Studying Social Inequality with Data Science

INFO 3370 / 5371
Spring 2023

Asking Research Questions

Learning goals for today

By the end of class, you will be able to

- articulate a clear research question
- use language appropriate for causal or descriptive questions

What makes a good research question?

Keys to a good research question

Keys to a good research question

1. a unit of analysis
    - a row of your dataset

Keys to a good research question

1. a unit of analysis
   - a row of your dataset
2. an outcome
   - a variable with a value for each unit

Keys to a good research question

1. a unit of analysis
   - a row of your dataset
2. an outcome
   - a variable with a value for each unit
3. a target population
   - a set of units about whom to infer
   - clear who is included and who is not

Keys to a good research question

1. a unit of analysis
   - a row of your dataset
2. an outcome
   - a variable with a value for each unit
3. a target population
   - a set of units about whom to infer
   - clear who is included and who is not
4. potential for surprising results

Describe a population

What is the proportion employed
among U.S. resident women ages 21–35?

Describe a population

What is the proportion employed
among U.S. resident women ages 21–35?

Woman 1
Woman 2
Woman 3
Woman 4

Describe a population

What is the proportion employed
among U.S. resident women ages 21–35?

	Employed?
Woman 1	1
Woman 2	0
Woman 3	1
Woman 4	1

Describe population subgroups

What is the proportion employed
among U.S. resident women ages 21–35,
comparing mothers to non-mothers?

Describe population subgroups

What is the proportion employed
among U.S. resident women ages 21–35,
comparing mothers to non-mothers?

	Employed?
Mother 1	0
Mother 2	0
Mother 3	0
Mother 4	1

	Employed?
Non-Mother 1	1
Non-Mother 2	0
Non-Mother 3	1
Non-Mother 4	1

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

Woman 1
Woman 2
Woman 3
Woman 4

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

	Would be employed if a mother? $Y(1)$
Woman 1	0
Woman 2	0
Woman 3	0
Woman 4	1

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

	Would be employed if a mother? $Y(1)$	Would be employed if a non-mother? $Y(0)$
Woman 1	0	1
Woman 2	0	0
Woman 3	0	1
Woman 4	1	1

Causal effect in a population

What is the causal effect of motherhood on employment among U.S. resident women ages 21–35?

	Would be employed if a mother? $Y(1)$	Would be employed if a non-mother? $Y(0)$	Causal effect $Y(1) - Y(0)$
Woman 1	0	1	-1
Woman 2	0	0	0
Woman 3	0	1	-1
Woman 4	1	1	0

Describe population subgroups

What is the proportion employed
among U.S. resident women ages 21–35,
comparing mothers to non-mothers?

	Employed?
Mother 1	0
Mother 2	0
Mother 3	0
Mother 4	1

	Employed?
Non-Mother 1	1
Non-Mother 2	0
Non-Mother 3	1
Non-Mother 4	1

Very **different** research goals

Causal effect in a population

What is the causal effect of motherhood on employment
among U.S. resident women ages 21–35?

	Would be employed if a mother? $Y(1)$	Would be employed if a non-mother? $Y(0)$	Causal effect $Y(1) - Y(0)$
Woman 1	0	1	-1
Woman 2	0	0	0
Woman 3	0	1	-1
Woman 4	1	1	0

Keywords: What kind of question is being asked?

Keywords: What kind of question is being asked?

Descriptive

among

across

difference

for those who

Keywords: What kind of question is being asked?

Descriptive	Causal
among	causes
across	affects
difference	produces
for those who	impacts

Keywords: What kind of question is being asked?

Descriptive	Murky Middle	Causal
among	associated with	causes
across	leads to	affects
difference	predicts	produces
for those who		impacts

Keywords: What kind of question is being asked?

Descriptive	Murky Middle	Causal
among	associated with	causes
across	leads to	affects
difference	predicts	produces
for those who		impacts

↑ verbs

Keywords: What kind of question is being asked?

Descriptive	Murky Middle	Causal
among	associated with	causes
across	leads to	affects
difference	predicts	produces
for those who		impacts
↑		↑
not verbs		**verbs**

Keywords: What kind of question is being asked?

Descriptive	Murky Middle	Causal
among	associated with	causes
across	leads to	affects
difference	predicts	produces
for those who		impacts
↑		↑
not verbs		**verbs**

Statements "predictor **verb** outcome" are often causal

(analysis needs a DAG!)

Keywords: What kind of question is being asked?

Descriptive	Murky Middle	Causal
among	associated with	causes
across	leads to	affects
difference	predicts	produces
for those who		impacts
↑		↑
not verbs		**verbs**

Statements "predictor **verb** outcome"
are often causal

(analysis needs a DAG!)

Statements "**among subpopulation**, mean outcome"
are often descriptive

A good project may have a very simple question

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

What proportion of children born in large U.S. cities in 1998–2000 was ever evicted from their home from birth to age 15?

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

What proportion of children born in large U.S. cities in 1998–2000 was ever evicted from their home from birth to age 15?

- unit of analysis
- target population
- outcome

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

What proportion of children born in large U.S. cities in 1998–2000 was ever evicted from their home from birth to age 15?

- unit of analysis
  - a child
- target population

- outcome

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

What proportion of children born in large U.S. cities in 1998–2000 was ever evicted from their home from birth to age 15?

- unit of analysis
  - a child
- target population
  - children born in large U.S. cities in 1998–2000
- outcome

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

What proportion of children born in large U.S. cities in 1998–2000 was ever evicted from their home from birth to age 15?

- unit of analysis
  - a child
- target population
  - children born in large U.S. cities in 1998–2000
  - (and subgroups by race and income)
- outcome

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

What proportion of children born in large U.S. cities in 1998–2000 was ever evicted from their home from birth to age 15?

- unit of analysis
  - a child
- target population
  - children born in large U.S. cities in 1998–2000
  - (and subgroups by race and income)
- outcome
  - evicted from home between birth and age 15

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

Future of Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

H19. We are also interested in some of the problems that families face making ends meet. In the past 12 months, did you do any of the following because there wasn't enough money?

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

Future of Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

H19. We are also interested in some of the problems that families face making ends meet. In the past 12 months, did you do any of the following because there wasn't enough money?

	YES	NO
H19E. (In the past 12 months), were you evicted from your home or apartment for not paying the rent or mortgage?	1	2

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

Future of Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

H19. We are also interested in some of the problems that families face making ends meet. In the past 12 months, did you do any of the following because there wasn't enough money?

		YES	NO
H19E.	(In the past 12 months), were you evicted from your home or apartment for not paying the rent or mortgage?	1	2

- we filled in missing values with regression

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

Future of Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

H19. We are also interested in some of the problems that families face making ends meet. In the past 12 months, did you do any of the following because there wasn't enough money?

		YES	NO
H19E.	(In the past 12 months), were you evicted from your home or apartment for not paying the rent or mortgage?	1	2

- we filled in missing values with regression
- we gathered responses across years

Example: Prevalence of housing eviction

Lundberg & Donnelly 2019

Example: Progress toward gender equality

From Homework 2: England, Levine, & Mishel 2020

Example: Progress toward gender equality

From Homework 2: England, Levine, & Mishel 2020

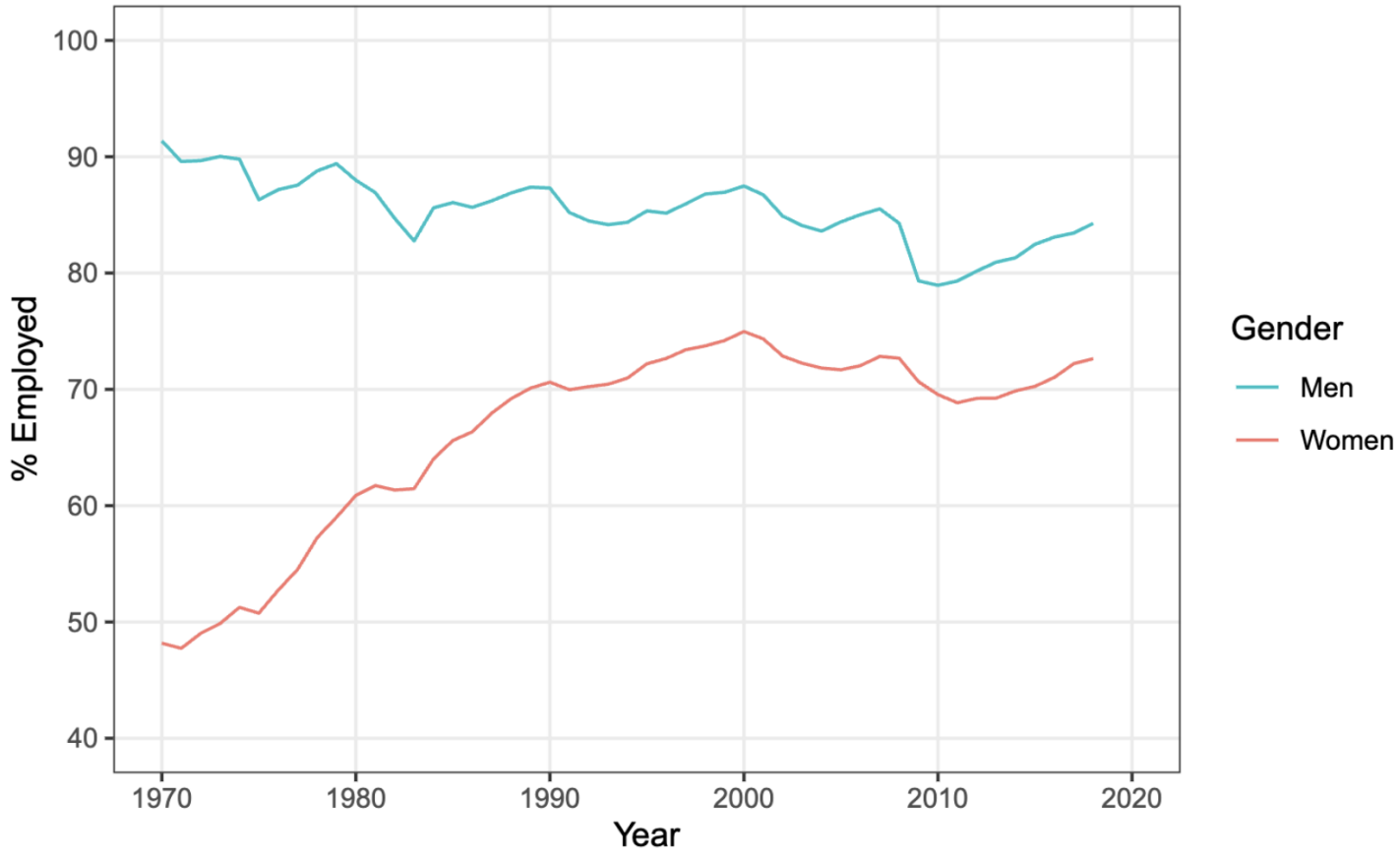

Fig. 1. Percentage of women and men, age 25 to 54, employed in the last week, 1970 to 2018. Source: Authors' computations from IPUMS CPS ASEC samples for 1970 to 2018.

Example: Progress toward gender equality

From Homework 2: England, Levine, & Mishel 2020

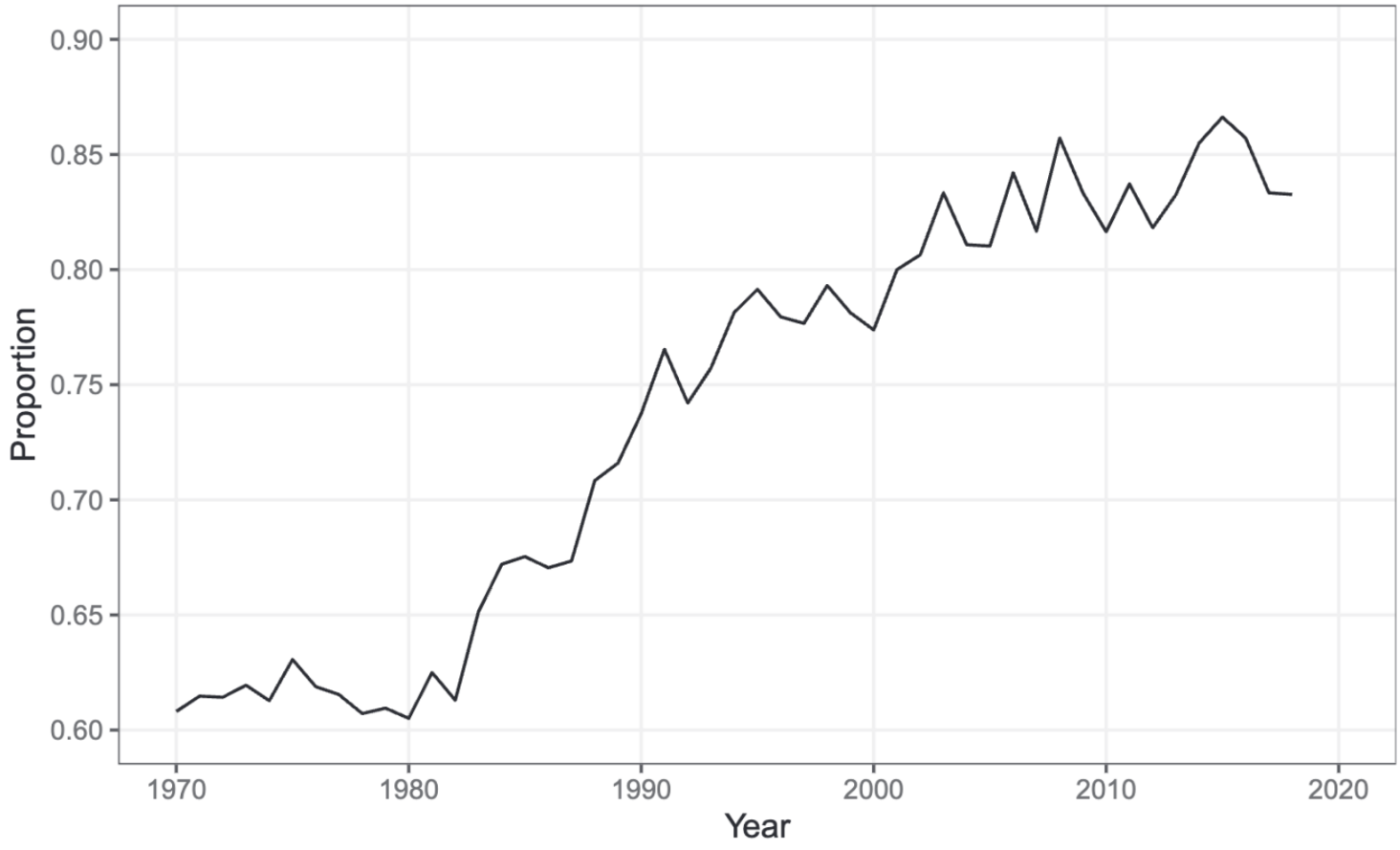

Fig. 9. Ratio of women's to men's median hourly wage among full-time workers employed in the last week, age 25 to 54, 1970 to 2018. Source: Authors' computations from IPUMS CPS ASEC samples for 1970 to 2018.

hypothetical examples of questions

Example: Not right for our class

Example: Not right for our class

Researcher uses a random forest

Example: Not right for our class

Researcher uses a random forest

- models hourly wage given many inputs

Example: Not right for our class

Researcher uses a random forest

- models hourly wage given many inputs
- reports a measure of variable importance
  - how important is each variable in predicting wage?

Example: Not right for our class

Researcher uses a random forest

- models hourly wage given many inputs
- reports a measure of variable importance
  - how important is each variable in predicting wage?

Why is this question not suitable for our class project?

Example: Not right for our class

Researcher uses a random forest

- models hourly wage given many inputs
- reports a measure of variable importance
  - how important is each variable in predicting wage?

Why is this question not suitable for our class project?

- we want a study of the world,
  and this is more a study of an algorithm

Example: Not right for our class

Researcher uses a random forest

- models hourly wage given many inputs
- reports a measure of variable importance
  - how important is each variable in predicting wage?

Why is this question not suitable for our class project?

- we want a study of the world,
  and this is more a study of an algorithm
- we want a clear target population,
  and there isn't one here

Example: Not right for our class

Researcher uses a random forest

- models hourly wage given many inputs
- reports a measure of variable importance
  - how important is each variable in predicting wage?

Why is this question not suitable for our class project?

- we want a study of the world,
  and this is more a study of an algorithm
- we want a clear target population,
  and there isn't one here
- we want an outcome aggregated within subgroups,
  and this is something else

Example: Could be improved (1 / 3)

A researcher studies the racial composition of those

- with college degrees
- without college degrees

among 25–50 year old American residents in 2022

Example: Could be improved (1 / 3)

A researcher studies the racial composition of those

- with college degrees
- without college degrees

among 25–50 year old American residents in 2022

What is your biggest concern about this study?

Example: Could be improved (1 / 3)

A researcher studies the racial composition of those

- with college degrees
- without college degrees

among 25–50 year old American residents in 2022

What is your biggest concern about this study?

- usually it is best of X precedes Y

Example: Could be improved (1 / 3)

A researcher studies the racial composition of those

- with college degrees
- without college degrees

among 25–50 year old American residents in 2022

What is your biggest concern about this study?

- usually it is best of X precedes Y
- better to study P(College | Race)
  instead of P(Race | College)

Example: Could be improved (2 / 3)

A researcher studies whether those with higher hourly wages have higher annual earnings

What is your biggest concern about this study?

Example: Could be improved (2 / 3)

A researcher studies whether those with higher hourly wages have higher annual earnings

What is your biggest concern about this study?

- needs a clearer target population

Example: Could be improved (2 / 3)

A researcher studies whether those with higher hourly wages have higher annual earnings

What is your biggest concern about this study?

- needs a clearer target population
- result is somewhat obvious:
  hard to argue that those with higher hourly wages would have lower annual earnings

Example: Could be improved (3 / 3)

A researcher uses a probability sample survey to estimate the share of total wealth held by the top 1% of households

Example: Could be improved (3 / 3)

A researcher uses a probability sample survey to estimate the share of total wealth held by the top 1% of households

- 30% of sampled individuals refuse to answer

Example: Could be improved (3 / 3)

A researcher uses a probability sample survey to estimate the share of total wealth held by the top 1% of households

- 30% of sampled individuals refuse to answer
- the researcher drops them

Example: Could be improved (3 / 3)

A researcher uses a probability sample survey to estimate the share of total wealth held by the top 1% of households

- 30% of sampled individuals refuse to answer
- the researcher drops them
- they say the target population is the top 1% wealth share among households willing to respond to the survey

Example: Could be improved (3 / 3)

A researcher uses a probability sample survey to estimate the share of total wealth held by the top 1% of households

- 30% of sampled individuals refuse to answer
- the researcher drops them
- they say the target population is the top 1% wealth share among households willing to respond to the survey

What is your biggest concern about this study?

Example: Could be improved (3 / 3)

A researcher uses a probability sample survey to estimate the share of total wealth held by the top 1% of households

- 30% of sampled individuals refuse to answer
- the researcher drops them
- they say the target population is the top 1% wealth share among households willing to respond to the survey

What is your biggest concern about this study?

- the target population is not very interesting if it excludes those who won't respond

Keys to a good research question

1. a unit of analysis
   - a row of your dataset
2. an outcome
   - a variable with a value for each unit
3. a target population
   - a set of units about whom to infer
   - clear who is included and who is not

Learning goals for today

By the end of class, you will be able to

- articulate a clear research question
- use language appropriate for causal or descriptive questions