Social Data Science

SOCIOL 114
Winter 2026

Sampling for Population Inference

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

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

- A) Front of the room
- B) Back of the room

Full count enumeration

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

Simple Random Sampling

Simple random sampling

Open R. Run this line

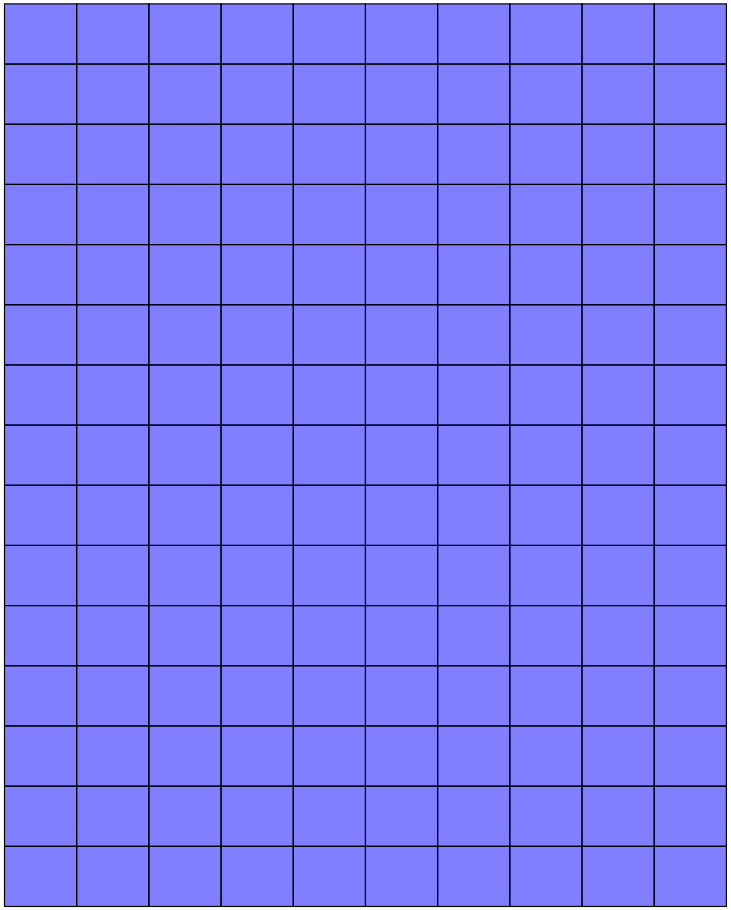
`runif(n = 1)`

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

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

Full Count Enumeration

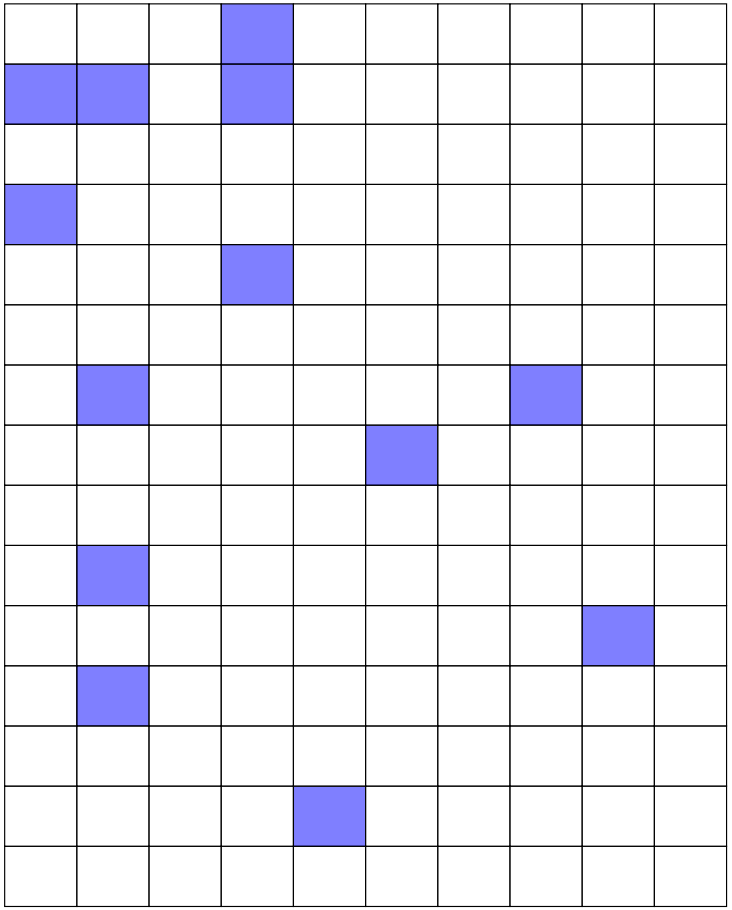
Back of Room



Front of Room

Probability Sample

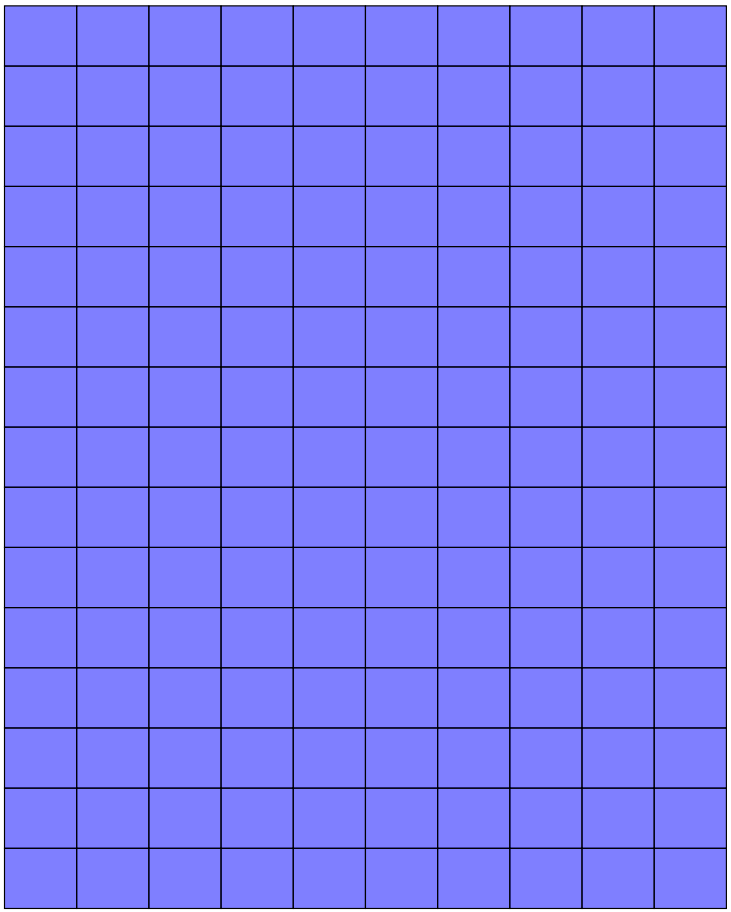
Back of Room



Front of Room

Full Count Enumeration

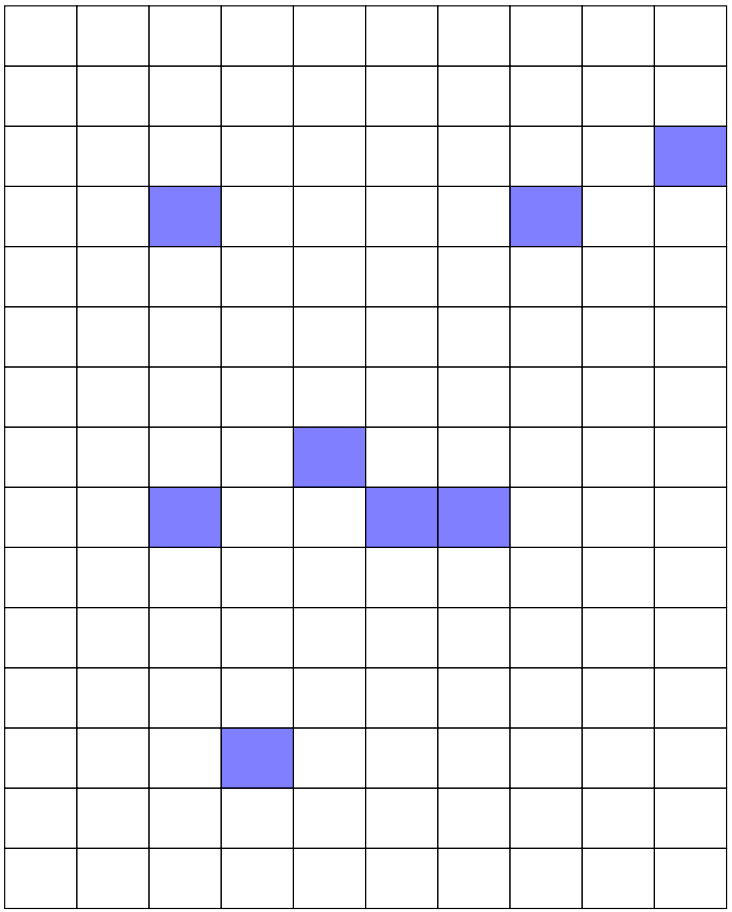
Back of Room



Front of Room

Probability Sample

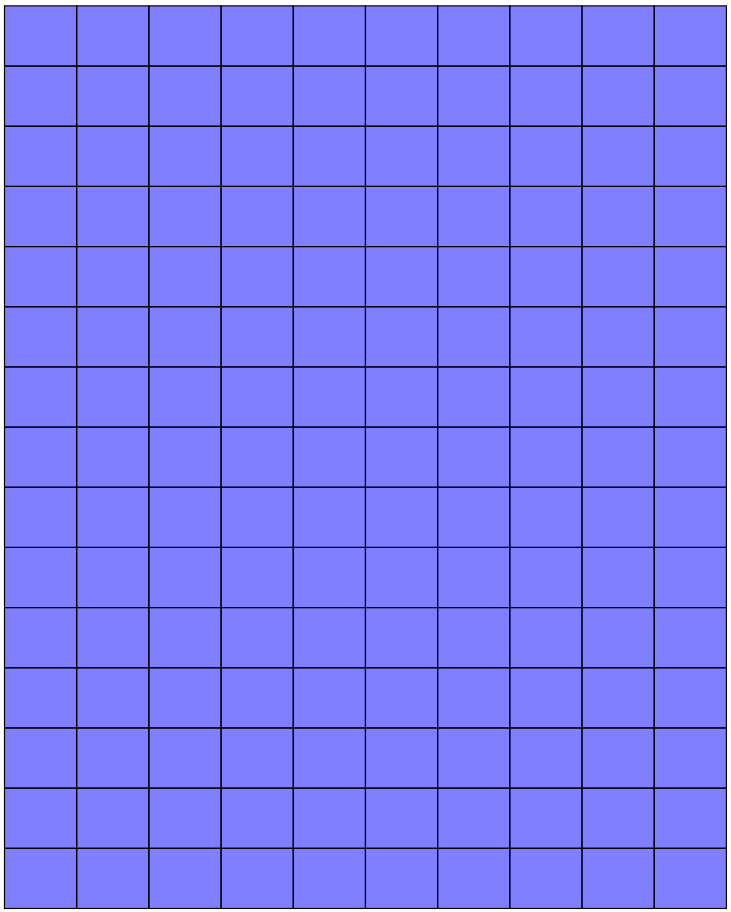
Back of Room



Front of Room

Full Count Enumeration

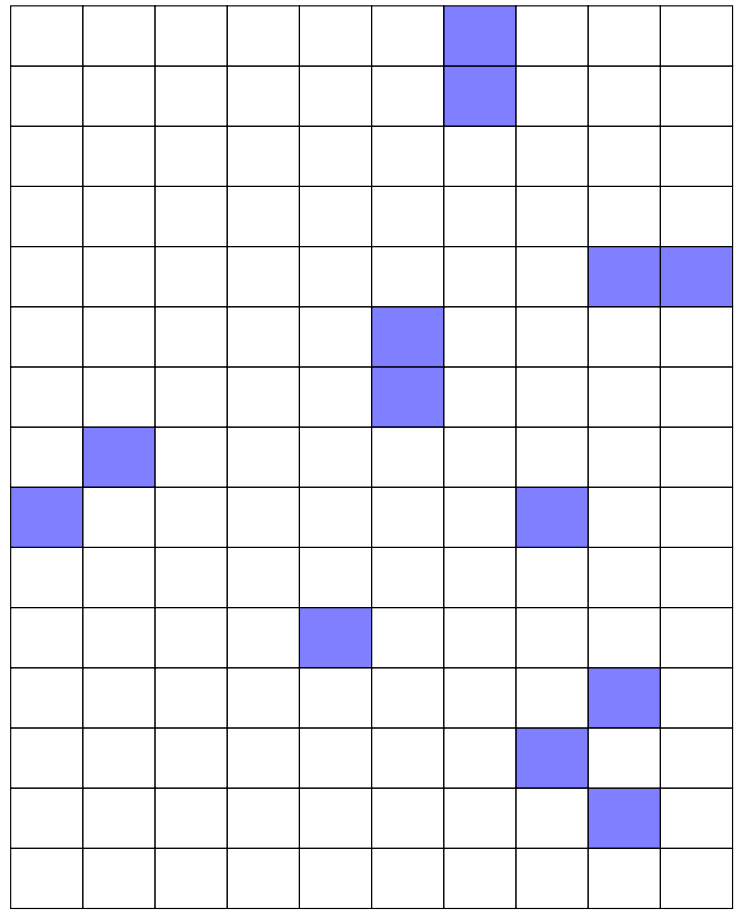
Back of Room



Front of Room

Probability Sample

Back of Room

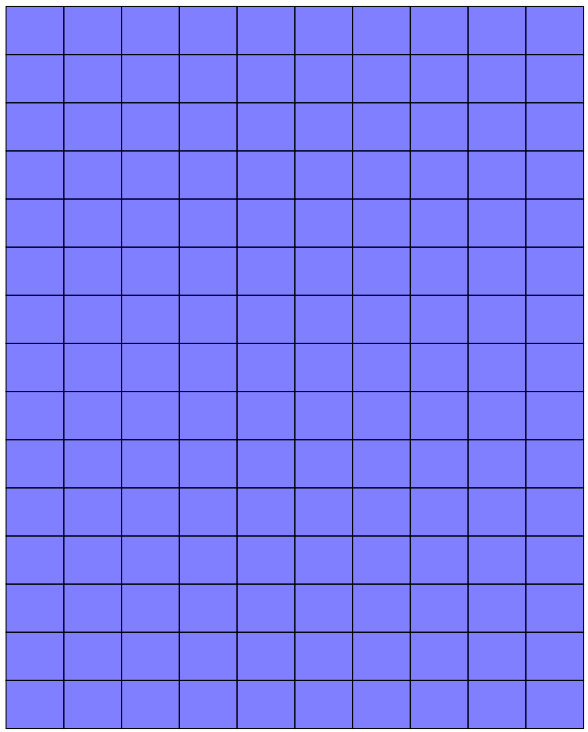


Front of Room

What are the advantages of each strategy?

Full Count Enumeration

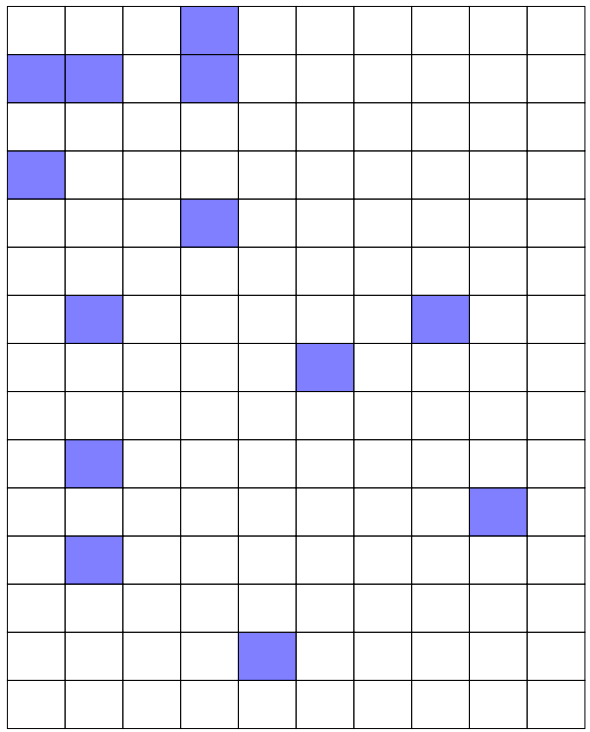
Back of Room



Front of Room

Probability Sample

Back of Room



Front of Room

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).

Unequal Probability Sampling

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

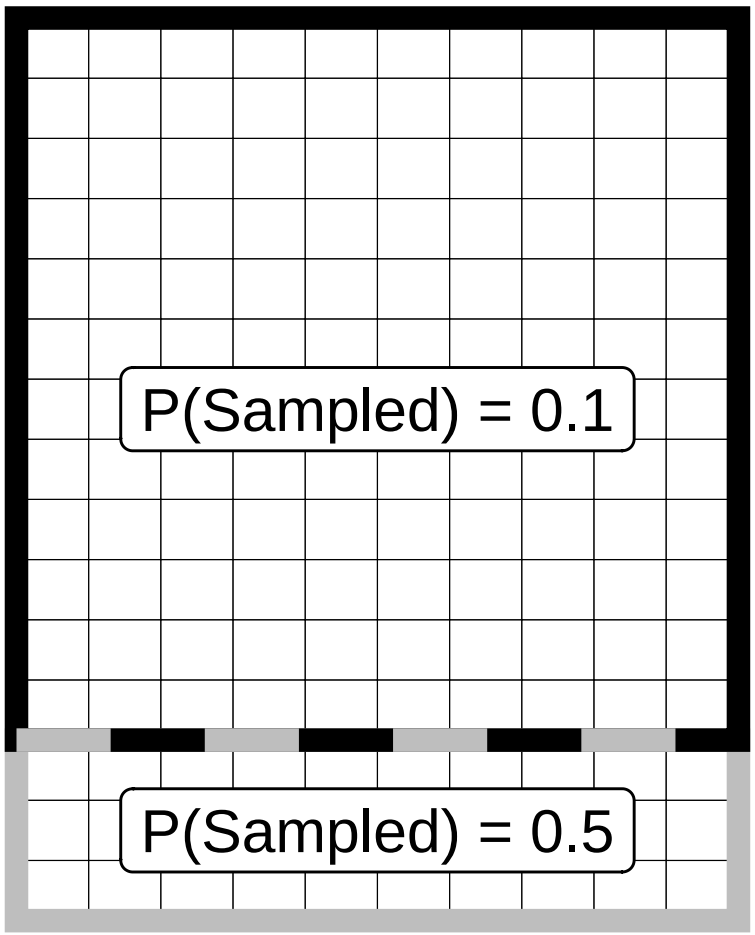
- everyone run `runif`
- everyone respond if $< .1$

Unequal probability sample

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

Sample Design

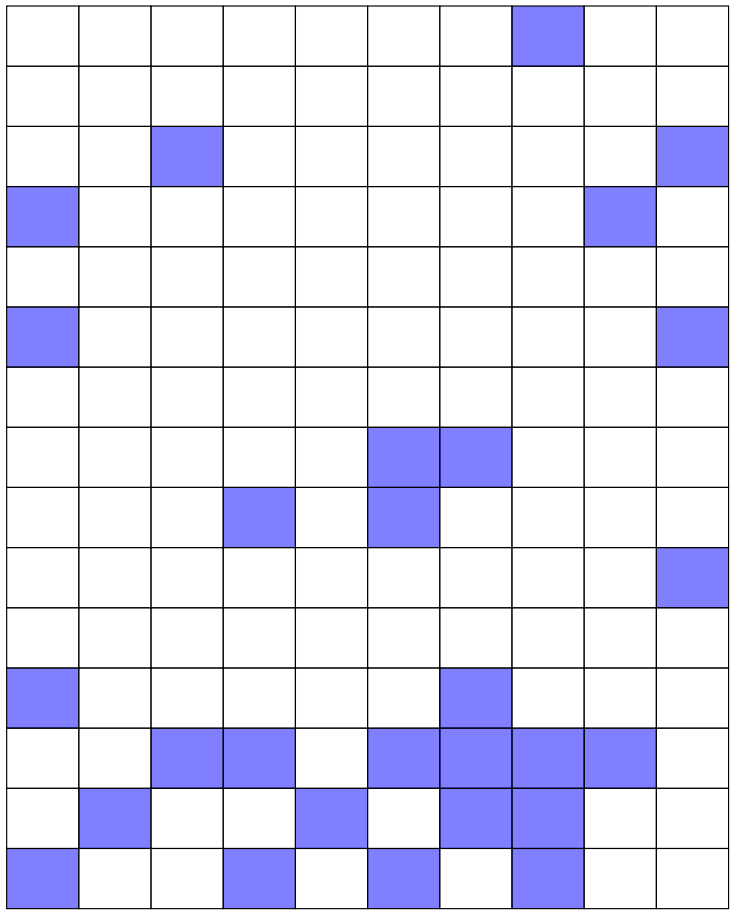
Back of Room



Front of Room

Sample

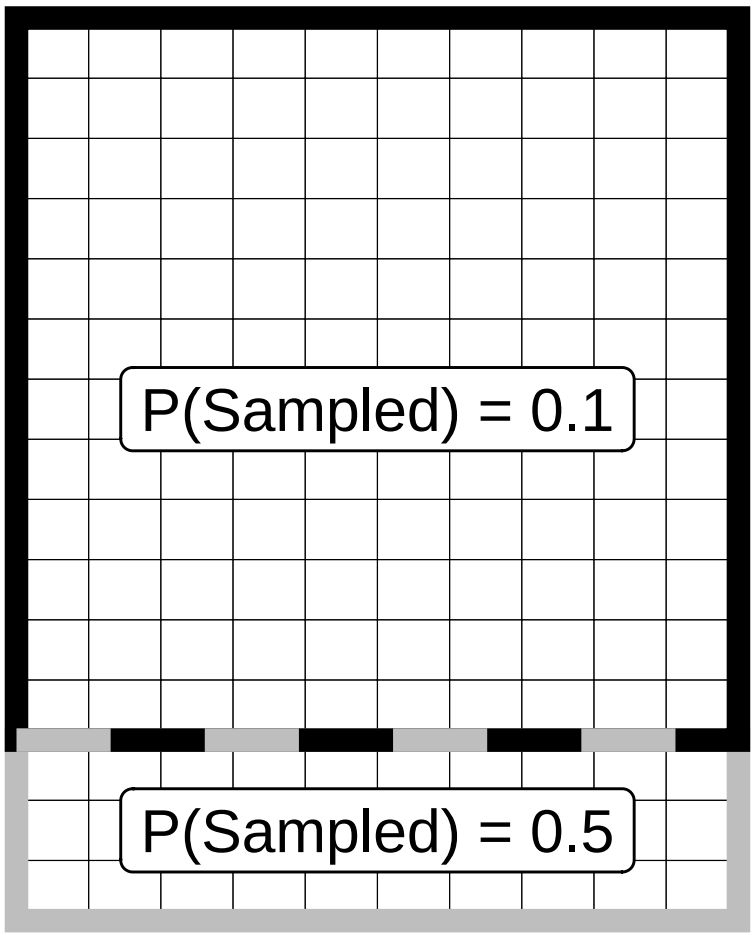
Back of Room



Front of Room

Sample Design

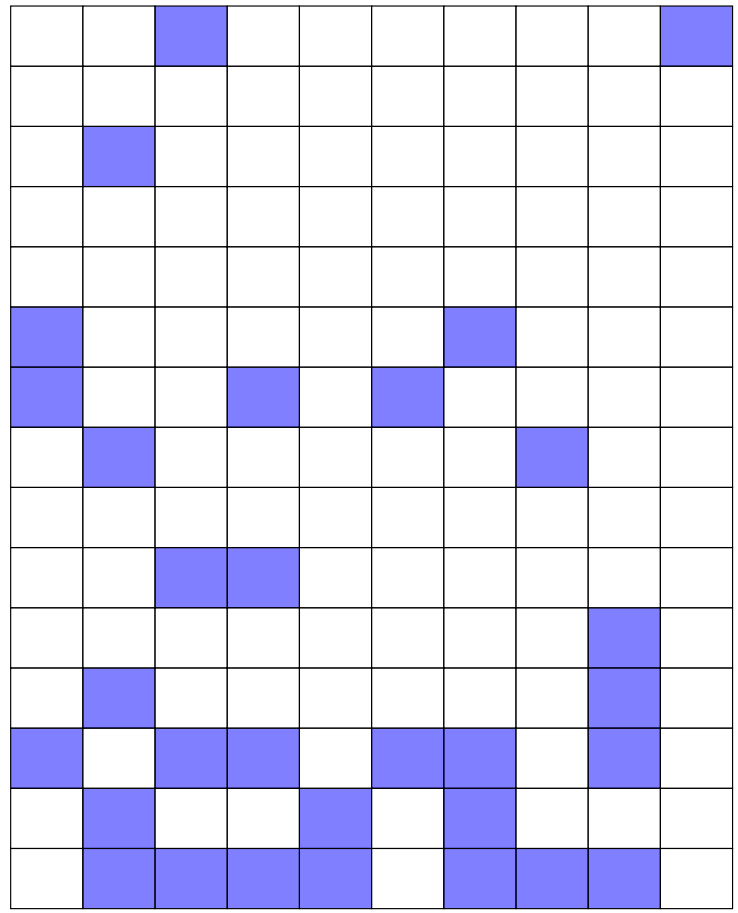
Back of Room



Front of Room

Sample

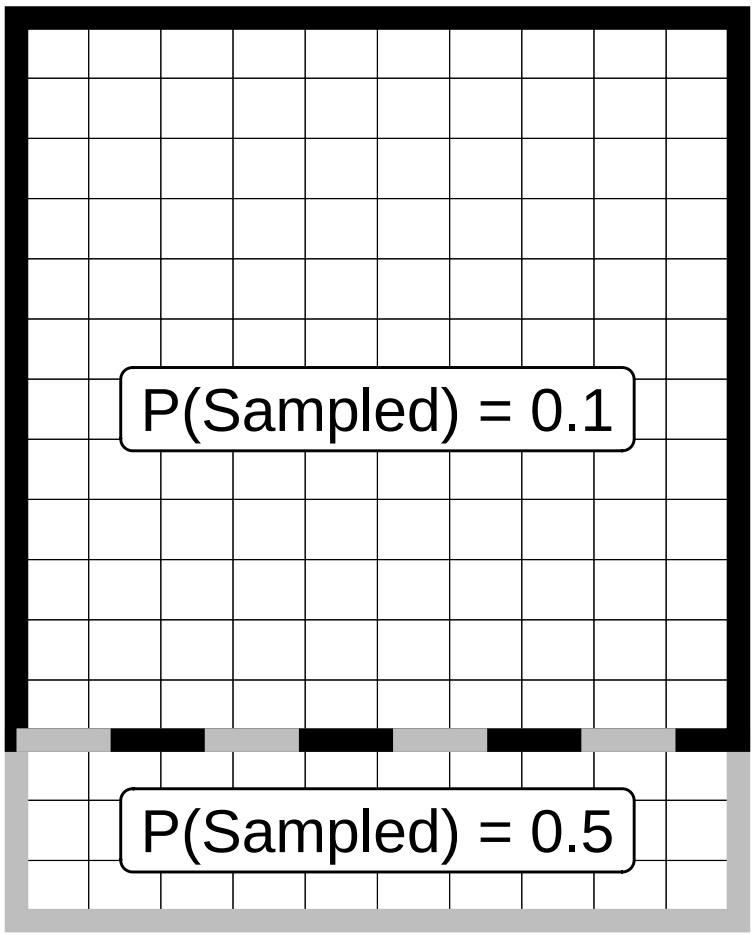
Back of Room



Front of Room

Sample Design

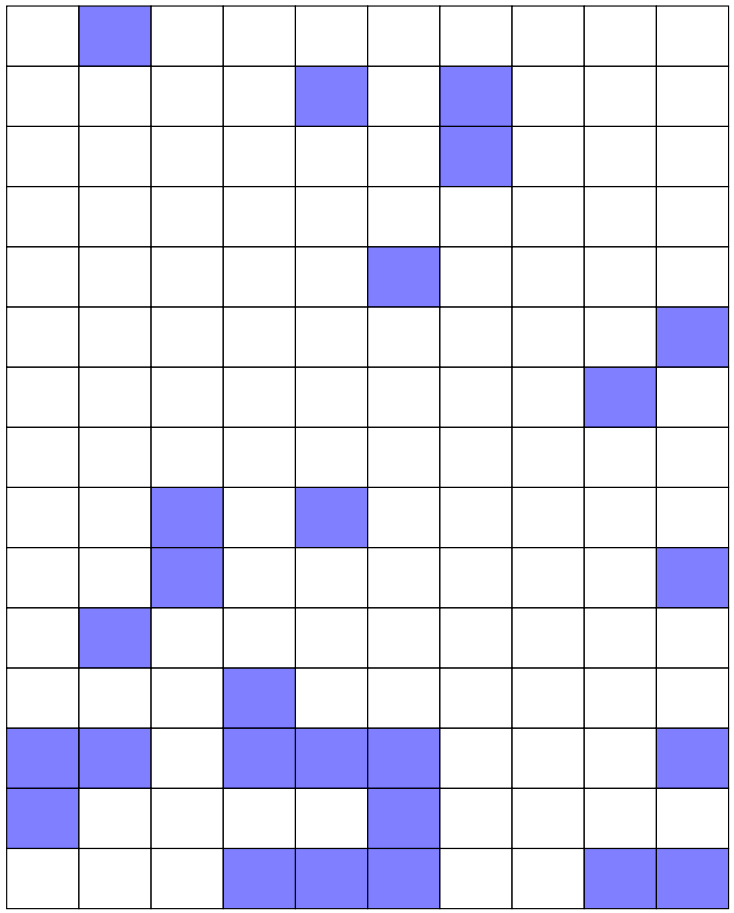
Back of Room



Front of Room

Sample

Back of Room



Front of Room

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)

Population total from unequal probability sample

- People indexed $i = 1, \ldots N$
- Sample $S_i = 1$ with probability π_i
- Observe Y_i for n units with $S_i = 1$

Question:
How to estimate the population total $\sum_{i=1}^{N} Y_i$ using the sample, all i such that $S_i = 1$?

Population total from unequal probability sample

Suppose $\pi_i = \frac{1}{10}$

Population total from unequal probability sample

Suppose $\pi_i = \frac{1}{10}$

- For each 1 sampled person there are 9 unsampled people

Population total from unequal probability sample

Suppose $\pi_i = \frac{1}{10}$

- For each 1 sampled person there are 9 unsampled people
- Each sampled i counts for 10 people

Population total from unequal probability sample

Suppose $\pi_i = \frac{1}{10}$

- For each 1 sampled person there are 9 unsampled people
- Each sampled i counts for 10 people

Suppose $\pi_i = \frac{1}{2}$

Population total from unequal probability sample

Suppose $\pi_i = \frac{1}{10}$

- For each 1 sampled person there are 9 unsampled people
- Each sampled i counts for 10 people

Suppose $\pi_i = \frac{1}{2}$

- For each 1 sampled person there is 1 unsampled person

Population total from unequal probability sample

Suppose $\pi_i = \frac{1}{10}$

- For each 1 sampled person there are 9 unsampled people
- Each sampled i counts for 10 people

Suppose $\pi_i = \frac{1}{2}$

- For each 1 sampled person there is 1 unsampled person
- Each sampled i counts for 2 people

Population total from unequal probability sample

Suppose $\pi_i = \frac{1}{10}$

- For each 1 sampled person there are 9 unsampled people
- Each sampled i counts for 10 people

Suppose $\pi_i = \frac{1}{2}$

- For each 1 sampled person there is 1 unsampled person
- Each sampled i counts for 2 people

General: Each sampled person counts for $\frac{1}{\pi_i}$ people.

Estimator of population total: $\underbrace{\sum_{i:S_i=1}}_{\substack{\text{Sum over} \\ \text{sampled} \\ \text{people}}} \underbrace{Y_i}_{\substack{\text{Observed} \\ \text{Outcome}}} \underbrace{\frac{1}{\pi_i}}_{\substack{\text{Number} \\ \text{of people} \\ \text{represented}}}$

Population mean from unequal probability sample

For population total $\tau = \sum_{i=1}^{N} Y_i$,
let our estimator be $\hat{\tau} = \sum_{i:S_i=1} \frac{Y_i}{\pi_i}$

Population mean from unequal probability sample

For population total $\tau = \sum_{i=1}^{N} Y_i$,
let our estimator be $\hat{\tau} = \sum_{i:S_i=1} \frac{Y_i}{\pi_i}$

Suppose we know the population size N.
How could we estimate the population mean?

Population mean from unequal probability sample

For population total $\tau = \sum_{i=1}^{N} Y_i$,
let our estimator be $\hat{\tau} = \sum_{i:S_i=1} \frac{Y_i}{\pi_i}$

Suppose we know the population size N.
How could we estimate the population mean?
The Horvitz-Thompson estimator:

$$\begin{aligned}\hat{\mu} &= \frac{1}{N}\hat{\tau} \\ &= \frac{1}{N} \sum_{i:S_i=1} \frac{Y_i}{\pi_i}\end{aligned}$$

Recap: Sampling strategies so far

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)**

Stratified and Clustered Sampling

Baseball salaries

BASEBALL

The New York Times

GIVE THE TIMES

Channeling the Old Steinbrenner Ways, Yankees Stepped Up for Judge

Aaron Judge, who hit 62 home runs in 2022, agreed to a nine-year, $360 million contract with the Yankees after meeting with at least two other teams.

Share full article 129

Aaron Judge set career highs in batting average (.311), home runs (62) and R.B.I. (131) in 2022. Chris Donovan for The New York Times

Sections

Los Angeles Times

SUBSCRIBE LOG IN

Dodgers news | Teoscar Hernández | California dreaming | Dodgers pitchers rising | $1 billion boon?

DODGERS

Complete coverage: Shohei Ohtani signs record deal with Dodgers

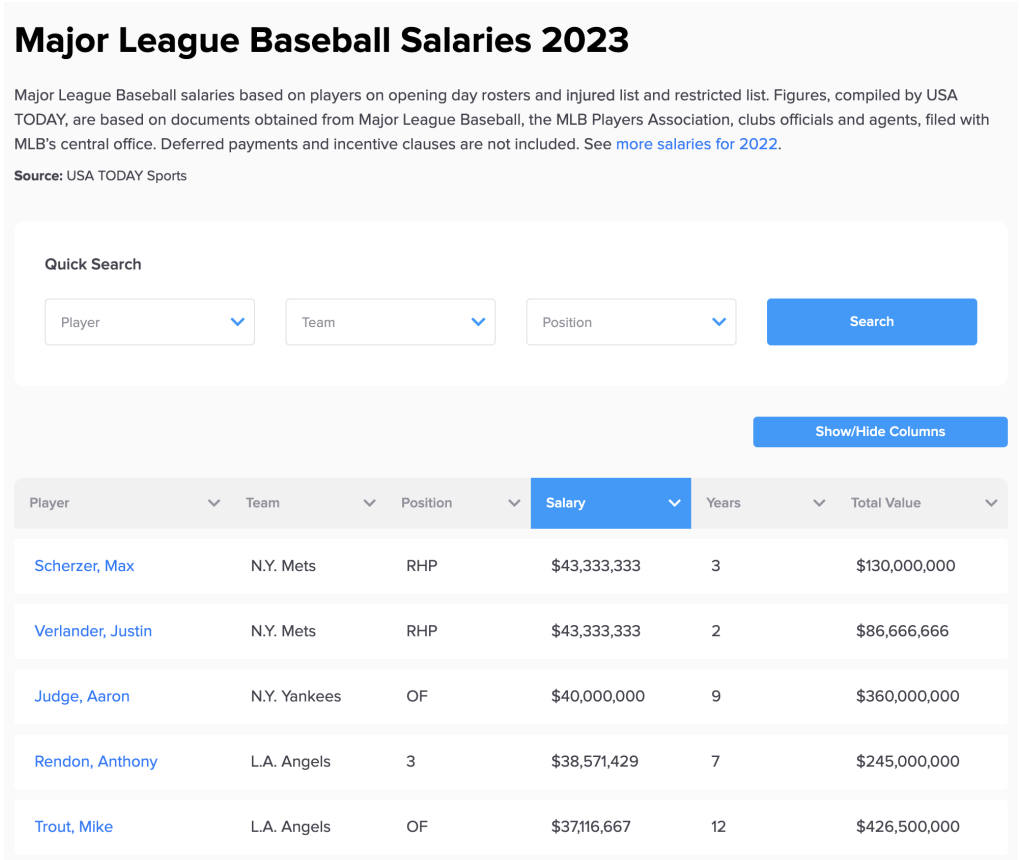
Shohei Ohtani speaks during his introductory Dodgers news conference at Dodger Stadium on Thursday. (Wally Skalij / Los Angeles Times)

BY LOS ANGELES TIMES STAFF

PUBLISHED DEC. 9, 2023 | **UPDATED** DEC. 22, 2023 8:54 AM PT

Baseball salaries

```
baseball <- read_csv("https://soc114.github.io/data/baseball.csv")
```



Major League Baseball Salaries 2023

Major League Baseball salaries based on players on opening day rosters and injured list and restricted list. Figures, compiled by USA TODAY, are based on documents obtained from Major League Baseball, the MLB Players Association, clubs officials and agents, filed with MLB's central office. Deferred payments and incentive clauses are not included. See more salaries for 2022.

Source: USA TODAY Sports

Quick Search

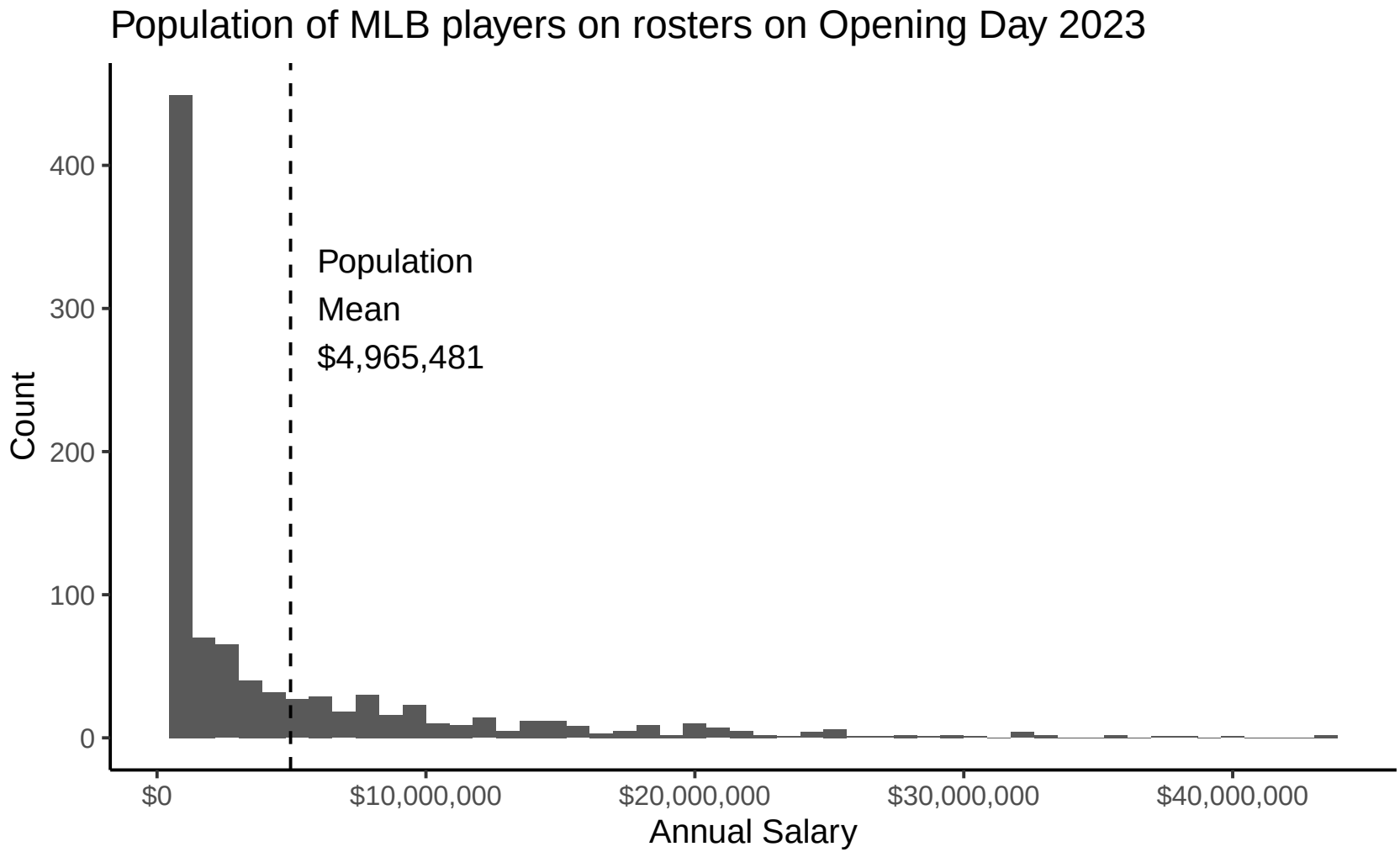
Player | Team | Position | Search

Show/Hide Columns

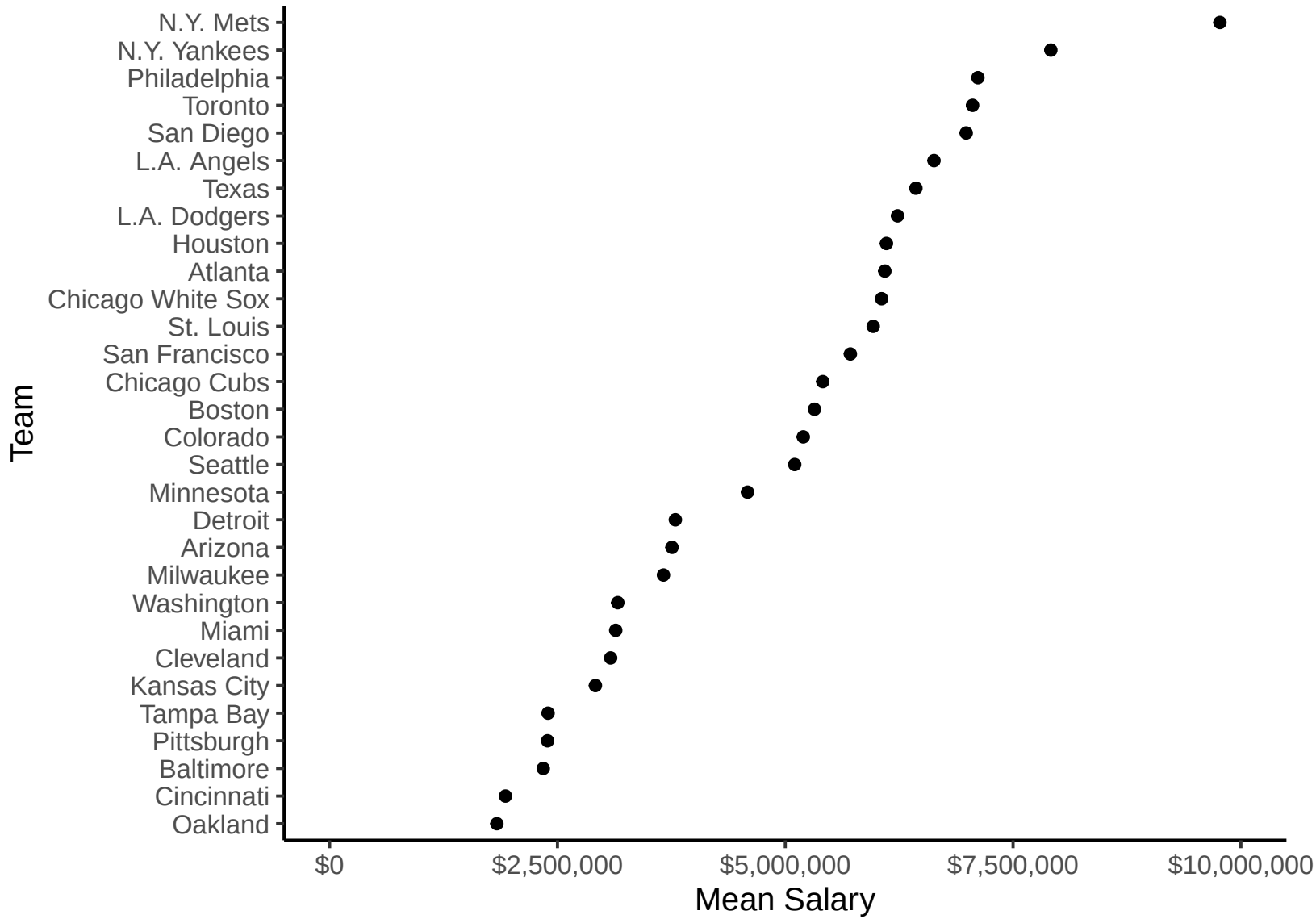
Player	Team	Position	Salary	Years	Total Value
Scherzer, Max	N.Y. Mets	RHP	$43,333,333	3	$130,000,000
Verlander, Justin	N.Y. Mets	RHP	$43,333,333	2	$86,666,666
Judge, Aaron	N.Y. Yankees	OF	$40,000,000	9	$360,000,000
Rendon, Anthony	L.A. Angels	3	$38,571,429	7	$245,000,000
Trout, Mike	L.A. Angels	OF	$37,116,667	12	$426,500,000

databases.usatoday.com/major-league-baseball-salaries-2023/

Baseball salaries



Baseball salaries



How to sample baseball players: Clustered

Players are grouped in 30 teams.

- Suppose it is costly to contact a team
- It is cheap to gather salary for many players on the team
- How would you draw a survey of 150 players?

How to sample baseball players: Stratified

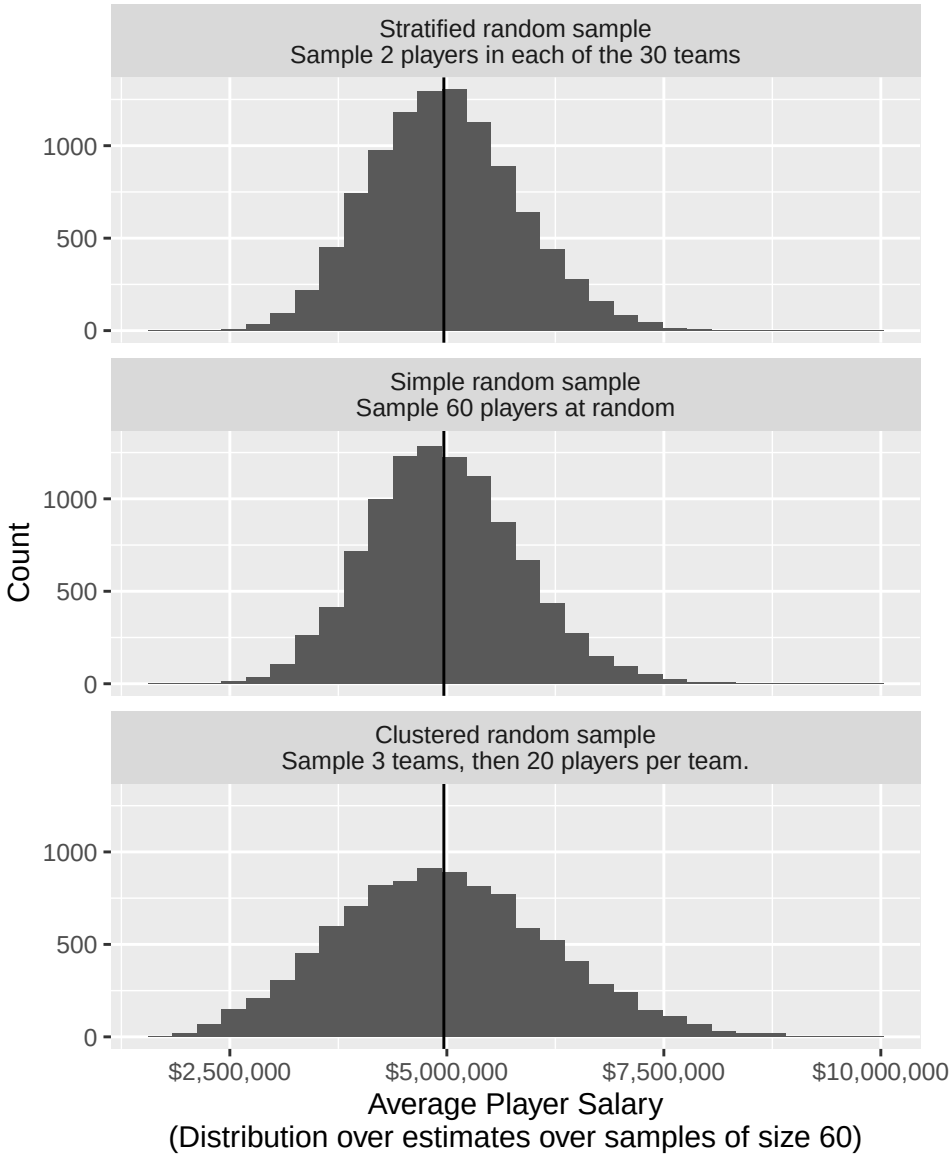
Players are grouped in 30 teams.

- Suppose salary varies a lot across teams
- You want a sample that represents the salary distribution well
- How would you draw a survey of 60 players?

Three sampling strategies

- Simple random: 60 players at random
- Stratified by team: 2 players per team
- Clustered by team: 20 players on each of 3 teams

Three sampling strategies



Three sampling strategies

- Simple random: 60 players at random
- Stratified by team: 2 players per team
  - less variable sample-to-sample
  - more costly
- Clustered by team: 20 players on each of 3 teams
  - more variable sample-to-sample
  - less costly

For reference: reading

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

The Future of Sample Surveys

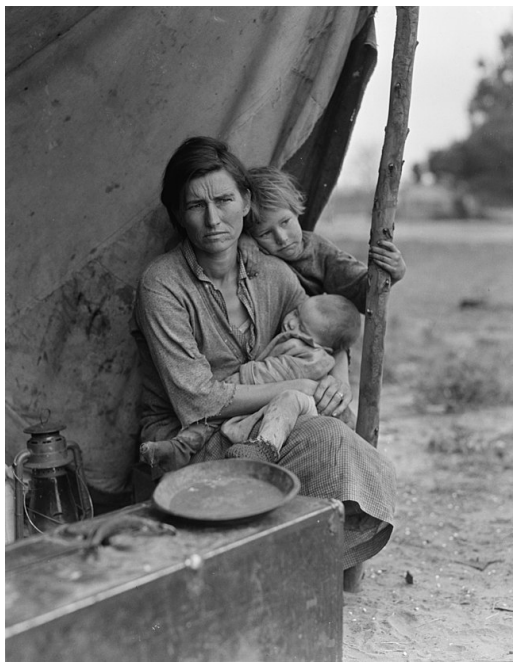
Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1930–1960: Era of Invention

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

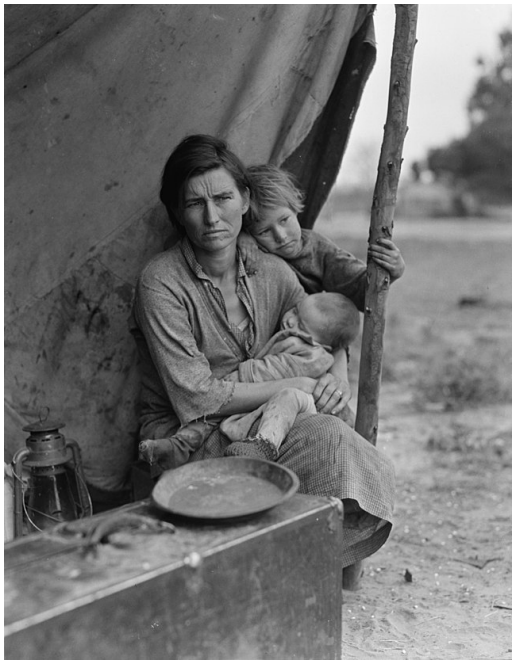
1930–1960: Era of Invention



The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1930–1960: Era of Invention



The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1930–1960: Era of Invention

sampling frame pieces of land

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1930–1960: Era of Invention

sampling frame	pieces of land
mode	face-to-face interviews

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1930–1960: Era of Invention

sampling frame	pieces of land
mode	face-to-face interviews
cost	high

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1930–1960: Era of Invention

sampling frame	pieces of land
mode	face-to-face interviews
cost	high
response rate	over 90 percent

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1960–1990: Era of Expansion

Technology helped: Telephones

Source: Wikimedia

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1960–1990: Era of Expansion

Technology helped: Telephones

— sampling frame

Source: Wikimedia

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1960–1990: Era of Expansion

Technology helped: Telephones

— sampling frame

— mode of data collection

Source: Wikimedia

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1960–1990: Era of Expansion

Technology helped: Telephones

— sampling frame

— mode of data collection

— falling costs

Source: Wikimedia

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1960–1990: Era of Expansion

Technology helped: Telephones

— sampling frame

— mode of data collection

— falling costs

— falling response rates

Source: Wikimedia

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present

Technology brought challenges

Technology brought opportunities

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present

Technology brought challenges — answering machines

Technology brought opportunities

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present

Technology brought challenges

— answering machines

— cell phones

Technology brought opportunities

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present

Technology brought challenges

— answering machines

— cell phones

— caller ID

Technology brought opportunities

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present

Technology brought challenges

— answering machines

— cell phones

— caller ID

— response rates plummeted

Technology brought opportunities

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present

Technology brought challenges

— answering machines

— cell phones

— caller ID

— response rates plummeted

Technology brought opportunities

— digital trace data

— internet panels

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present: Designed and Organic Data

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present: Designed and Organic Data

Designed data

Organic data

Example
Census age distribution

Example
Web histories

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present: Designed and Organic Data

Designed data
— high cost

Organic data
— almost free

Example
Census age distribution

Example
Web histories

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present: Designed and Organic Data

Designed data

— high cost

— becoming scarce

Organic data

— almost free

— becoming abundant

Example
Census age distribution

Example
Web histories

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present: Designed and Organic Data

Designed data

— high cost

— becoming scarce

— speak to population

Example
Census age distribution

Organic data

— almost free

— becoming abundant

— iffy for population

Example
Web histories

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present: Designed and Organic Data

Designed data
— high cost
— becoming scarce
— speak to population

Organic data
— almost free
— becoming abundant
— iffy for population

Example
Census age distribution

Example
Web histories

future of **organic data**

future of **designed data**

The Future of Sample Surveys

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1990–Present: Designed and Organic Data

Designed data

— high cost

— becoming scarce

— speak to population

Organic data

— almost free

— becoming abundant

— iffy for population

Example
Census age distribution

Example
Web histories

the future is **together**

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