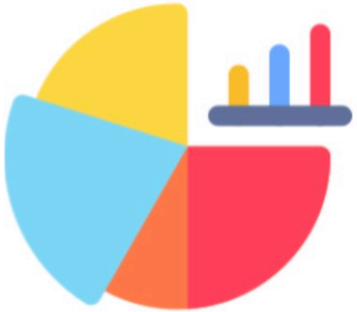
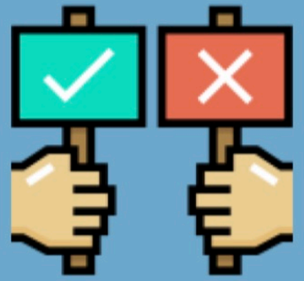


Statistics Session-1

CLARUSWAY©
WAY TO REINVENT YOURSELF

Did you finish Statistics (Data Types & Patterns & Graphs) pre-class activity?



Students choose an option

Pear Deck Interactive Slide
Do not remove this bar

NEXT SLIDE

No Multiple Choice Response
You didn't answer this question


SUCCESS NEEDS
PREPARATION

Urgent & Important *do it now*	Important not urgent *decide when to do it*
Urgent not important *delegate it*	Not important not urgent *delete it*

Prioritize


Right on time...

Course Info

Lesson Plan

STATISTICS BASICS

The goal of this course is to provide a comprehensive overview of the basics of statistics you will need to start your data science journey.

Custodian : Jason-Acad.Coord. (jason@clarusway.com)

In-class Sessions : 7 In-classes / 21 hours (*Part-1 → 3 In-classes | Part-2 → 4 In-classes*)

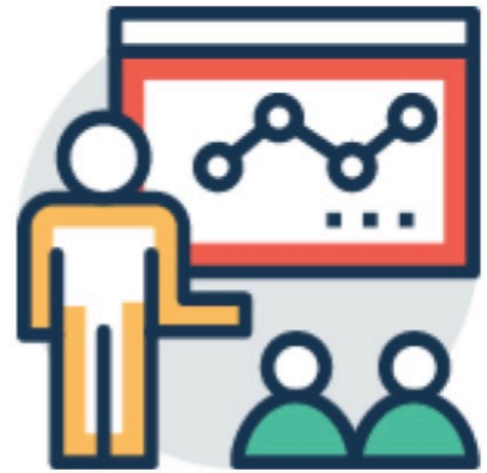
Lab Sessions : 3 Labs / 3 hours (*Part-1 → 1 Lab | Part-2 → 2 Labs*)

Certification Requirements:

1. Attend at least 70% of in-class sessions (at least 5 sessions of attendance)
2. Successfully complete and submit assignments (at least 2 assignments)

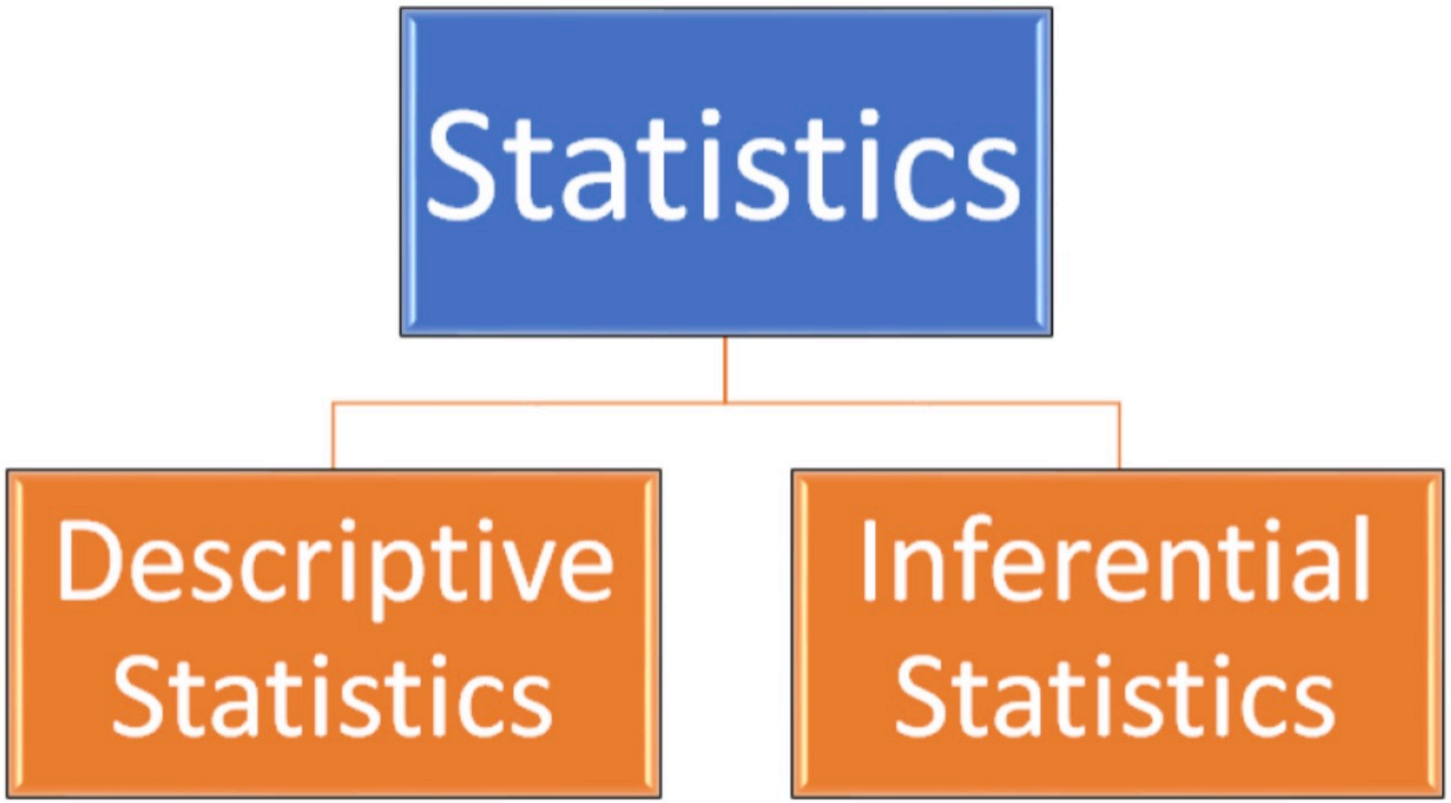
Scope of the Course

- Data Types & Patterns & Graphs
- Central Tendency & Dispersion
- Correlation & Normal Distribution



- Central Limit Theorem and Confidence Intervals
- Basic Concepts of Hypothesis Testing
- Hypothesis Tests about Means
- Analysis of Categorical Variables

Descriptive vs Inferential Statistics



Presenting, organizing and summarizing data

Drawing conclusions about a population based on data observed in a sample

Sources

Clarusway LMS (Pre-class Activities)

- Content
- Let's Practice, Check Yourself
- Assignments

Textbooks

- Brownlee J., Statistical Methods for Machine Learning.
- Gedeck P., Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python
- Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2014). Mathematical statistics with applications. Cengage Learning

WWWs

- http://onlinestatbook.com/
- StatQuest with Josh Starmer
- https://www.khanacademy.org/math/statistics-probability
- https://gelecegiyazanlar.turkcell.com.tr/konu/veri-bilimi-icin-istatistik

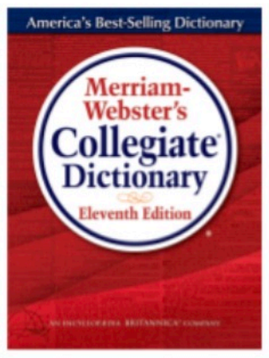
Table of Contents

1 What is “Statistics”?

What is "Statistics"?

WIKIPEDIA

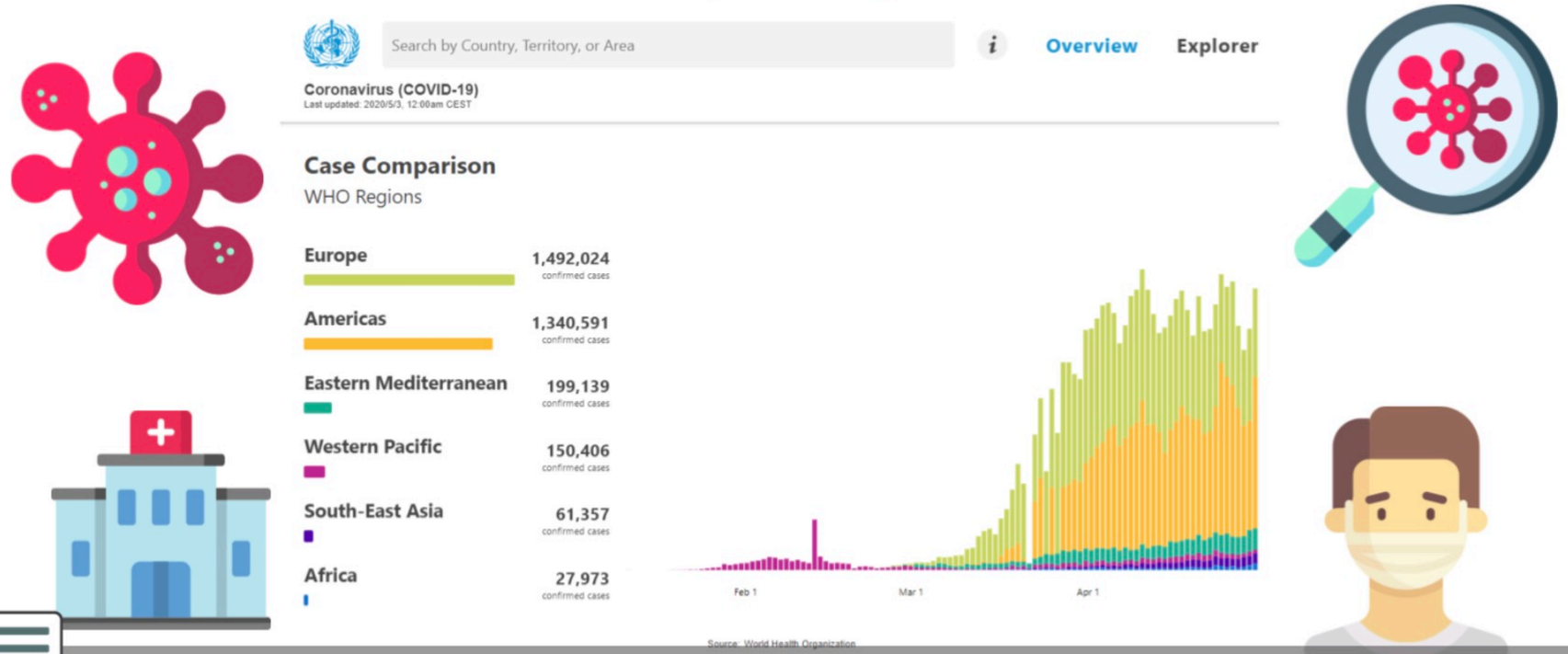
Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data.



A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.

All the authors imply that statistics is a theory of information, with inference making as its objective.

What are some examples of statistics in everyday life?



Students, write your response!

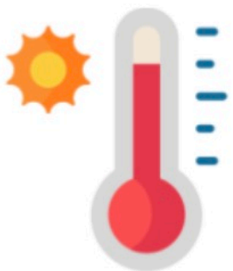
REINVENT YOURSELF

Pear Deck Interactive Slide
Do not remove this bar

No Text Response

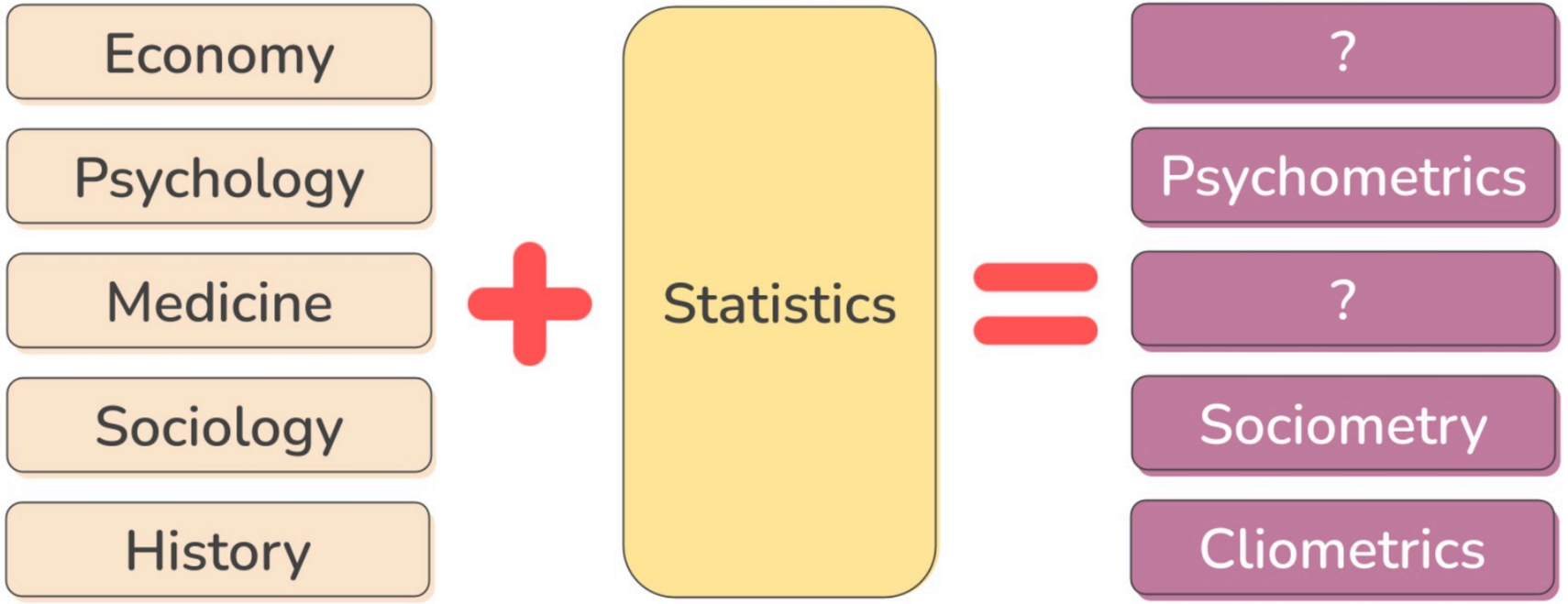
You didn't answer this question

What are some examples of statistics in everyday life?



- Weather Forecasts
- Stock Market
- Predicting Disease
- Medical Studies
- Insurance
- Consumer Goods

Relation of Statistics with other Sciences

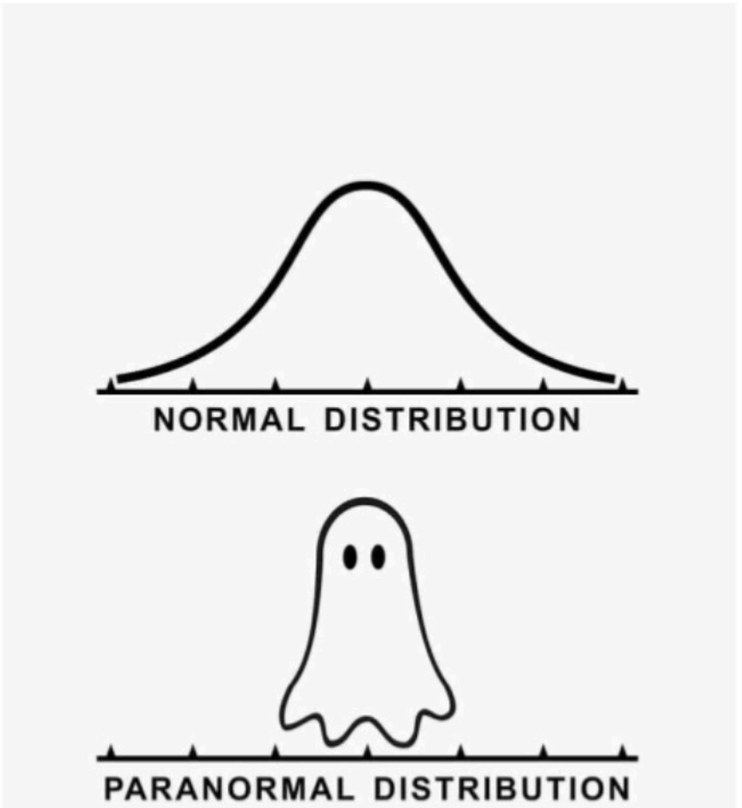


Students, write your response!

No Text Response
You didn't answer this question

Funny Statistics

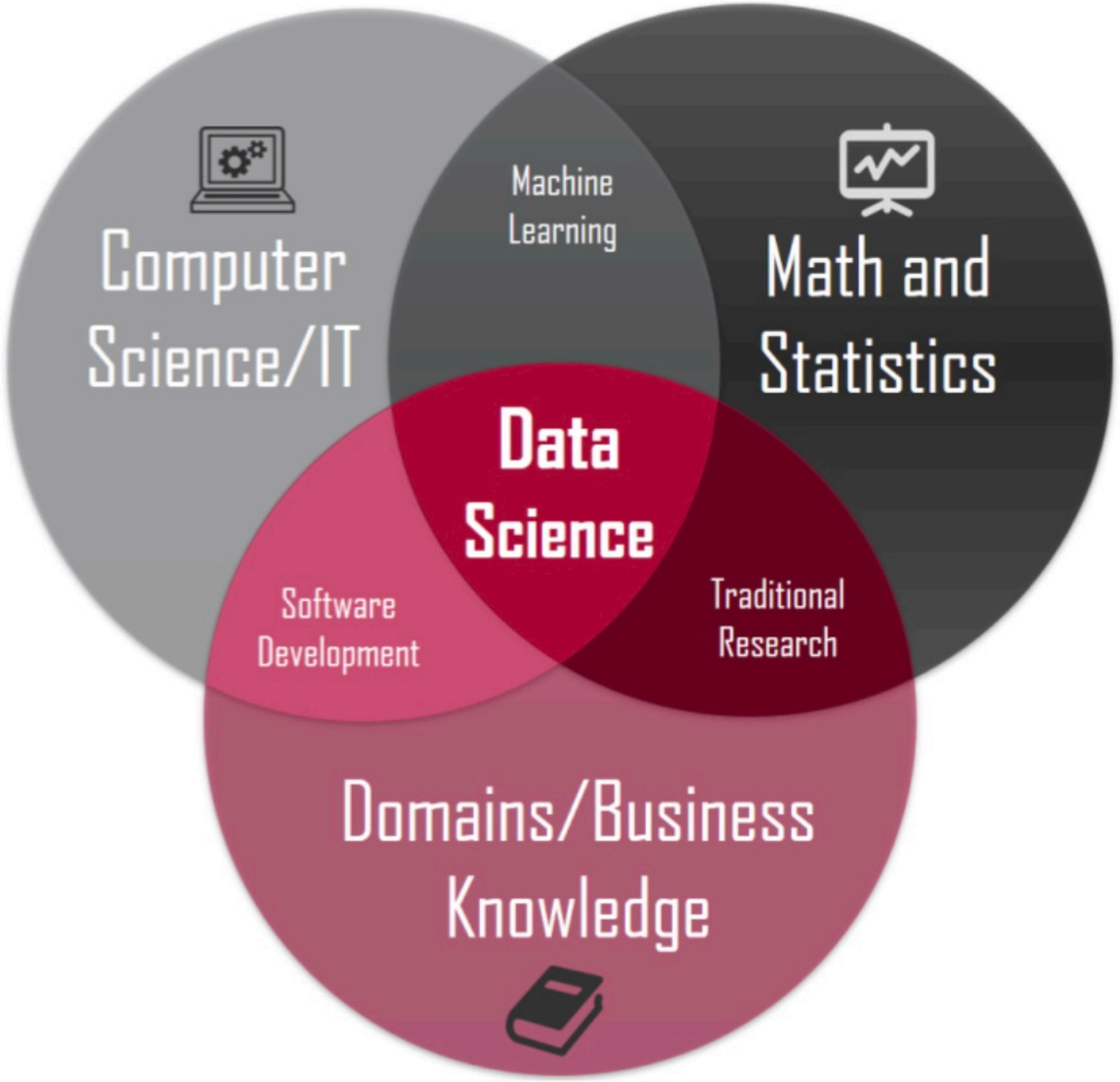
"Excellent health statistics - smokers are less likely to die of age related illness"



2

Why Should You Learn Statistics?

Data Science vs. Statistics



"A Data Scientist is one who knows more statistics than a programmer and more programming than a statistician"

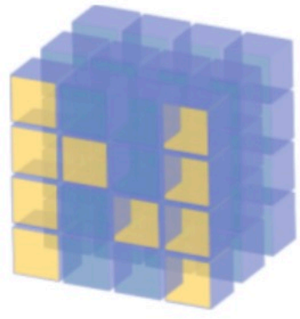
3

Stats with Python

Stats with Python

python™

SciPy

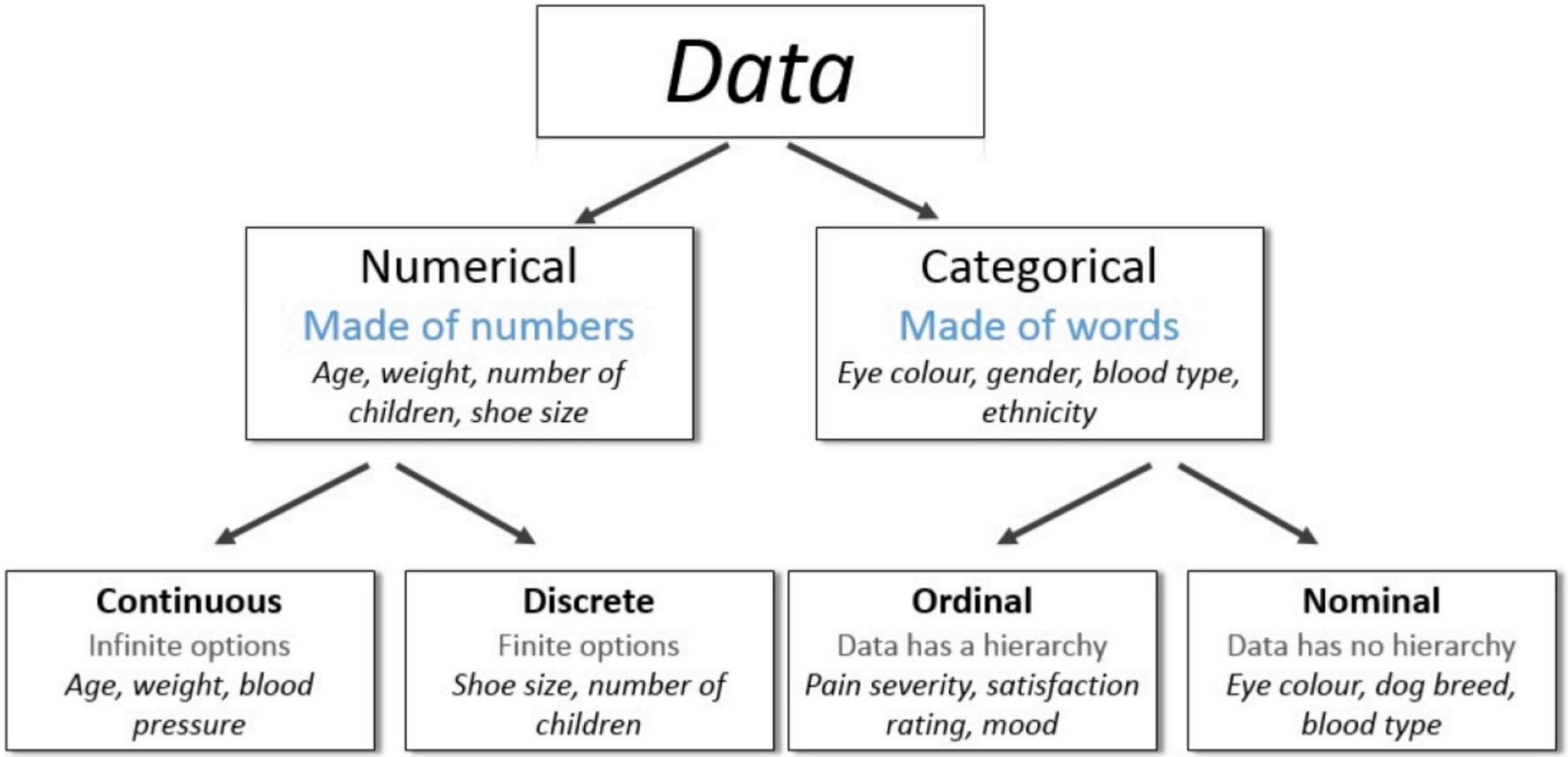


NumPy

matplotlib

4 Types of Data

Types of Data



Numerical Data

Continuous Data

- Continuous data can have an infinite continuum of possible values.
  - height
  - weight
  - age
  - the amount of time it takes to complete an assignment

Discrete Data

- Any variable with a finite number of possible values is discrete.
  - the number of pets in a household
  - the number of children in a family
  - the number of foreign languages in which a person is fluent

Types of Data

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
339	7	178	84	0	0	39.9	0.331	41
403	9	72	78	25	0	31.6	0.280	38
551	3	84	68	30	106	31.9	0.591	25
197	3	107	62	13	48	22.9	0.678	23
563	6	99	60	19	54	26.9	0.497	32
239	0	104	76	0	0	18.4	0.582	27
141	5	106	82	30	0	39.5	0.286	38
523	9	130	70	0	0	34.2	0.652	45
696	3	169	74	19	125	29.9	0.268	31
238	9	164	84	21	0	30.8	0.831	32

Categorical Data

Ordinal Data

- Ordinal data requires an order
  - small medium, large
  - good, average, poor
  - strongly agree, agree, disagree

- The distance between ordered categories is not measurable.
- No arithmetic can be done with the ordinal data as they show sequence only.

Nominal Data

- Nominal data simply names something without an order being given.
  - employee's status
  - color
  - race

- Data obtained on nominal scale is in terms of frequency.

Types of Data

	gender	race/ethnicity	parental level of education	lunch
0	female	group B	bachelor's degree	standard
1	female	group C	some college	standard
2	female	group B	master's degree	standard
3	male	group A	associate's degree	free/reduced
4	male	group C	some college	standard

Types of Data

	division	level of education	training level	work experience	salary	sales
0	printers	some college	2	6	91684	372302
1	printers	associate's degree	2	10	119679	495660
2	peripherals	high school	0	9	82045	320453
3	office supplies	associate's degree	2	5	92949	377148
4	office supplies	high school	1	5	71280	312802

Categorical Data in ML Models

Raw Data

ID	Country	Population
1	Japan	127185332
2	U.S	326766748
3	India	1354051854
4	China	1415045928
5	U.S	326766748
6	India	1354051854

Encoding

Ready for Model

ID	Country	Population
1	0	127185332
2	1	326766748
3	2	1354051854
4	3	1415045928
5	1	326766748
6	2	1354051854

Encoding Method: Label Encoder

Not appropriate. Because Countries are nominal, not ordinal.

Categorical Data in ML Models

Raw Data

ID	Country
1	Japan
2	U.S
3	India
4	China
5	U.S
6	India

Encoding

Ready for Model

ID	Country_Japan	Country_U.S	Country_India	Country_China
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	0	1	0	0
6	0	0	1	0

Encoding Method: One Hote Encoder

Appropriate for nominal data. (No order)

Which variable is categorical?

Students choose an option

Pear Deck Interactive Slide
Do not remove this bar

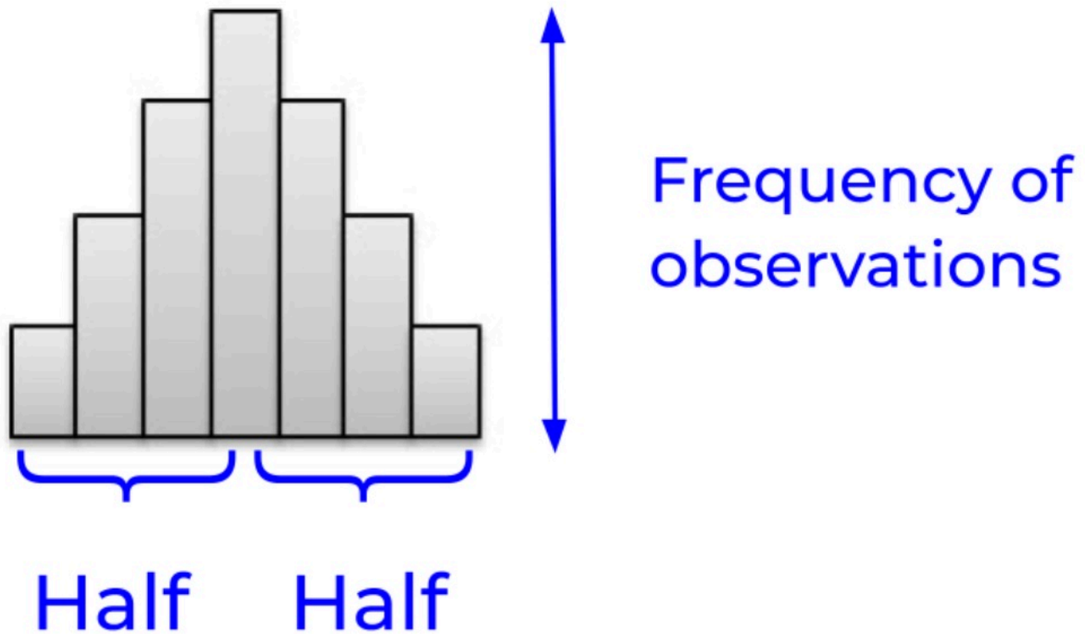
No Multiple Choice Response
You didn't answer this question

5 Data Patterns

Data Patterns in Statistics

Center | Spread | Shape | Unusual Features

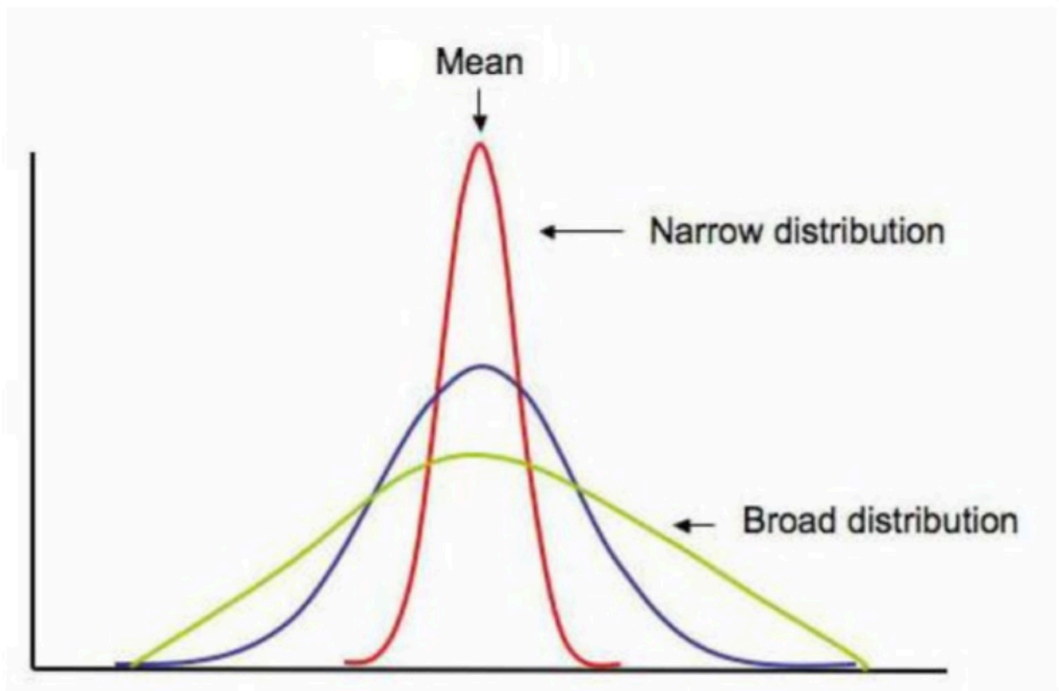
The center of a distribution, graphically, is located at the median of the distribution.



Data Patterns in Statistics

Center | **Spread** | Shape | Unusual Features

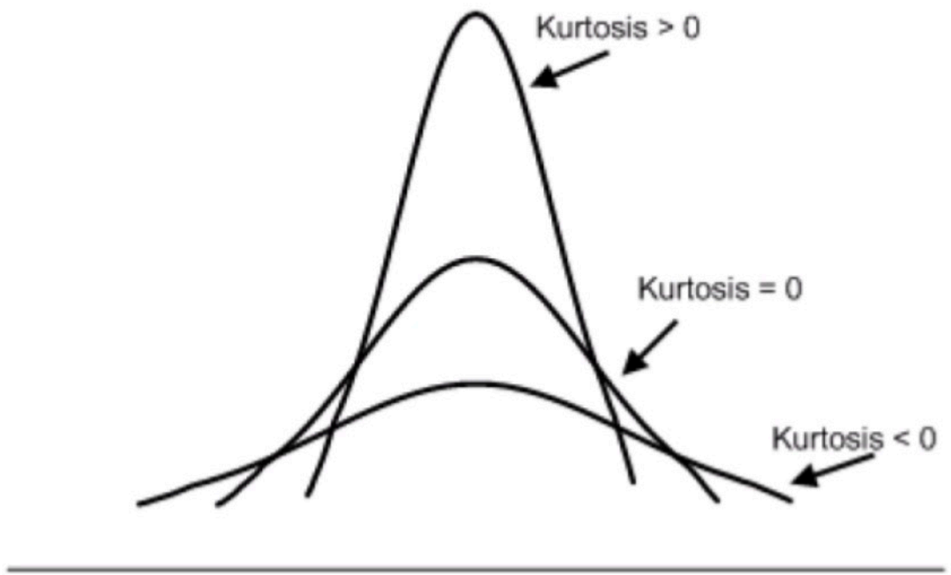
The spread of a distribution refers to the variation of the data.



Data Patterns in Statistics

Center | Spread | **Shape** | Unusual Features

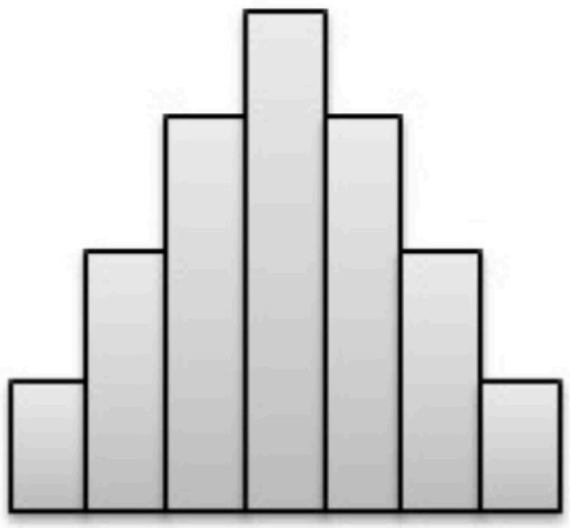
Kurtosis - Some distributions may have multiple observations on one side of the graph than the other side.



Data Patterns in Statistics

Center | Spread | **Shape** | Unusual Features

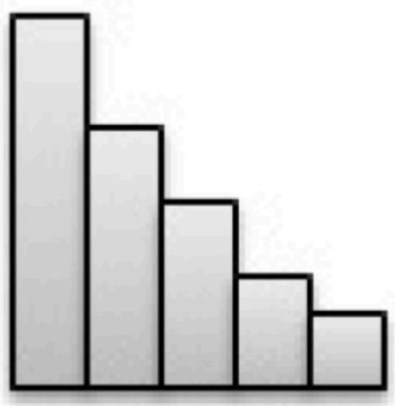
Symmetry - In symmetric distribution, graph can be divided at the center in such a way that each half is a mirror image of the other.



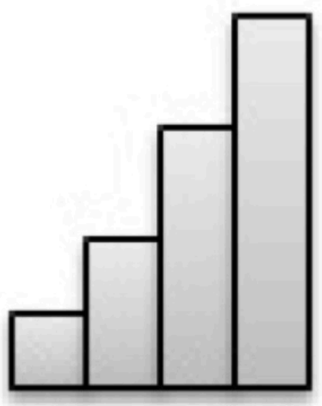
Data Patterns in Statistics

Center | Spread | **Shape** | Unusual Features

Skewness - Some distributions may have multiple observations on one side of the graph than the other side.



Skewed Right

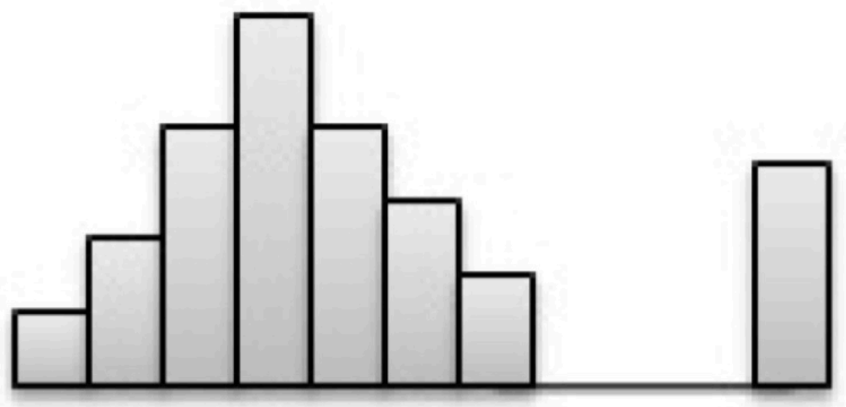


Skewed Left

Data Patterns in Statistics

Center | Spread | Shape | **Unusual Features**

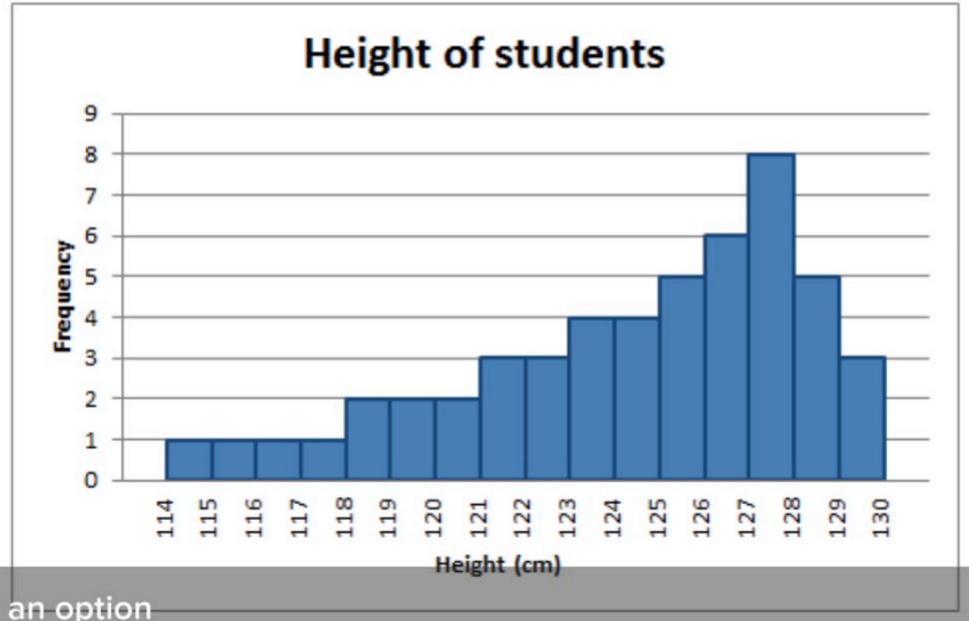
Outliers - Distributions may be characterized by extreme values that differ greatly from the other set of observation data.



Outliers

Let's Practice

Which of the following statements is true about the figure?



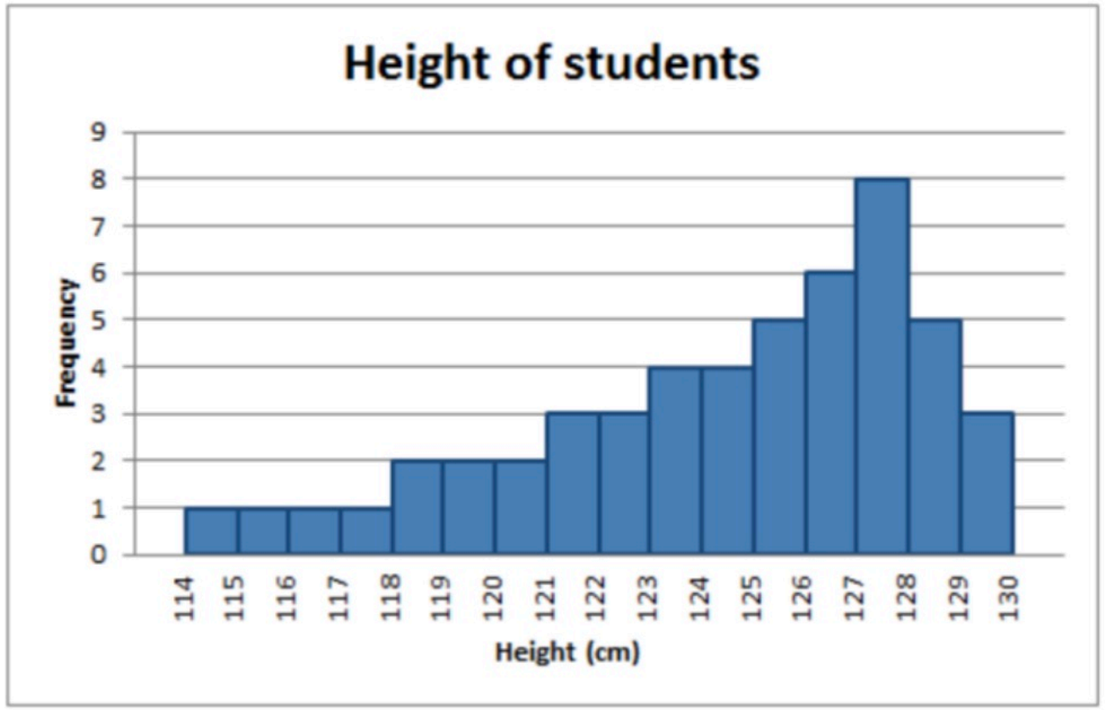
Students choose an option

No Multiple Choice Response

You didn't answer this question

Let's Practice Answer

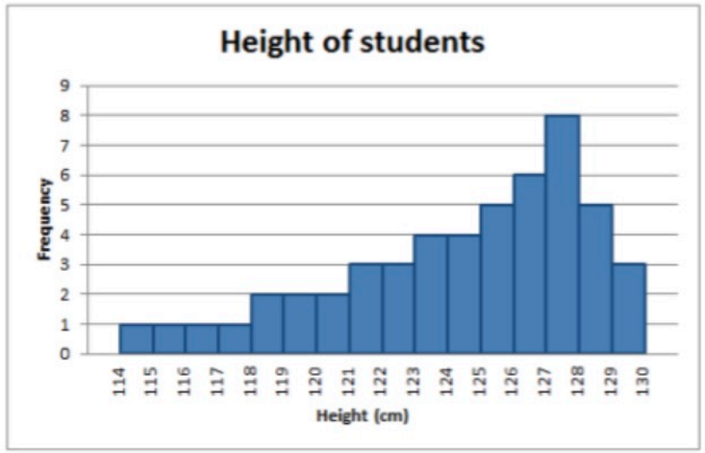
Which of the following statements is true about the figure?

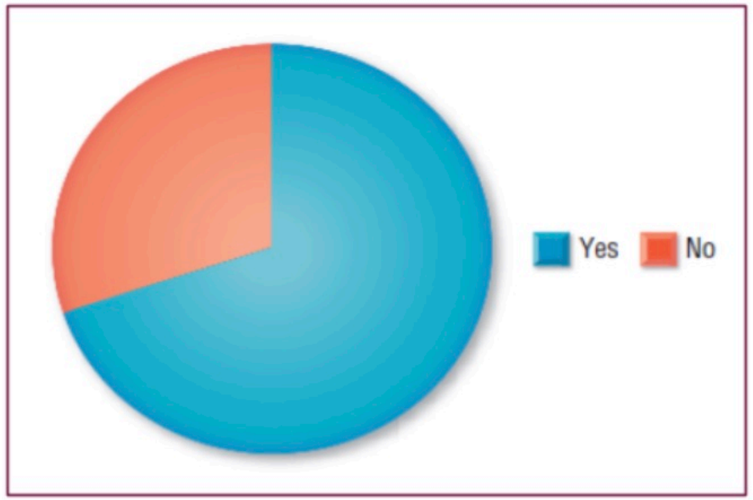


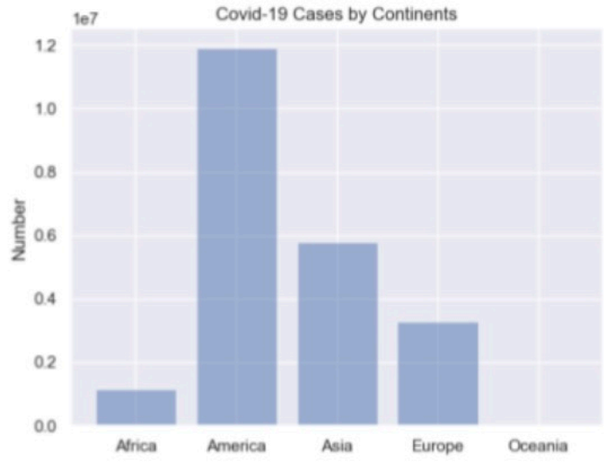
C: The distribution is left-skewed with no outliers.

6

Graphical Representation of Data



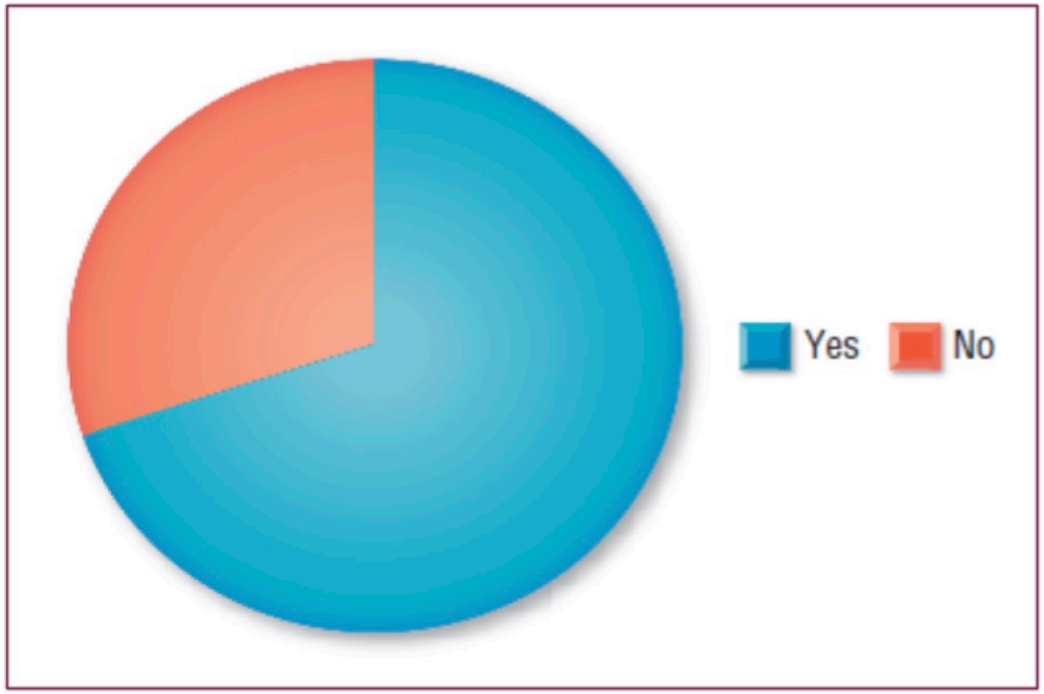




Pie Charts

- Often used with nominal and ordinal variables.
- Circle cut into “pie slices” that add up to 100%.
- Each pie slice represents a category

Did you find the course challenging?



Pie Charts

The following table shows the numbers of hours spent by a child on different events on a working day.

Represent the data on a pie chart

Activity	No. of Hours
School	6
Sleep	8
Playing	2
Study	4
T. V.	1
Others	3

Pie Charts

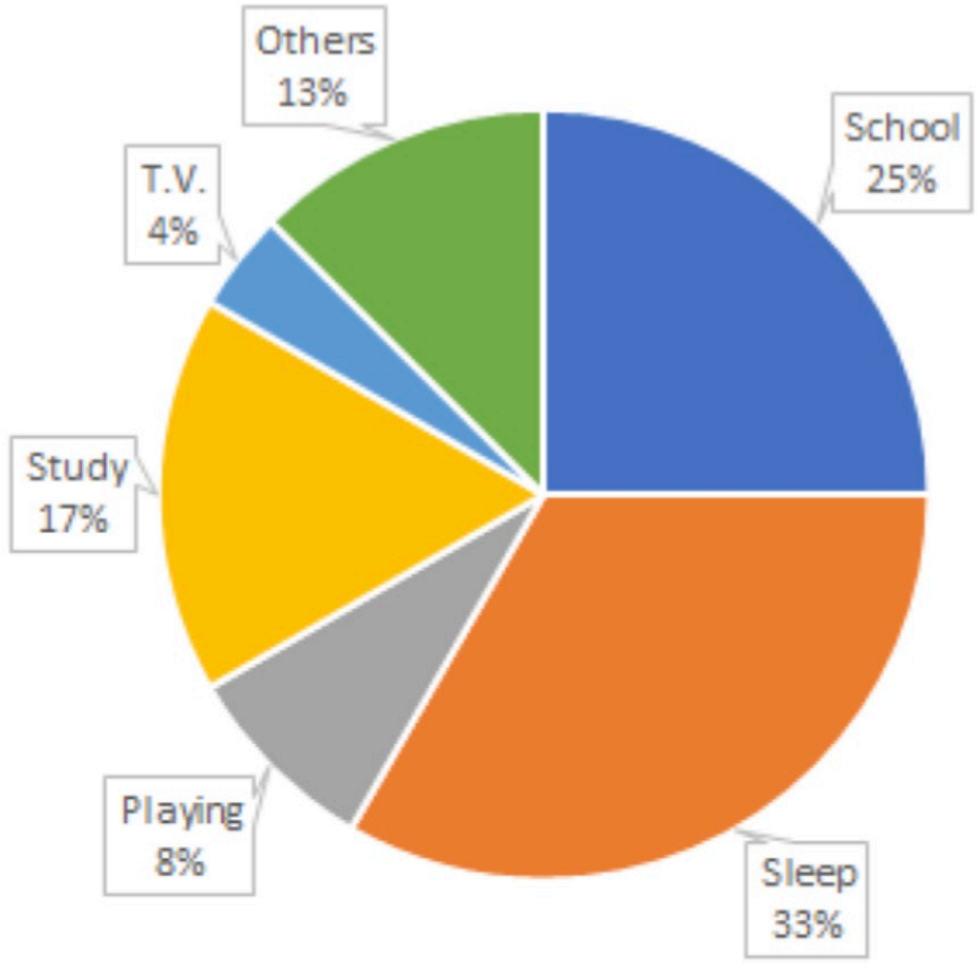
The central angles for various observations can be calculated as:

Activity	No. of Hours	Measure of central angle
School	6	$({}^{6}/_{24} \times 360)° = 90°$
Sleep	8	$({}^{8}/_{24} \times 360)° = 120°$
Playing	2	$({}^{2}/_{24} \times 360)° = 30°$
Study	4	$({}^{4}/_{24} \times 360)° = 60°$
T. V.	1	$({}^{1}/_{24} \times 360)° = 15°$
Others	3	$({}^{3}/_{24} \times 360)° = 45°$

Pie Charts

Now, we shall represent these angles within the circle as different sectors.

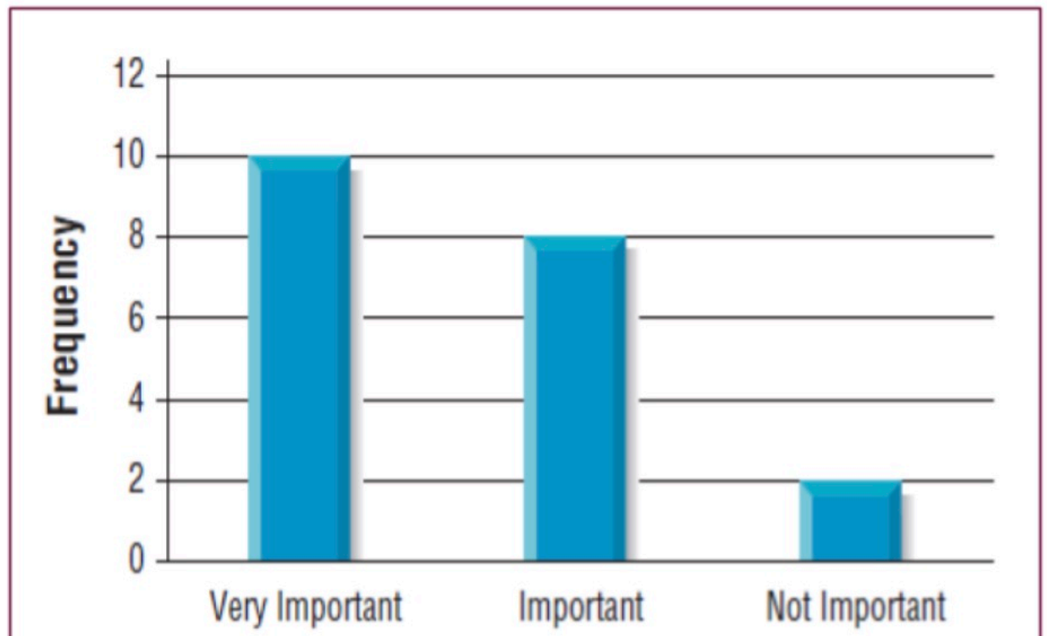
Activity	No. of Hours
School	6
Sleep	8
Playing	2
Study	4
T. V.	1
Others	3



Bar Charts

- Often used with nominal and ordinal variables.
- A series of bars represent the different attributes of a variable.
- The height of each bar reflects frequencies for each attribute.

How important is *Data Science* to you?



Bar Charts

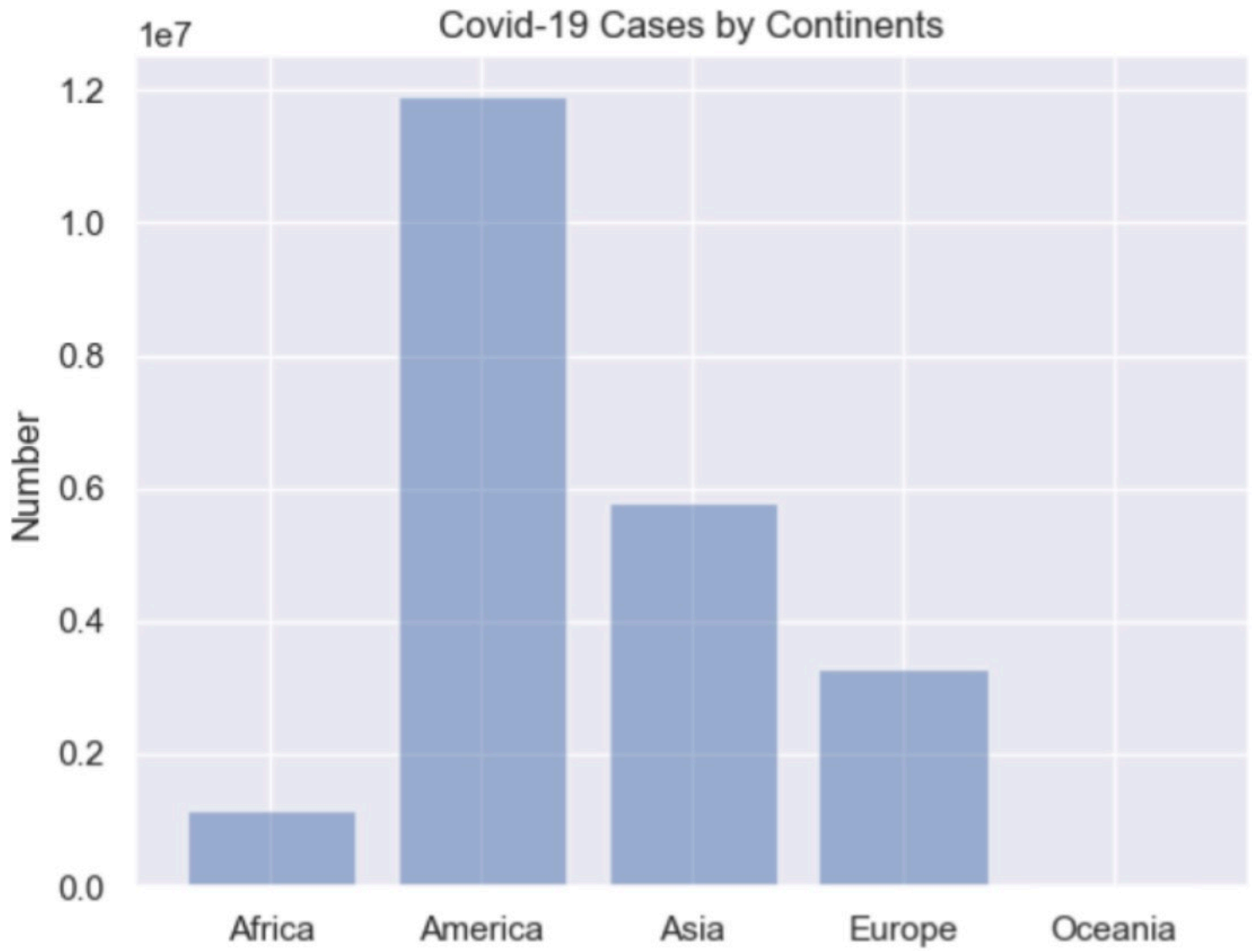
The following table shows the numbers of Covid-19 data for continents.

Represent the data on a bar chart

continent	cases	deaths
Africa	1119579	26260
America	11698368	427207
Asia	5606210	122034
Europe	3239237	205144
Oceania	25742	471

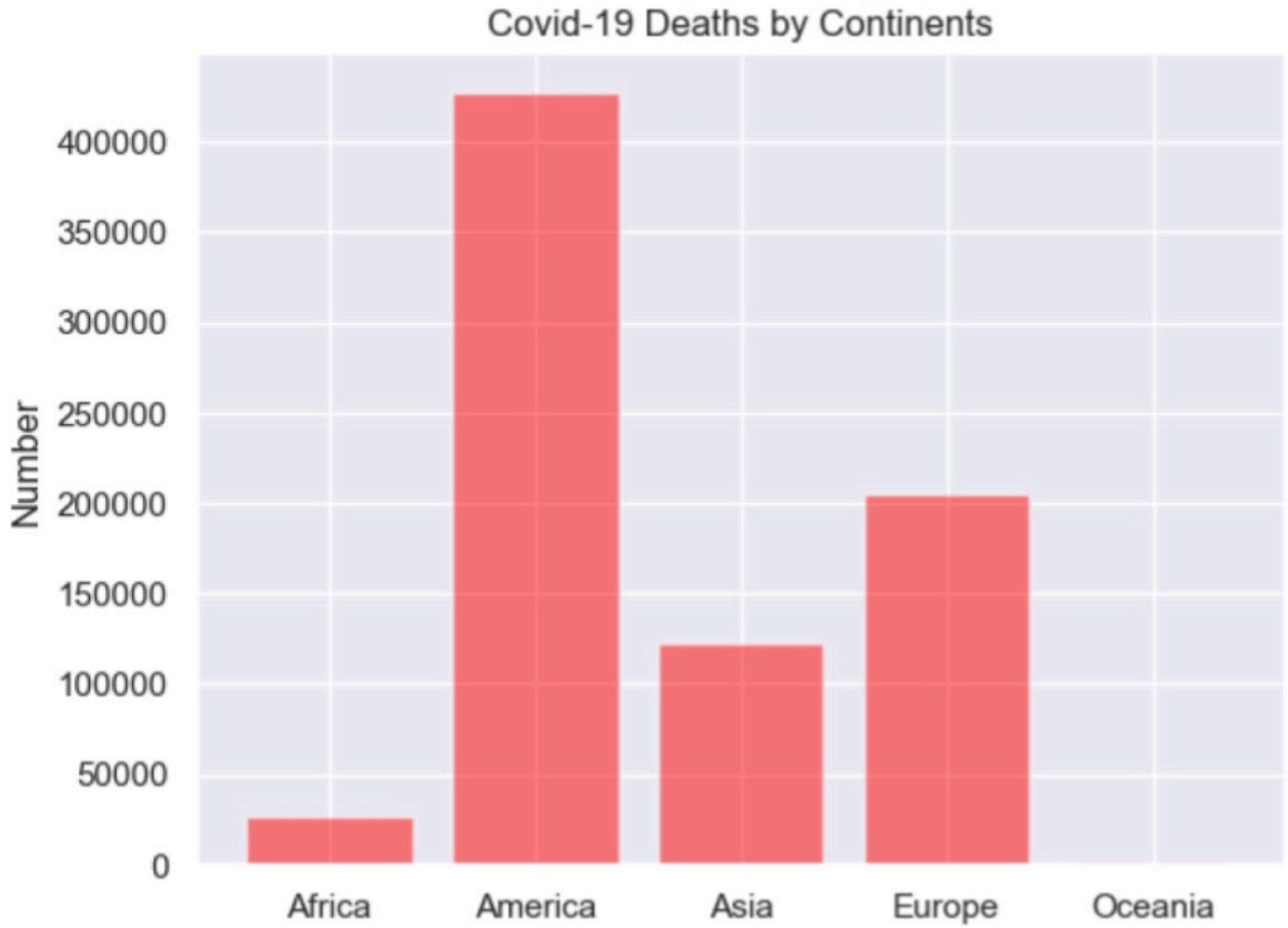
Bar Charts

continent	cases	deaths
Africa	1119579	26260
America	11698368	427207
Asia	5606210	122034
Europe	3239237	205144
Oceania	25742	471



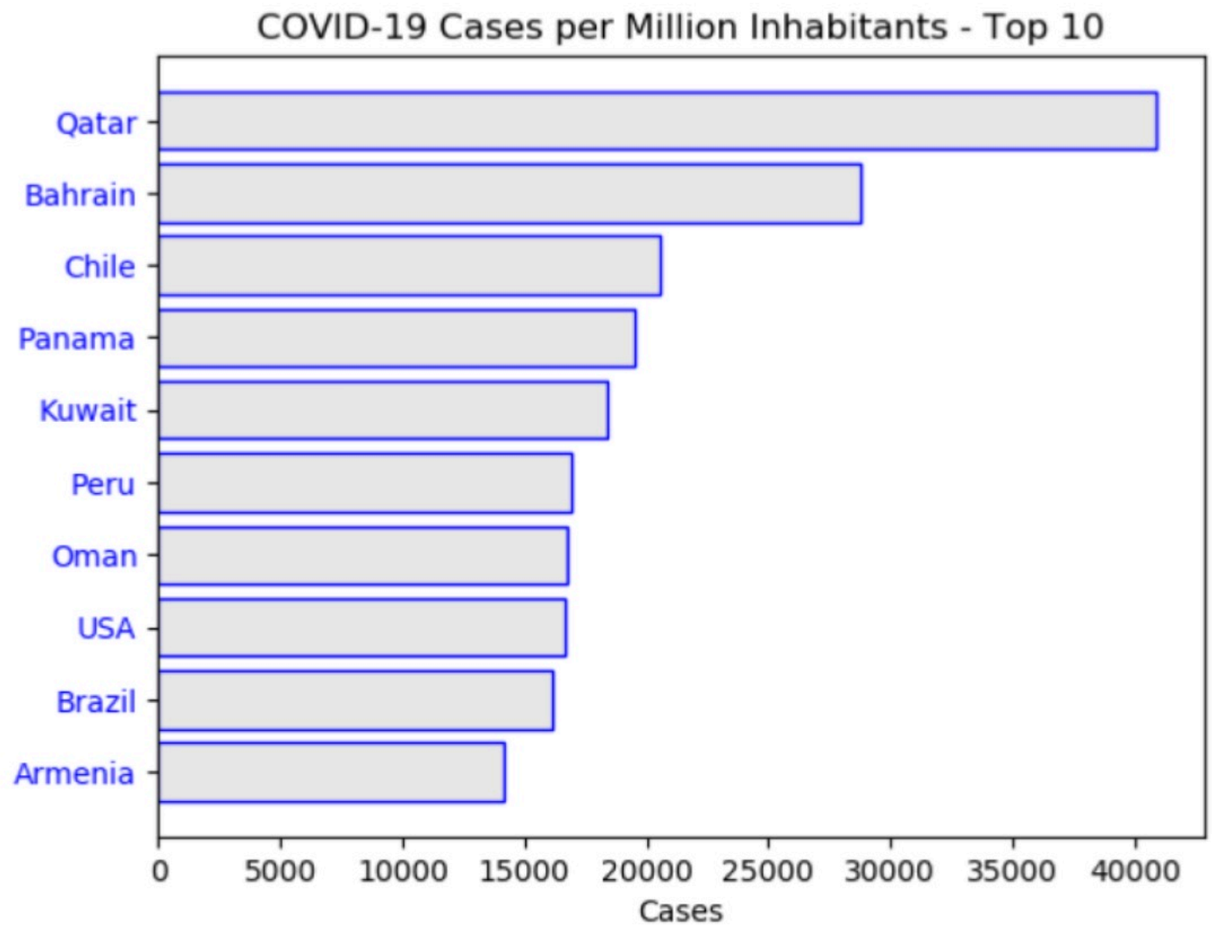
Bar Charts

continent	cases	deaths
Africa	1119579	26260
America	11698368	427207
Asia	5606210	122034
Europe	3239237	205144
Oceania	25742	471



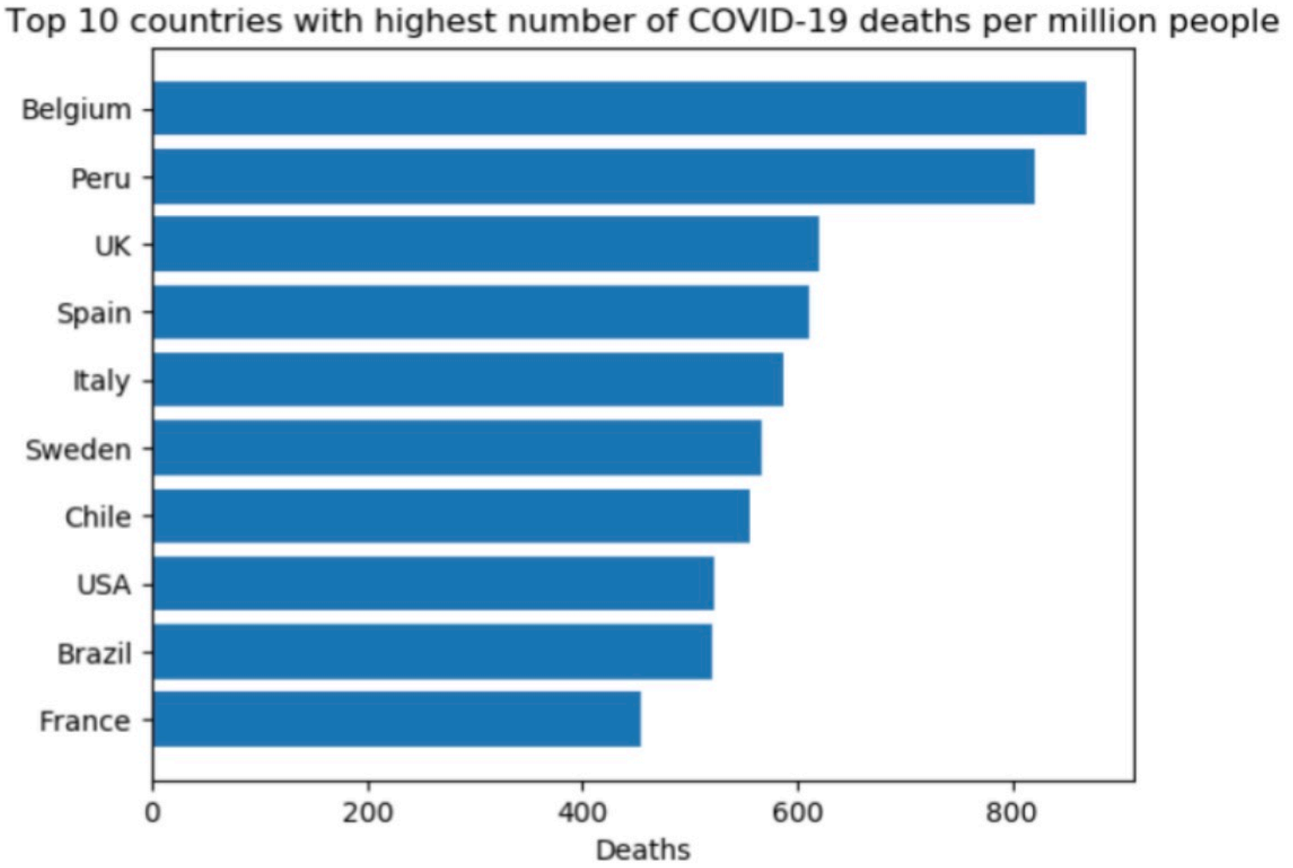
Bar Charts

countriesAndTerritories	cases	deaths	popData2019	casesPer1M
Qatar	115661	193	2832071.0	40839.724710
Bahrain	47185	175	1641164.0	28750.935312
Chile	388855	10546	18952035.0	20517.849402
Panama	82790	1809	4246440.0	19496.331044
Kuwait	77470	505	4207077.0	18414.210151
Peru	549321	26658	32510462.0	16896.745423
Oman	83418	597	4974992.0	16767.464149
USA	5482416	171821	329064917.0	16660.591016
Brazil	3407354	109888	211049519.0	16144.808177
Armenia	41846	832	2957728.0	14148.021725



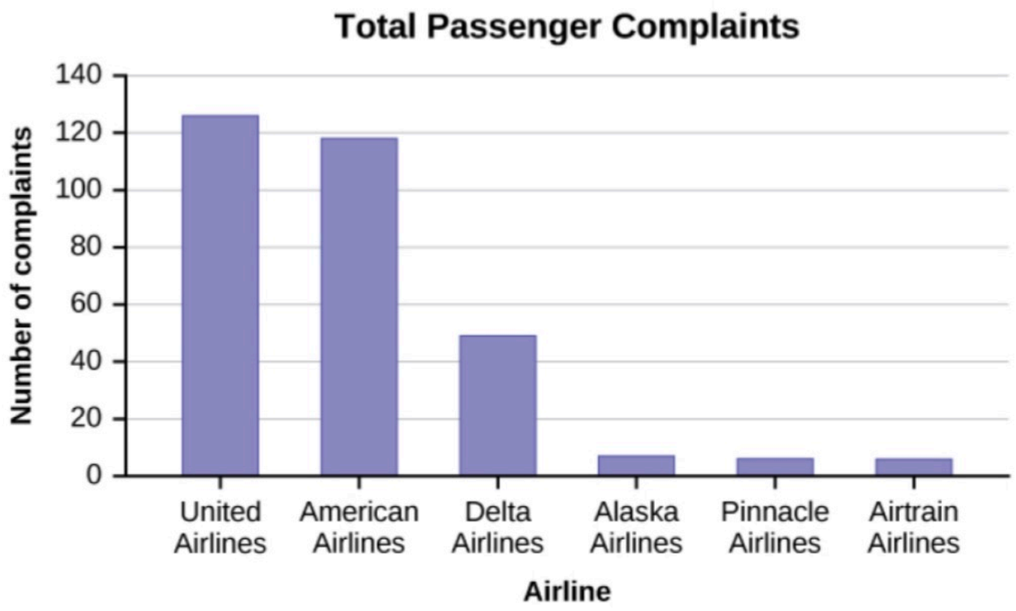
Bar Charts

countriesAndTerritories	cases	deaths	popData2019	casesPer1M	deathsPer1M
Belgium	78804	9959	11455519.0	6879.129614	869.362619
Peru	549321	26658	32510462.0	16896.745423	819.982195
UK	320286	41381	66647112.0	4805.699608	620.897122
Spain	364196	28670	46937060.0	7759.241844	610.817976
Italy	254636	35405	60359546.0	4218.653334	586.568362
Sweden	85219	5790	10230185.0	8330.152387	565.972170
Chile	388855	10546	18952035.0	20517.849402	556.457394
USA	5482416	171821	329064917.0	16660.591016	522.149251
Brazil	3407354	109888	211049519.0	16144.808177	520.674013
France	221267	30451	67012883.0	3301.857644	454.405163



Let's Practice

The graph shows the number of complaints for six different airlines as reported to the US Department of Transportation in February 2013. Alaska, Pinnacle, and Airtran Airlines have far fewer complaints reported than American, Delta, and United.



Can we conclude that American, Delta, and United are the worst airline carriers since they have the most complaints?

Students, write your response!

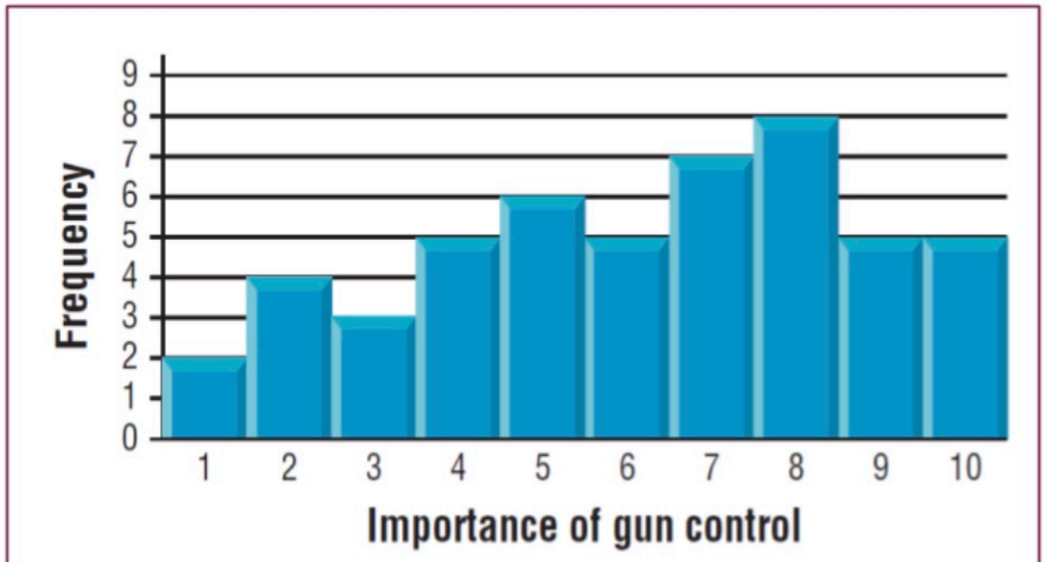
No Text Response

You didn't answer this question

Histograms

- Used with numerical variables.
- Represent the frequency of each attribute for a variable.
- Good overview of the **distribution** of your data

On a Scale of 1 to 10, How Important Is Gun Control to You? (N = 50)

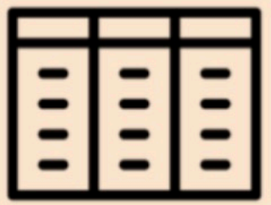


Histogram link

Histograms

1 Divide the range of the data into intervals of equal width. For a discrete variable with few values, use the actual possible values.

2 Count the number of observations (the frequency) in each interval, forming a frequency table.



3 Draw a bar over each value or interval with height equal to its frequency (or percentage), values of which are marked on the vertical axis.

7 Populations & Samples

Populations

The study of statistics revolves around the study of data sets.

Populations → Include each element from the set of observations that can be made.

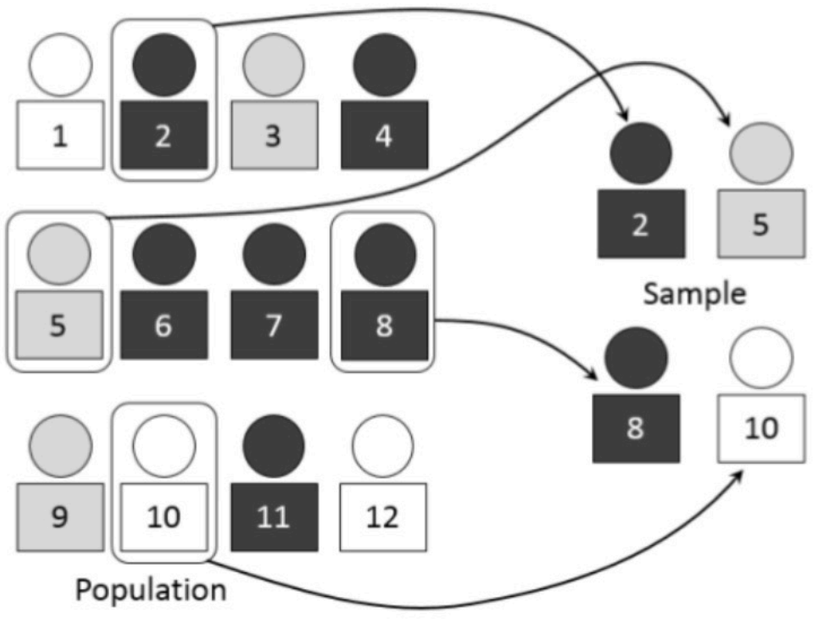
The set of all stars within the Milky Way galaxy

All people living in the USA

Samples

Populations and samples are data sets.

Samples → Include one or more observations from the population.



The elements of a sample are known as sample points, or observations.

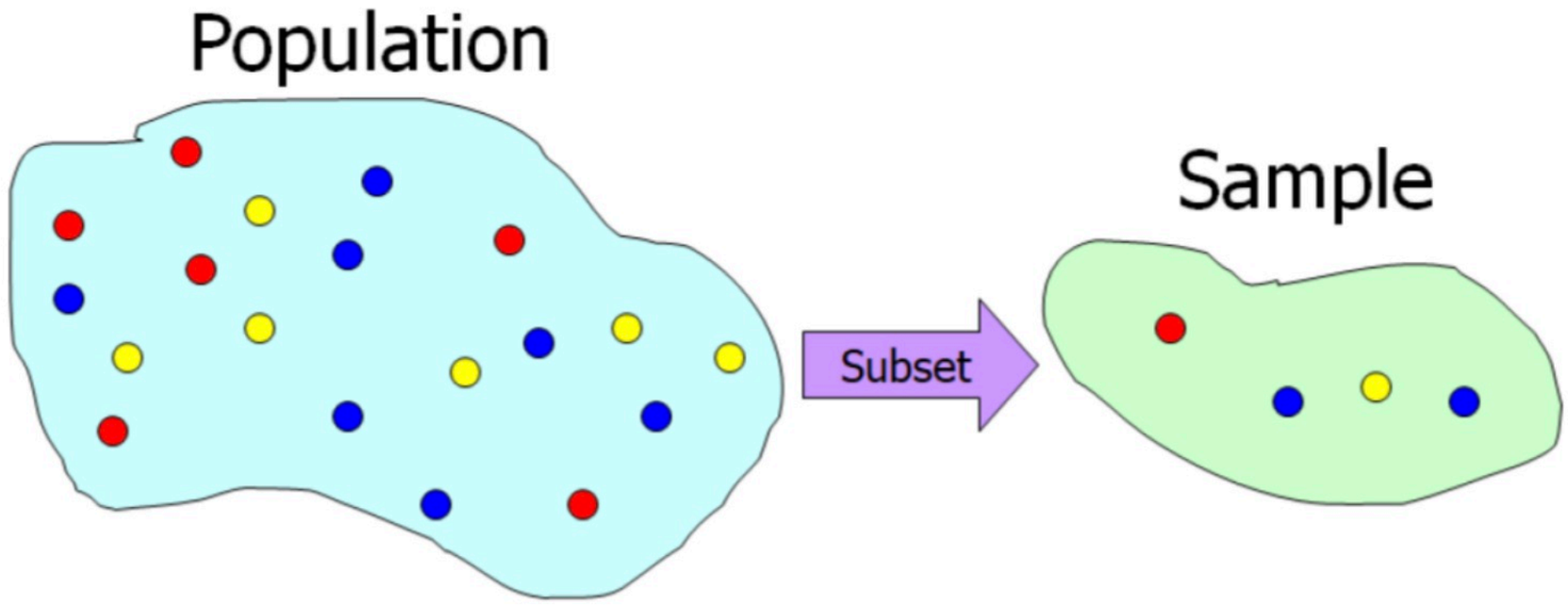
Parameters & Statistics

Population attributes → **Parameters**

Sample attributes → **Statistics**

Sample statistics are often used to estimate population parameters

Parameters & Statistics



- Populations have Parameters (like μ, σ^2, θ, p)
- Samples have Statistics, functions of observed data, like $\bar{x}, \tilde{x}, s^2, \hat{\theta}, \hat{p}$

7 Sampling Techniques

Sampling Techniques

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole.

Probability sampling — involves random selection, allowing you to make strong statistical inferences about the whole group.

Non-probability sampling — involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Probability Sampling Methods

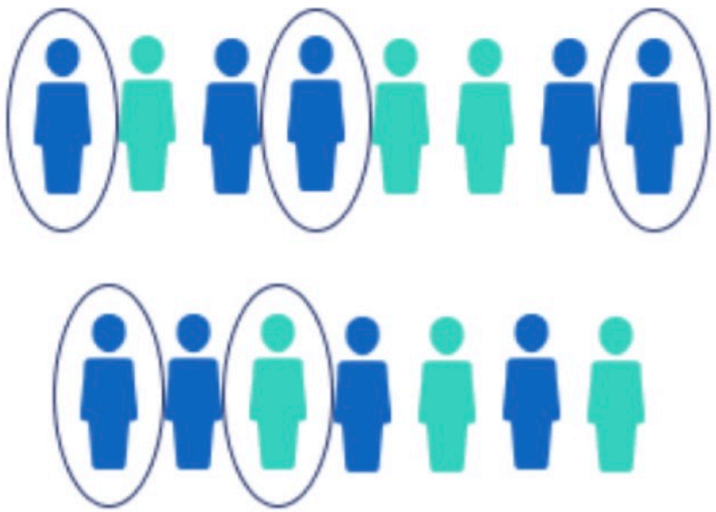
Probability sampling means that every member of the population has a chance of being selected.

- Simple random sample
- Systematic sample
- Stratified sample
- Cluster sample

Simple Random Sample

In a simple random sample, every member of the population has an equal chance of being selected.

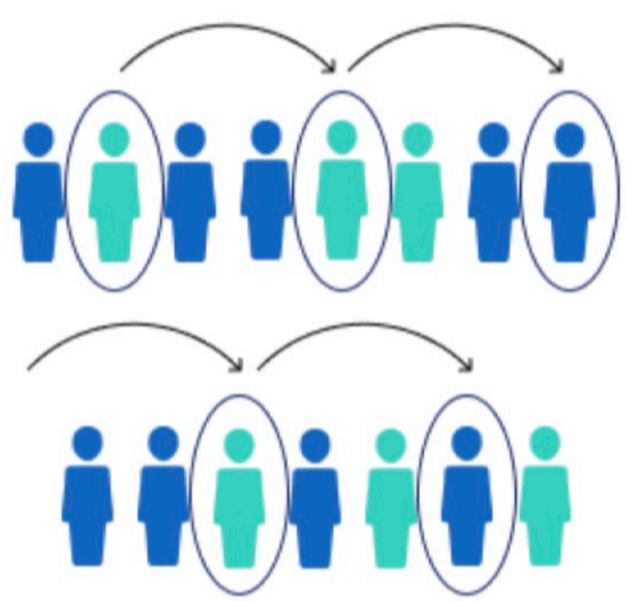
Simple random sample



Systematic Sample

Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

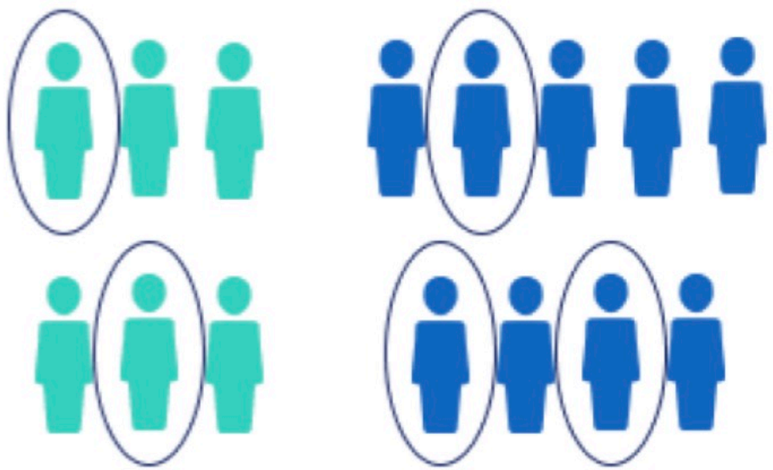
Systematic sample



Stratified Sample

Stratified sampling involves dividing the population into subpopulations that may differ in important ways.

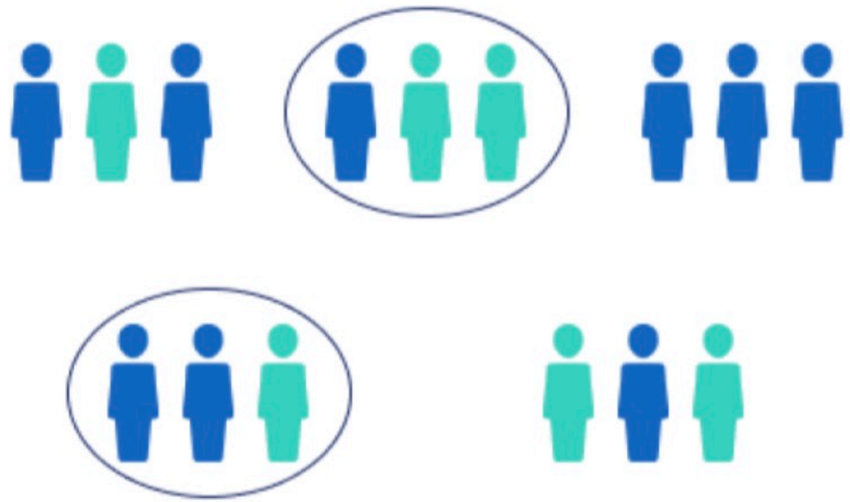
Stratified sample



Cluster Sample

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample.

Cluster sample



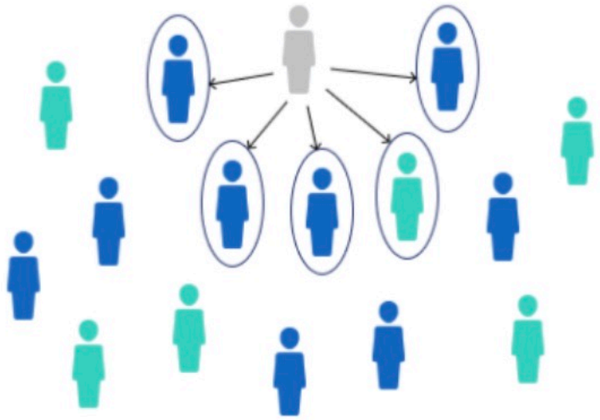
Non-probability Sampling Methods

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

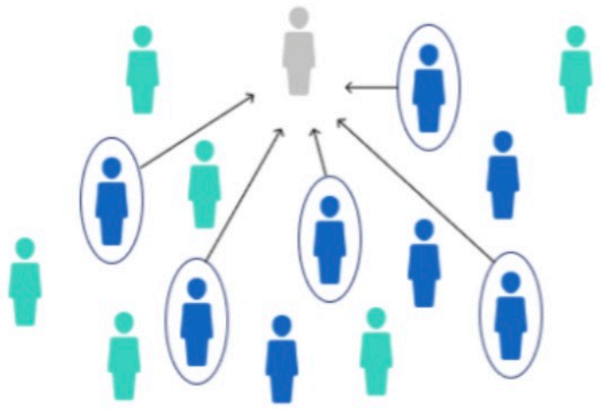
- Convenience sample
- Voluntary response sample
- Purposive sample
- Snowball sample

Non-probability Sampling Methods

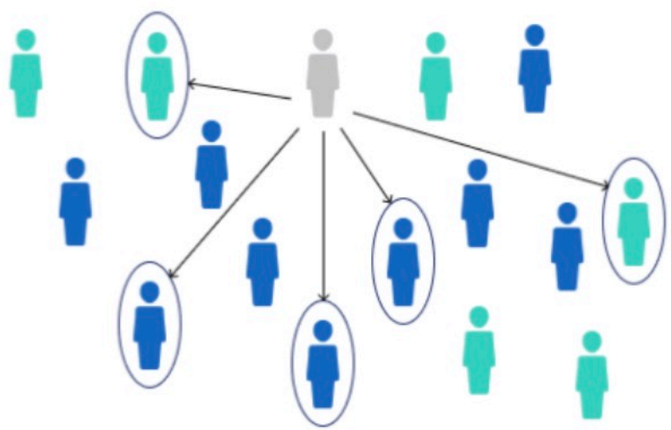
Convenience sample



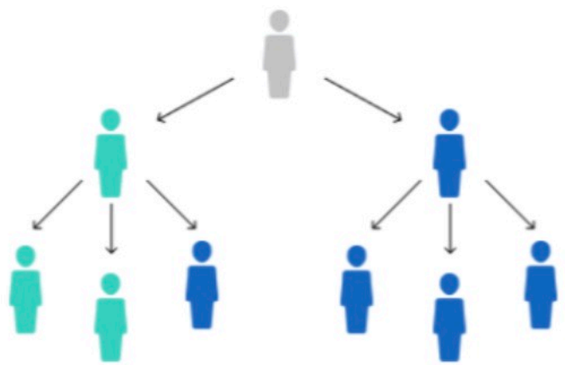
Voluntary response sample



Purposive sample



Snowball sample

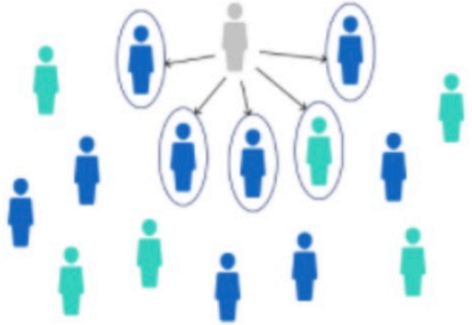


Researchers use this method in studies where it is impossible to draw random probability sampling due to time or cost considerations.

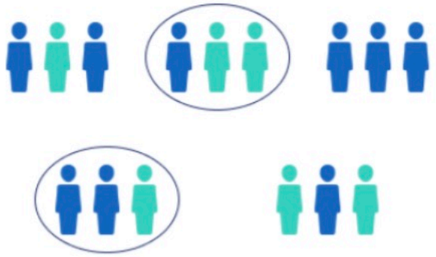
Let's Practice

Which of the following will give a more "accurate" representation of the population from which a sample has been taken?

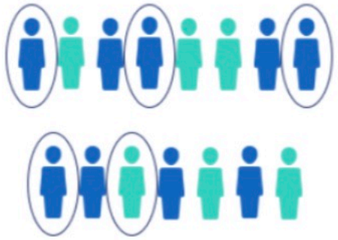
Convenience sample



Cluster sample



Simple random sample



Students choose an option

No Multiple Choice Response

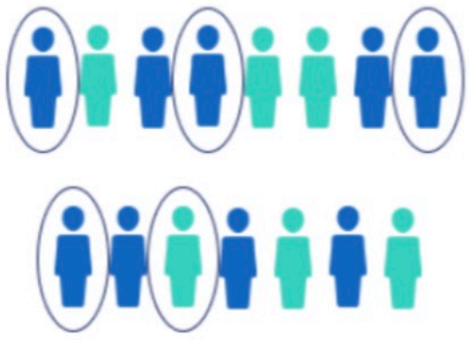
You didn't answer this question

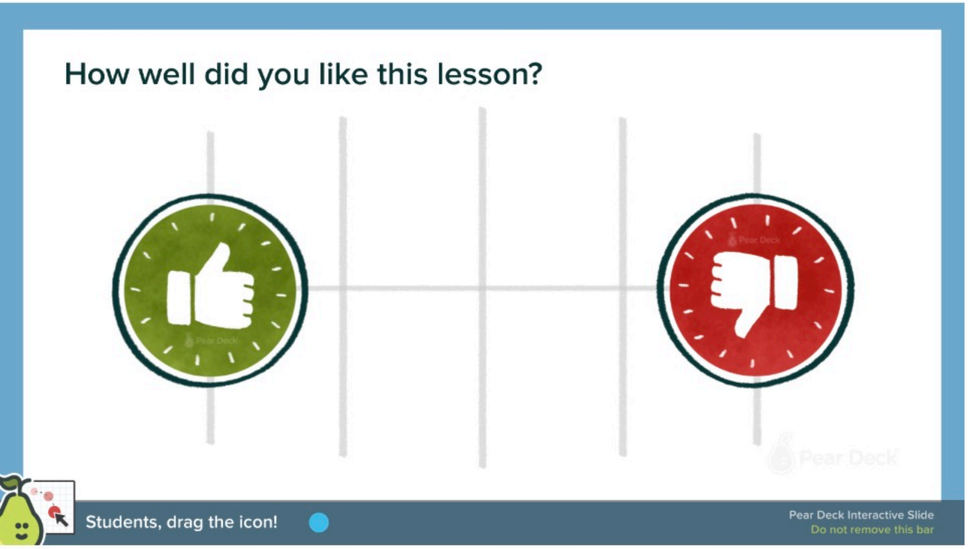
Let's Practice Answer

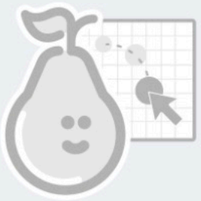
Which of the following will give a more "accurate" representation of the population from which a sample has been taken?

A large sample based on simple random sampling

Simple random sample







No Draggable™ Response

You didn't answer this question