Social Data Science

Course Recap

Goals of the course

- connect theories about inequality to quantitative empirical evidence
- evaluate the effects of hypothetical interventions to reduce inequality
- conduct data analysis using the R programming language

Structure of the Course

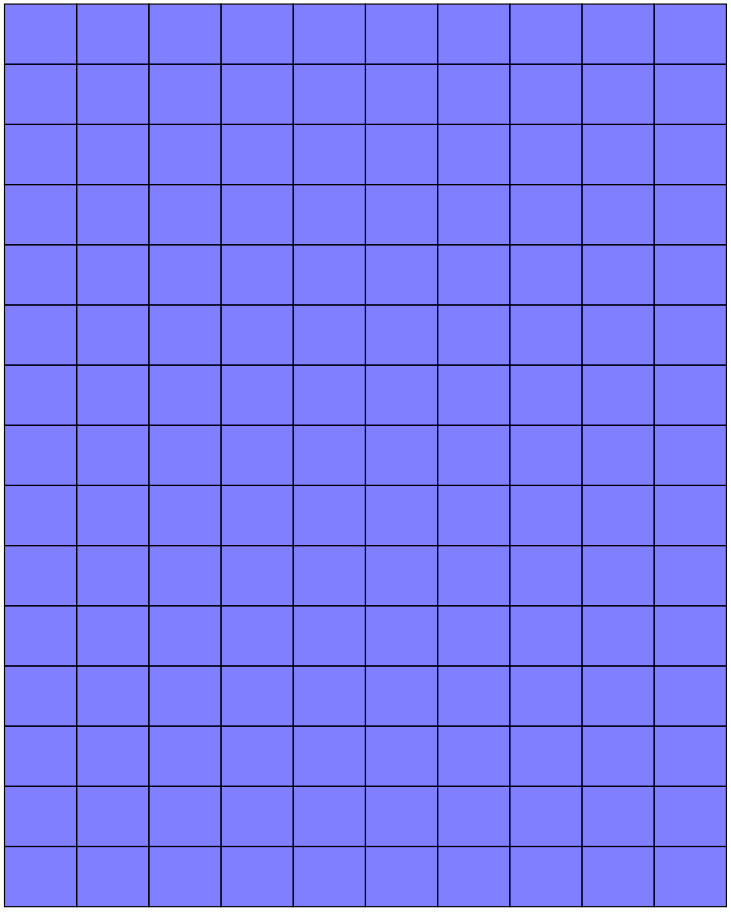
We learned key concepts in social data science.

1. Population Sampling
2. Models for Subgroup Summaries
3. Causal Inference

Population Sampling

Full Count Enumeration

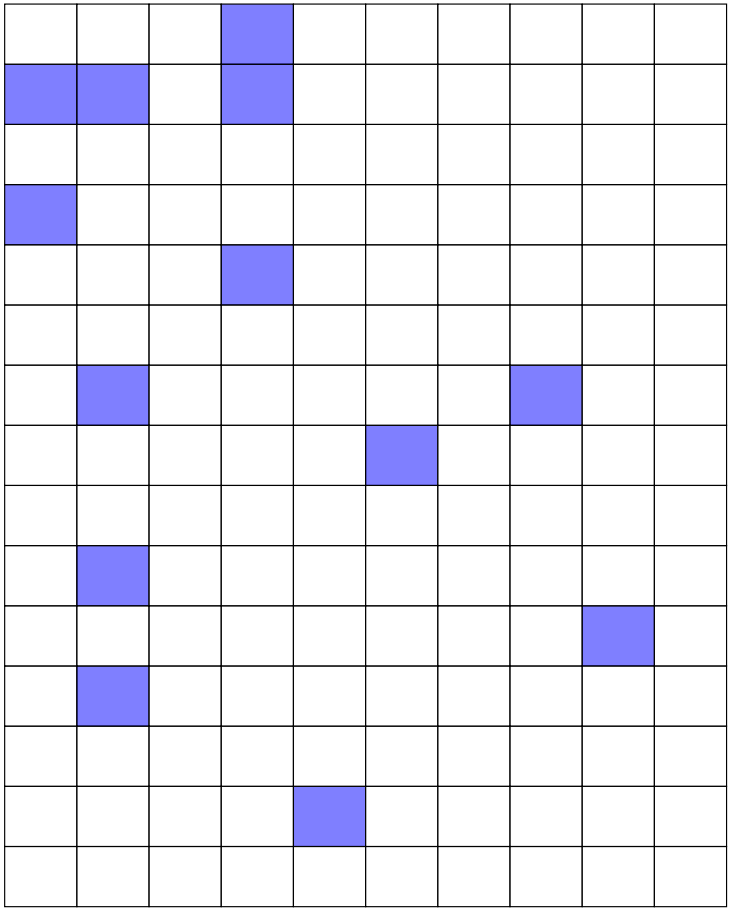
Back of Room



Front of Room

Probability Sample

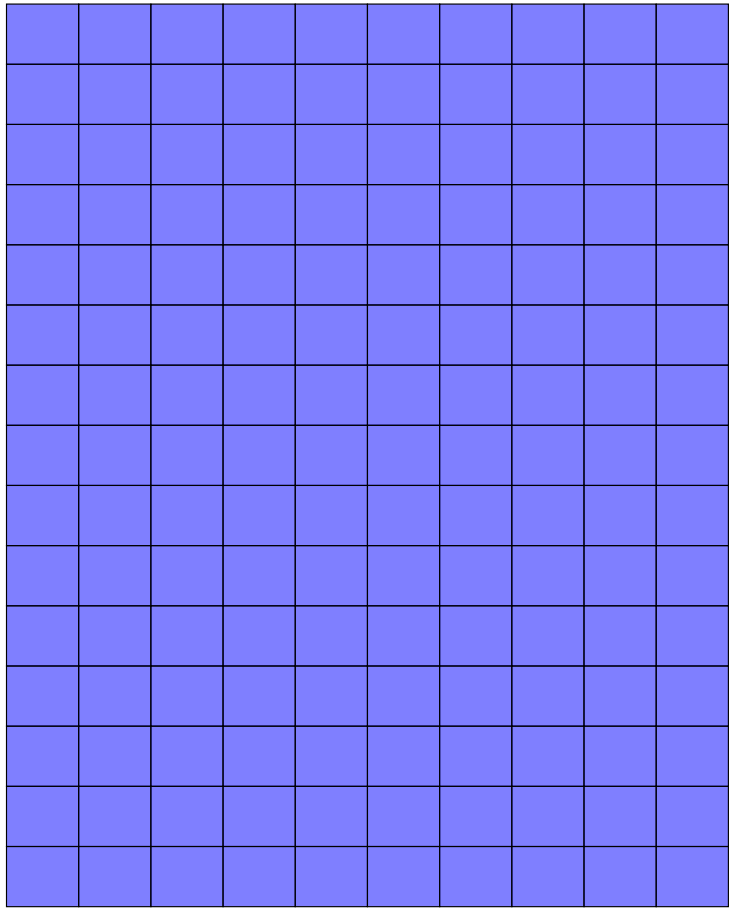
Back of Room



Front of Room

Full Count Enumeration

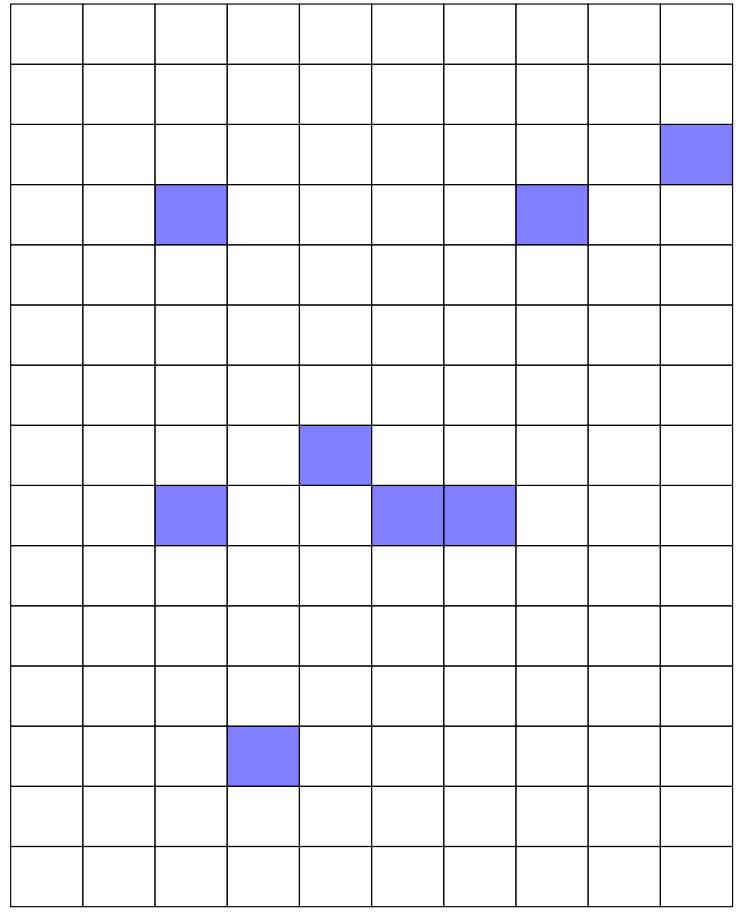
Back of Room



Front of Room

Probability Sample

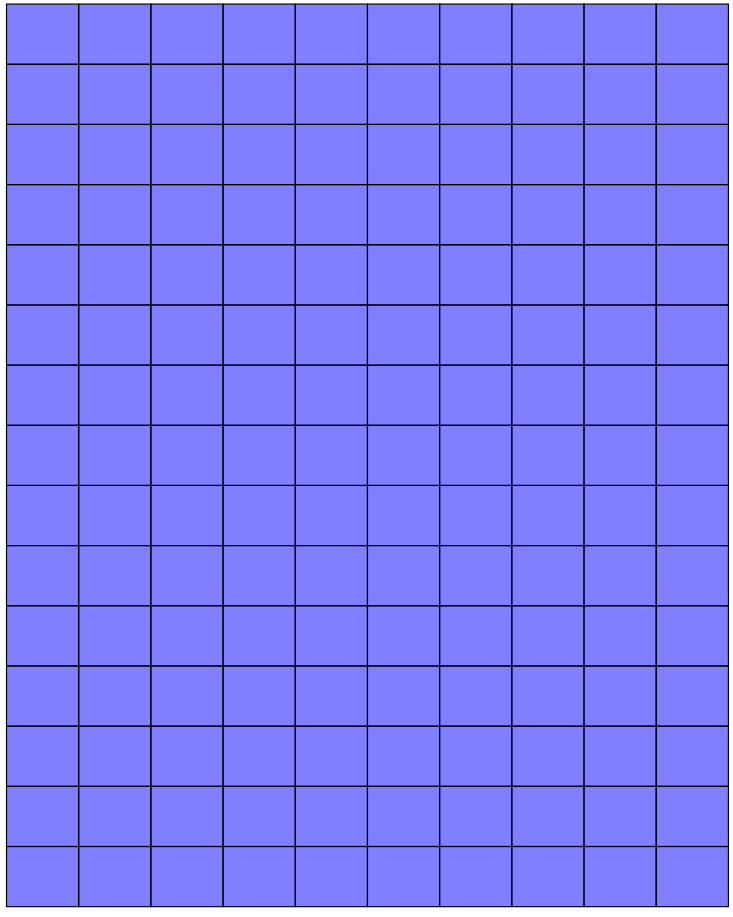
Back of Room



Front of Room

Full Count Enumeration

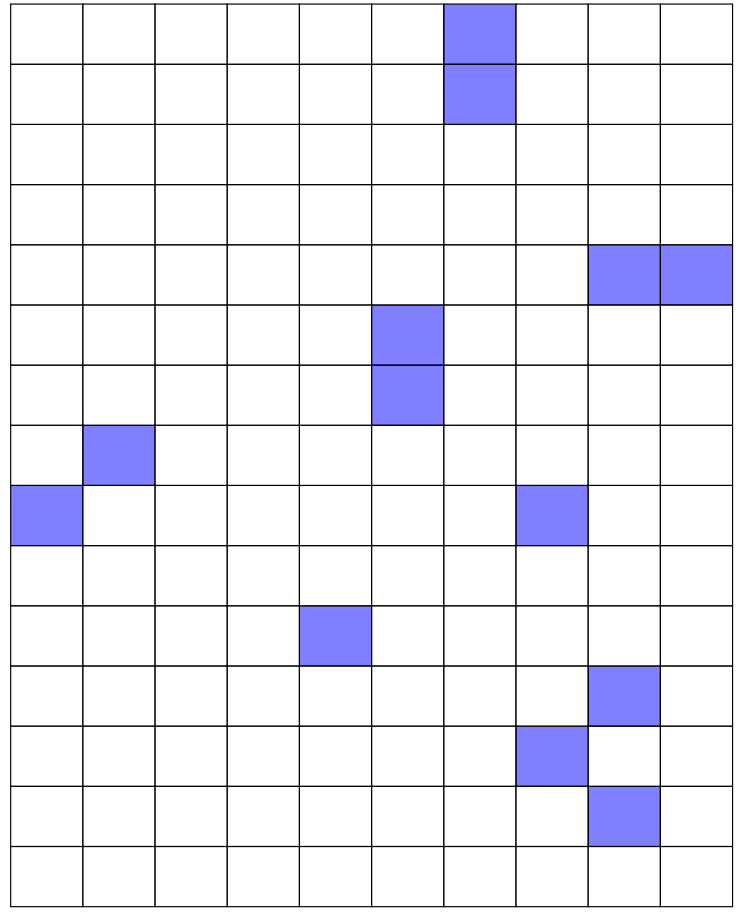
Back of Room



Front of Room

Probability Sample

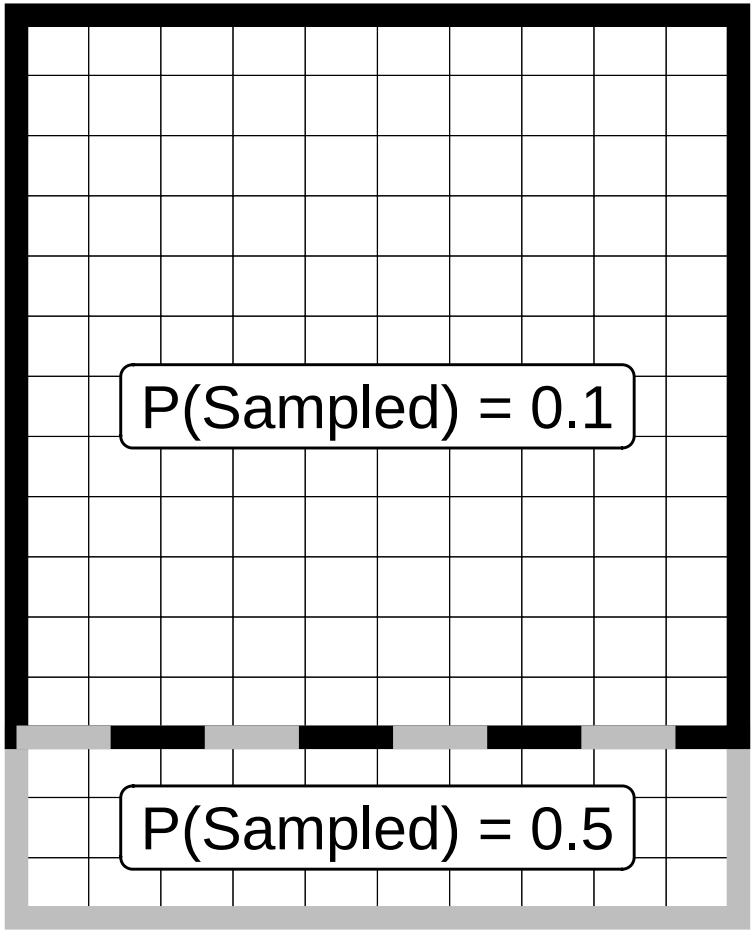
Back of Room



Front of Room

Sample Design

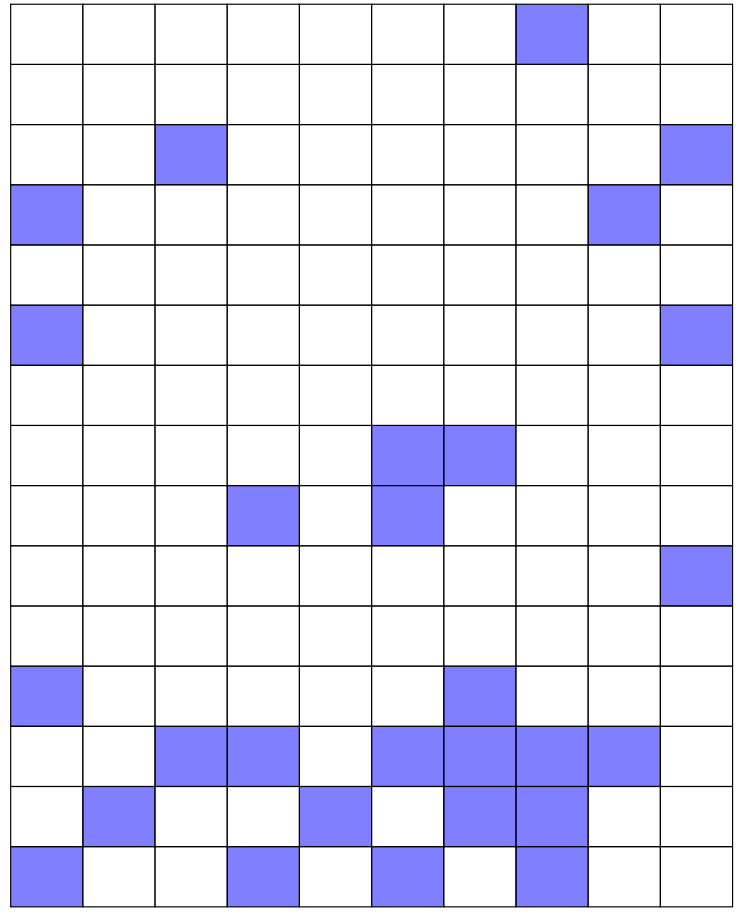
Back of Room



Front of Room

Sample

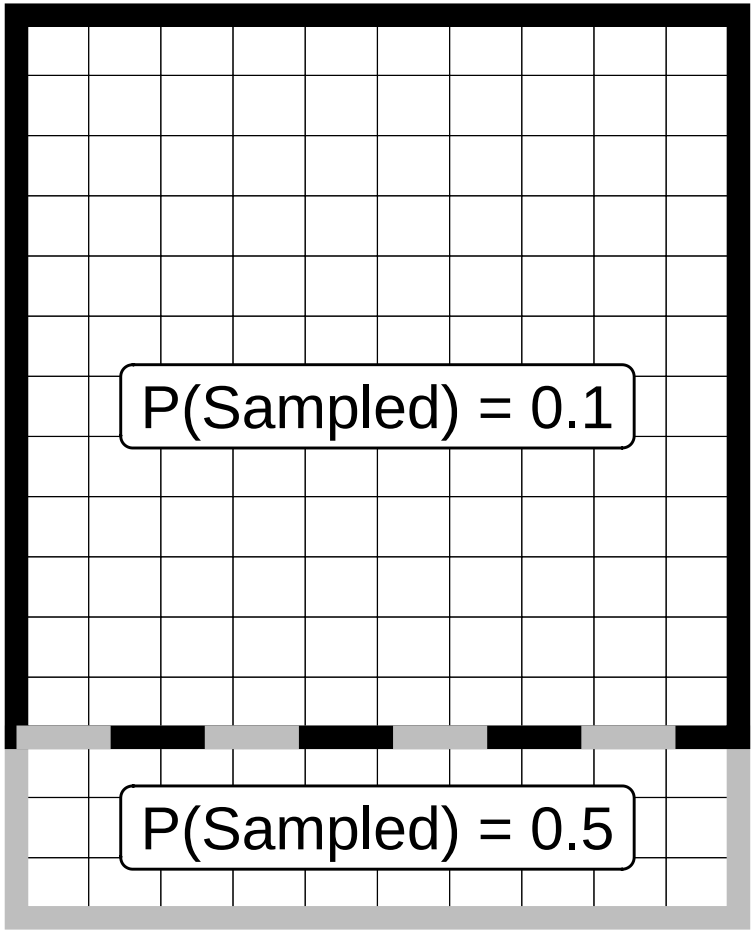
Back of Room



Front of Room

Sample Design

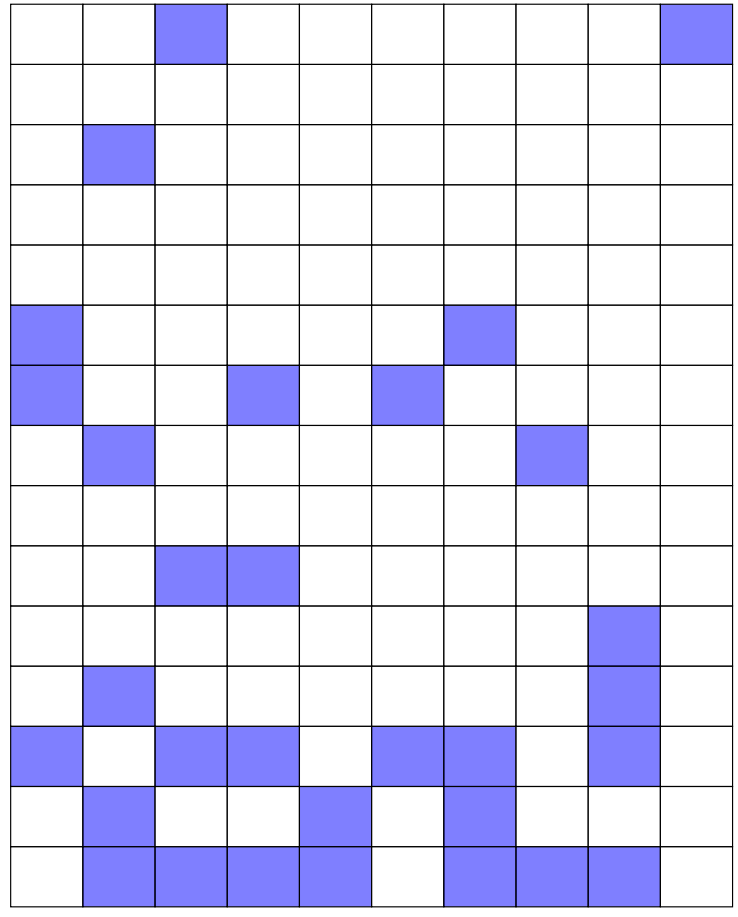
Back of Room



Front of Room

Sample

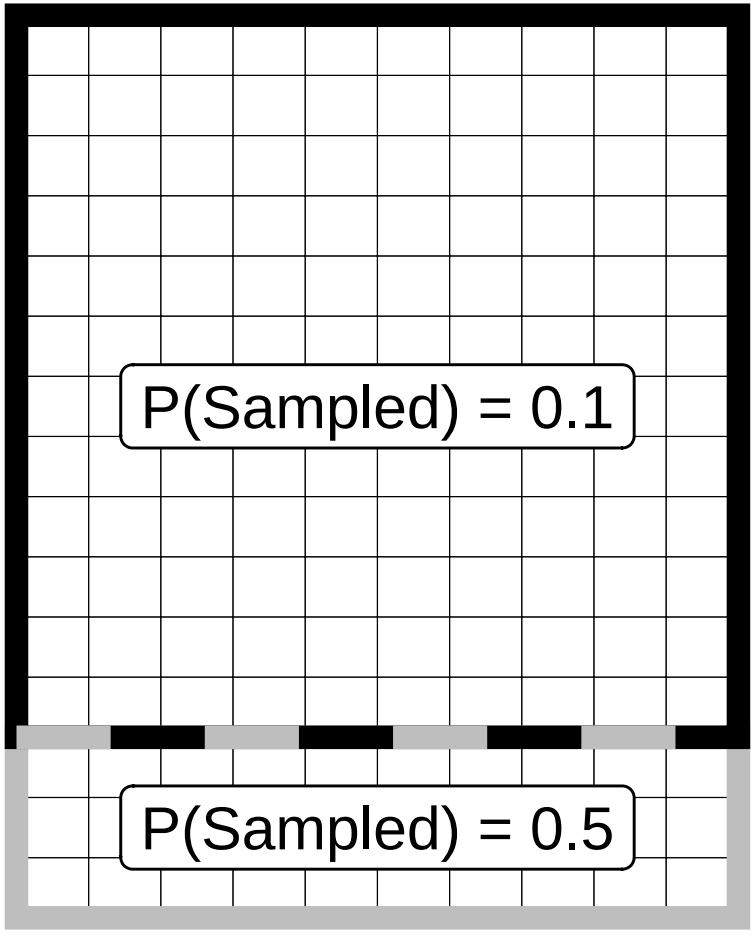
Back of Room



Front of Room

Sample Design

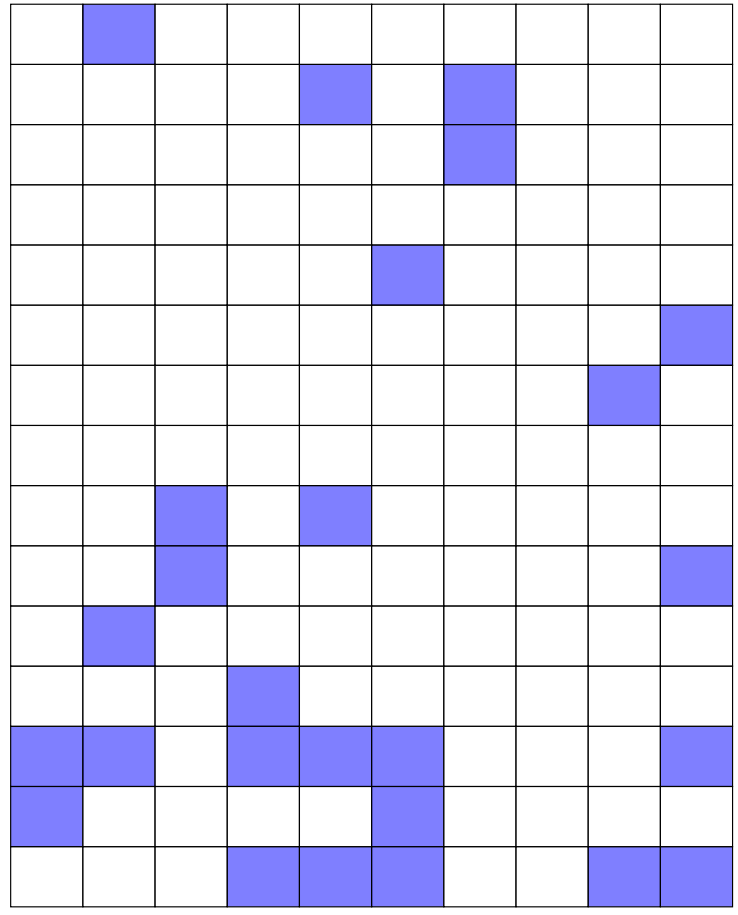
Back of Room



Front of Room

Sample

Back of Room



Front of Room

Working with data

- Creating objects in R
- Piping data into functions
- Summarizing a sample
- Visualizing with ggplot

Working with data

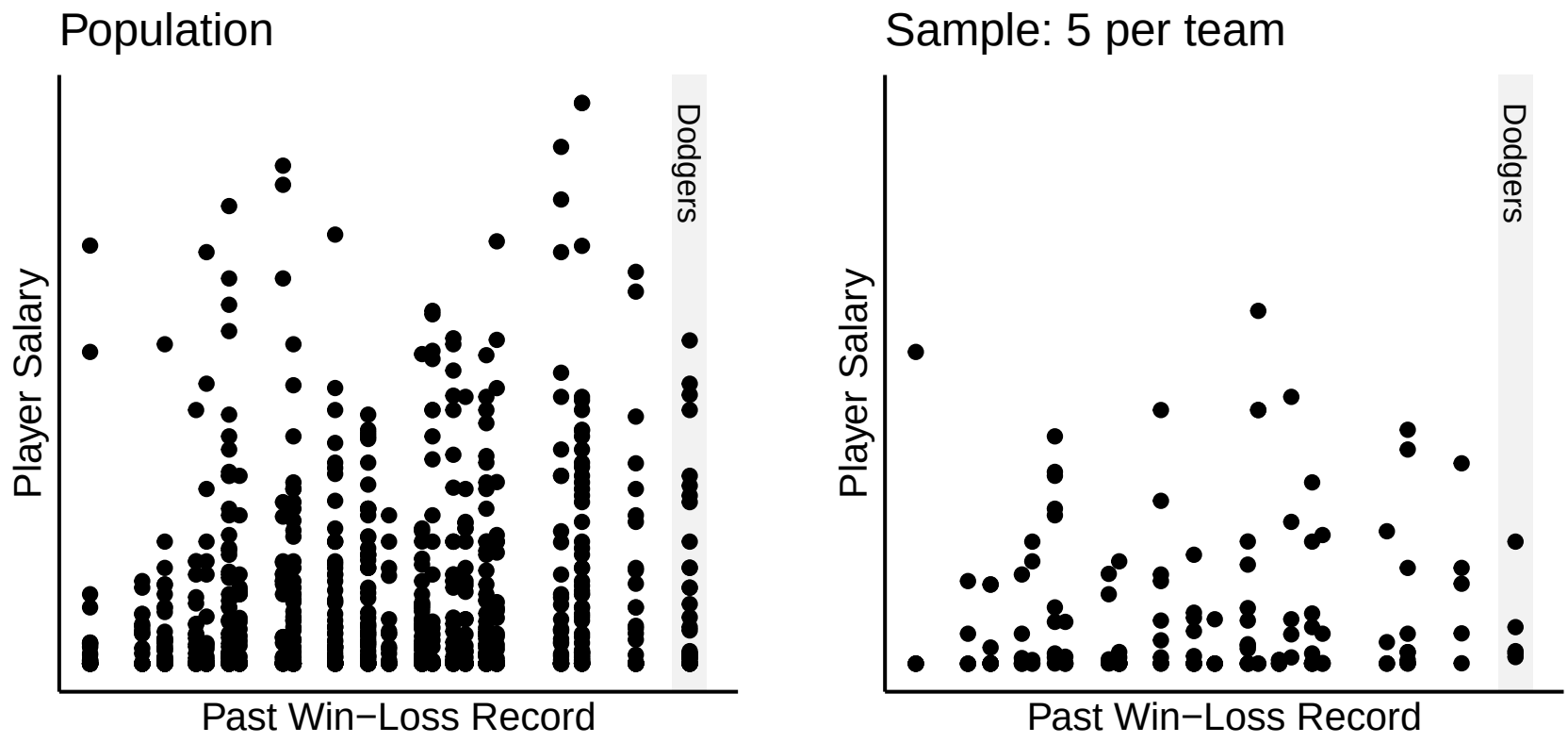
Questions of the form:

Among _____, what is the mean of _____?

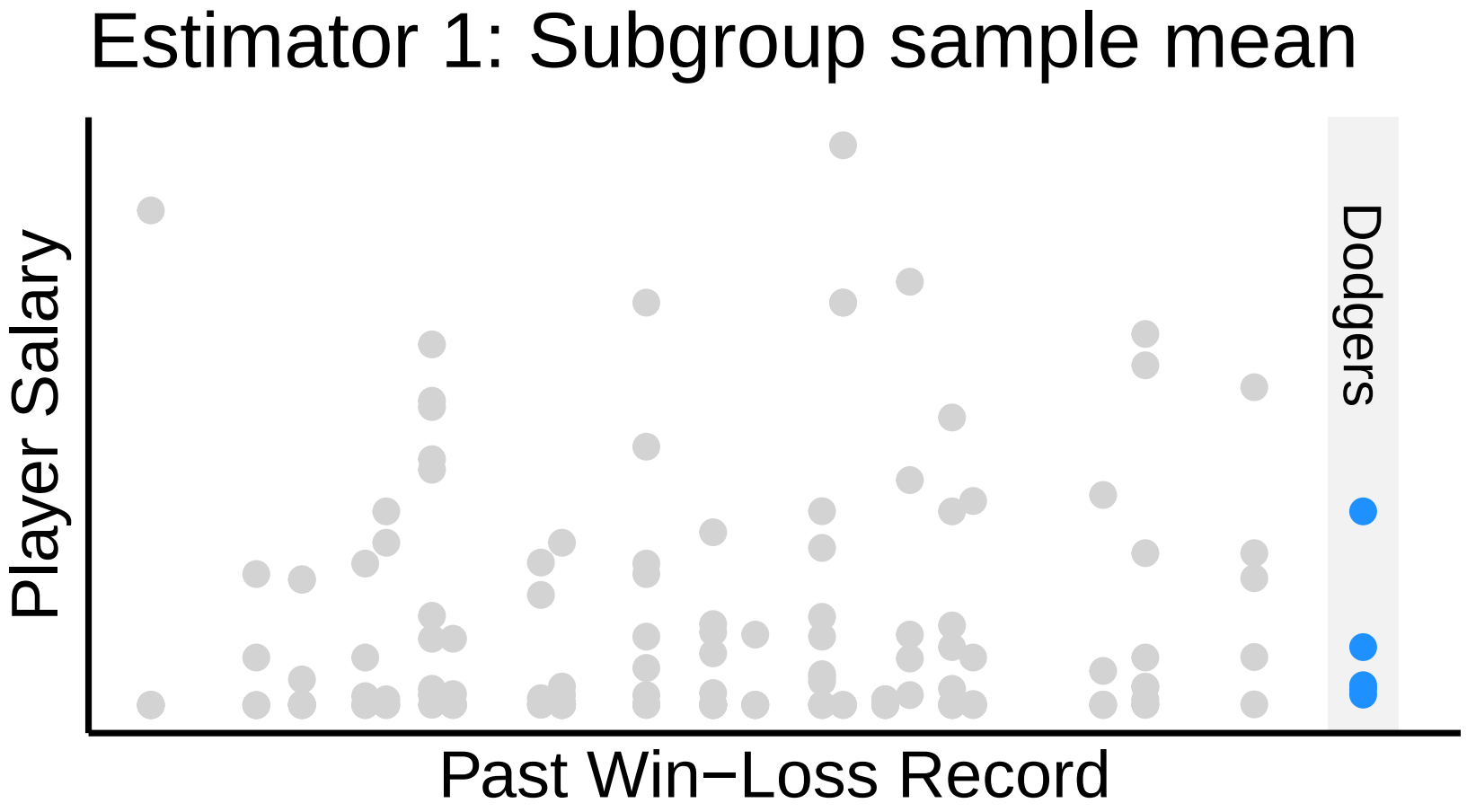
(population inference from a sample)

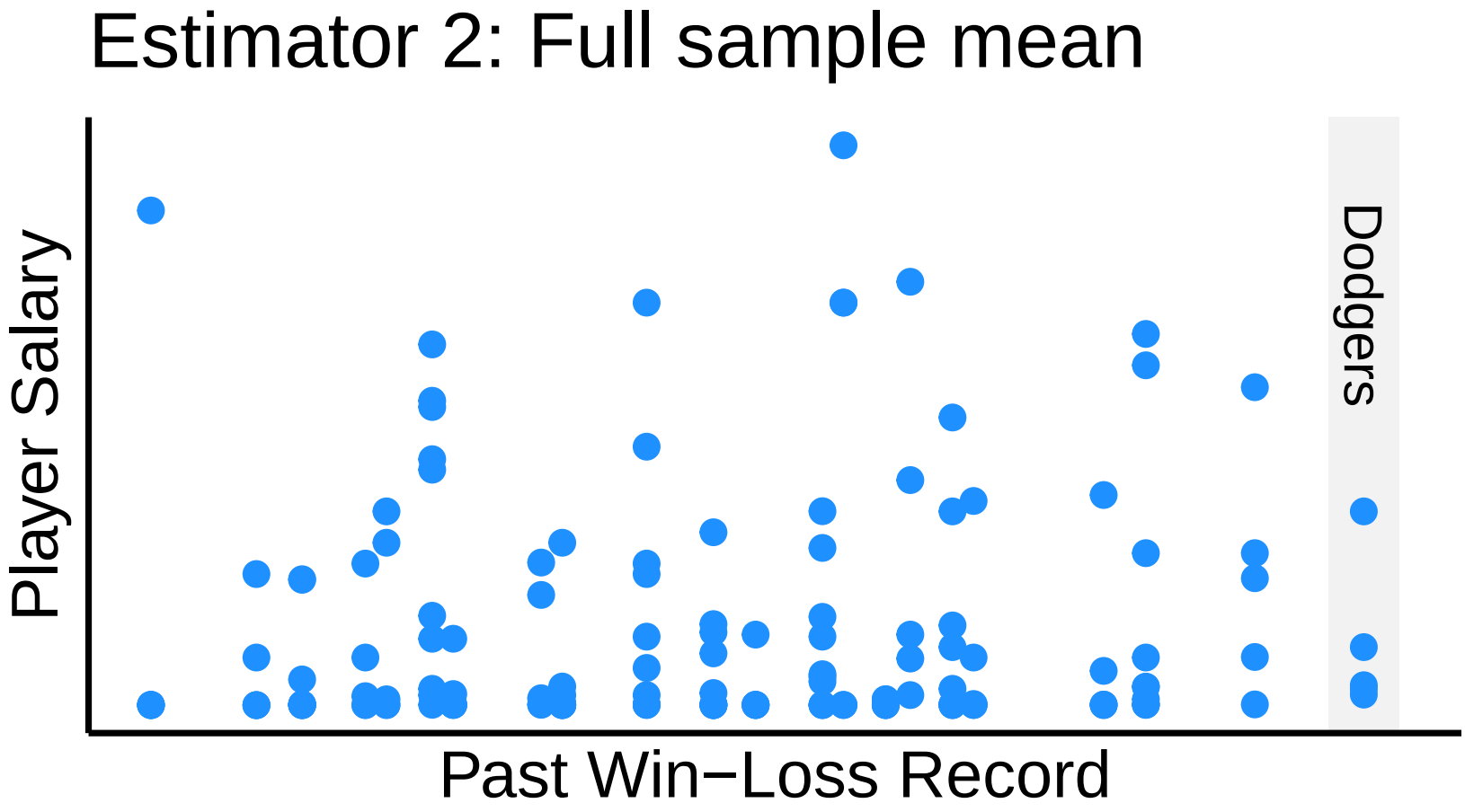
Models for Subgroup Summaries

With only the sample, estimate the mean salary of the Dodgers

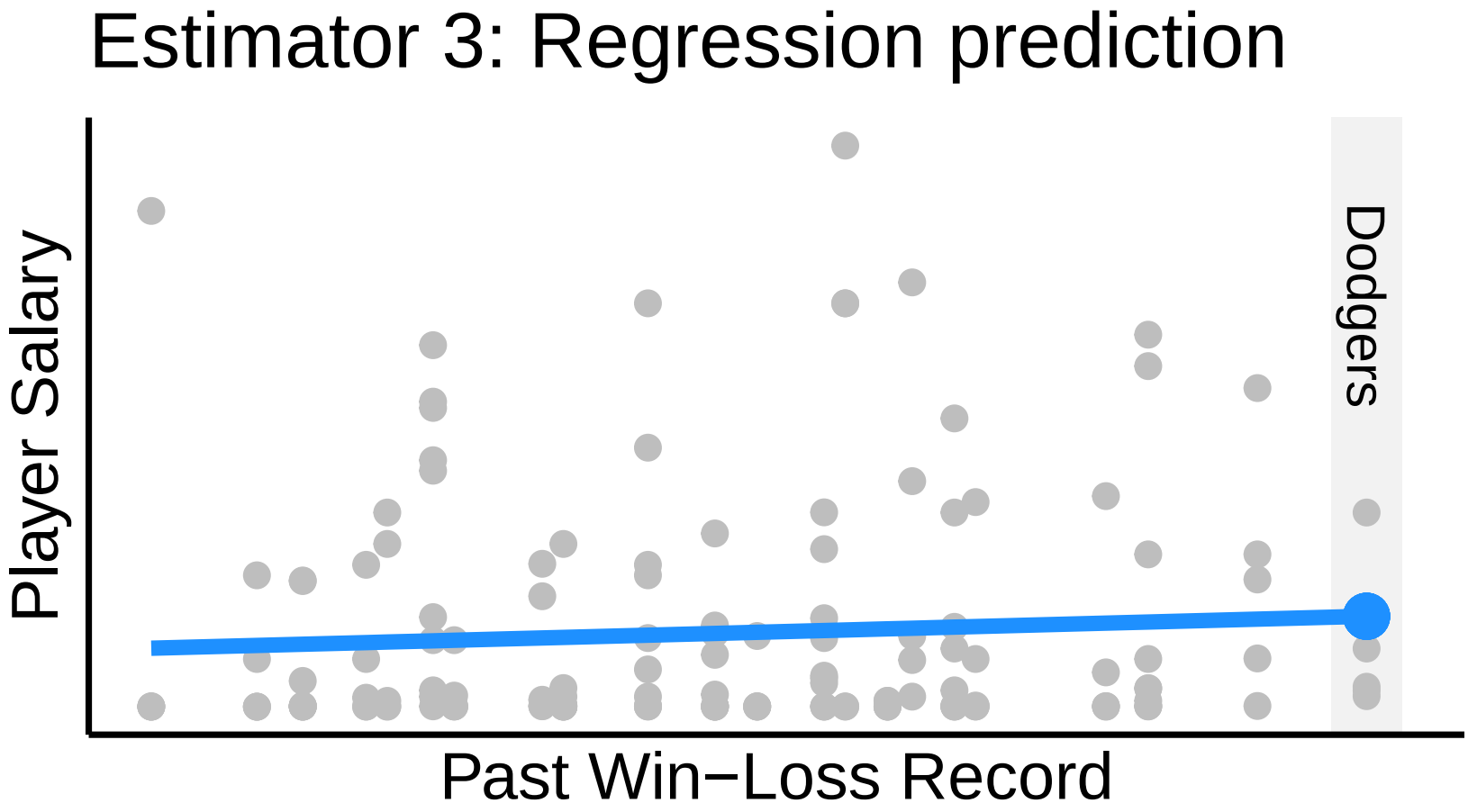


Working with Data: Statistical Learning

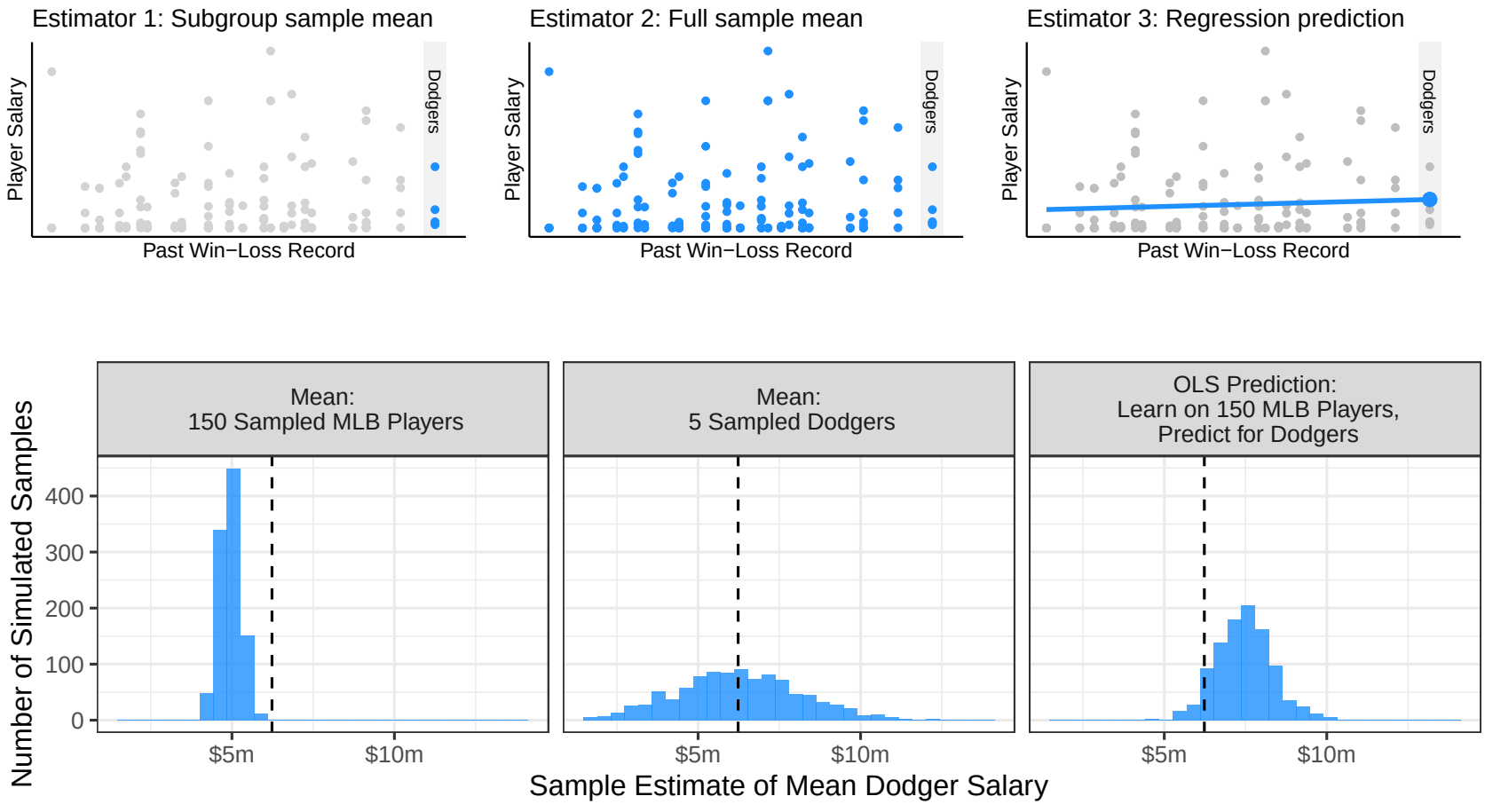



Estimator 2: Full sample mean
Player Salary
Past Win−Loss Record
Dodgers

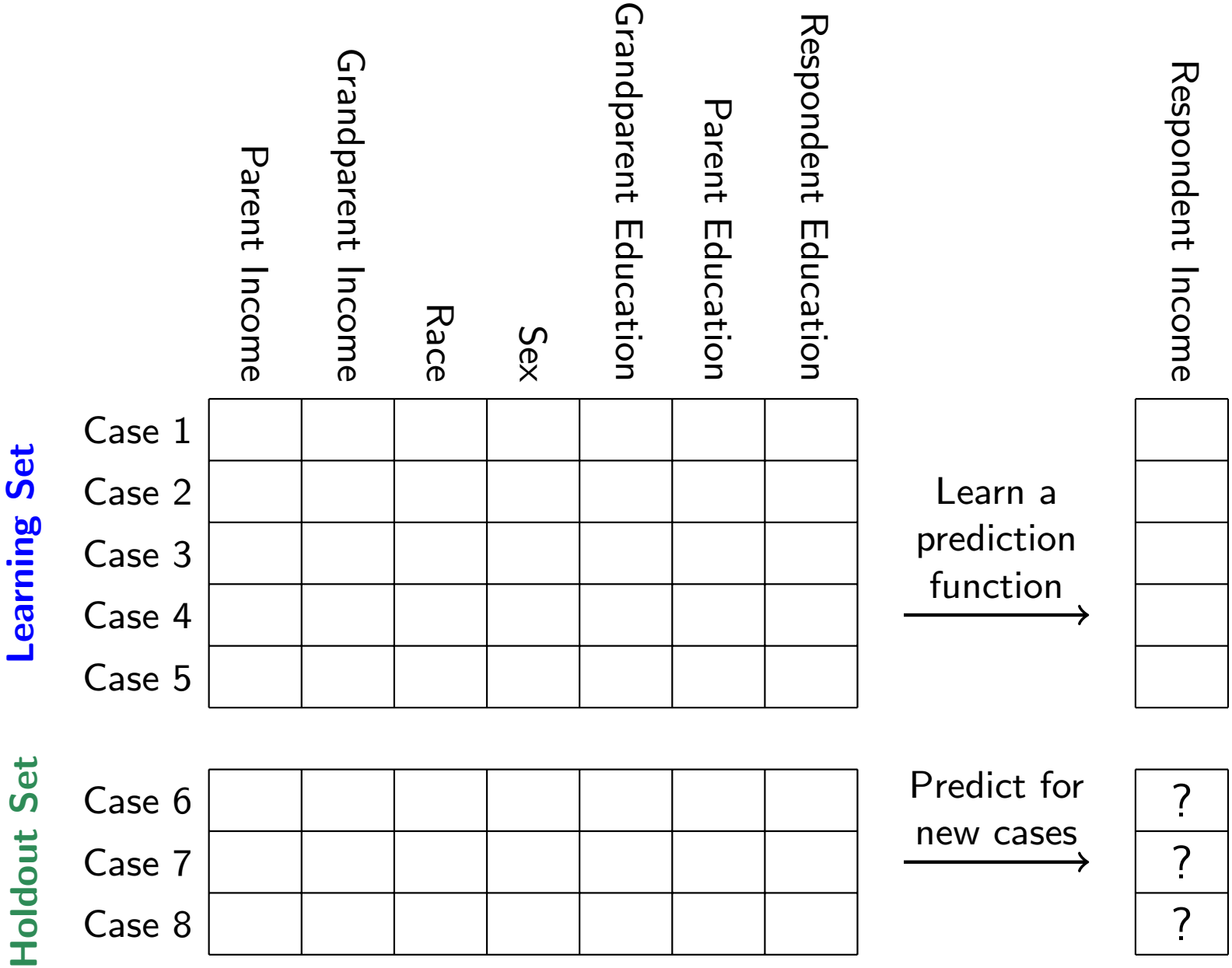
Working with Data: Statistical Learning



Working with Data: Statistical Learning



Working with Data: Statistical Learning



Causal inference


olio
olio
olio
olio

Fundamental problem of causal inference

Holland 1986

Descriptive evidence

average lifespan − average lifespan

Causal claim

average lifespan − average lifespan

Causal inference is a **missing data** problem

	Outcome under Mediterranean diet	Outcome under standard diet
Person 1	lifespan	missing
Person 2	missing	lifespan
Person 3	lifespan	missing
Person 4	missing	lifespan
Person 5	lifespan	missing
Person 6	lifespan	missing
Person 7	missing	lifespan
Person 8	lifespan	missing

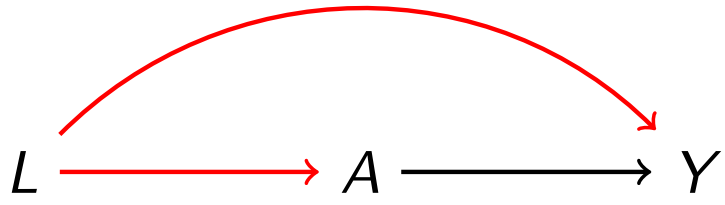
Outcome under Mediterranean diet	Outcome under standard diet
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan

Potential outcomes

$$Y_i^a$$

the outcome Y
of person i
if exposed to treatment $A = a$

Causal identification with DAGs



Non-causal path starts with an edge pointing in to A and ends at Y

Identified cusal effects by an adjustment set that blocks all non-causal paths.

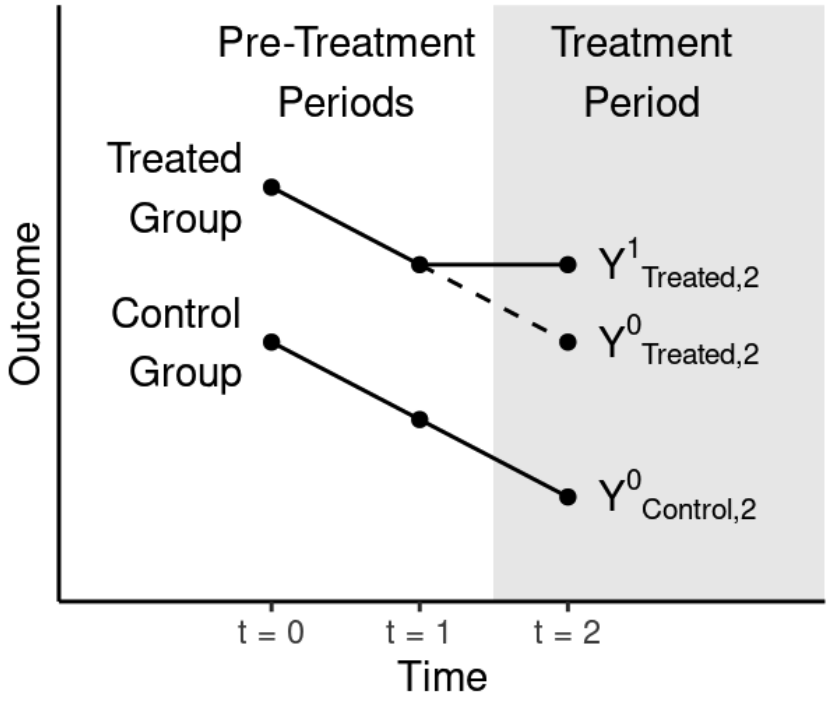
Estimation

- regression
- inverse probability weighting
- matching

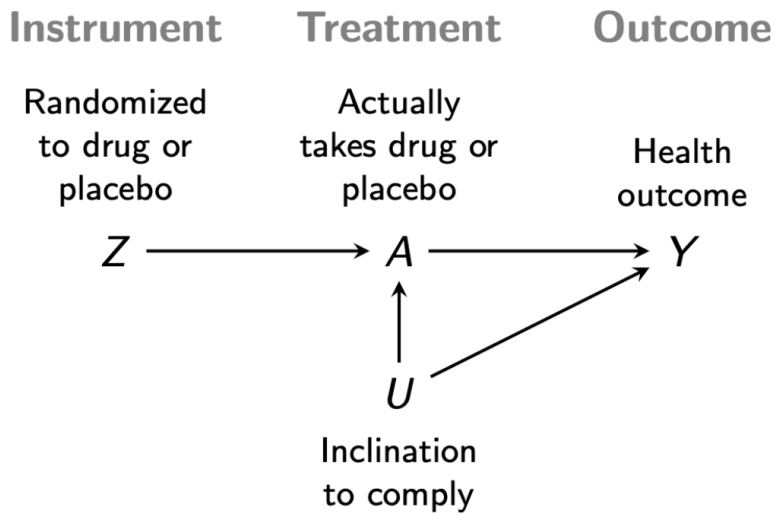
Causal inference with unmeasured confounding

Causal inference with unmeasured confounding

Causal inference with unmeasured confounding



Causal inference with unmeasured confounding



Course structure

- concepts introduced in lecture
- conceptual practice with quizzes
- coding practice with problem sets
- support in office hours and on Piazza

Goals of the course

- connect theories about inequality to quantitative empirical evidence
- evaluate the effects of hypothetical interventions to reduce inequality
- conduct data analysis using the R programming language

Your thoughts

- What could we do to make this course better?
- What is your favorite thing you learned?
- What parts do you anticipate being most useful for your future work?

Course evaluations

- Specific examples are especially helpful
  - Comments on course content
  - Comments on course delivery
- Specific positive experiences with your TA!