	$\hat{\sigma}$	$s = \hat{\sigma}$	$\sigma_{\bar{x}}$	$\frac{s}{\sqrt{n-1}} = \hat{\sigma}_{\bar{x}}$

a. We don't know the values of these parameters, but we want to estimate them.

b. The value of the mean of a sampling distribution of means is equal to the value of the population mean.

c. The standard deviation of the sampling distribution for means is called the standard error of the mean.

- **What is Standard Deviation?**
  - **A number that represents how one group differs from the mean value of entire group or data set**



What is the Variance?

- **Measures how far the set of (random) numbers are spread out from mean (their average value)**
- **This tells us the measure of numbers from our mean**
  - **Two Types of Variance**
    - **Population**
      - **Variance from all data**
    - **Sample**
      - **Variance from a sample of the data**
        - **Variance could be higher because it is compensating without all data**

Variance has different use cases. For this specific one above variance is just the variance from the mean.

If the mean is 100 and we have several different test scores the highest score would be 100 but the mean would be 70 and the lowest score would be 30. Our variance would be between 30 to 100.

- **Probability distribution** is a function used in Data Science to describe all possible values or outcomes with random variables.
  - Steps to determine information about a population to figure out probability distribution of a **population:**
    - First we want to take a random sample from a population or group to get probability distribution (this is called **Sampling Distribution**)
    - We use this **sample** to summarize the entire population called **Sample Statistic or parameter estimate**
    - The remaining untested population is called the **parameter**
    - The parameter Estimate is used to find out info about the overall parameter or population
    - The issues with Sample Statistic is the values can change because chosen samples are random

- **Probability distribution** is a function used in Data Science to describe all possible values or outcomes with random variables.
    - **Four Types of Sampling Distribution**

Probability Sampling Methods

- **Simple Random Sampling**
- Sampling with or without replacement
- **Systematic Random Sampling**
  - Total number of cases (M) divided by the sample (N), this is your sampling interval K. M/N=K
  - Use random start. Select each Kth case

Stratified Random Sampling

  - Create homogenous groups (strata)
  - Sample randomly from each separately

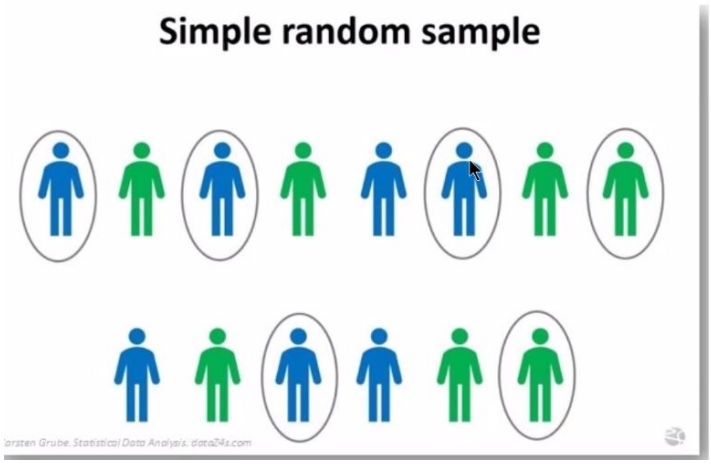
Cluster Sampling

Pick groups (clusters) randomly (weight groups by size)

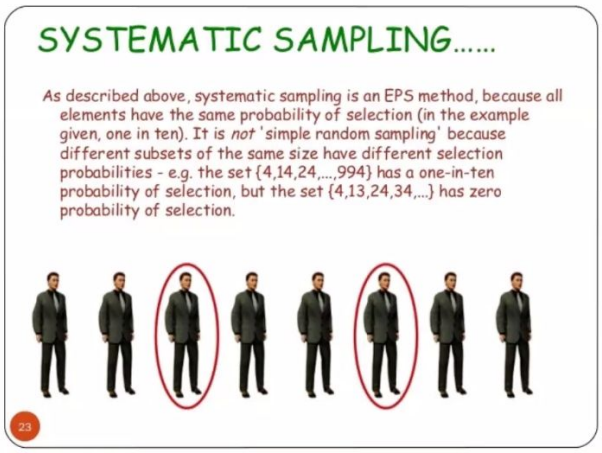
Interview/observe every member in the group

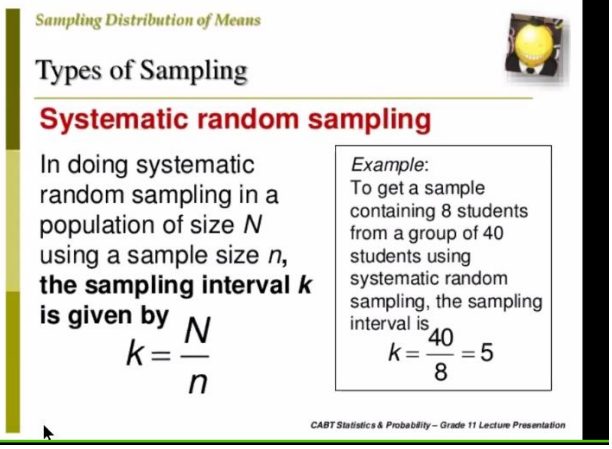
- **Probability distribution** is a function used in Data Science to describe all possible values or outcomes with random variables.
    - **Four Types of Sampling Distribution**

        - **Simple Random Sampling**



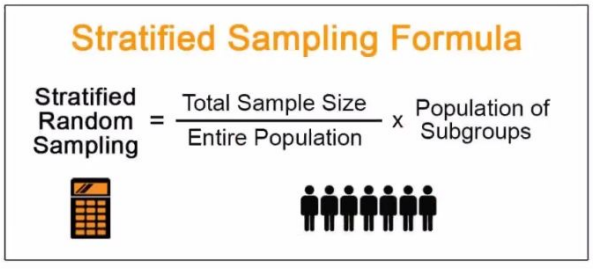
- **Probability distribution** is a function used in Data Science to describe all possible values or outcomes with random variables.
    - **Four Types of Sampling Distribution**
        - **Simple Random Sampling**
        - **Systematic Sampling (Sample Interval)**

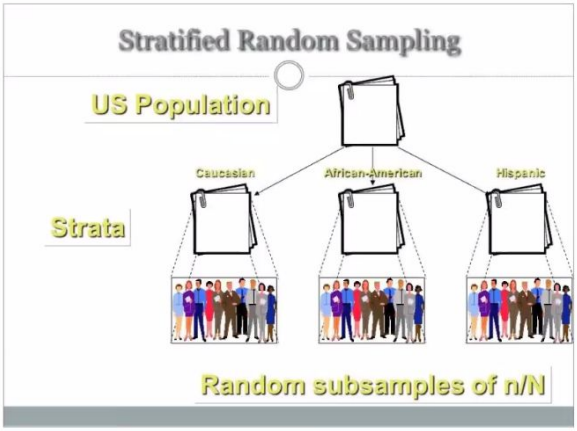




Systematic sampling is sampling used with a system.

- **Probability distribution** is a function used in Data Science to describe all possible values or outcomes with random variables.
    - **Four Types of Sampling Distribution**
        - **Simple Random Sampling**
        - **Systematic Sampling (Sample Interval)**
        - **Stratified Random Sample**

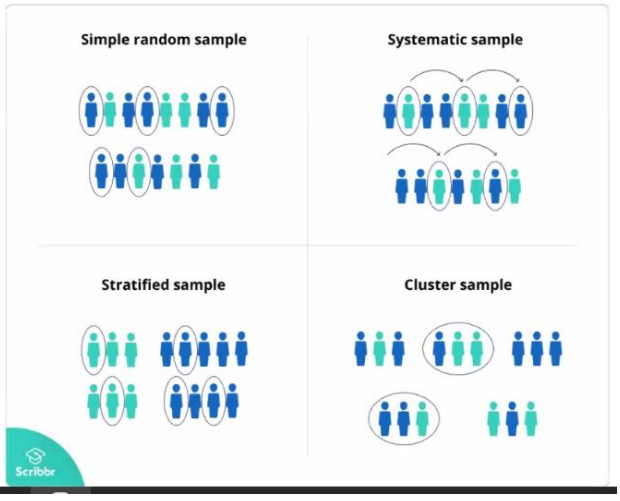




Random samples within a specific group

- **Probability distribution** is a function used in Data Science to describe all possible values o outcomes with random variables.
    - **Four Types of Sampling Distribution**
        - **Simple Random Sampling**
        - **Systematic Sampling (Sample Interval)**
        - **Stratified Random Sample**
        - **Cluster Sampling**

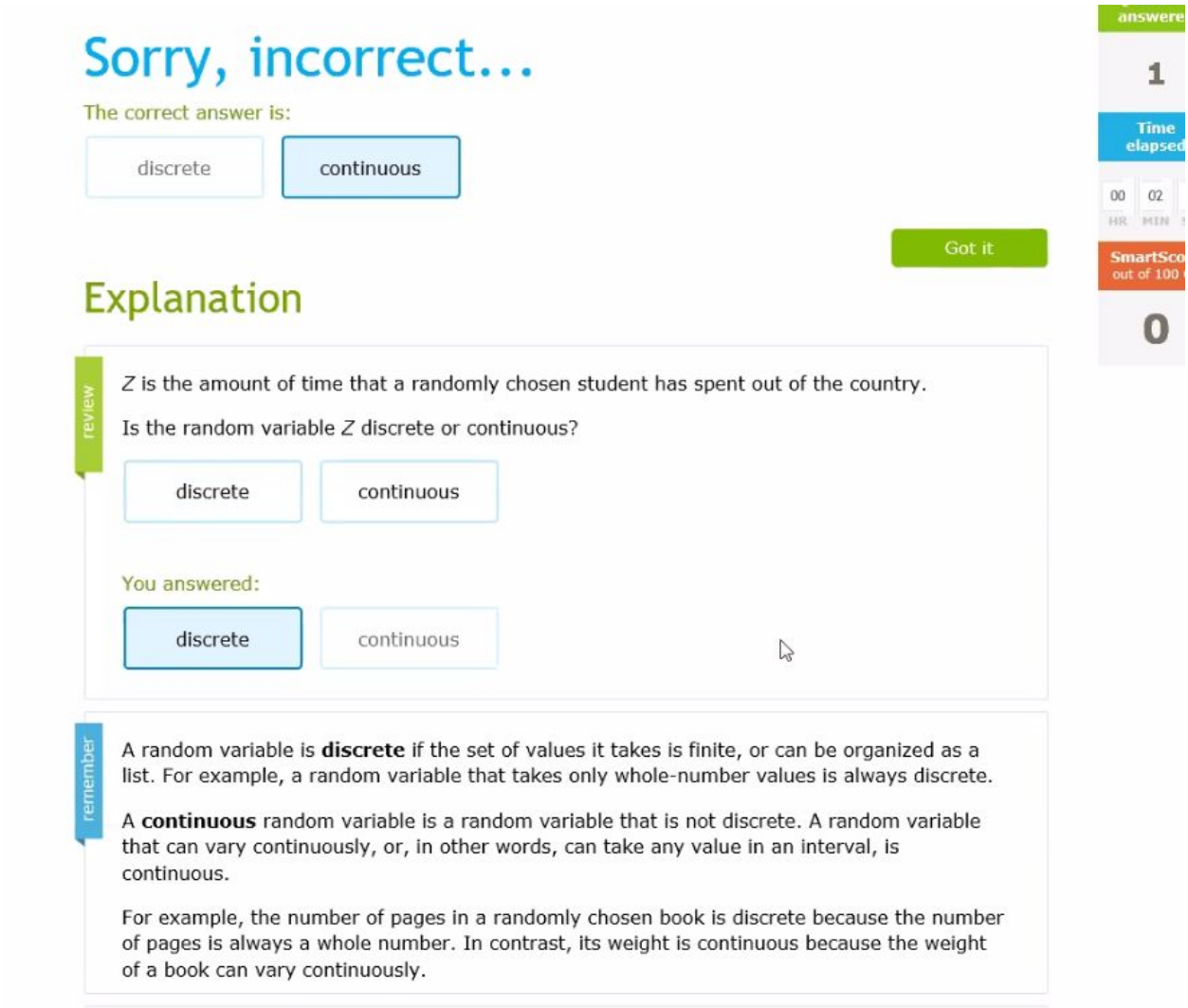
- **Probability distribution** is a function used in Data Science to describe all possible values or outcomes with random variables.
    - **Four Types of Sampling Distribution**
        - **Simple Random Sampling**
        - **Systematic Sampling (Sample Interval)**
        - **Stratified Random Sample**
        - **Cluster Sampling**



We started watching a Khan Academy video:

1st: Techniques for generating a simple random sample:

- Talked about 80 people being in your grade at school
- Not being able to get the height of all 80 so you random sample 30
- Associate everyone in school with a piece of paper and randomly picking out 30
- Don't put the paper back in the bowl or you may end up with the same person twice
- Other way a random generator by computer or calculator
- Other way is random digit table (old school but still a way)



SmartScore out of 100

28

Explanation

review

Z is the volume of water used by a randomly chosen household in a month.

Is the random variable Z discrete or continuous?

discrete | continuous

You answered:

discrete | continuous

remember

A random variable is **discrete** if the set of values it takes is finite, or can be organized as a list. For example, a random variable that takes only whole-number values is always discrete.

A **continuous** random variable is a random variable that is not discrete. A random variable that can vary continuously, or, in other words, can take any value in an interval, is continuous.

For example, the number of pages in a randomly chosen book is discrete because the number of pages is always a whole number. In contrast, its weight is continuous because the weight of a book can vary continuously.

solve

Z is a continuous random variable. The volume of water used in a household can vary continuously. All possible volumes cannot be listed.

Got it