Introduction to Social Data Analytics

Class 4

Arushi Kaushik

arkaushi@ucsd.edu

Today's Learning Objectives

From Pre Class 4 Exercise:

- resizing columns, pasting values, using MATCH and VLOOKUP

Today's Learning Objectives

From Pre Class 4 Exercise:

- resizing columns, pasting values, using MATCH and VLOOKUP

After today, you should be able to:

- Classify kinds of variables (numerical, (un)ordered categorical, logical)
- Identify when a sample contains sampling bias and implications for external validity
- Use the RAND function to conduct a simple random sample

Today's Learning Objectives

From Pre Class 4 Exercise:

- resizing columns, pasting values, using MATCH and VLOOKUP

After today, you should be able to:

- Classify kinds of variables (numerical, (un)ordered categorical, logical)
- Identify when a sample contains sampling bias and implications for external validity
- Use the RAND function to conduct a simple random sample

Please download and open class4.xlsx if you haven't already.

Kinds of Variables

- Numerical (Quantitative): Represented by a number

Kinds of Variables

- Numerical (Quantitative): Represented by a number
- Categorical: Data in categories
  - Unordered: Unordered categories
  - Ordinal: Ordered categories

Kinds of Variables

- Numerical (Quantitative): Represented by a number
- Categorical: Data in categories
  - Unordered: Unordered categories
  - Ordinal: Ordered categories
- Logical: True/False variables

Which kinds of variables are the following:

Numerical, unordered categorical, ordered categorical, or logical

Which kinds of variables are the following:

Numerical, unordered categorical, ordered categorical, or logical

- Rating of High, Medium, or Low of a student's confidence with Excel

Which kinds of variables are the following:

Numerical, unordered categorical, ordered categorical, or logical

- Rating of High, Medium, or Low of a student's confidence with Excel
- Whether or not a student turned in an assignment on time

Which kinds of variables are the following:

Numerical, unordered categorical, ordered categorical, or logical

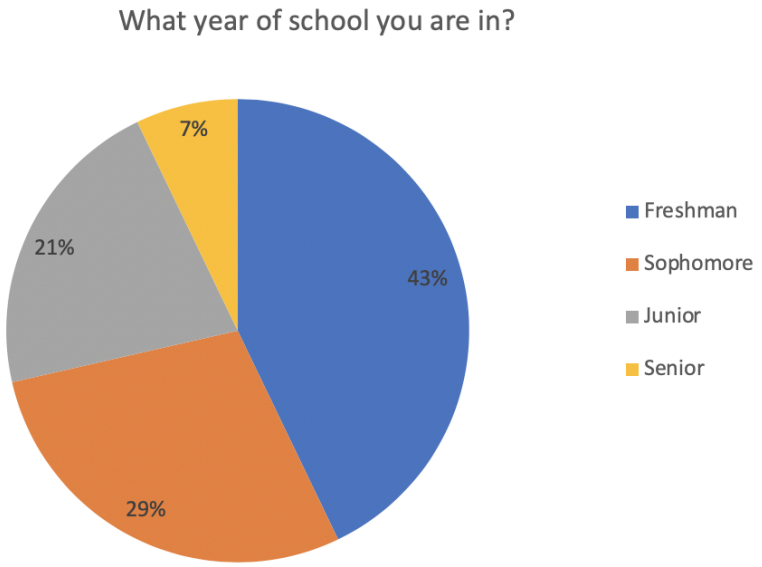
- Rating of High, Medium, or Low of a student's confidence with Excel
- Whether or not a student turned in an assignment on time
- A student's score on an exam

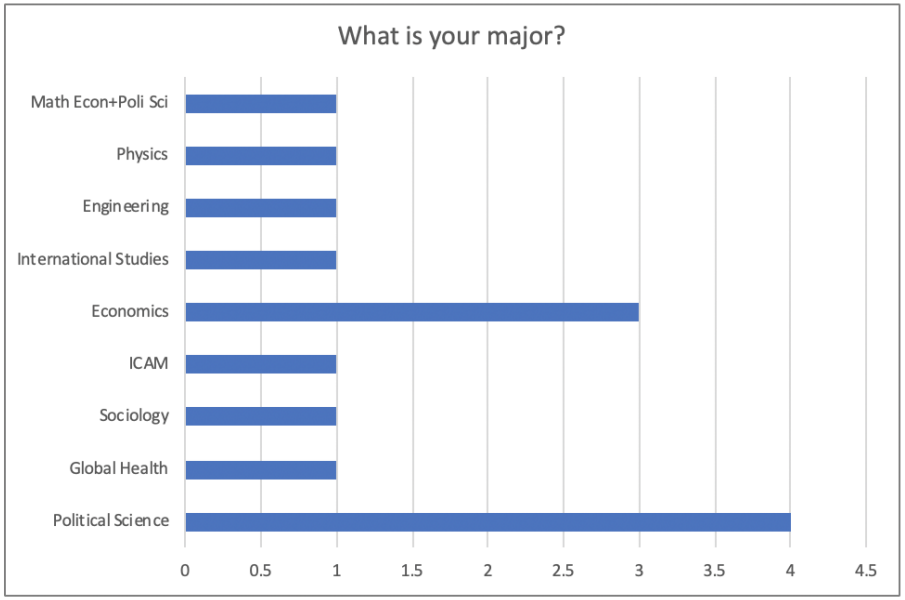
Which kinds of variables are the following:

Numerical, unordered categorical, ordered categorical, or logical

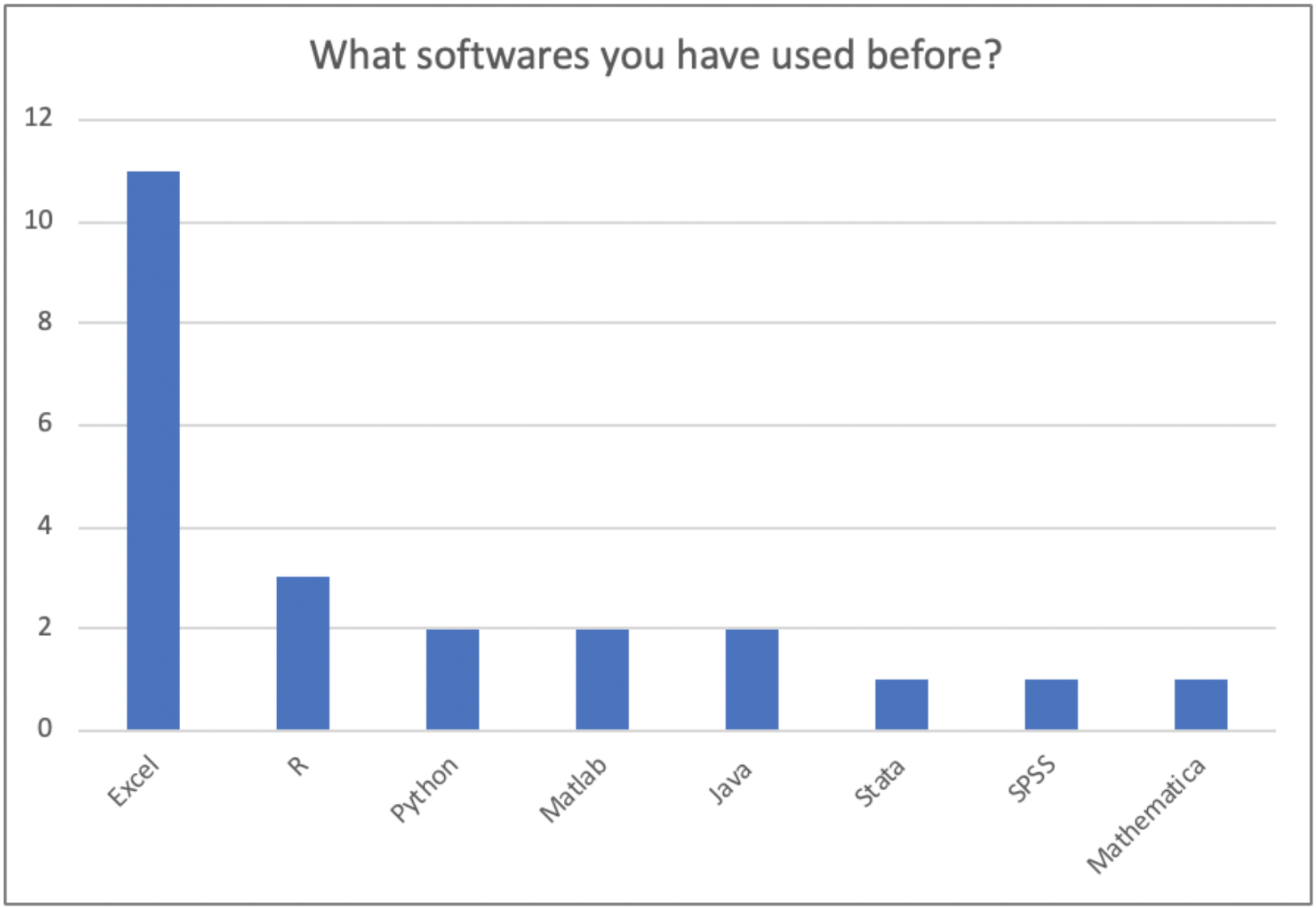
- Rating of High, Medium, or Low of a student's confidence with Excel
- Whether or not a student turned in an assignment on time
- A student's score on an exam
- A student's college at UCSD

What year are you?





Check all software you've used before.



The kind of variable depends on the question

Classify the following variables:

- Do you know how to use Excel?

The kind of variable depends on the question

Classify the following variables:

- Do you know how to use Excel?
- On a scale from 1 to 10 how well do you know Excel?

The kind of variable depends on the question

Classify the following variables:

- Do you know how to use Excel?
- On a scale from 1 to 10 how well do you know Excel?
- Across sections, what % have used Excel before?

The kind of variable depends on the question

Classify the following variables:

- Do you know how to use Excel?
- On a scale from 1 to 10 how well do you know Excel?
- Across sections, what % have used Excel before?
- Which software package do you know best?

How can we learn from a sample?

We are often interested in population averages (e.g. poverty rates, election turnout, etc.), but polling the entire population is typically too costly.

How can we learn from a sample?

We are often interested in population averages (e.g. poverty rates, election turnout, etc.), but polling the entire population is typically too costly.

We settle for a sample and use statistics instead.

How can we learn from a sample?

We are often interested in population averages (e.g. poverty rates, election turnout, etc.), but polling the entire population is typically too costly.

We settle for a sample and use statistics instead.

How can we learn from a sample?

- External validity: when relationships found within the sample data apply to the population

How can we learn from a sample?

We are often interested in population averages (e.g. poverty rates, election turnout, etc.), but polling the entire population is typically too costly.

We settle for a sample and use statistics instead.

How can we learn from a sample?

- External validity: when relationships found within the sample data apply to the population
- To be externally valid, the sample must be representative of the population

How can we learn from a sample?

We are often interested in population averages (e.g. poverty rates, election turnout, etc.), but polling the entire population is typically too costly.

We settle for a sample and use statistics instead.

How can we learn from a sample?

- External validity: when relationships found within the sample data apply to the population
- To be externally valid, the sample must be representative of the population
    - One technique is simple random sampling

How can we learn from a sample?

We are often interested in population averages (e.g. poverty rates, election turnout, etc.), but polling the entire population is typically too costly.

We settle for a sample and use statistics instead.

How can we learn from a sample?

- External validity: when relationships found within the sample data apply to the population
- To be externally valid, the sample must be representative of the population
  - One technique is simple random sampling
  - Difficult in practice - what happens when some of the population cannot be contacted (e.g. no phone or email)?

How can we learn from a sample?

We are often interested in population averages (e.g. poverty rates, election turnout, etc.), but polling the entire population is typically too costly.

We settle for a sample and use statistics instead.

How can we learn from a sample?

- External validity: when relationships found within the sample data apply to the population
- To be externally valid, the sample must be representative of the population
  - One technique is simple random sampling
  - Difficult in practice - what happens when some of the population cannot be contacted (e.g. no phone or email)?
  - Non-responses result in sampling bias

Another kind of bias: response bias

Often in surveys, respondents provide answers that do not wholly reflect the truth.

Another kind of bias: response bias

Often in surveys, respondents provide answers that do not wholly reflect the truth.

Consider how respondents might be inclined to provide biased answers to the following:

- How often do you make racist comments?

Another kind of bias: response bias

Often in surveys, respondents provide answers that do not wholly reflect the truth.

Consider how respondents might be inclined to provide biased answers to the following:

- How often do you make racist comments? (Social pressure to answer a certain way)

Another kind of bias: response bias

Often in surveys, respondents provide answers that do not wholly reflect the truth.

Consider how respondents might be inclined to provide biased answers to the following:

- How often do you make racist comments? (Social pressure to answer a certain way)
- How much income do you earn per year?

Another kind of bias: response bias

Often in surveys, respondents provide answers that do not wholly reflect the truth.

Consider how respondents might be inclined to provide biased answers to the following:

- How often do you make racist comments? (Social pressure to answer a certain way)
- How much income do you earn per year? (Too much effort to report exact income $\Rightarrow$ rounding)

Another kind of bias: response bias

Often in surveys, respondents provide answers that do not wholly reflect the truth.

Consider how respondents might be inclined to provide biased answers to the following:

- How often do you make racist comments? (Social pressure to answer a certain way)
- How much income do you earn per year? (Too much effort to report exact income $\Rightarrow$ rounding)
- How likely are you to recommend my company's product to a friend?

Another kind of bias: response bias

Often in surveys, respondents provide answers that do not wholly reflect the truth.

Consider how respondents might be inclined to provide biased answers to the following:

- How often do you make racist comments? (Social pressure to answer a certain way)
- How much income do you earn per year? (Too much effort to report exact income $\Rightarrow$ rounding)
- How likely are you to recommend my company's product to a friend? (Don't want to hurt surveyor's feelings)

Example: wages after graduation

Universities often ask alumni for information about their careers over time.

- What is the population of interest?

Example: wages after graduation

Universities often ask alumni for information about their careers over time.

- What is the population of interest?
- What is the sample?

Example: wages after graduation

Universities often ask alumni for information about their careers over time.

- What is the population of interest?
- What is the sample?
- Any concerns with phone or mail based surveys?

Example: wages after graduation

Universities often ask alumni for information about their careers over time.

- What is the population of interest?
- What is the sample?
- Any concerns with phone or mail based surveys?
- Why might the sample *not* be representative (contain sampling bias)?

Example: wages after graduation

Universities often ask alumni for information about their careers over time.

- What is the population of interest?
- What is the sample?
- Any concerns with phone or mail based surveys?
- Why might the sample *not* be representative (contain sampling bias)?
- Why might the data contain response bias?

class4.xlsx

	A	B	C	D	E	F	G	H	I	J	K
1	id	phone	actual salary	salary provided to surveyor	random number	selected?	selected and answered phone?		Sample	Average Salary	Difference from population mean
2	1	Y	112820.87	115000					Entire Population	$ 57,868.98	-
3	2	Y	39371.62	40000					Simple random sample w/ response bias		
4	3	Y	63118.67	65000					Sample with response and sampling bias		
5	4	Y	95692.48	100000							
6	5	N	36642.72	40000							
7	6	Y	84664.83	85000							
8	7	Y	11981.42	15000							
9	8	Y	86193.65	90000							
10	9	Y	43013.69	45000							
11	10	Y	73771.39	75000							

- What is the unit of analysis?

class4.xlsx

	A	B	C	D	E	F	G	H	I	J	K
1	id	phone	actual salary	salary provided to surveyor	random number	selected?	selected and answered phone?		Sample	Average Salary	Difference from population mean
2	1	Y	112820.87	115000					Entire Population	$ 57,868.98	-
3	2	Y	39371.62	40000					Simple random sample w/ response bias		
4	3	Y	63118.67	65000					Sample with response and sampling bias		
5	4	Y	95692.48	100000							
6	5	N	36642.72	40000							
7	6	Y	84664.83	85000							
8	7	Y	11981.42	15000							
9	8	Y	86193.65	90000							
10	9	Y	43013.69	45000							
11	10	Y	73771.39	75000							

- What is the unit of analysis?
- Use the RAND() function to assign a random number (column E)

class4.xlsx

	A	B	C	D	E	F	G	H	I	J	K
1	id	phone	actual salary	salary provided to surveyor	random number	selected?	selected and answered phone?		Sample	Average Salary	Difference from population mean
2	1	Y	112820.87	115000					Entire Population	$ 57,868.98	-
3	2	Y	39371.62	40000					Simple random sample w/ response bias		
4	3	Y	63118.67	65000					Sample with response and sampling bias		
5	4	Y	95692.48	100000							
6	5	N	36642.72	40000							
7	6	Y	84664.83	85000							
8	7	Y	11981.42	15000							
9	8	Y	86193.65	90000							
10	9	Y	43013.69	45000							
11	10	Y	73771.39	75000							

- What is the unit of analysis?
- Use the RAND() function to assign a random number (column E)
- If the random number > 0.5, they are selected for the random sample (column F)

class4.xlsx

	A	B	C	D	E	F	G	H	I	J	K
1	id	phone	actual salary	salary provided to surveyor	random number	selected?	selected and answered phone?		Sample	Average Salary	Difference from population mean
2	1	Y	112820.87	115000					Entire Population	$ 57,868.98	-
3	2	Y	39371.62	40000					Simple random sample w/ response bias		
4	3	Y	63118.67	65000					Sample with response and sampling bias		
5	4	Y	95692.48	100000							
6	5	N	36642.72	40000							
7	6	Y	84664.83	85000							
8	7	Y	11981.42	15000							
9	8	Y	86193.65	90000							
10	9	Y	43013.69	45000							
11	10	Y	73771.39	75000							

- What is the unit of analysis?
- Use the RAND() function to assign a random number (column E)
- If the random number > 0.5, they are selected for the random sample (column F)
- If the random number > 0.5 but phone $=$ "N", they are NOT included in the random sample with sampling bias (column G)

class4.xlsx

	A	B	C	D	E	F	G	H	I	J	K
1	id	phone	actual salary	salary provided to surveyor	random number	selected?	selected and answered phone?		Sample	Average Salary	Difference from population mean
2	1	Y	112820.87	115000					Entire Population	$ 57,868.98	-
3	2	Y	39371.62	40000					Simple random sample w/ response bias		
4	3	Y	63118.67	65000					Sample with response and sampling bias		
5	4	Y	95692.48	100000							
6	5	N	36642.72	40000							
7	6	Y	84664.83	85000							
8	7	Y	11981.42	15000							
9	8	Y	86193.65	90000							
10	9	Y	43013.69	45000							
11	10	Y	73771.39	75000							

- What is the unit of analysis?
- Use the RAND() function to assign a random number (column E)
- If the random number $>$ 0.5, they are selected for the random sample (column F)
- If the random number $>$ 0.5 but phone $=$ "N", they are NOT included in the random sample with sampling bias (column G)
- Determine the average reported salary amongst those selected (J3) and amongst those selected who answered their phones (J4).

Next class

1. We will be plotting in Excel.
2. Submit your pre class exercise before class.