MAT1372, Classwork1, Fall2025

ID:_____ Name:_____

1.1 Case Study: using stents to prevent strokes

1. Consider an experiment that studies effectiveness of stents in treating of strokes. We start by writing the principle question the researchers hope to answer:

Does the use of stents reduce the risk of strokes?

2. Two groups that apply when conducting an experiment with patients:

Treatment group: patients in this group get both medical management and treatment.

Control group: patients in this group get only medical management but no treatment.

3. To conduct the experiment to answer the question in 1., how to apply the two groups setup?

Treatment group: patients received a stent and medical management.

Control group: Patients recived only medical managment.

4. After **randomly** assigning patients into two groups, we have the effectiveness result of stents:

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
⋮	⋮	⋮	
450	control	no event	no event
451	control	no event	no event

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Figure 1.2: Descriptive statistics for the stent study.

(1) How many of patients are in the treatment group? 224 (33+191 or 45+179)

(2) How many of patients are in the control group? 227

(3) How many of patients had stroke by the end of the first year are in the treatment group? 45

(4) How many of patients have gotten a stroke by the end of the first year? 45+28 = 73

(5) After a year, the proportion who had a stroke in the treatment group is 45/224 = 20% = 0.2

(6) After a year, the proportion who had a stroke in the control group is 28/227 = 12% = 0.12

5. Summary statistic: A summary statistic is a single number summarizing a large amount of date.

which (5) and (6) in 4. are summary statistics.

1.2 Data Basics

1. Given a data set. We have

Cases : The number of the units in this data set. It is also called **observational unit**.

Variables : The characteristics of the cases.

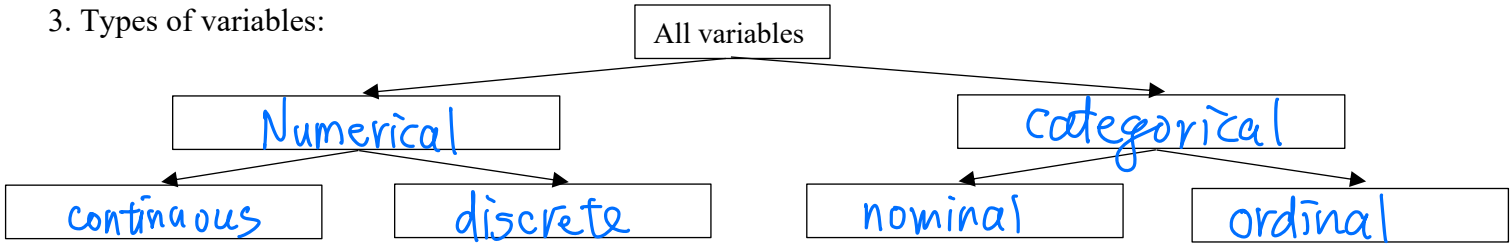
2. Given the data set for 3142 counties in the United State:

	name	state	pop	pop_change	poverty	homeownership	multi_unit	unemp_rate	metro	median_edu	median_hh_income
1	Autauga	Alabama	55504	1.48	13.7	77.5	7.2	3.86	yes	some_college	55317
2	Baldwin	Alabama	212628	9.19	11.8	76.7	22.6	3.99	yes	some_college	52562
3	Barbour	Alabama	25270	-6.22	27.2	68.0	11.1	5.90	no	hs_diploma	33368
4	Bibb	Alabama	22668	0.73	15.2	82.9	6.6	4.39	yes	hs_diploma	43404
5	Blount	Alabama	58013	0.68	15.6	82.0	3.7	4.02	yes	hs_diploma	47412
6	Bullock	Alabama	10309	-2.28	28.5	76.9	9.9	4.93	no	hs_diploma	29655
7	Butler	Alabama	19825	-2.69	24.4	69.0	13.7	5.49	no	hs_diploma	36326
8	Calhoun	Alabama	114728	-1.51	18.6	70.7	14.3	4.93	yes	some_college	43686
9	Chambers	Alabama	33713	-1.20	18.8	71.4	8.7	4.08	no	hs_diploma	37342
10	Cherokee	Alabama	25857	-0.60	16.1	77.5	4.3	4.05	no	hs_diploma	40041
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Weston	Wyoming	6927	-2.93	14.4	77.9	6.5	3.98	no	some_college	59605

Cases: all the counties in the U.S.

Variables: name, state, population, ..., median household income

3. Types of variables:



4. In the table from 2., classify the variables by the types:

(1) continuous numerical: pop. change, poverty, homeownership.

(2) discrete numerical: population.

(3) nominal: county name.

(4) ordinal: median education.

5. Relationships between variables.

Based on the data set for 3142 counties in the United State, we can ask questions like:

(1) If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county tend to above or below the national average?

(2) Does a higher-than-average increase in county population tend to correspond to counties with higher or lower median household incomes?