Studying Social Inequality with Data Science

INFO 3370 / 5371
Spring 2024

Sampling for Population Inference

When you think of data science,
what kind of data do you think of?

Learning goals for today

By the end of class, you will be able to

- explain key ideas of data collection
  - target population
  - sampling frame
  - undercoverage
  - simple random sample
  - unequal probability sample
- access survey data online

Do you prefer the front or the back of the room?

Full count enumeration

- find everyone in the target population
- ask them all the question

Probability sampling

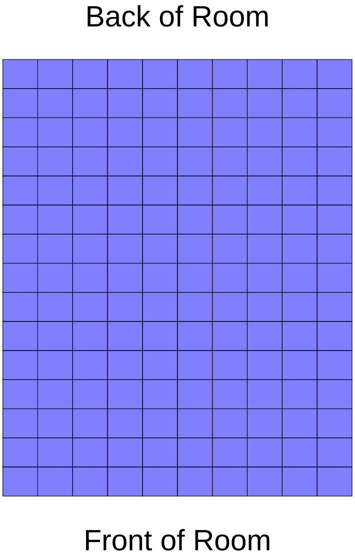
Open R. Run this line

```
runif(n = 1)
```

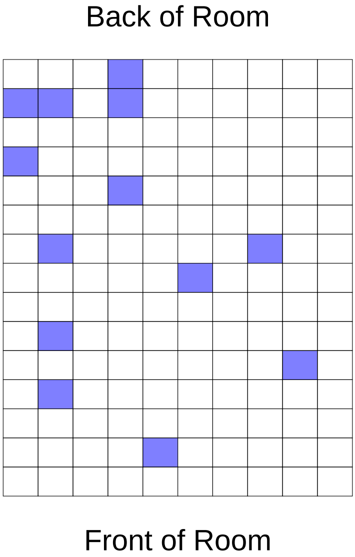
If answer $< .1$, then answer the question

- Do you prefer the front or the back of the room?

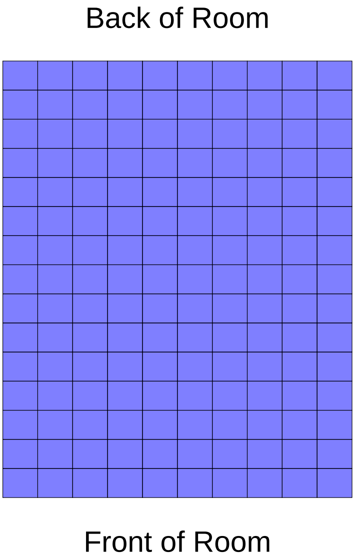
Full Count Enumeration



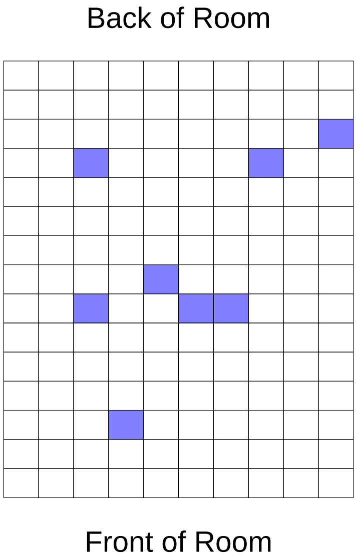
Probability Sample



Full Count Enumeration

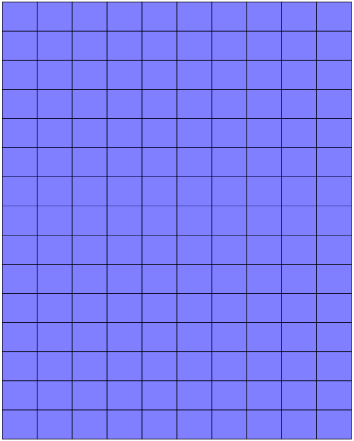


Probability Sample



Full Count Enumeration

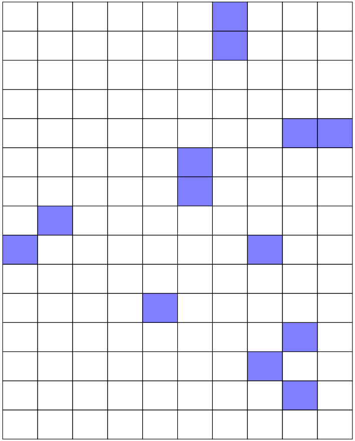
Back of Room



Front of Room

Probability Sample

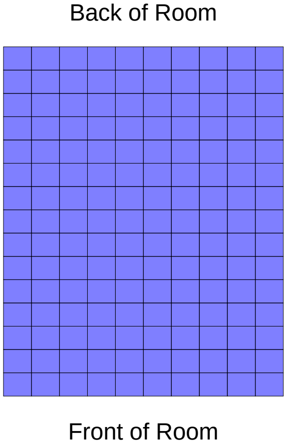
Back of Room



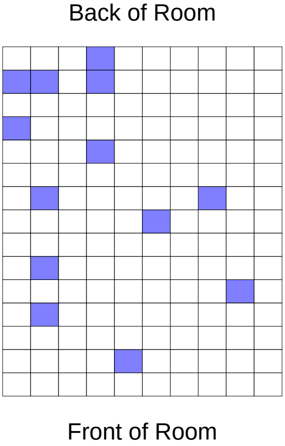
Front of Room

What are the advantages of each strategy?

Full Count Enumeration



Probability Sample



Probability sampling

What you need

Probability sampling

What you need	
target population	who you want to study

Probability sampling

What you need

target population	who you want to study
sampling frame	list of those people

Probability sampling

What you need	
target population	who you want to study
sampling frame	list of those people
sampling probability	e.g. 10%

Probability sampling

What you need

target population	who you want to study
sampling frame	list of those people
sampling probability	e.g. 10%
people you sampled	

Probability sampling

What you need	
target population	who you want to study
sampling frame	list of those people
sampling probability	e.g. 10%
people you sampled	
people who responded	

Probability sampling

Sources of error	What you need	
	target population	who you want to study
	sampling frame	list of those people
	sampling probability	e.g. 10%
	people you sampled	
	people who responded	

Probability sampling

Sources of error	**What you need**	
	target population	who you want to study
undercoverage	sampling frame	list of those people
	sampling probability	e.g. 10%
	people you sampled	
	people who responded	

Probability sampling

Sources of error	**What you need**	
	target population	who you want to study
undercoverage	sampling frame	list of those people
	sampling probability	e.g. 10%
sampling variability	people you sampled	
	people who responded	

Probability sampling

Sources of error	**What you need**	
	target population	who you want to study
undercoverage	sampling frame	list of those people
	sampling probability	e.g. 10%
sampling variability	people you sampled	
nonresponse	people who responded	

Probability sampling

Sources of error	**What you need**	
	target population	who you want to study
undercoverage	sampling frame	list of those people
	sampling probability	e.g. 10%
sampling variability	people you sampled	
nonresponse	people who responded	

Groves & Lyberg. 2010.
Total Survey Error: Past, Present, and Future.
Public Opinion Quarterly 74(5).

Subgroup estimates

Do the people in the first 3 rows prefer the front?

Subgroup estimates

Do the people in the first 3 rows prefer the front?

Simple random sample

- everyone run `runif`
- everyone respond if $< .1$

Subgroup estimates

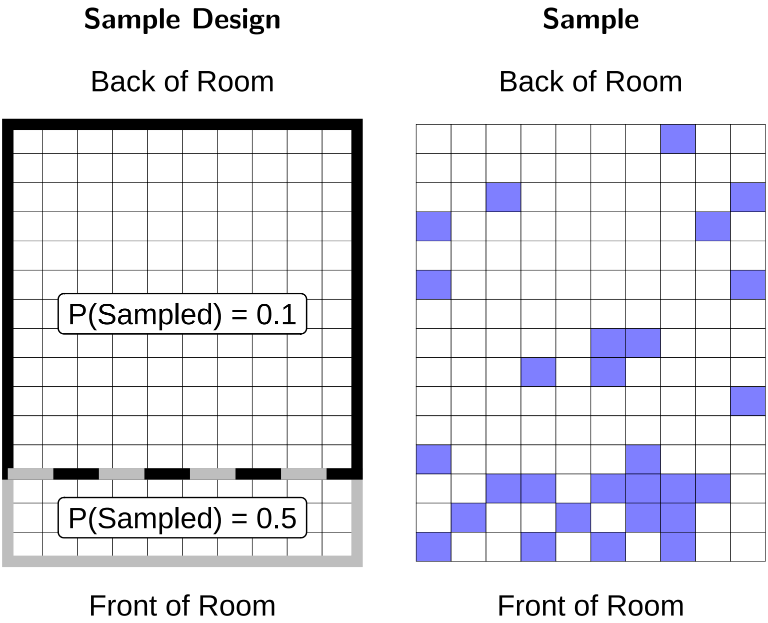
Do the people in the first 3 rows prefer the front?

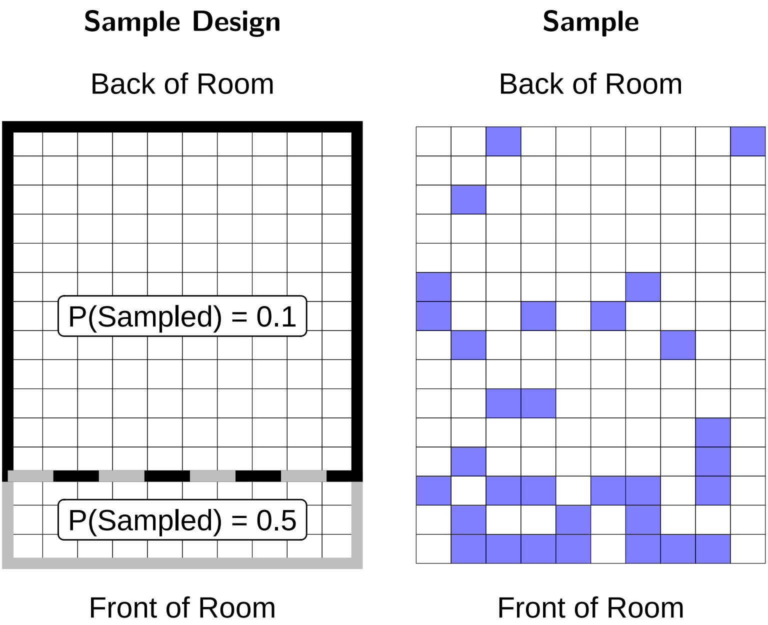
Simple random sample

- everyone run `runif`
- everyone respond if $< .1$

Unequal probability sample

- everyone run `runif`
- first 3 rows: respond if $< .5$
- others: respond if $< .1$


Sample Design
Sample
Back of Room
Back of Room
P(Sampled) = 0.1
P(Sampled) = 0.5
Front of Room
Front of Room

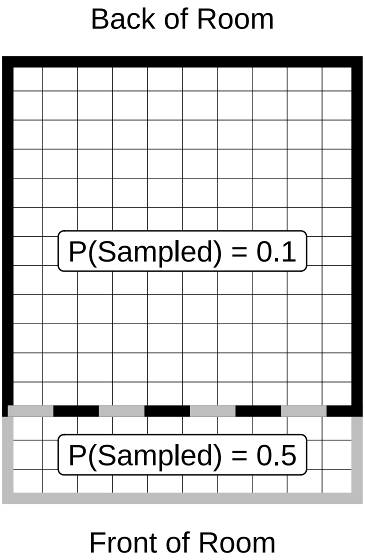


Sample Design

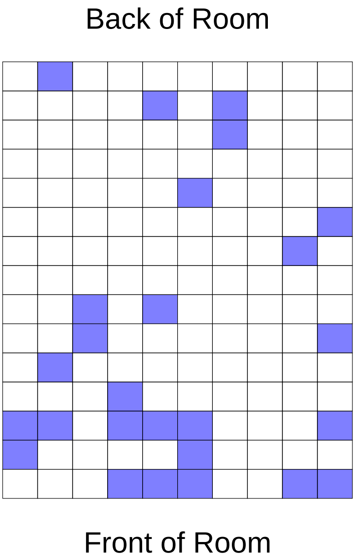
Sample

		🟦							🟦
	🟦								
🟦						🟦			
🟦			🟦		🟦				
	🟦						🟦		
		🟦	🟦						
								🟦	
	🟦							🟦	
🟦		🟦	🟦		🟦	🟦		🟦	
	🟦			🟦		🟦			
	🟦	🟦	🟦	🟦		🟦	🟦	🟦	

Sample Design



Sample



	■								
				■		■			
						■			
					■				
									■
								■	
		■		■					
		■							■
	■								
			■						
■	■		■	■	■				■
■					■				
			■	■	■			■	■

full count enumeration	talk to everyone
simple random sample	sampling frame known, equal probabilities
unequal probability sample	sampling frame known, unequal probabilities

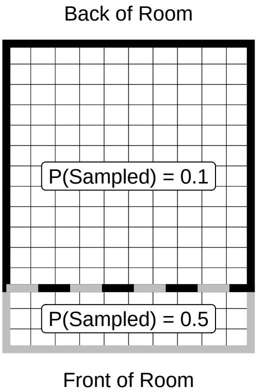
full count enumeration	talk to everyone (ideal but costly!)
simple random sample	sampling frame known, equal probabilities
unequal probability sample	sampling frame known, unequal probabilities

full count enumeration	talk to everyone (ideal but costly!)
simple random sample	sampling frame known, equal probabilities (good for population average)
unequal probability sample	sampling frame known, unequal probabilities

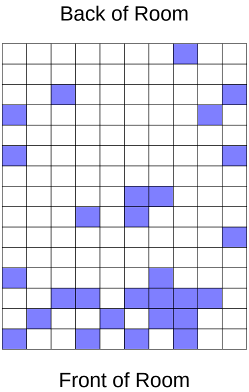
full count enumeration	talk to everyone (ideal but costly!)
simple random sample	sampling frame known, equal probabilities (good for population average)
unequal probability sample	sampling frame known, unequal probabilities (good for subgroups)

What if we want to estimate the population average from an unequal probability sample?

Sample Design



Sample



Sampling weights: Population mean estimator

Among those sampled if `runif` $< .1$,
on average 1 in 10 people sampled.

Sampling weights: Population mean estimator

Among those sampled if `runif` $< .1$,
on average 1 in 10 people sampled.
Each person represents 10 people.

Sampling weights: Population mean estimator

Among those sampled if `runif` $< .1$,
on average 1 in 10 people sampled.
Each person represents 10 people.
$w_i = \frac{1}{\text{P(Sampled)}} = \frac{1}{.1} = 10$

Sampling weights: Population mean estimator

Among those sampled if `runif` $< .1$,
on average 1 in 10 people sampled.
Each person represents 10 people.
$w_i = \frac{1}{\text{P(Sampled)}} = \frac{1}{.1} = 10$

Among those who sampled if `runif` $< .5$,
on average 1 in 2 people sampled.
Each person represents 2 people.
$w_i = \frac{1}{\text{P(Sampled)}} = \frac{1}{.5} = 2$

Sampling weights: Population mean estimator

Among those sampled if `runif` $< .1$,
on average 1 in 10 people sampled.
Each person represents 10 people.
$w_i = \frac{1}{\text{P(Sampled)}} = \frac{1}{.1} = 10$

Among those who sampled if `runif` $< .5$,
on average 1 in 2 people sampled.
Each person represents 2 people.
$w_i = \frac{1}{\text{P(Sampled)}} = \frac{1}{.5} = 2$

Unweighted estimator

$$\hat{\text{E}}_{\text{Unweighted}}(Y) = \frac{\sum_i y_i}{n}$$

(easily misleading!)

Weighted estimator

$$\hat{\text{E}}_{\text{Weighted}}(Y) = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

(correct)

full count enumeration	talk to everyone (ideal but costly!)
simple random sample	sampling frame known, equal probabilities (good for population average)
unequal probability sample	sampling frame known, unequal probabilities (good for subgroups) **(weight for population average)**

A real question:
The unemployment rate

A real question: The unemployment rate

Imagine you are the Bureau of Labor Statistics.
How would you design a sample to estimate unemployment?

1. What would be your sampling frame?
2. How would you define sampling probabilities?
3. What mode of data collection?
    - Mail, phone, web, in person, etc.
4. What if people didn't respond?

Current Population Survey: Sample Design

U.S. BUREAU OF LABOR STATISTICS

Current Population Survey: Sample Design

U.S. BUREAU OF LABOR STATISTICS

Begin with a **sampling frame**: all housing units in the U.S.

Current Population Survey: Sample Design

U.S. BUREAU OF LABOR STATISTICS

Begin with a **sampling frame**: all housing units in the U.S.

- 1,987 Primary Sampling Units (PSUs)
    - County or contiguous counties within a state

Current Population Survey: Sample Design

U.S. BUREAU OF LABOR STATISTICS

Begin with a **sampling frame**: all housing units in the U.S.

- 1,987 Primary Sampling Units (PSUs)
  - County or contiguous counties within a state
- Stratified (grouped) within states
  - Stratum: Group of PSUs with similar characteristics
  - One PSU always chosen per stratum
  - **Why?** Ensure representation across strata

Current Population Survey: Sample Design

Begin with a **sampling frame**: all housing units in the U.S.

- 1,987 Primary Sampling Units (PSUs)
  - County or contiguous counties within a state
- Stratified (grouped) within states
  - Stratum: Group of PSUs with similar characteristics
  - One PSU always chosen per stratum
  - **Why?** Ensure representation across strata
- Within PSU, sample geographic clusters of housing units
  - **Why?** Reduce travel costs for field representatives

Current Population Survey: Sample

More than 75,000 households are sampled

Current Population Survey: Contacting respondents

U.S. BUREAU OF LABOR STATISTICS

United States™ Census Bureau

Current Population Survey: Contacting respondents

1. Send a letter

Current Population Survey: Contacting respondents

1. Send a letter
2. Call or visit in person

Current Population Survey: Contacting respondents

1. Send a letter
2. Call or visit in person
3. Try many times if needed

Current Population Survey: Contacting respondents

U.S. BUREAU OF LABOR STATISTICS

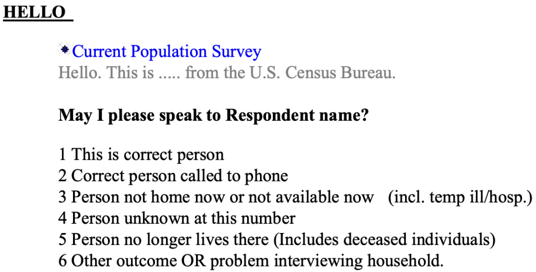
United States™ Census Bureau

1. Send a letter
2. Call or visit in person
3. Try many times if needed

Learn about the experience for participants here

Current Population Survey: Mode of Data Collection

Computer-assisted telephone interview



HELLO

Current Population Survey

Hello. This is ..... from the U.S. Census Bureau.

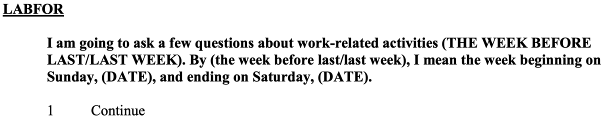
May I please speak to Respondent name?

1 This is correct person
2 Correct person called to phone
3 Person not home now or not available now (incl. temp ill/hosp.)
4 Person unknown at this number
5 Person no longer lives there (Includes deceased individuals)
6 Other outcome OR problem interviewing household.

census.gov/programs-surveys/cps/technical-documentation/questionnaires.html

Current Population Survey: Mode of Data Collection

Computer-assisted telephone interview



LABFOR

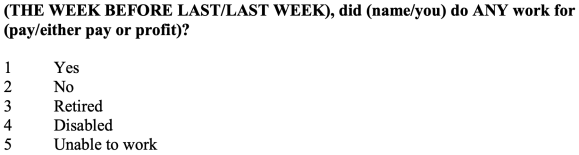
I am going to ask a few questions about work-related activities (THE WEEK BEFORE LAST/LAST WEEK). By (the week before last/last week), I mean the week beginning on Sunday, (DATE), and ending on Saturday, (DATE).

1 Continue

census.gov/programs-surveys/cps/technical-documentation/questionnaires.html

Current Population Survey: Mode of Data Collection

Computer-assisted telephone interview



(THE WEEK BEFORE LAST/LAST WEEK), did (name/you) do ANY work for (pay/either pay or profit)?

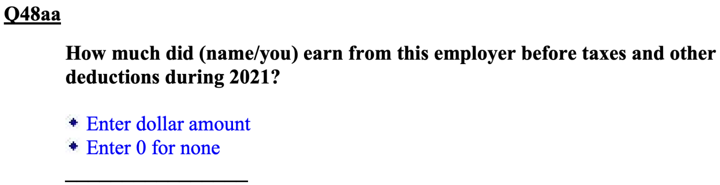
1 Yes
2 No
3 Retired
4 Disabled
5 Unable to work

census.gov/programs-surveys/cps/technical-documentation/questionnaires.html

Current Population Survey

Annual Social and Economic Supplement

- Extended survey
- Conducted each March



Q48aa

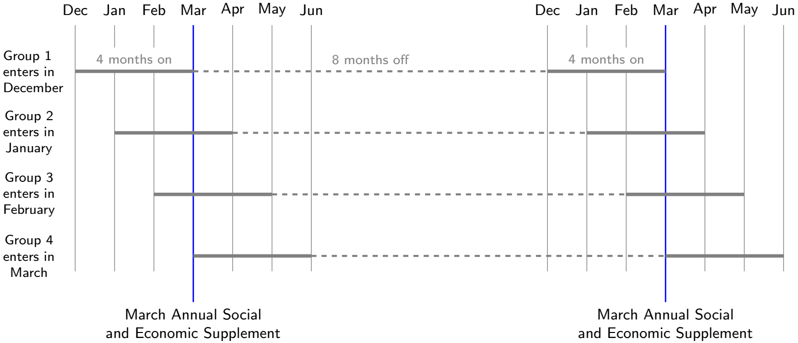
How much did (name/you) earn from this employer before taxes and other deductions during 2021?

- Enter dollar amount
- Enter 0 for none

____

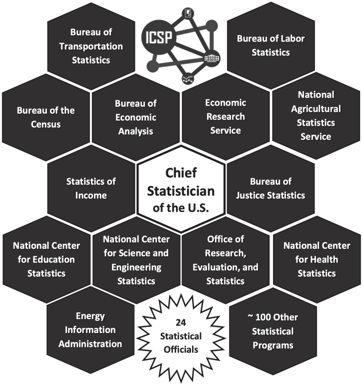
Questionnaire from 2022

Rotating panels


Dec
Jan
Feb
Mar
Apr
May
Jun
Dec
Jan
Feb
Mar
Apr
May
Jun
Group 1 enters in December
4 months on
8 months off
4 months on
Group 2 enters in January
Group 3 enters in February
Group 4 enters in March
March Annual Social and Economic Supplement
March Annual Social and Economic Supplement

The Current Population Survey (CPS) is only one of many surveys in the federal statistical system

Chart 15-1. THE DECENTRALIZED FEDERAL STATISTICAL SYSTEM



ICSP
Bureau of Transportation Statistics
Bureau of Labor Statistics
Bureau of the Census
Bureau of Economic Analysis
Economic Research Service
National Agricultural Statistics Service
Statistics of Income
Chief Statistician of the U.S.
Bureau of Justice Statistics
National Center for Education Statistics
National Center for Science and Engineering Statistics
Office of Research, Evaluation, and Statistics
National Center for Health Statistics
Energy Information Administration
24 Statistical Officials
~ 100 Other Statistical Programs

source

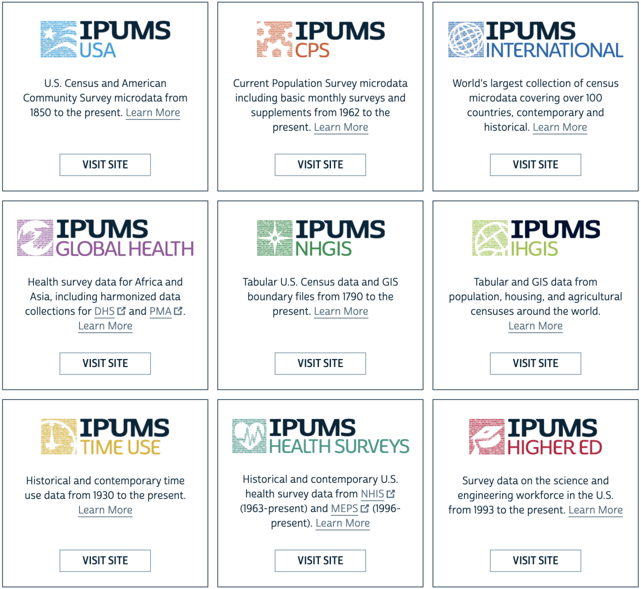
IPUMS CPS

The Integrated Public Use Microdata Series (IPUMS) distributes these data and more

- Easy to access
- Harmonized documentation
- Select the variables you want
- Compare over history

UNIVERSITY OF MINNESOTA

Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren and Michael Westberry. Integrated Public Use Microdata Series, Current Population Survey: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2022. https://doi.org/10.18128/D030.V10.0



IPUMS USA

U.S. Census and American Community Survey microdata from 1850 to the present. Learn More

VISIT SITE

IPUMS CPS

Current Population Survey microdata including basic monthly surveys and supplements from 1962 to the present. Learn More

VISIT SITE

IPUMS INTERNATIONAL

World's largest collection of census microdata covering over 100 countries, contemporary and historical. Learn More

VISIT SITE

IPUMS GLOBAL HEALTH

Health survey data for Africa and Asia, including harmonized data collections for DHS and PMA. Learn More

VISIT SITE

IPUMS NHGIS

Tabular U.S. Census data and GIS boundary files from 1790 to the present. Learn More

VISIT SITE

IPUMS IHGIS

Tabular and GIS data from population, housing, and agricultural censuses around the world. Learn More

VISIT SITE

IPUMS TIME USE

Historical and contemporary time use data from 1930 to the present. Learn More

VISIT SITE

IPUMS HEALTH SURVEYS

Historical and contemporary U.S. health survey data from NHIS (1963-present) and MEPS (1996-present). Learn More

VISIT SITE

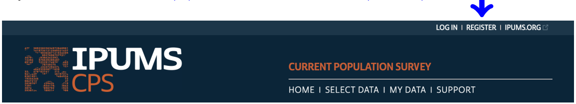
IPUMS HIGHER ED

Survey data on the science and engineering workforce in the U.S. from 1993 to the present. Learn More

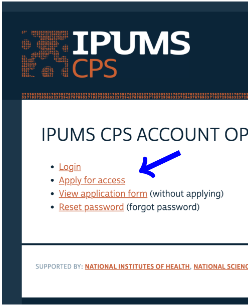
VISIT SITE

How to access IPUMS-CPS

1) Visit https://cps.ipums.org/cps/. Click **Register**



2) Click **Apply for access**

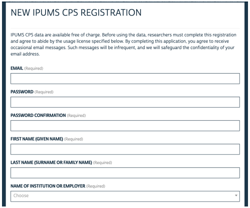


IPUMS CPS ACCOUNT OP

- Login
- Apply for access
- View application form (without applying)
- Reset password (forgot password)

SUPPORTED BY: NATIONAL INSTITUTES OF HEALTH, NATIONAL SCIENC

3) Complete the form



NEW IPUMS CPS REGISTRATION

IPUMS CPS data are available free of charge. Before using the data, researchers must complete this registration and agree to abide by the usage license specified below. By completing this application, you agree to receive occasional email messages. Such messages will be infrequent, and we will safeguard the confidentiality of your email address.

EMAIL (Required)

PASSWORD (Required)

PASSWORD CONFIRMATION (Required)

FIRST NAME (GIVEN NAME) (Required)

LAST NAME (SURNAME OR FAMILY NAME) (Required)

NAME OF INSTITUTION OR EMPLOYER (Required)

Choose

General research statement: I am in a class using these data to study socioeconomic inequality in America.

Learning goals for today

By the end of class, you will be able to

- explain key ideas of data collection
  - target population
  - sampling frame
  - undercoverage
  - simple random sample
  - unequal probability sample
- access survey data online