

Section 1.1 The Structure of Data

E. Nordmoe

Math 260 Applied Statistics

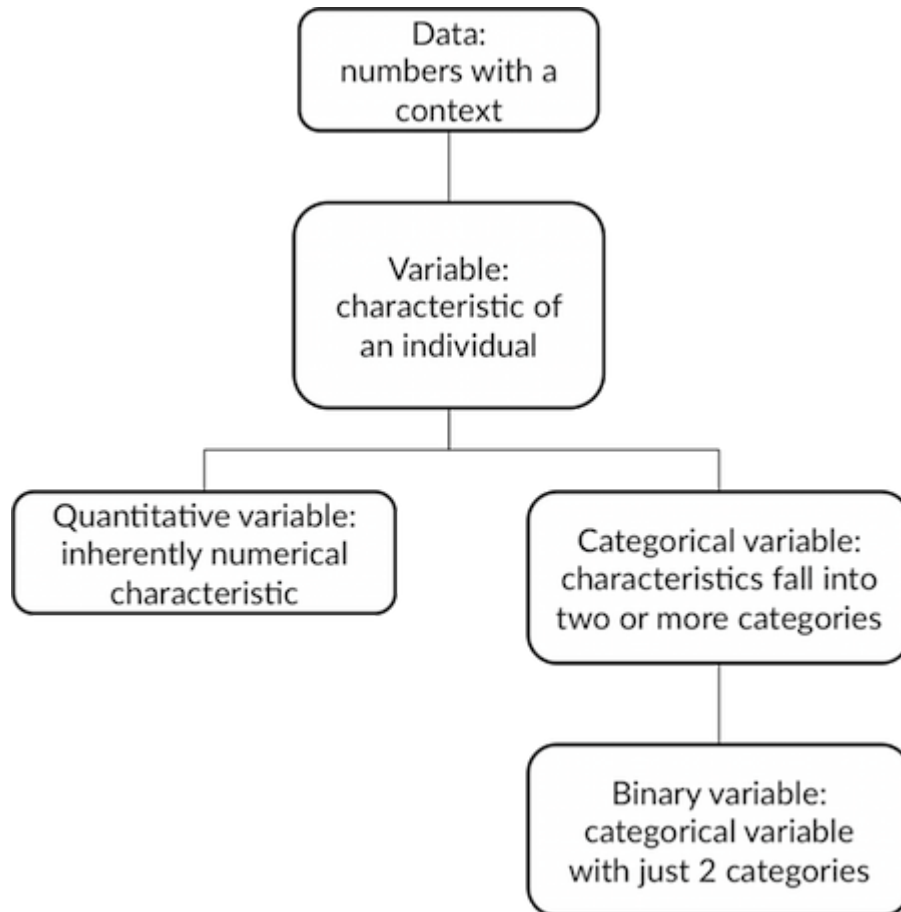
Outline

- Data
- Cases and variables
- Categorical and quantitative variables
- Explanatory and response variables
- Using data to answer a question

Data

- Data are a set of measurements taken on a set of individual units
- Data is typically stored and presented in a *dataset*, comprised of variables measured on cases

The Data Family Tree



Cautions

- Race is a social construct
 - ⇒ How to measure the effects of systemic racism?
 - ⇒ How to study disease risk factors?
- The Gender Unicorn
 - ⇒ How to measure salary discrimination by sex/gender?
 - ⇒ Estimate incidence of cancer?

The American Community Survey

Sex	Age	Married	Income	HoursWk	Race	USCitizen	HealthInsurance	Language
0	58	1	35.4	40	white	1	1	1
1	73	1	23.0	49	white	1	1	1
1	22	0	36.0	53	black	1	1	1
0	61	1	42.0	40	white	1	1	1
1	65	1	70.0	45	white	1	1	1
1	43	1	14.4	40	other	0	1	0
0	55	0	24.0	25	white	1	1	1
1	33	1	30.0	43	white	1	1	1
1	27	0	35.0	60	white	1	0	0
0	35	0	30.2	38	white	1	1	0
0	34	1	20.0	20	other	0	1	0
0	32	0	18.0	36	asian	0	1	0
0	31	0	12.0	50	other	1	1	1

- Rows are cases
- Columns are variables

COVID-19 Data by County

 COVID-19 alert

Coronavirus disease

Michigan

Overview

Testing

Health Info

Coping

News

Statistics

 Share

Each day shows new cases reported since the previous day · Updated less than 3 hours ago · Source: [Wikipedia](#) · [About this data](#)

Total cases

Total ▾



United States ▾

Michigan ▾

Confirmed

99,963

+515

Deaths

6,556

+9

Location	Confirmed ↓	Deaths
Wayne County	28,715	2,835
Oakland County	16,083	1,136
Macomb County	11,195	954
Kent County	7,677	157
Genesee County	3,699	299

"+" shows new cases reported yesterday · Updated less than 1 day ago · Sources: [Wikipedia](#) and [The New York Times](#). · [About this data](#)

Flights Data

Source: [US Bureau of Transportation Statistics](#)

flight	tailnum	origin	month	day	distance	air_time
4626	N8EGMQ	LGA	6	6	479	74
4304	N14105	EWR	6	22	1167	147
71	N607JB	JFK	5	3	1069	154
27	N854VA	JFK	10	11	2586	338
393	N526JB	LGA	1	13	950	133
404	N374DA	JFK	12	22	1069	156

Categorical - tailnum, origin, flight

Quantitative - distance, air_time, month, day

Careful: Dates need to be handled with care and are not typical quantitative variables. Context is everything!

Flights Data Sets: Part 2

faa	name	lat	lon	alt	tz	dst	tzone
04G	Lansdowne Airport	41.13	-80.62	1044	-5	A	America/New_York
06A	Moton Field Municipal Airport	32.46	-85.68	264	-6	A	America/Chicago
06C	Schaumburg Regional	41.99	-88.10	801	-6	A	America/Chicago
06N	Randall Airport	41.43	-74.39	523	-5	A	America/New_York
09J	Jekyll Island Airport	31.07	-81.43	11	-5	A	America/New_York
0A9	Elizabethton Municipal Airport	36.37	-82.17	1593	-5	A	America/New_York

Explanatory and Reponse Variables

If we are using one variable to help us understand or predict values of another variable, we call the former the **explanatory** variable and the latter the **response** variable.

Examples:

- Does meditation help reduce stress?



Explanatory and Response Variables

- Do multivitamin supplements increase longevity?



Popular

Latest

The Atlantic

Sign In



SOCIAL DISTANCE

Listen: The Empty Promise of Vitamins

Should we all be taking something during the pandemic? Supplements and vitamins make big claims, but their benefits are doubtful.

JUNE 17, 2020

Data is Everywhere!

Examples

- [US News and World Report National University Rankings]<https://www.usnews.com/best-colleges/rankings/national-liberal-arts-colleges>)
 - No place like home
- US Government's Open Data
- CDC COVID Data Tracker