MATH 4720 Introduction to Statistics

Overview of Statistics and Data 📖

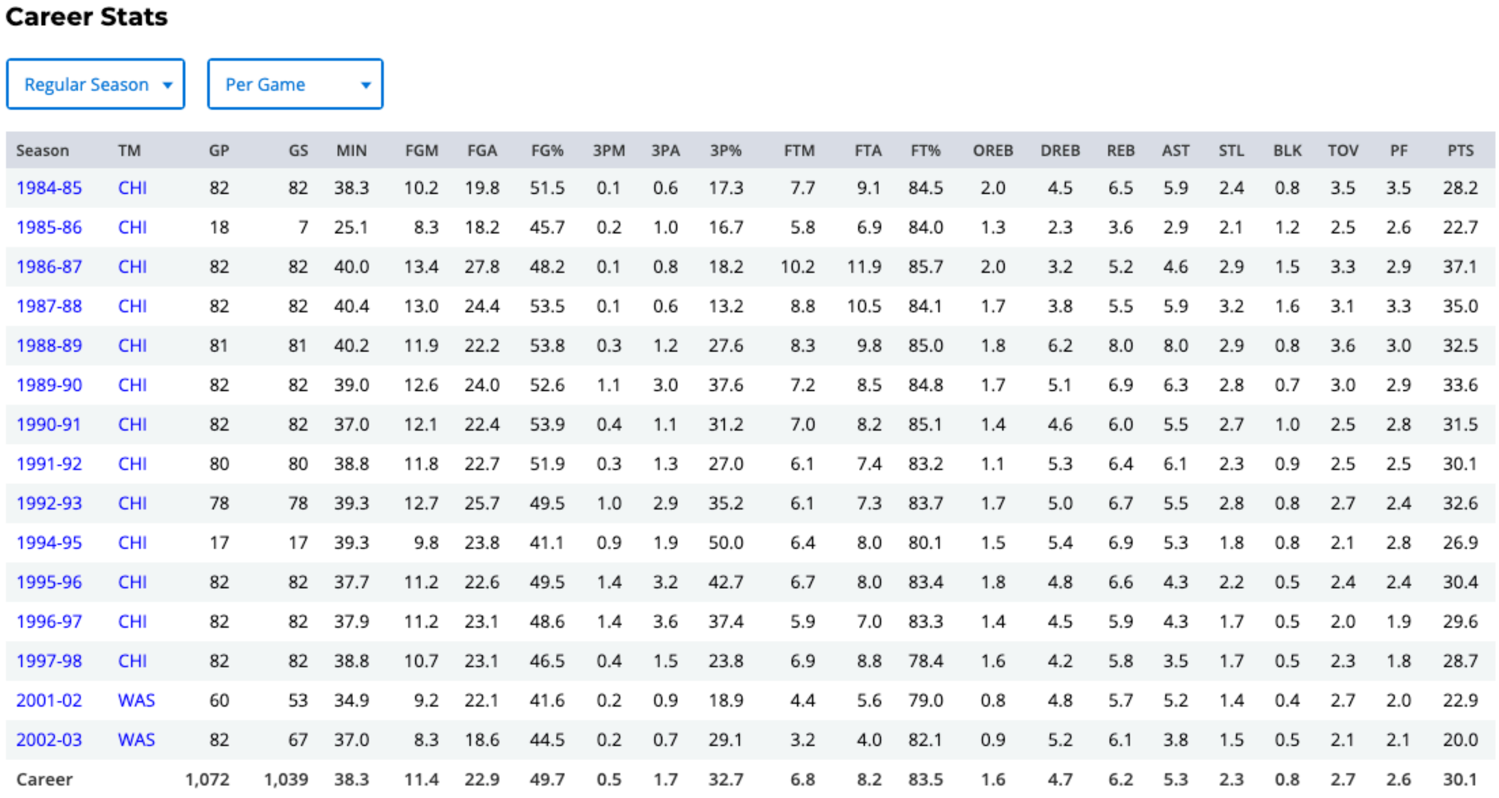
Dr. Cheng-Han Yu
Department of Mathematical and Statistical Sciences
Marquette University

August 31 2021

What is Statisitcs

Statistics as Numeric Records

- In ordinary conversations, the word **statistics** is used as a term to indicate a set or collection of **numeric records**.



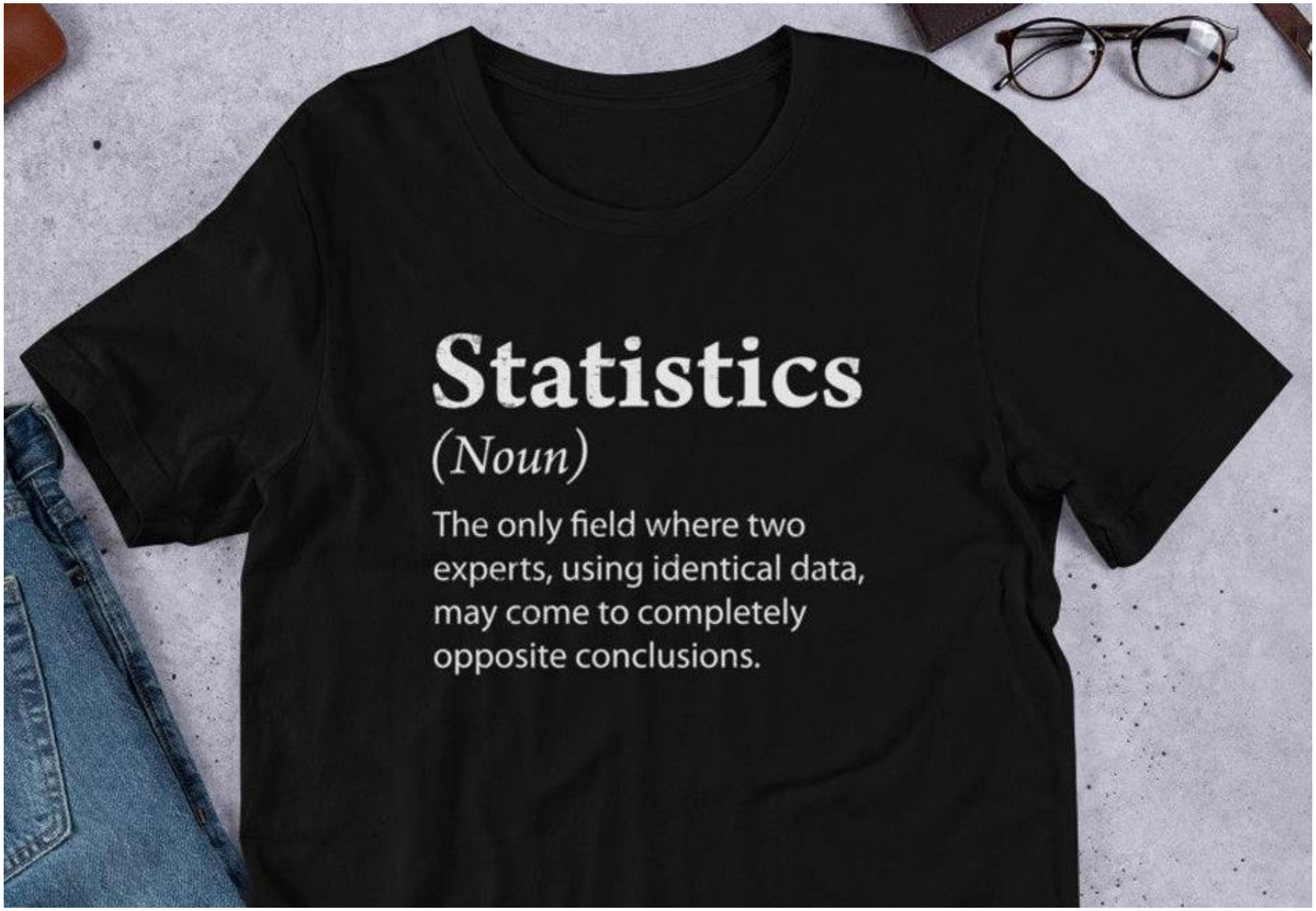
Career Stats

Season	TM	GP	GS	MIN	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	STL	BLK	TOV	PF	PTS
1984-85	CHI	82	82	38.3	10.2	19.8	51.5	0.1	0.6	17.3	7.7	9.1	84.5	2.0	4.5	6.5	5.9	2.4	0.8	3.5	3.5	28.2
1985-86	CHI	18	7	25.1	8.3	18.2	45.7	0.2	1.0	16.7	5.8	6.9	84.0	1.3	2.3	3.6	2.9	2.1	1.2	2.5	2.6	22.7
1986-87	CHI	82	82	40.0	13.4	27.8	48.2	0.1	0.8	18.2	10.2	11.9	85.7	2.0	3.2	5.2	4.6	2.9	1.5	3.3	2.9	37.1
1987-88	CHI	82	82	40.4	13.0	24.4	53.5	0.1	0.6	13.2	8.8	10.5	84.1	1.7	3.8	5.5	5.9	3.2	1.6	3.1	3.3	35.0
1988-89	CHI	81	81	40.2	11.9	22.2	53.8	0.3	1.2	27.6	8.3	9.8	85.0	1.8	6.2	8.0	8.0	2.9	0.8	3.6	3.0	32.5
1989-90	CHI	82	82	39.0	12.6	24.0	52.6	1.1	3.0	37.6	7.2	8.5	84.8	1.7	5.1	6.9	6.3	2.8	0.7	3.0	2.9	33.6
1990-91	CHI	82	82	37.0	12.1	22.4	53.9	0.4	1.1	31.2	7.0	8.2	85.1	1.4	4.6	6.0	5.5	2.7	1.0	2.5	2.8	31.5
1991-92	CHI	80	80	38.8	11.8	22.7	51.9	0.3	1.3	27.0	6.1	7.4	83.2	1.1	5.3	6.4	6.1	2.3	0.9	2.5	2.5	30.1
1992-93	CHI	78	78	39.3	12.7	25.7	49.5	1.0	2.9	35.2	6.1	7.3	83.7	1.7	5.0	6.7	5.5	2.8	0.8	2.7	2.4	32.6
1994-95	CHI	17	17	39.3	9.8	23.8	41.1	0.9	1.9	50.0	6.4	8.0	80.1	1.5	5.4	6.9	5.3	1.8	0.8	2.1	2.8	26.9
1995-96	CHI	82	82	37.7	11.2	22.6	49.5	1.4	3.2	42.7	6.7	8.0	83.4	1.8	4.8	6.6	4.3	2.2	0.5	2.4	2.4	30.4
1996-97	CHI	82	82	37.9	11.2	23.1	48.6	1.4	3.6	37.4	5.9	7.0	83.3	1.4	4.5	5.9	4.3	1.7	0.5	2.0	1.9	29.6
1997-98	CHI	82	82	38.8	10.7	23.1	46.5	0.4	1.5	23.8	6.9	8.8	78.4	1.6	4.2	5.8	3.5	1.7	0.5	2.3	1.8	28.7
2001-02	WAS	60	53	34.9	9.2	22.1	41.6	0.2	0.9	18.9	4.4	5.6	79.0	0.8	4.8	5.7	5.2	1.4	0.4	2.7	2.0	22.9
2002-03	WAS	82	67	37.0	8.3	18.6	44.5	0.2	0.7	29.1	3.2	4.0	82.1	0.9	5.2	6.1	3.8	1.5	0.5	2.1	2.1	20.0
Career		1,072	1,039	38.3	11.4	22.9	49.7	0.5	1.7	32.7	6.8	8.2	83.5	1.6	4.7	6.2	5.3	2.3	0.8	2.7	2.6	30.1

Statistics as Numeric Records

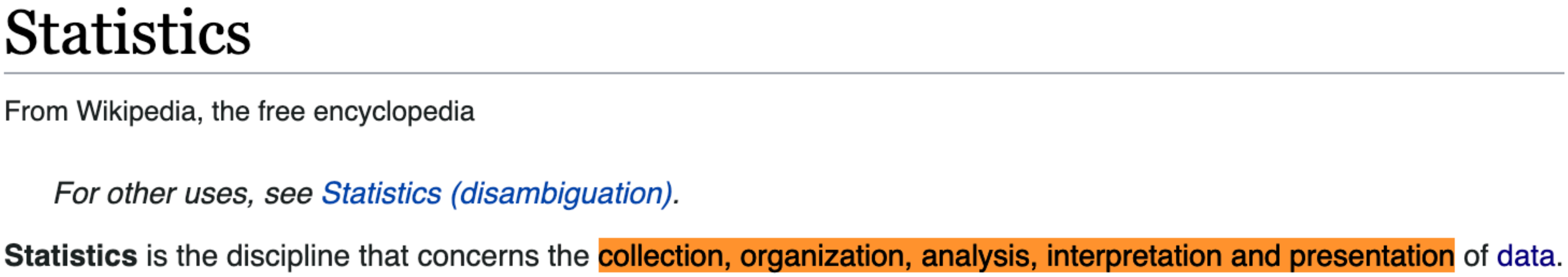
- In ordinary conversations, the word **statistics** is used as a term to indicate a set or collection of **numeric records**.

https://slamgoods.com/products/jordan-collectors-issue


Statistics
(Noun)
The only field where two experts, using identical data, may come to completely opposite conclusions.

shorturl.at/huyLS

Statistics as a Discipline



Statistics

From Wikipedia, the free encyclopedia

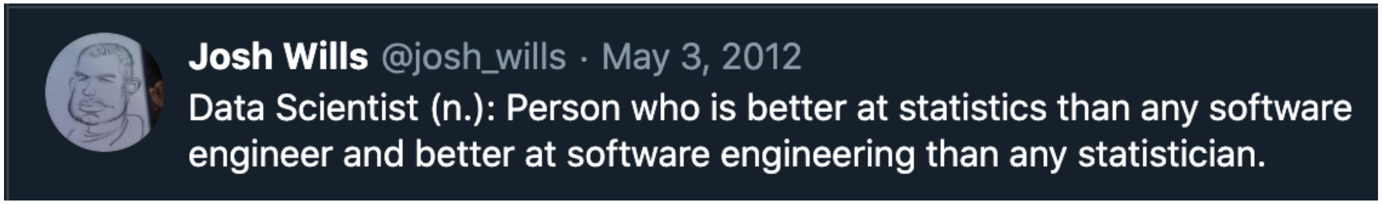
For other uses, see Statistics (disambiguation).

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

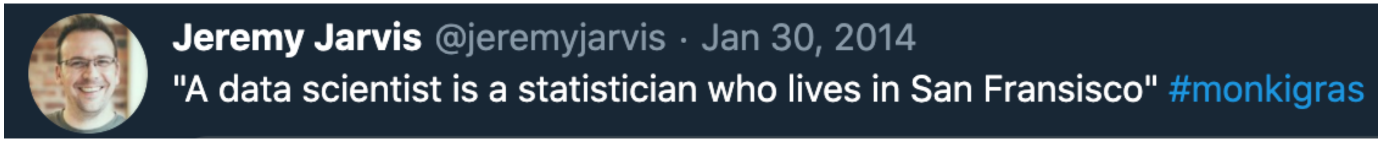
- **Statistics** is a **Science of Data**.
- A ***science of data*** using **statistical thinking, methods and models**.

🤔 But wait, then what is **DATA SCIENCE** ❓

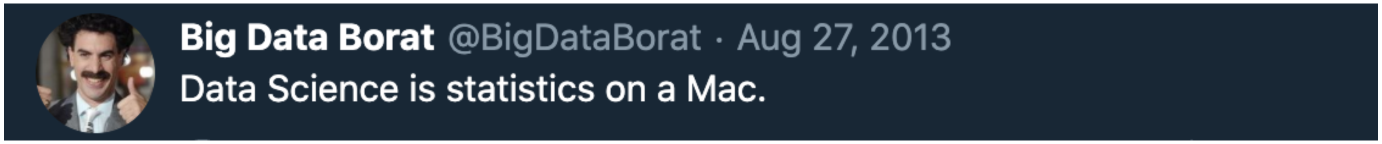
Difference between Statistics and Data Science



Josh Wills @josh_wills · May 3, 2012
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

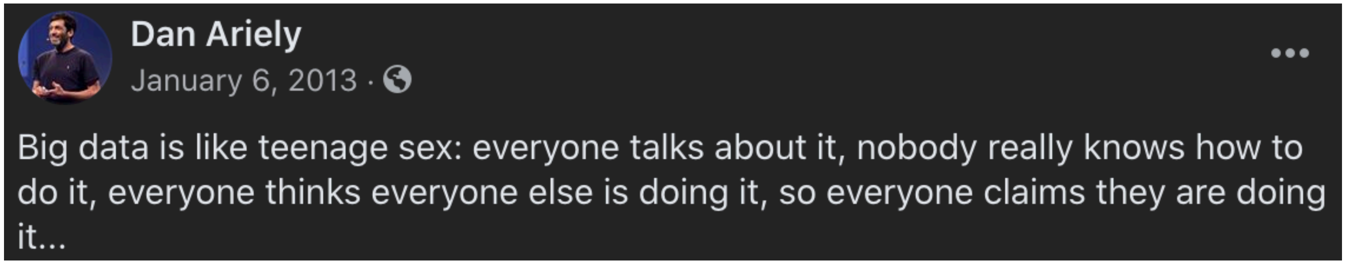


Jeremy Jarvis @jeremyjarvis · Jan 30, 2014
"A data scientist is a statistician who lives in San Fransisco" #monkigras



Big Data Borat @BigDataBorat · Aug 27, 2013
Data Science is statistics on a Mac.

- Investopedia defines Data Science as a field of **Big Data** which seeks to provide meaningful information from large amounts of complex data.



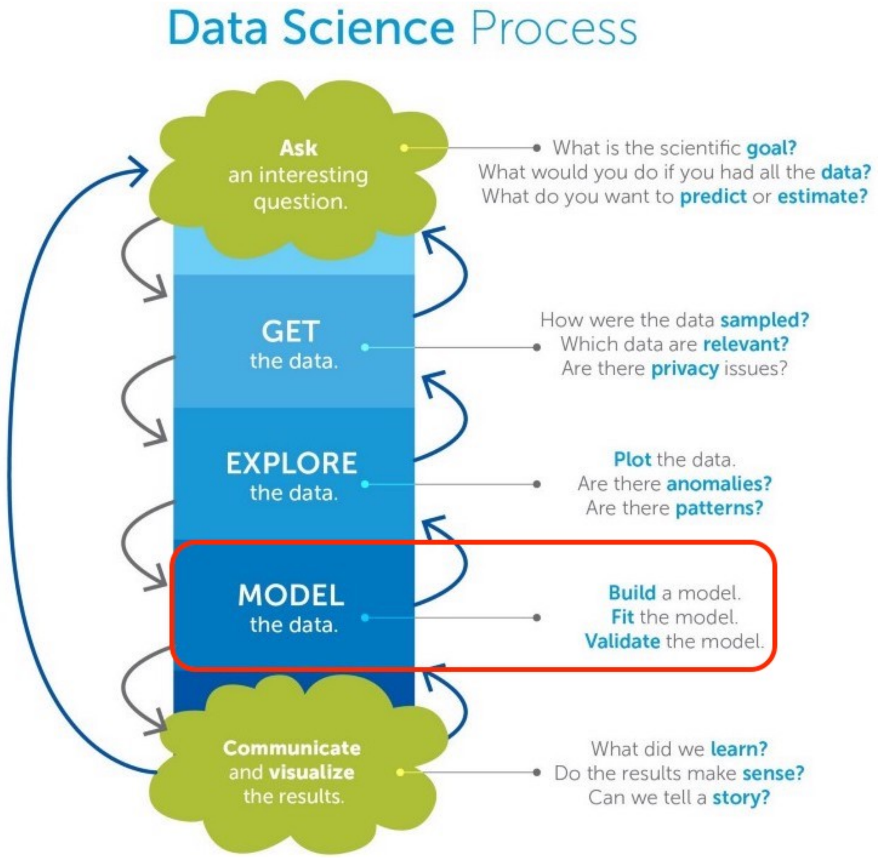
Dan Ariely
January 6, 2013 ·

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

UC Santa Cruz Department of Statistics Courses

Data Science Is Now a Broader View of Statistics

- Collection, organization, analysis, interpretation and presentation of data.


Data Science Process
Ask an interesting question.
GET the data.
EXPLORE the data.
MODEL the data.
Communicate and visualize the results.

What is the scientific **goal?**
What would you do if you had all the **data?**
What do you want to **predict** or **estimate?**

How were the data **sampled?**
Which data are **relevant?**
Are there **privacy** issues?

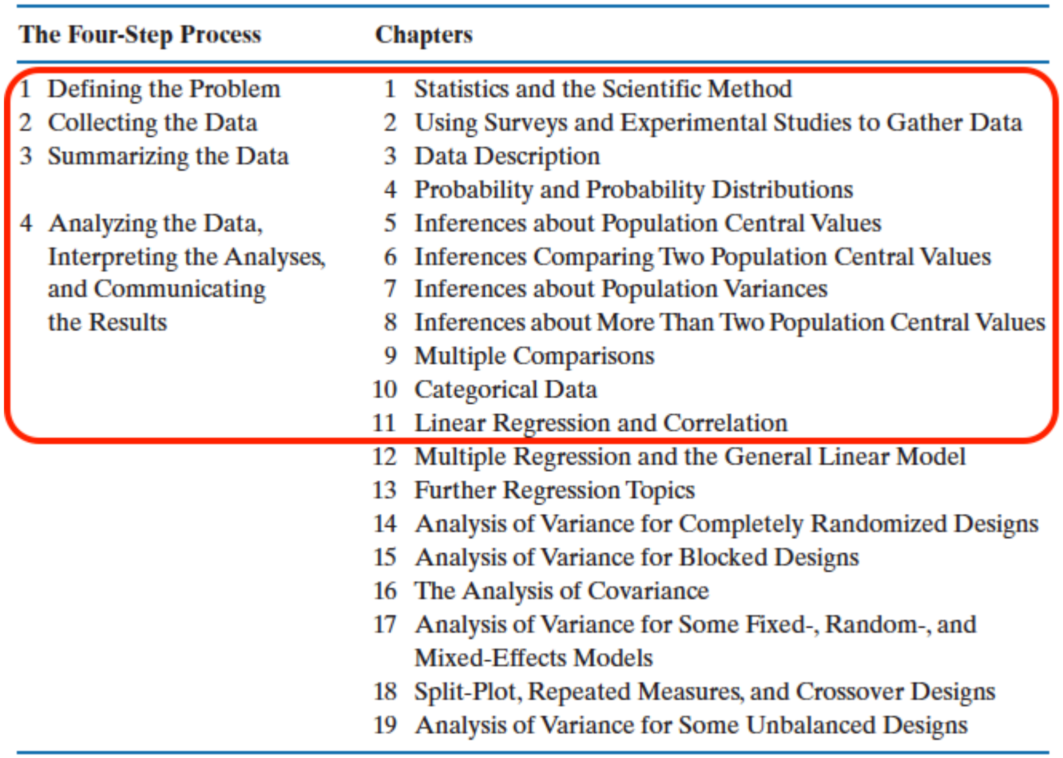
Plot the data.
Are there **anomalies?**
Are there **patterns?**

Build a model.
Fit the model.
Validate the model.

What did we **learn?**
Do the results make **sense?**
Can we tell a **story?**

What Do We Learn In this Course

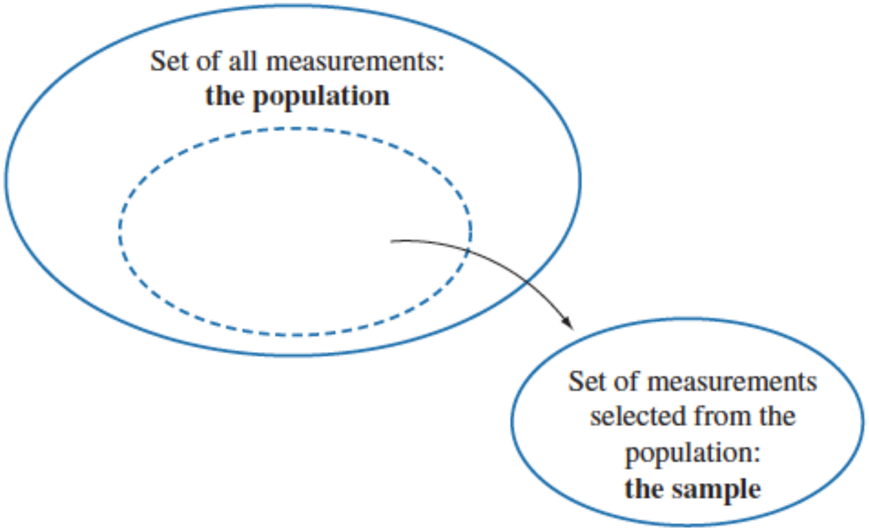
- We will discuss most of materials in Chapter 1 to Chapter 11.



The Four-Step Process	Chapters
1 Defining the Problem	1 Statistics and the Scientific Method
2 Collecting the Data	2 Using Surveys and Experimental Studies to Gather Data
3 Summarizing the Data	3 Data Description
	4 Probability and Probability Distributions
4 Analyzing the Data, Interpreting the Analyses, and Communicating the Results	5 Inferences about Population Central Values
	6 Inferences Comparing Two Population Central Values
	7 Inferences about Population Variances
	8 Inferences about More Than Two Population Central Values
	9 Multiple Comparisons
	10 Categorical Data
	11 Linear Regression and Correlation
	12 Multiple Regression and the General Linear Model
	13 Further Regression Topics
	14 Analysis of Variance for Completely Randomized Designs
	15 Analysis of Variance for Blocked Designs
	16 The Analysis of Covariance
	17 Analysis of Variance for Some Fixed-, Random-, and Mixed-Effects Models
	18 Split-Plot, Repeated Measures, and Crossover Designs
	19 Analysis of Variance for Some Unbalanced Designs

We Focus On Statistical Inference

- We spend most of time on various statistical methods for analyzing data. (Chapter 4 to 11)
- Learn useful information
  - about the **population** we are interested
  - from our **sample data**
  - through **statistical inferential** methods, including **estimation** and **testing**

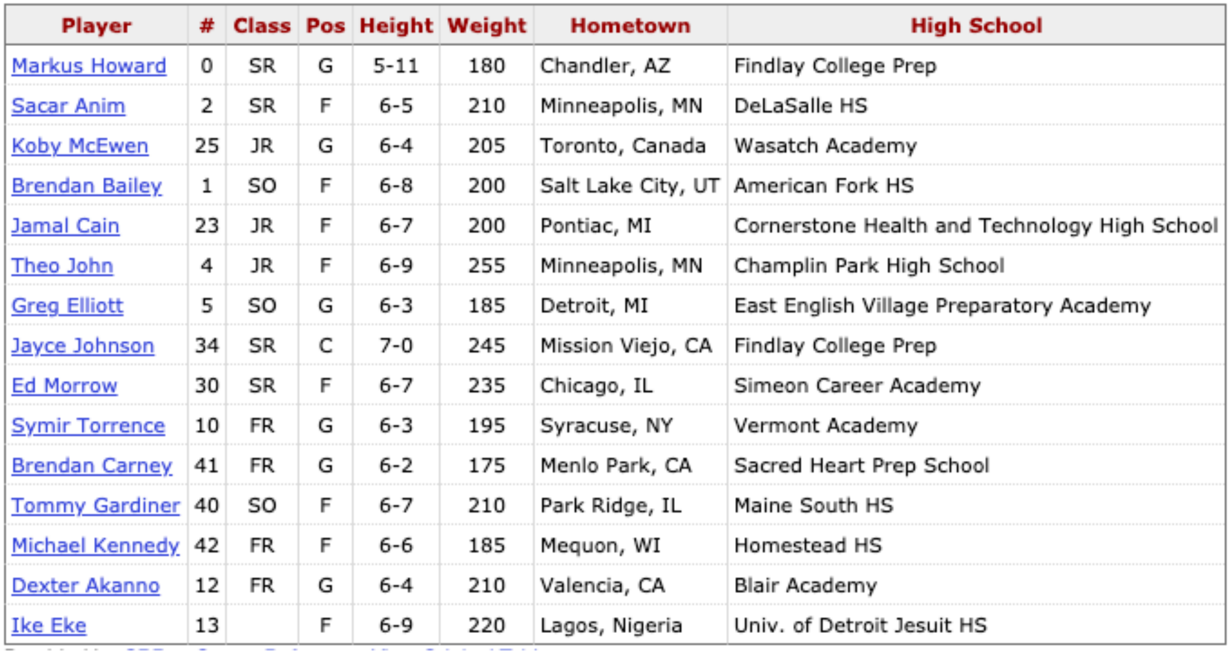


Statistics is a Science of Data, so What is Data?

- **Data**: A set of **objects** on which we observe or measure one or more **characteristics**.
- Objects are individuals, observations, subjects or cases in statistical studies.
- A characteristic or attribute is called a **variable** because it *varies* from one to another.

Data Matrix

- Each row corresponds to a unique case or observational unit.
- Each column represents a characteristic or variable.
- This structure allows new cases to be added as rows or new variables as new columns.



Player	#	Class	Pos	Height	Weight	Hometown	High School
Markus Howard	0	SR	G	5-11	180	Chandler, AZ	Findlay College Prep
Sacar Anim	2	SR	F	6-5	210	Minneapolis, MN	DeLaSalle HS
Koby McEwen	25	JR	G	6-4	205	Toronto, Canada	Wasatch Academy
Brendan Bailey	1	SO	F	6-8	200	Salt Lake City, UT	American Fork HS
Jamal Cain	23	JR	F	6-7	200	Pontiac, MI	Cornerstone Health and Technology High School
Theo John	4	JR	F	6-9	255	Minneapolis, MN	Champlin Park High School
Greg Elliott	5	SO	G	6-3	185	Detroit, MI	East English Village Preparatory Academy
Jayce Johnson	34	SR	C	7-0	245	Mission Viejo, CA	Findlay College Prep
Ed Morrow	30	SR	F	6-7	235	Chicago, IL	Simeon Career Academy
Symir Torrence	10	FR	G	6-3	195	Syracuse, NY	Vermont Academy
Brendan Carney	41	FR	G	6-2	175	Menlo Park, CA	Sacred Heart Prep School
Tommy Gardiner	40	SO	F	6-7	210	Park Ridge, IL	Maine South HS
Michael Kennedy	42	FR	F	6-6	185	Mequon, WI	Homestead HS
Dexter Akanno	12	FR	G	6-4	210	Valencia, CA	Blair Academy
Ike Eke	13		F	6-9	220	Lagos, Nigeria	Univ. of Detroit Jesuit HS

Population and Sample

Target Population

- The first step in conducting a study is to *identify questions* to be investigated.
- A clear research question is helpful in identifying
  - what *cases* should be studied (row)
  - what *variables* are important (column)
- Target **Population**: The **complete** collection of data we'd like to make inference about.

- *What is the average GPA of currently enrolled Marquette undergrads?*

- All Marquette undergrads that are currently enrolled.

Target Population

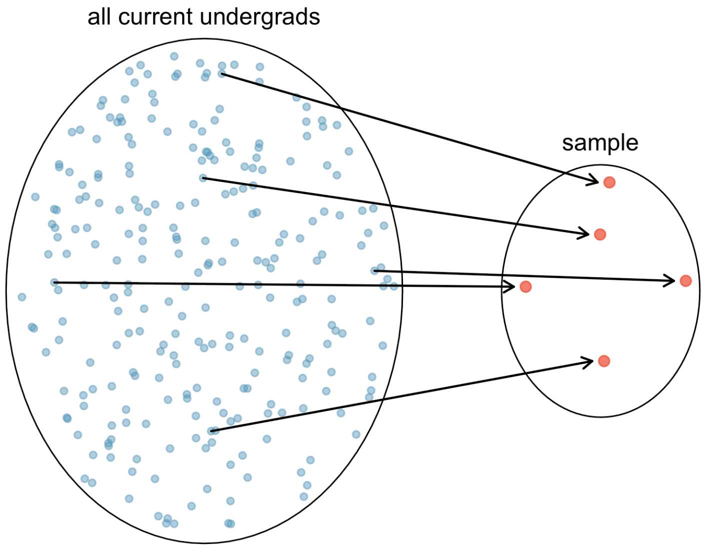
- The first step in conducting a study is to *identify questions* to be investigated.
- A clear research question is helpful in identifying
  - what *cases* should be studied (row)
  - what *variables* are important (column)
- Target **Population**: The **complete** collection of data we'd like to make inference about.

- *Does a new drug reduce mortality in patients with severe heart disease?*

- All people with severe heart disease.

Sample Data

- Sometimes, it's possible to collect data of all cases we are interested.
- Most of the time, it is too expensive to collect data for every case in a population.
- What about the average GPA of undergrads in Illinois? the U.S.? the world? 😱 😱 😱



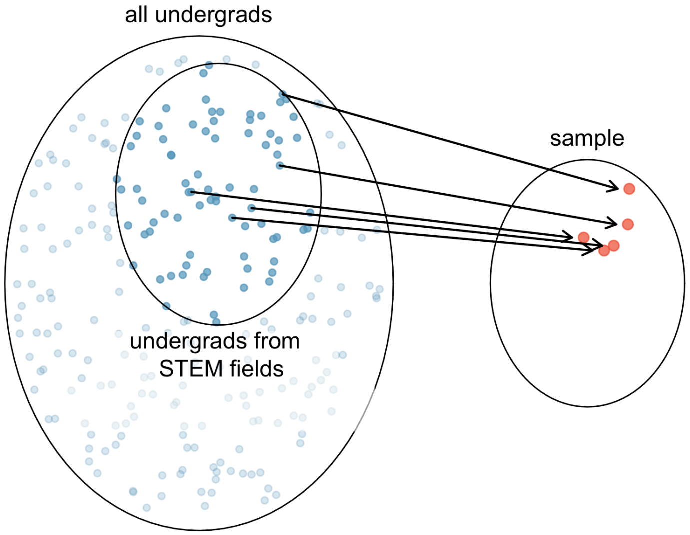
- **Sample**: A **subset** of cases selected from a population.
- Compute the average GPA of the sample data
- Hope sample avg GPA $\approx$ population avg GPA. 🙏

Good Sample vs. Bad Sample

> Is **this 4720 class** a sample data of the target population Marquette undergrads?

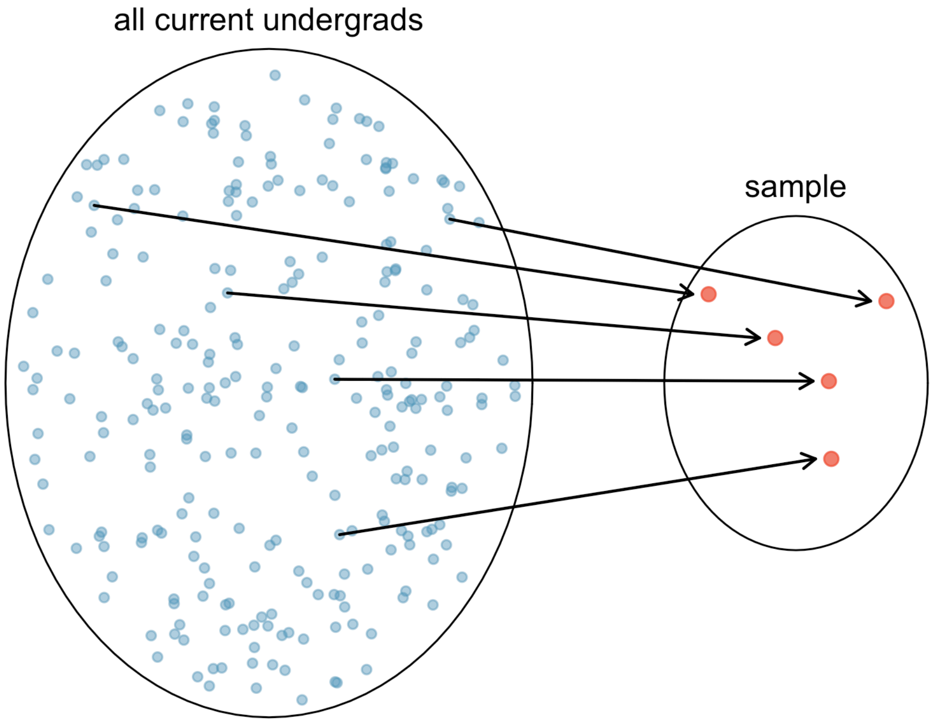
> Is **this 4720 class** a *"good"* sample of the target population?

- The sample is convenient to be collected, but it is NOT **representative** of the population.
- This 4720 class is a **biased** sample. The average GPA of you guys may not be close to the average GPA of all Marquette undergrads.



How and Why a Representative Sample?

- We always seek to **randomly** select a sample from a population.
- Almost all statistical methods are based on randomness assumption.



Data Collection

Two Types of Studies to Collect Sample Data

- **Observational Study**: Observe and measure characteristics/variables, and do **NOT** attempt to modify or intervene with the subjects being studied.
    - Sample from 1️⃣ the heart disease population and 2️⃣ heart disease-free population and record the fat content of the diets for the two groups.

- **Experimental Study**: Apply some **treatment(s)** and then proceed to observe its responses or effects on the individuals (experimental units).
    - Assign volunteers to one of several diets with different levels of dietary fat (treatments) and compare the treatments with respect to the incidence of hear disease after a period of time.

Limitation of Observational Studies: Confounding

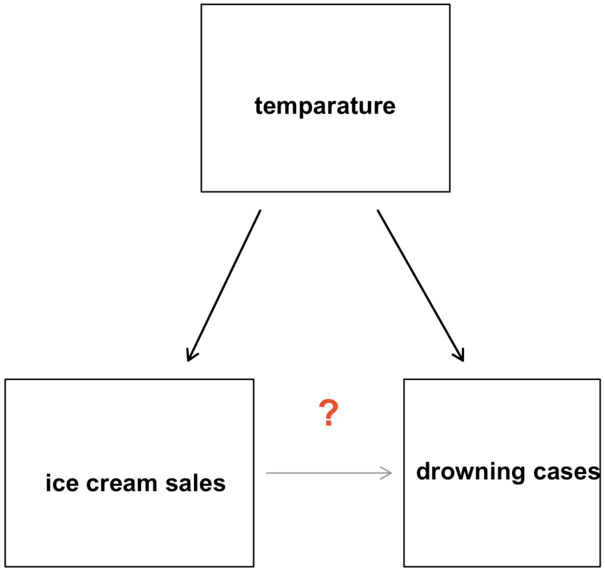
- **Confounder**: A variable NOT included in a study but affects the variables in the study.
- Observe past data show that increases in ice cream sales are associated with increases in drownings, and we conclude that **ice cream causes drownings**. 😱 😕 ⁉️

> What is the confounder that is not in the data, but affects ice cream sales and the number of drownings?

Temperature: as temperature increases, ice cream sales increase and the number of drownings goes up because more people swim.

Causal Relationship

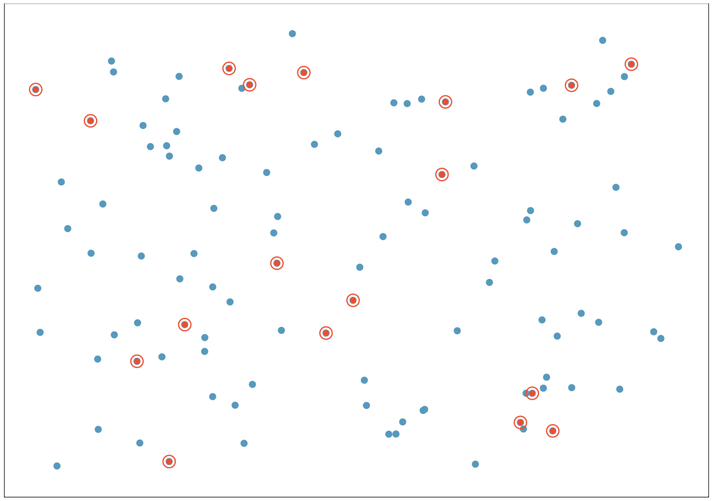
- Making causal conclusions based on *experiments* is often more reasonable than making the same causal conclusions based on observational data.
- Observational studies are generally only sufficient to show **associations, not causality**.



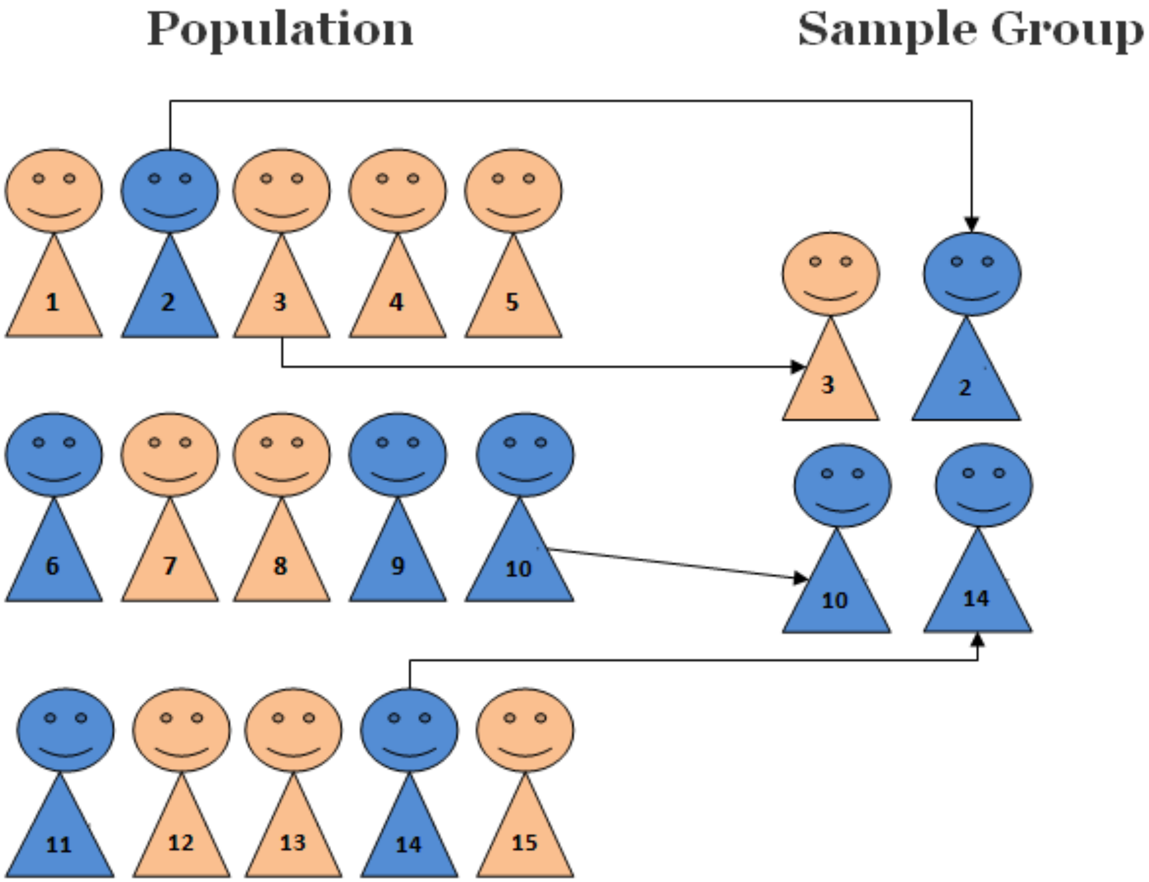
Sampling Methods

Simple Random Sample

- **Random Sample**: Each member of a population is **equally likely** to be selected.
- **Simple Random Sample (SRS)**: Every possible sample of sample size n has the same chance to be chosen.
- **Example**: If I want to sample 100 students from all, say 10,000 Marquette students, I would randomly assign each student a number (from 1 to 10,000), then randomly select 100 numbers.



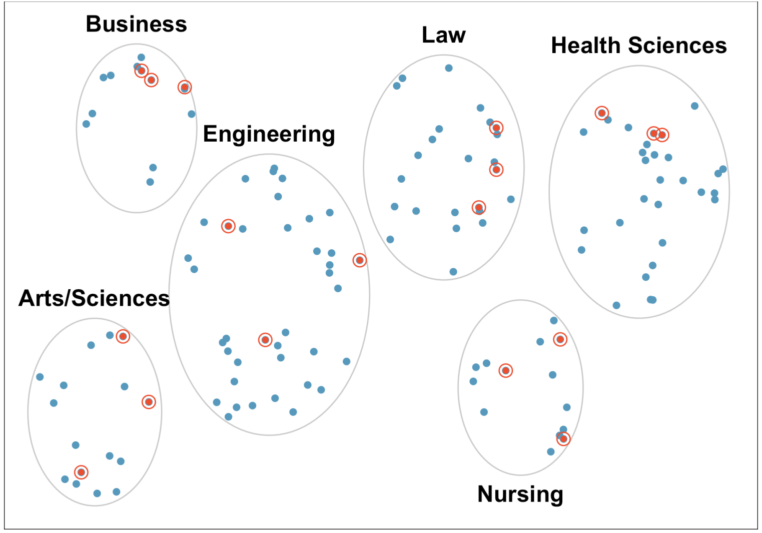
Simple Random Sample



https://research-methodology.net/sampling-in-primary-data-collection/random-sampling/

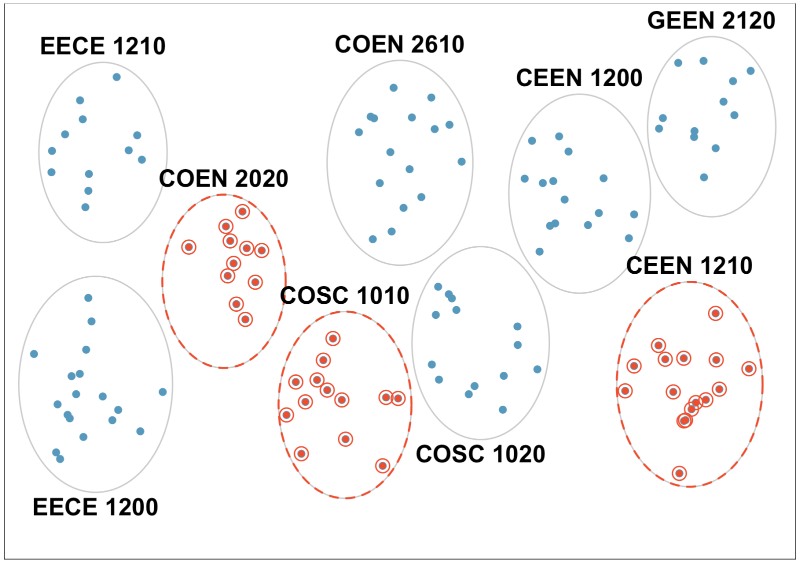
Stratified Random Sample

- **Stratified Sampling**: Subdivide the population into different subgroups (strata) that share the **same** characteristics, then draw a simple random sample from each subgroup.
- **Homogeneous within strata; Non-homogeneous between strata**
- **Example**: I divide Marquette students into groups by colleges/schools, and then do an SRS for each group.

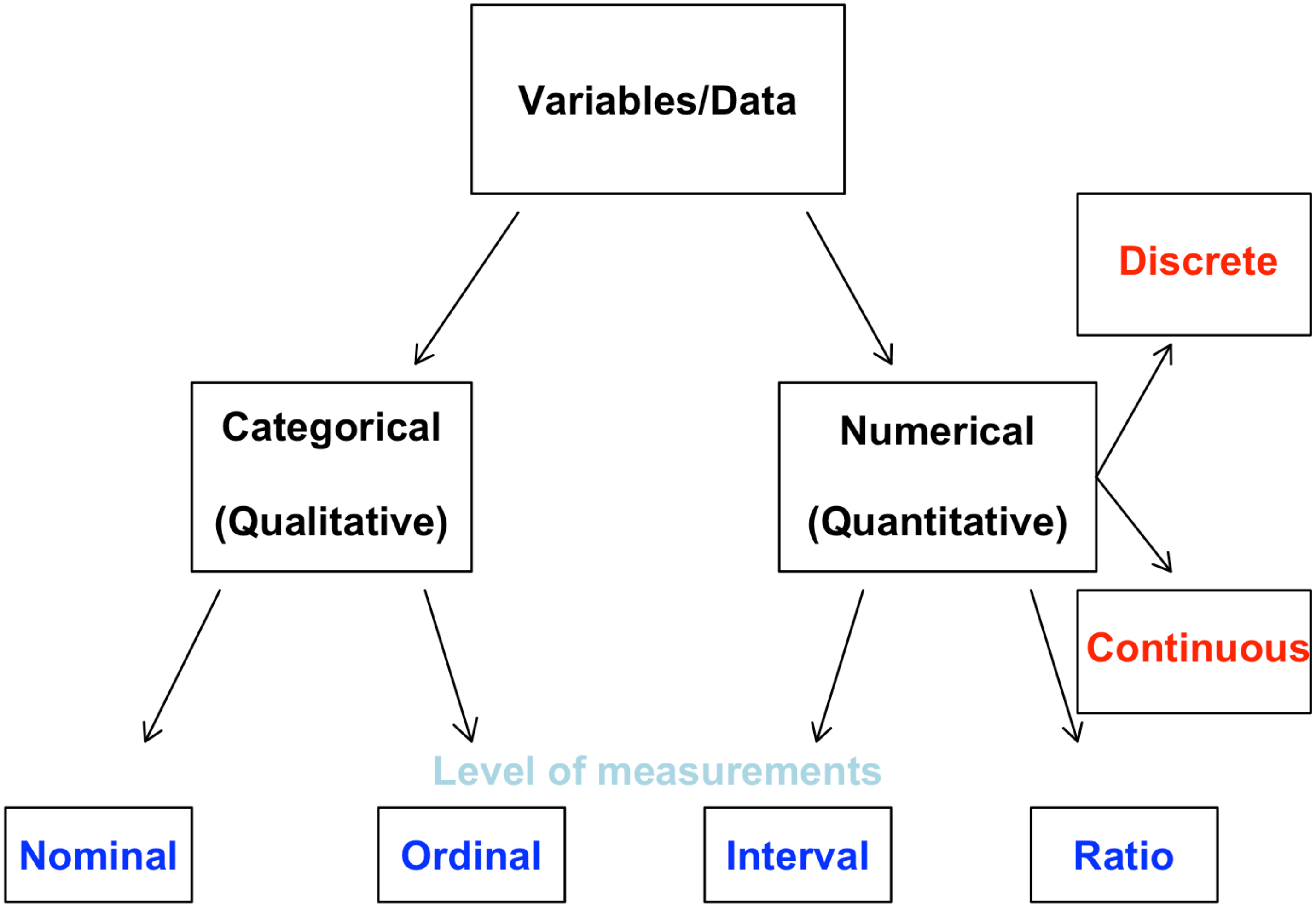


Cluster Sampling

- **Cluster Sampling**: Divide the population into sections (clusters), then randomly select some of those clusters, and then choose **all** the members from those selected clusters.
- **Homogeneous between clusters; Non-homogeneous within clusters**
- **Example**: Conducting a study of STEM student drinking habit by randomly selecting 10 different STEM classes and interviewing all of the students in each of those classes.



Data Type


Variables/Data
Categorical
(Qualitative)
Numerical
(Quantitative)
Discrete
Continuous
Level of measurements
Nominal
Ordinal
Interval
Ratio

Categorical vs. Numerical Variables

- A **categorical (qualitative)** variable provides *non-numerical* information which can be placed in **one (and only one)** category from two or more categories.
  - Gender (Male 👨, Female 👩, Other 🏳️‍🌈)
  - Class (Freshman, Sophomore, Junior, Senior, Graduate)
  - Country (USA 🇺🇸, Canada 🇨🇦, UK 🇬🇧, Germany 🇩🇪, Japan 🇯🇵, Korea 🇰🇷)
- A **numerical (quantitative)** variable is recorded in a *numerical* value representing counts or measurements.
  - GPA
  - The number of relationships you've had
  - Height

Numerical Variables can be Discrete or Continuous

- A **discrete** variable takes on values of a **finite** or **countable** number.
- A **continuous** variable takes on values **anywhere** over a particular range *without gaps or jumps*.
    - GPA is **continuous** because it can be any value between 0 and 4.
    - The number of relationships you've had is **discrete** because you can count the number and it is finite.
    - Height is **continuous** because it can be any number within a range.

Categorical Variables are Usually Recorded as Numbers

- Gender (Male = 0, Female = 1, Others = 2)
- Class (Freshman = 1, Sophomore = 2, Junior = 3, Senior = 4, Graduate = 5)
- Country (USA = 100, Canada = 101, UK = 200, Germany = 201, Japan = 300, Korea = 301)
- United Airlines boarding groups
- **The numbers represent categories only; differences between them are meaningless.**
  - Canada - USA = 101 - 100 = 1?
  - Graduate - Sophomore = 5 - 2 = 3 = Junior?
- We need to learn the **level of measurements** to know whether or which arithmetic operations are meaningful.

Levels of Measurements: Nominal and Ordinal for Categorical Variables

- **Nominal**: The data can *NOT be ordered* in a meaningful or natural way.
  - Gender (Male = 0, Female = 1, Others = 2) is **nominal** because Male, Female and Other cannot be ordered.
  - Country (USA = 100, Canada = 101, UK = 200, Germany = 201, Japan = 300, Korea = 301) is **nominal**.

- **Ordinal**: The data can be arranged in some meaningful order, but differences between data values can NOT be determined or are meaningless.
  - Class (Freshman = 1, Sophomore = 2, Junior = 3, Senior = 4, Graduate = 5) is **ordinal** because Sophomore is one class higher than Freshman.

Levels of Measurements: Interval and Ratio for Numerical Variables

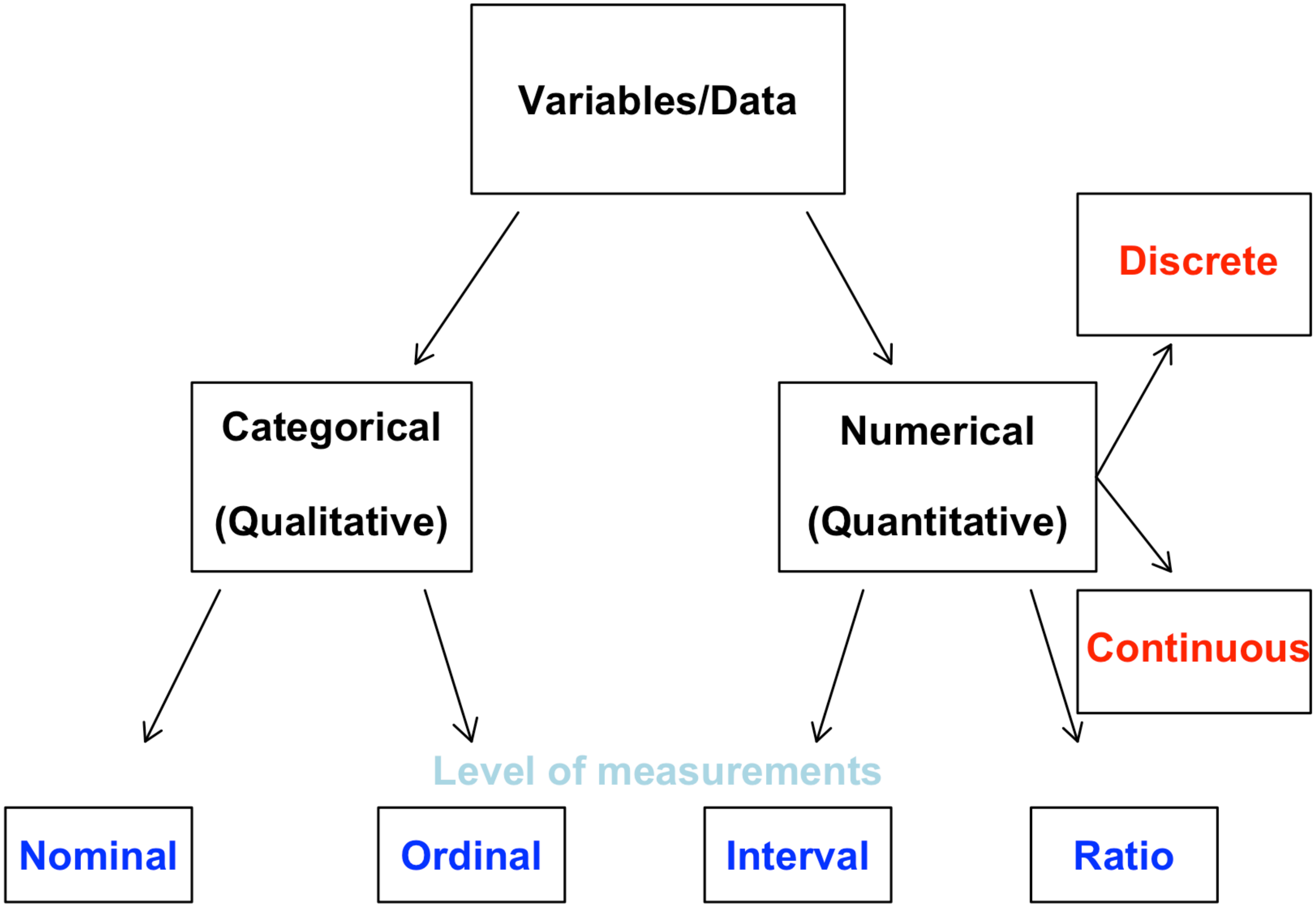
- **Interval**: The data have meaningful difference between any two values. But the data do NOT have a **natural zero or starting point**. The data can do $+$ and $-$, but can't reasonably do $\times$ and $\div$.
  - Temperature is **interval** because 80°F is 40 degrees higher than 40°F $(80 - 40 = 40)$, but 0° does not mean NO heat and 80°F is NOT twice as hot as 40°F.

- **Ratio**: The data have both meaningful differences and ratios, and there is a natural zero starting point that indicates none of the quantity. The data can do $+$, $-$, $\times$ and $\div$.
  - Distance is **ratio** because 80 miles is twice as far as 40 miles $(80/40 = 2)$, and 0 mile means no distance.

Converting Numerical to Categorical

- Yes, you've already seen an example.

Grade	Percentage
A	[93, 100]
A-	[90, 93)
B+	[87, 90)
B	[83, 87)
B-	[80, 83)
C+	[77, 80)
C	[73, 77)
C-	[70, 73)
D+	[65, 70)
D	[60, 65)
F	[0, 60)



Variables/Data

Categorical

(Qualitative)

Numerical

(Quantitative)

Discrete

Continuous

Level of measurements

Nominal

Ordinal

Interval

Ratio