Studying Social Inequality with Data Science

Soc 114
Winter 2025

Sampling: Stratified, Clustered, and the Future

Learning goals for today

By the end of class, you will be able to

- explain a stratified sample
- explain a clustered sample
- connect sampling to the replication crisis
- discuss the future of 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

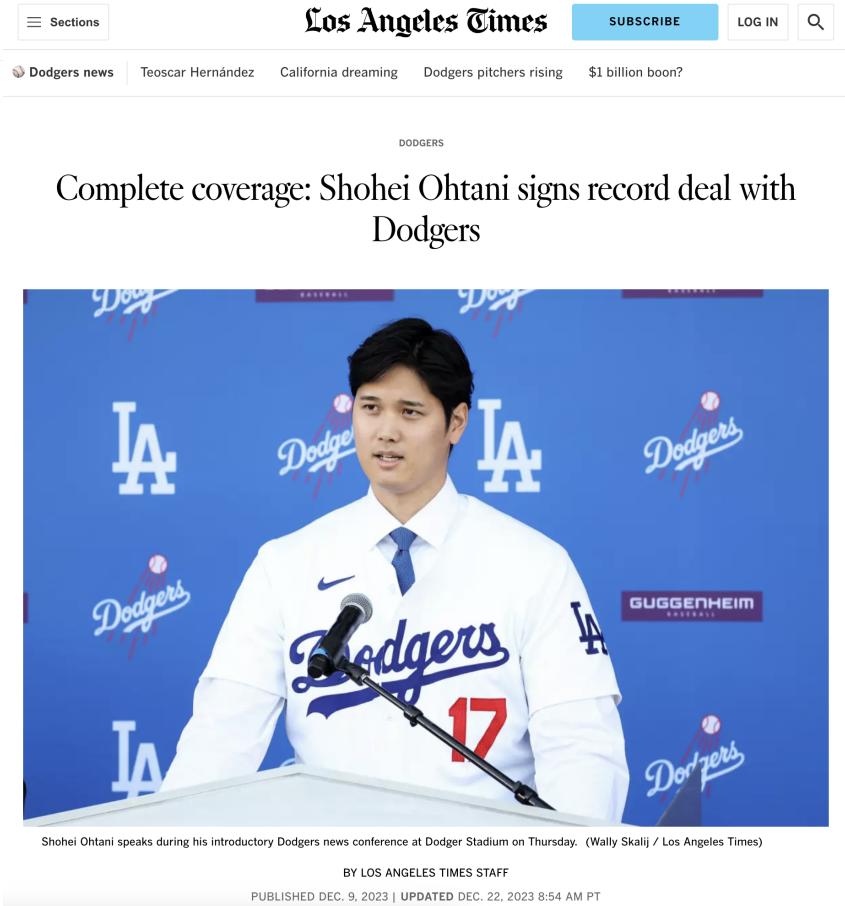
Complete coverage: Shohei Ohtani signs record deal with Dodgers

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

BY LOS ANGELES TIMES STAFF

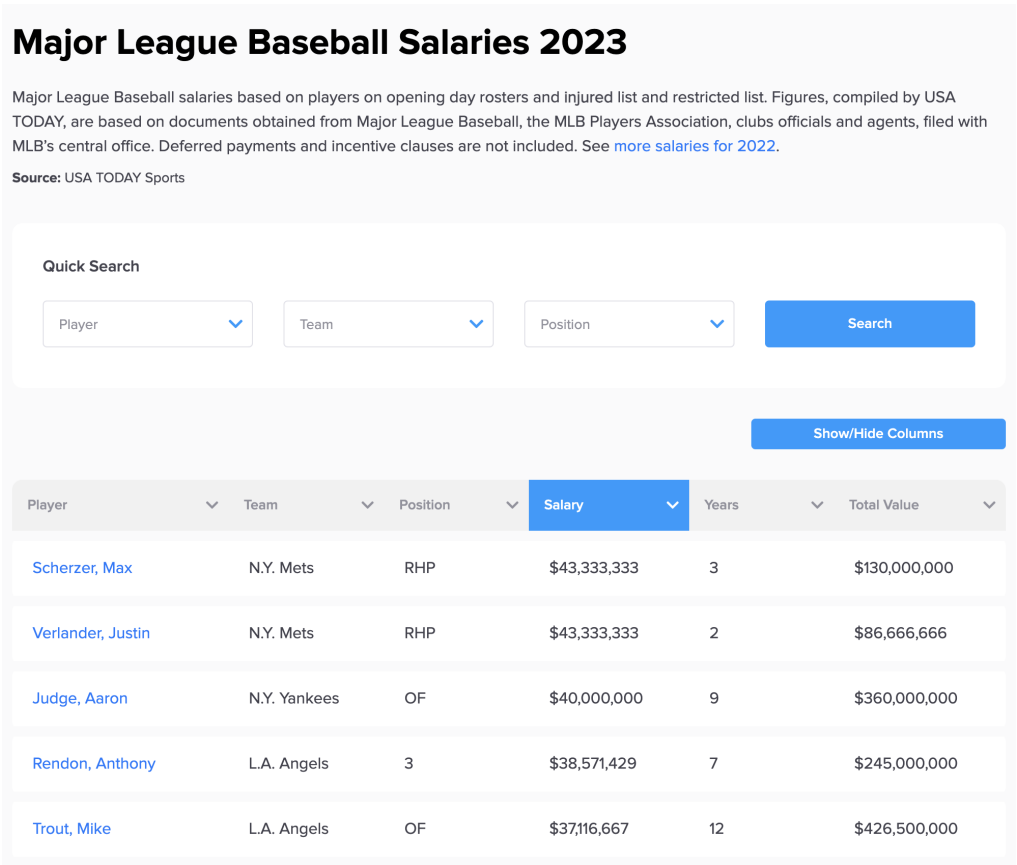
PUBLISHED DEC. 9, 2023 | **UPDATED** DEC. 22, 2023 8:54 AM PT

Baseball salaries



Major League Baseball Minimum: **$720,000**

Baseball salaries



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

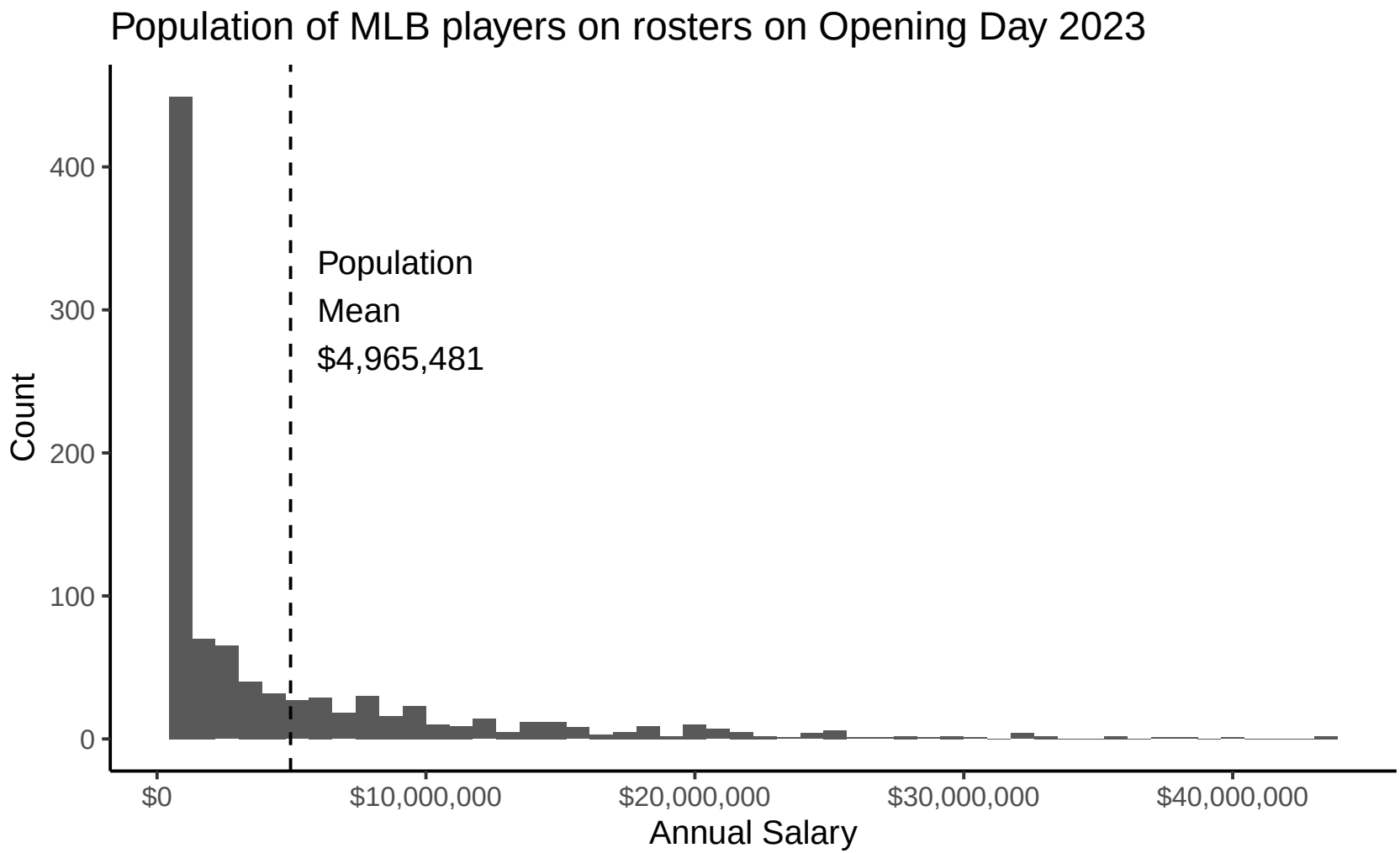
Player | Team | Position | Search

Show/Hide Columns

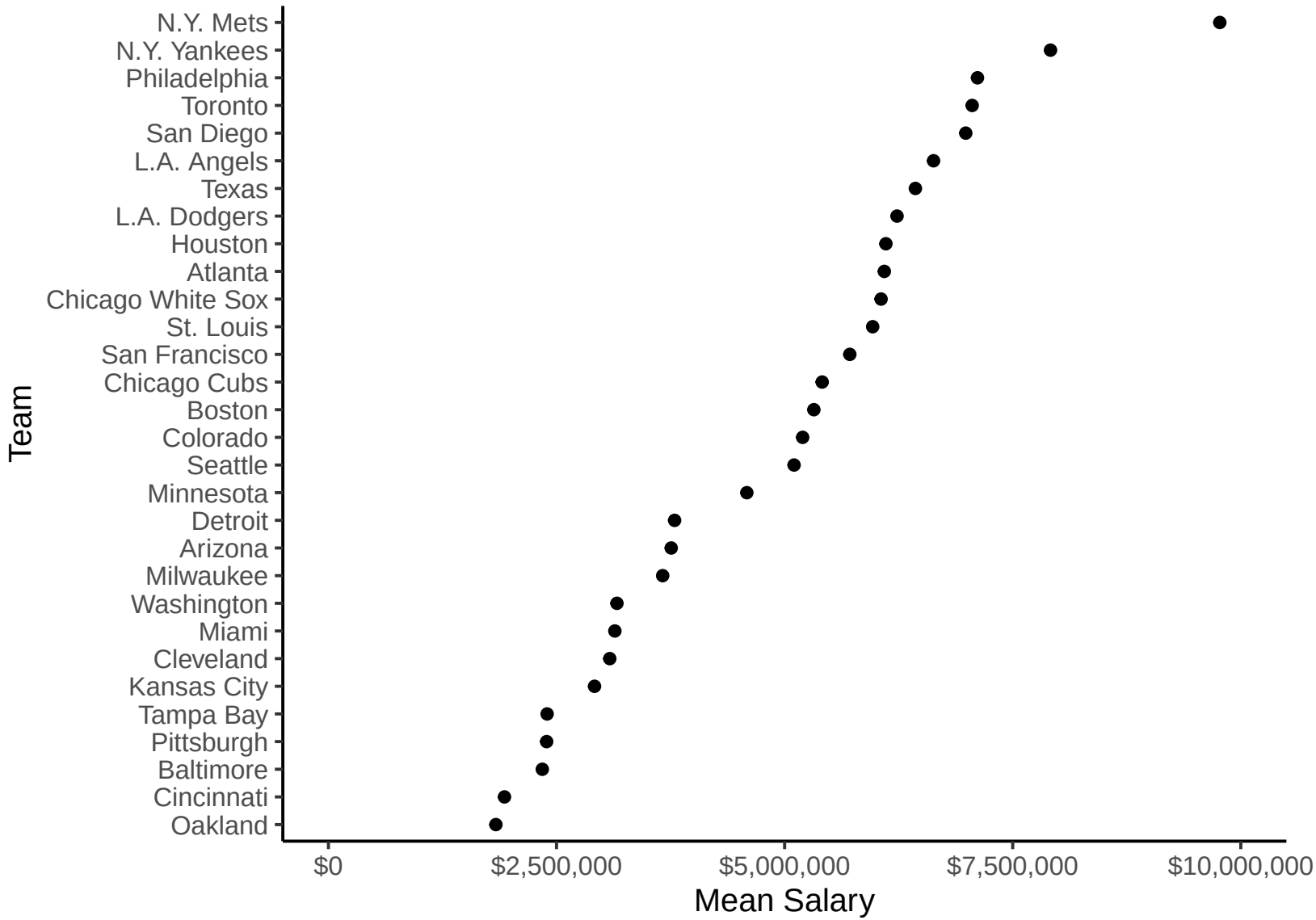
Player	Team	Position	Salary	Years	Total Value
Scherzer, Max	N.Y. Mets	RHP	$43,333,333	3	$130,000,000
Verlander, Justin	N.Y. Mets	RHP	$43,333,333	2	$86,666,666
Judge, Aaron	N.Y. Yankees	OF	$40,000,000	9	$360,000,000
Rendon, Anthony	L.A. Angels	3	$38,571,429	7	$245,000,000
Trout, Mike	L.A. Angels	OF	$37,116,667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

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

Baseball salaries



Baseball salaries



Draw a Sample to Estimate the Mean Salary

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

How to sample baseball players

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

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 types of sampling

Three types of sampling

- Simple random sample: 60 players at random

Three types of sampling

- Simple random sample: 60 players at random
- Stratified sampling by team: 2 players per team

Three types of sampling

- Simple random sample: 60 players at random
- Stratified sampling by team: 2 players per team

- Random sample clustered by team: 20 players on each of 3 sampled teams

For reference: reading

Three types of sampling

- Simple random sample: 60 players at random
- Stratified sampling by team: 2 players per team
  - stratification makes our sample better
  - rules out unlucky bad draws that miss whole teams
- Random sample clustered by team: 20 players on each of 3 sampled teams

For reference: reading

Three types of sampling

- Simple random sample: 60 players at random
- Stratified sampling by team: 2 players per team
  - stratification makes our sample better
  - rules out unlucky bad draws that miss whole teams
- Random sample clustered by team: 20 players on each of 3 sampled teams
  - clustering makes our sample cheaper
  - sample is not as high quality—the 3 teams may be unusual

For reference: reading

Apply an Estimator

Write a function that I like to call `estimator()`

- input is a sample
- output is an estimate

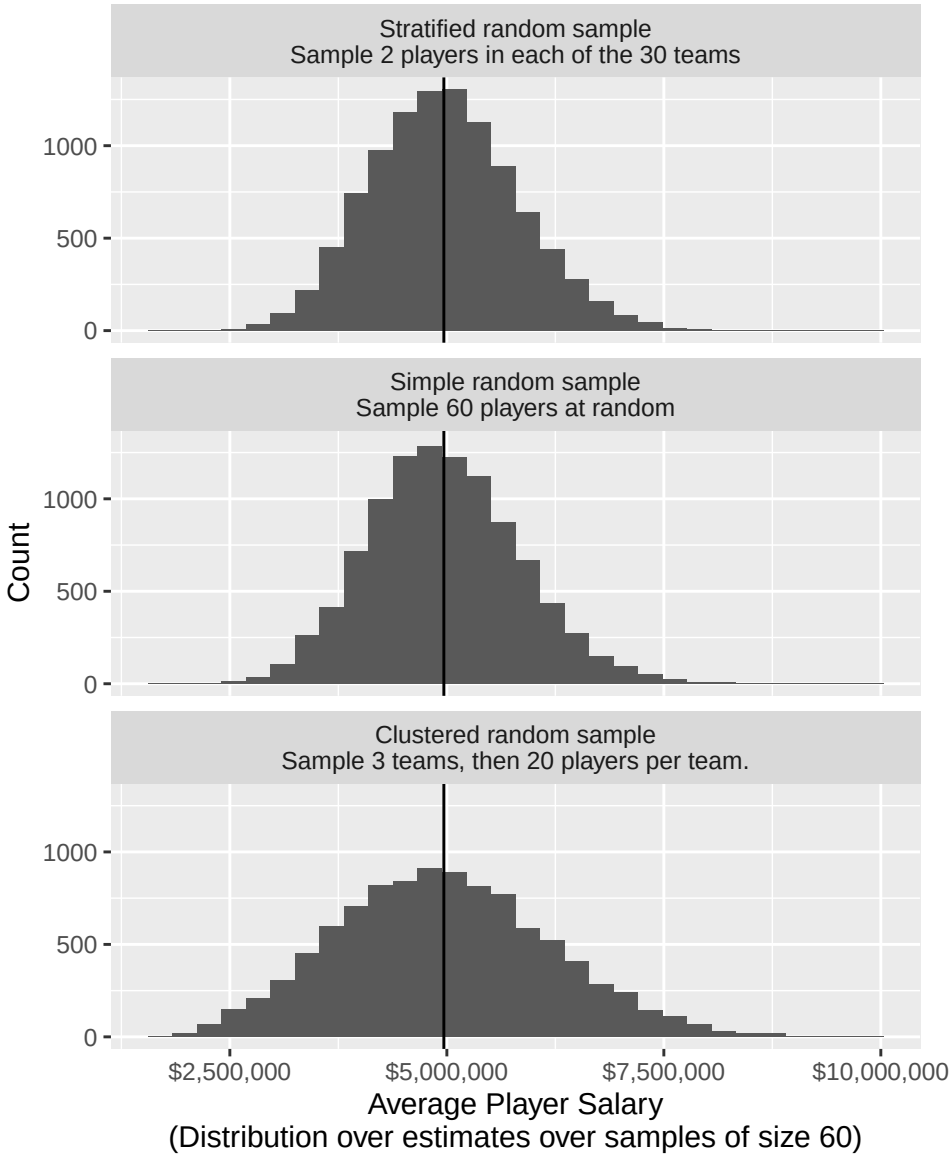
Evaluate performance

We will first calculate the population mean

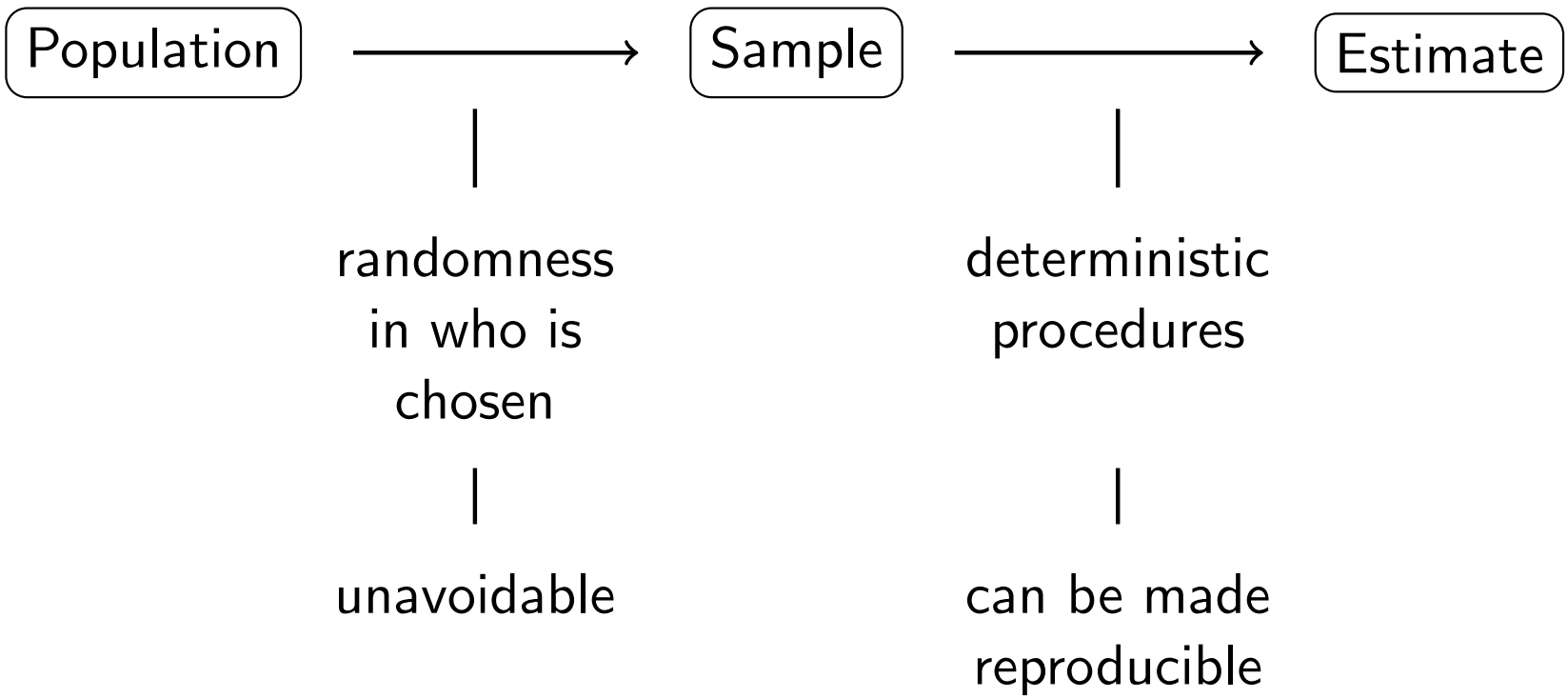
Then we will repeatedly

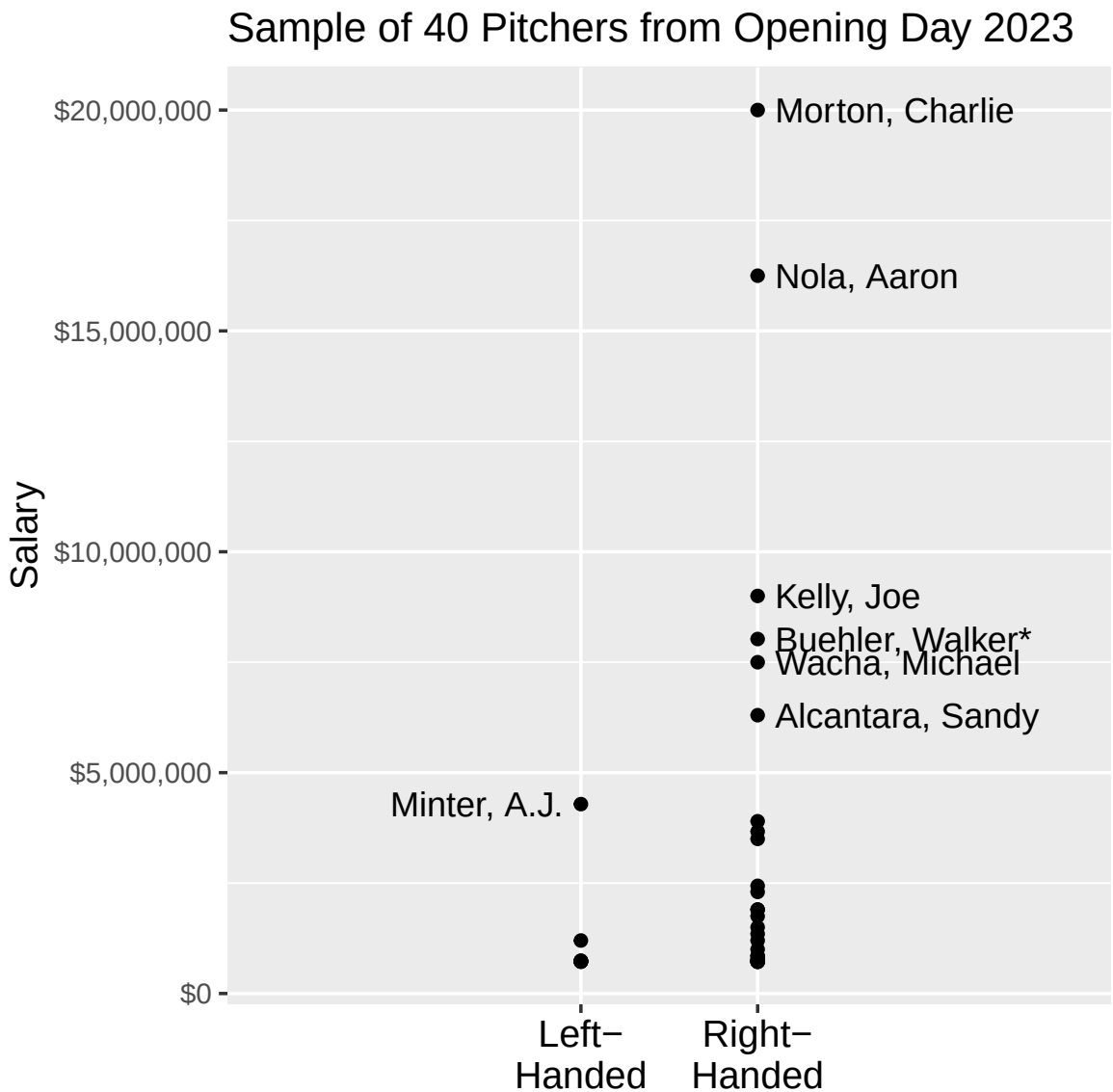
- draw a sample
- apply the estimator
- store the result

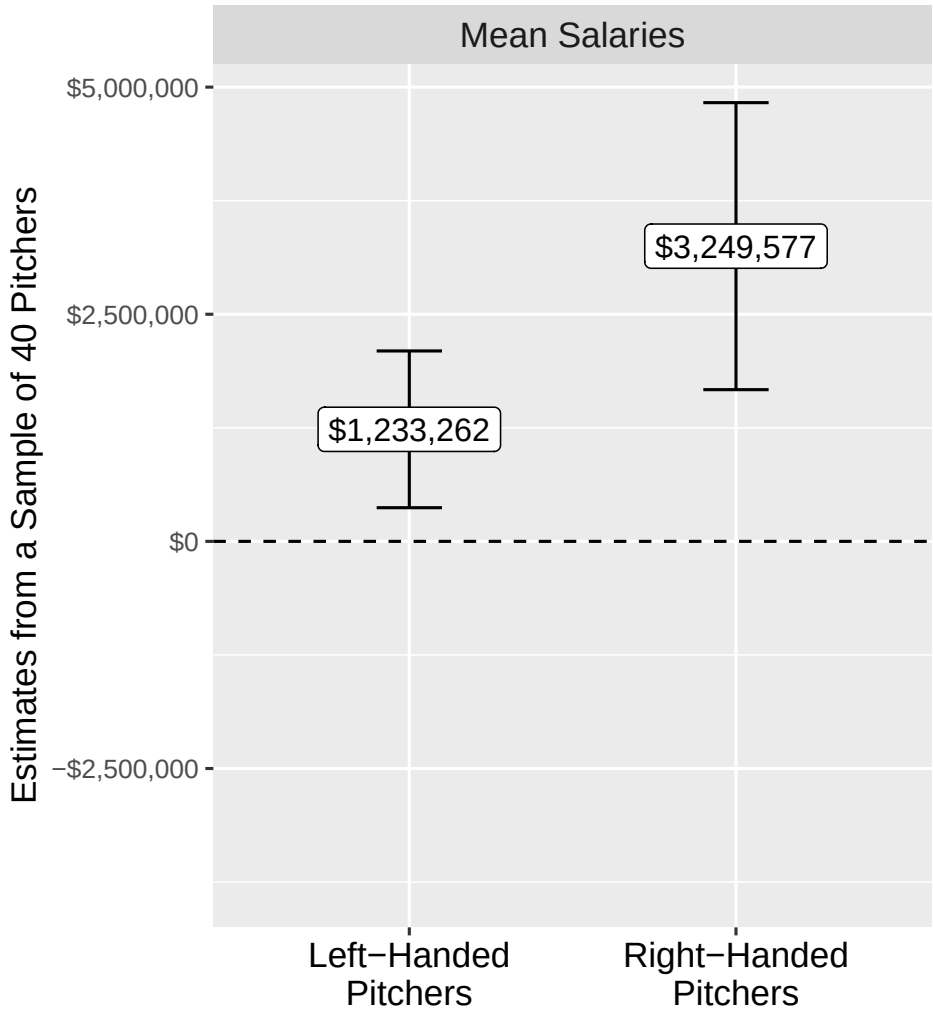
Three sampling strategies

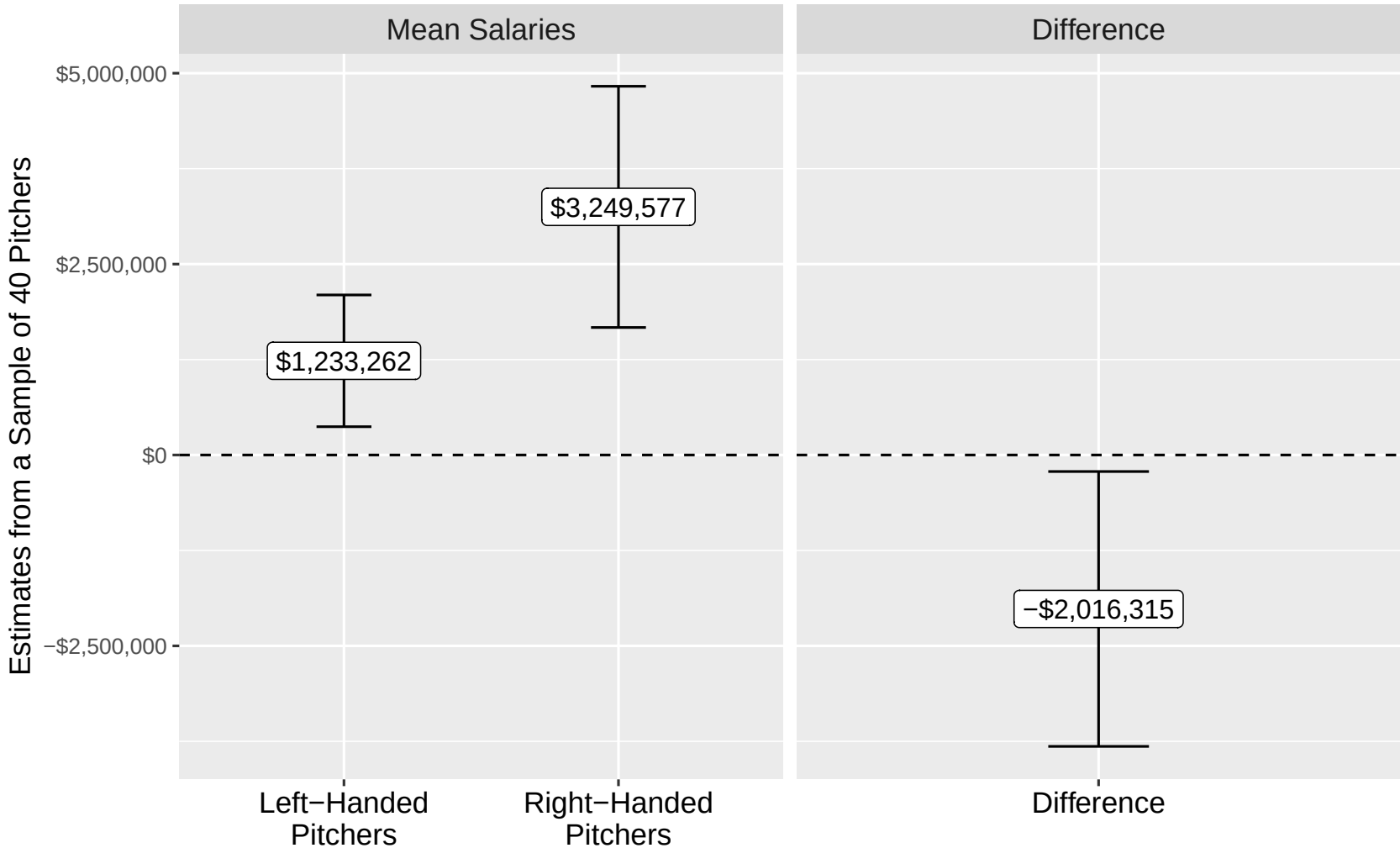


Danger of One Sample

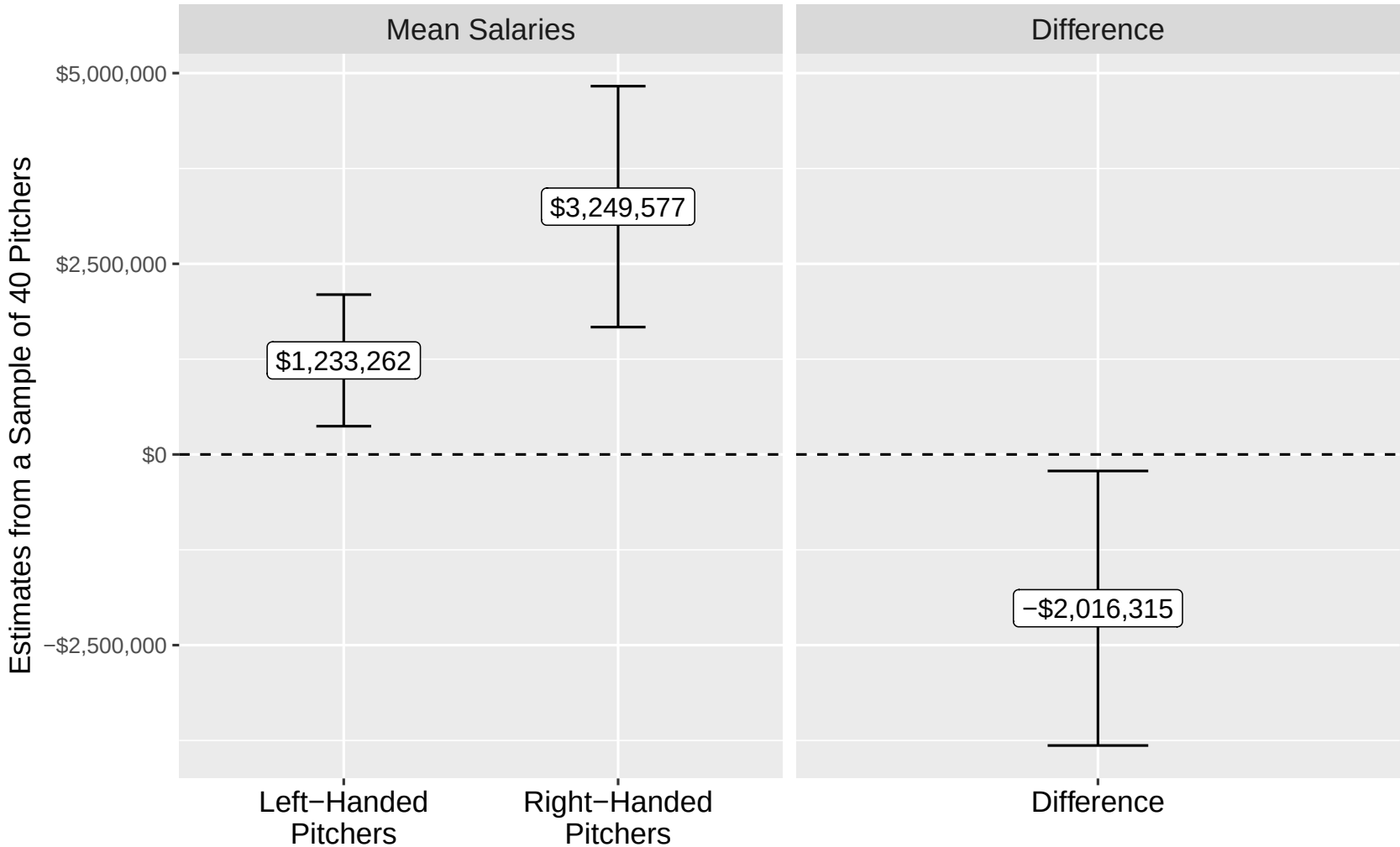



Sample of 40 Pitchers from Opening Day 2023
Salary
$20,000,000
$15,000,000
$10,000,000
$5,000,000
$0
Morton, Charlie
Nola, Aaron
Kelly, Joe
Buehler, Walker*
Wacha, Michael
Alcantara, Sandy
Minter, A.J.
Left−
Handed
Right−
Handed


Mean Salaries
Estimates from a Sample of 40 Pitchers
$5,000,000
$2,500,000
$0
−$2,500,000
$3,249,577
$1,233,262
Left−Handed Pitchers
Right−Handed Pitchers


Mean Salaries
Difference
Estimates from a Sample of 40 Pitchers
$5,000,000
$2,500,000
$0
−$2,500,000
$3,249,577
$1,233,262
−$2,016,315
Left−Handed Pitchers
Right−Handed Pitchers
Difference

Why might right-handed pitchers earn more?



I did this 1,000 times

Distribution of Sample Estimates

6% of sample estimates are statistically significant

The replication crisis

The replication crisis

- unless we see the population, all estimates involve noise

The replication crisis

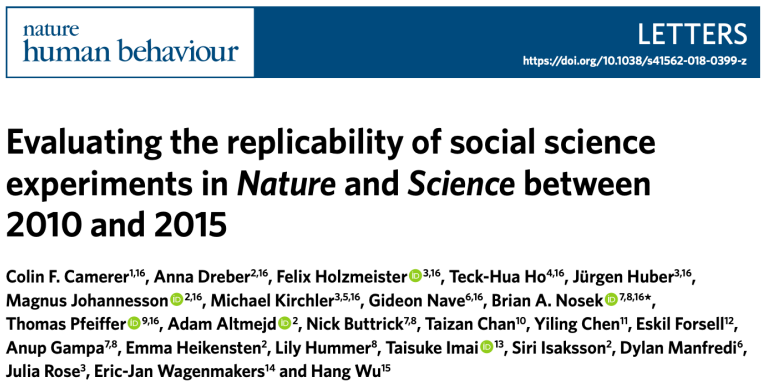
- unless we see the population, all estimates involve noise
- surprising findings yield big rewards

The replication crisis

- unless we see the population, all estimates involve noise
- surprising findings yield big rewards
- unsurprising findings get ignored

The replication crisis

- unless we see the population, all estimates involve noise
- surprising findings yield big rewards
- unsurprising findings get ignored
- science is just discovering noise



nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-018-0399-z

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer[1,16], Anna Dreber[2,16], Felix Holzmeister[3,16], Teck-Hua Ho[4,16], Jürgen Huber[3,16], Magnus Johannesson[2,16], Michael Kirchler[3,5,16], Gideon Nave[6,16], Brian A. Nosek[7,8,16*], Thomas Pfeiffer[9,16], Adam Altmejd[2], Nick Buttrick[7,8], Taizan Chan[10], Yiling Chen[11], Eskil Forsell[12], Anup Gampa[7,8], Emma Heikensten[2], Lily Hummer[8], Taisuke Imai[13], Siri Isaksson[2], Dylan Manfredi[6], Julia Rose[3], Eric-Jan Wagenmakers[14] and Hang Wu[15]

Camerer et al. in Nature Human Behavior.

The New York Times

SCIENCE TIMES AT 40

Essay: The Experiments Are Fascinating. But Nobody Can Repeat Them.

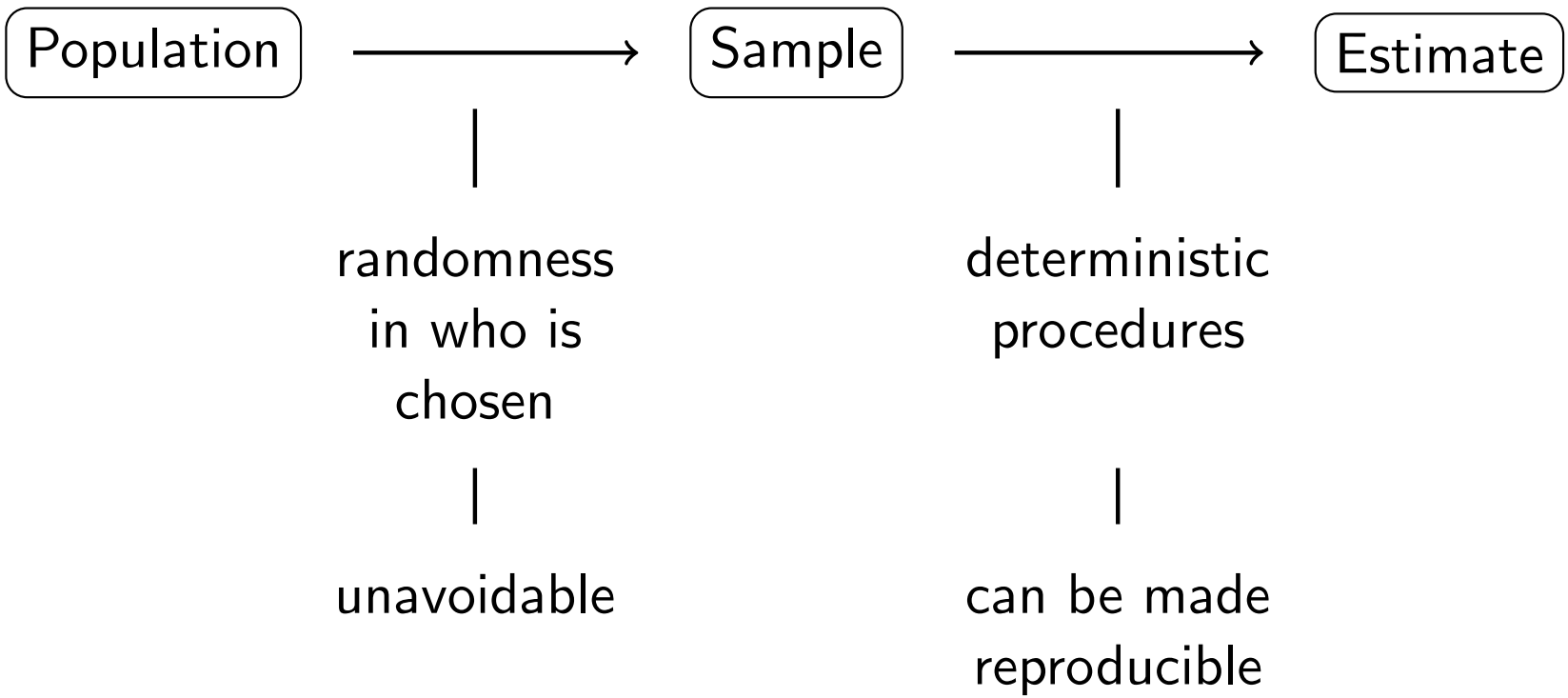
Science is mired in a "replication" crisis. Fixing it will not be easy.

Gelman in NYTimes.

The replication crisis

- unless we see the population, all estimates involve noise
- surprising findings yield big rewards
- unsurprising findings get ignored
- science is just discovering noise

Danger of One Sample



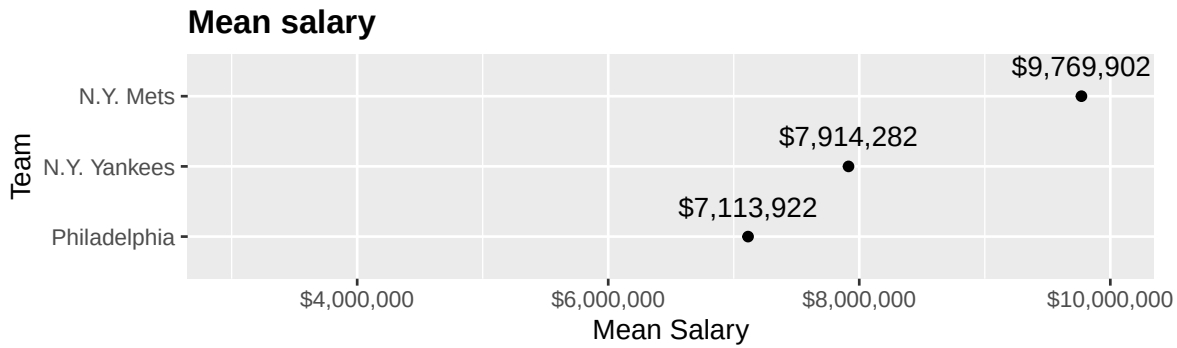
Reproducibility

What is a typical salary in the three highest-paying teams in American baseball?

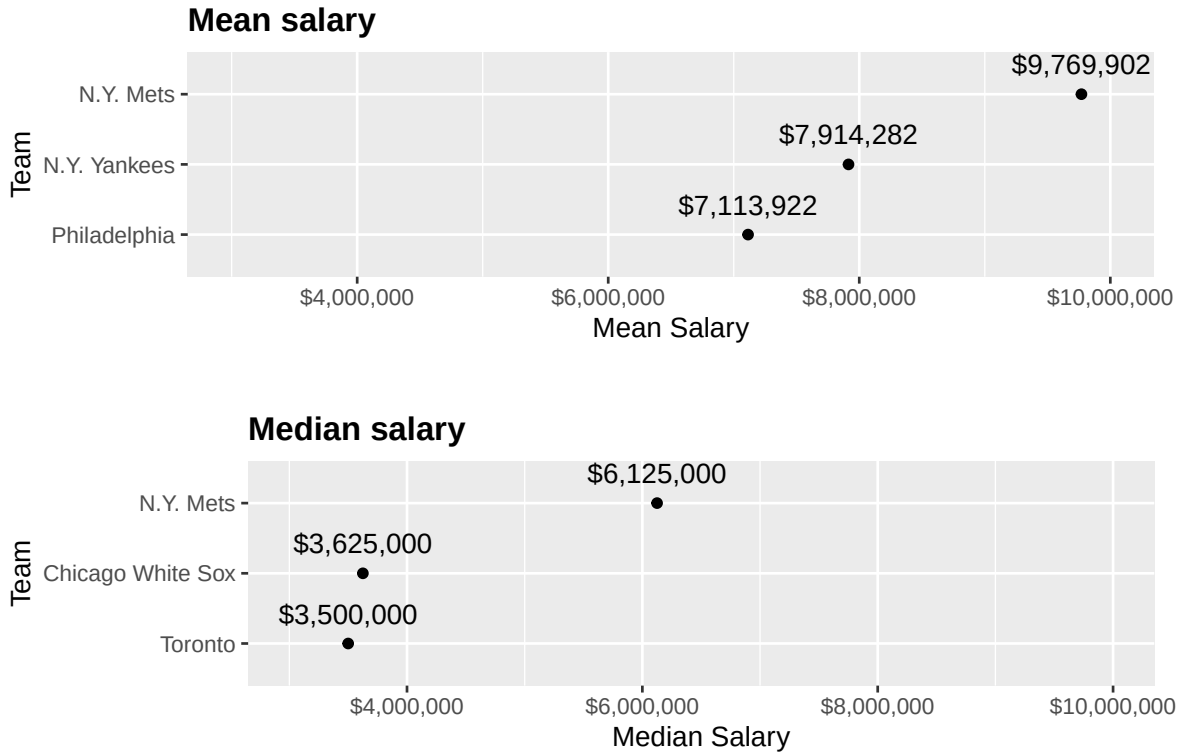
- how would you answer this question with data?

What is a typical salary in the three highest-paying teams in American baseball?

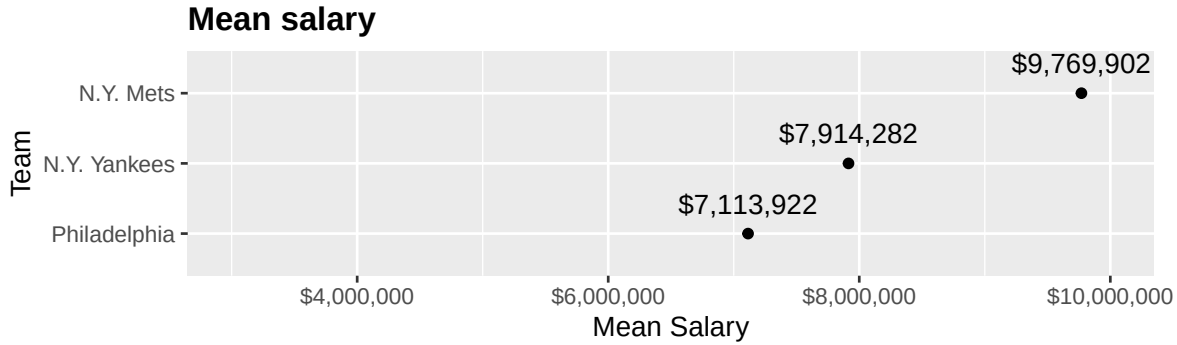
What is a typical salary in the three highest-paying teams in American baseball?

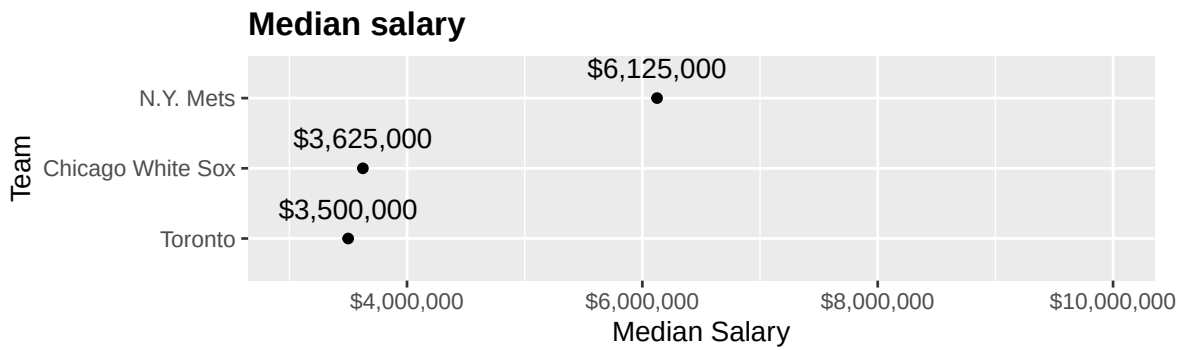


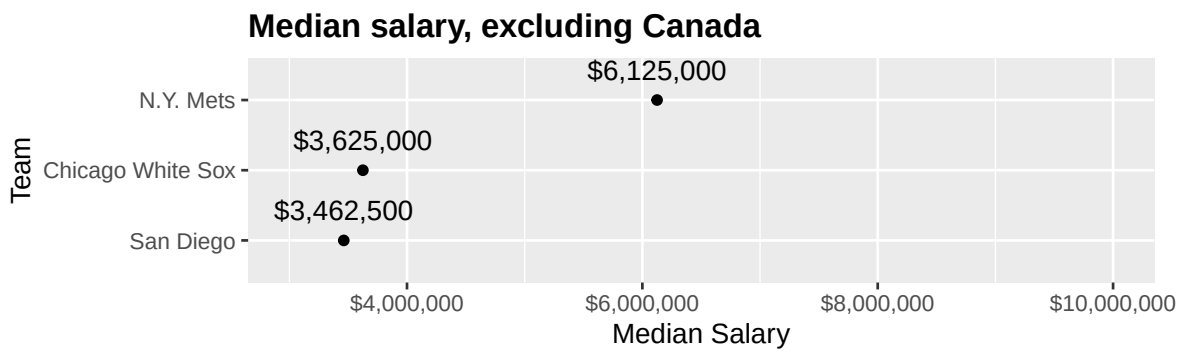
What is a typical salary in the three highest-paying teams in American baseball?



What is a typical salary in the three highest-paying teams in American baseball?







quarto

```
---
title: "Problem Set 1: Visualization"
format: pdf
---

**Due: 5pm on Wednesday, January 31.**

Student identifer: [type your anonymous identifying number here]

- Use this template to complete the problem set
- In Canvas, you will upload the PDF produced by your .qmd file
- Put your identifier above, not your name! We want anonymous grading to be
possible

This problem set involves both data analysis and reading.

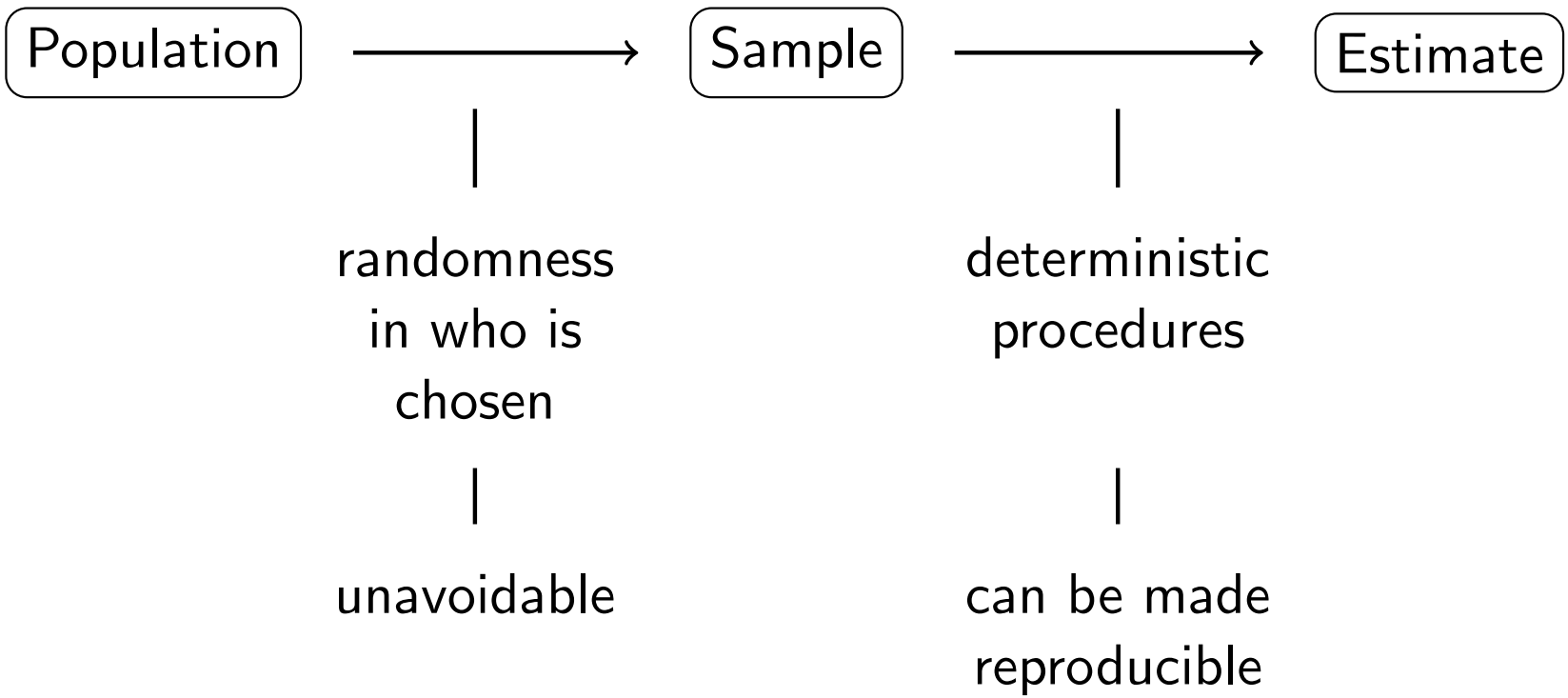
### Data analysis

This problem set uses the data
[`lifeCourse.csv`](https://info3370.github.io/data/lifeCourse.csv).

```{r, comment = F, message = F}
library(tidyverse)
library(scales)
lifeCourse <- read_csv("https://info3370.github.io/data/lifeCourse.csv")
```

The data contain life course earnings profiles for four cohorts of American
workers: those born in 1940, 1950, 1960, and 1970. Each row contains a
```

Danger of One Sample



The Future of Sample Surveys

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

The Future of Sample Surveys

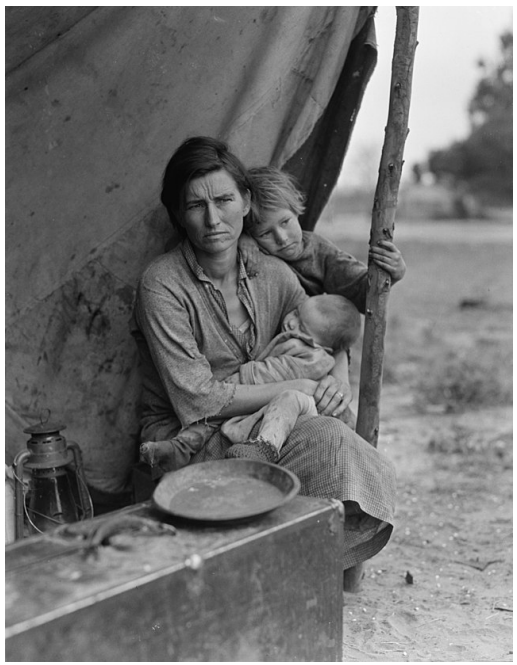
Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly.

1930–1960: Era of Invention

The Future of Sample Surveys

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

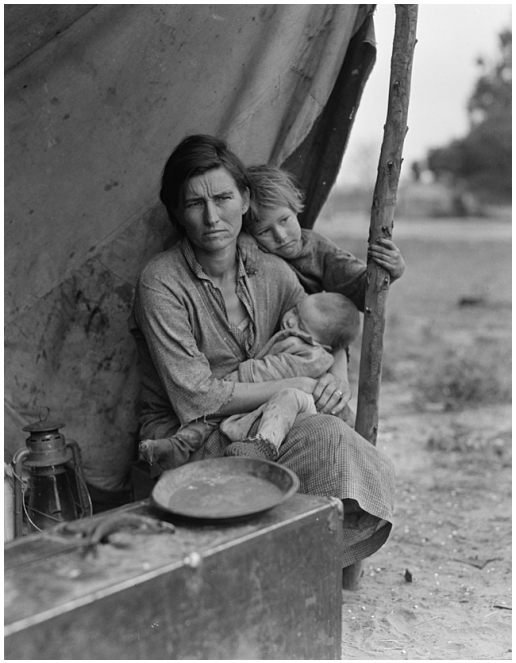
1930–1960: Era of Invention



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 a stratified sample
- explain a clustered sample
- connect sampling to the replication crisis
- discuss the future of sampling