Adjusting for Covariate Misclassification to Quantify the Relationship Between Diabetes and Local Access to Healthy Food

Ashley E. Mullan

April 8, 2024

Wake Forest University
Department of Statistical Sciences

Follow along with me!

https://bit.ly/ashley_talks

SCAN ME

Roadmap

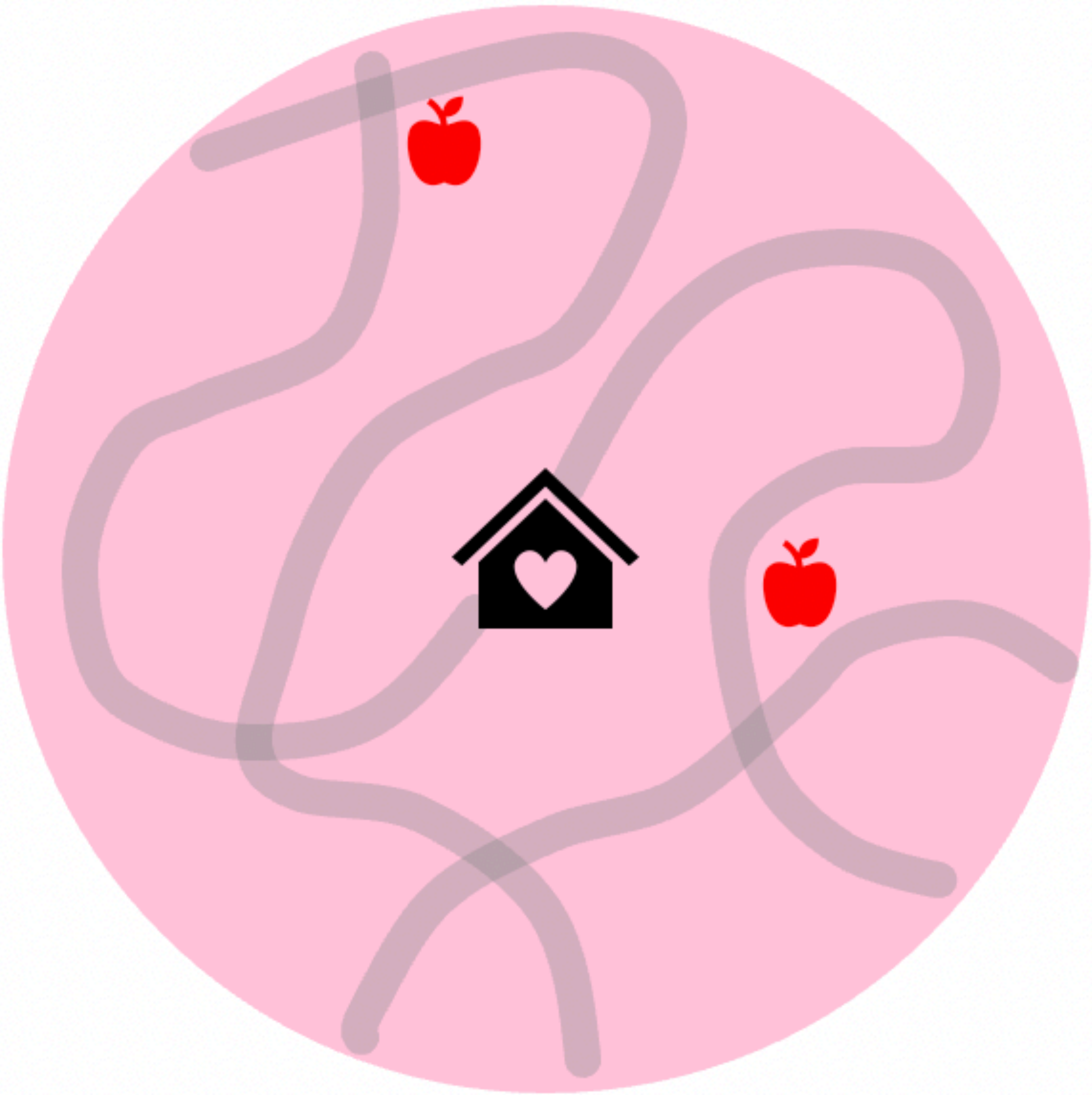
1. Motivation
2. Methods
3. Simulations
4. Case Study
5. Wrap Up

Motivation 🤔

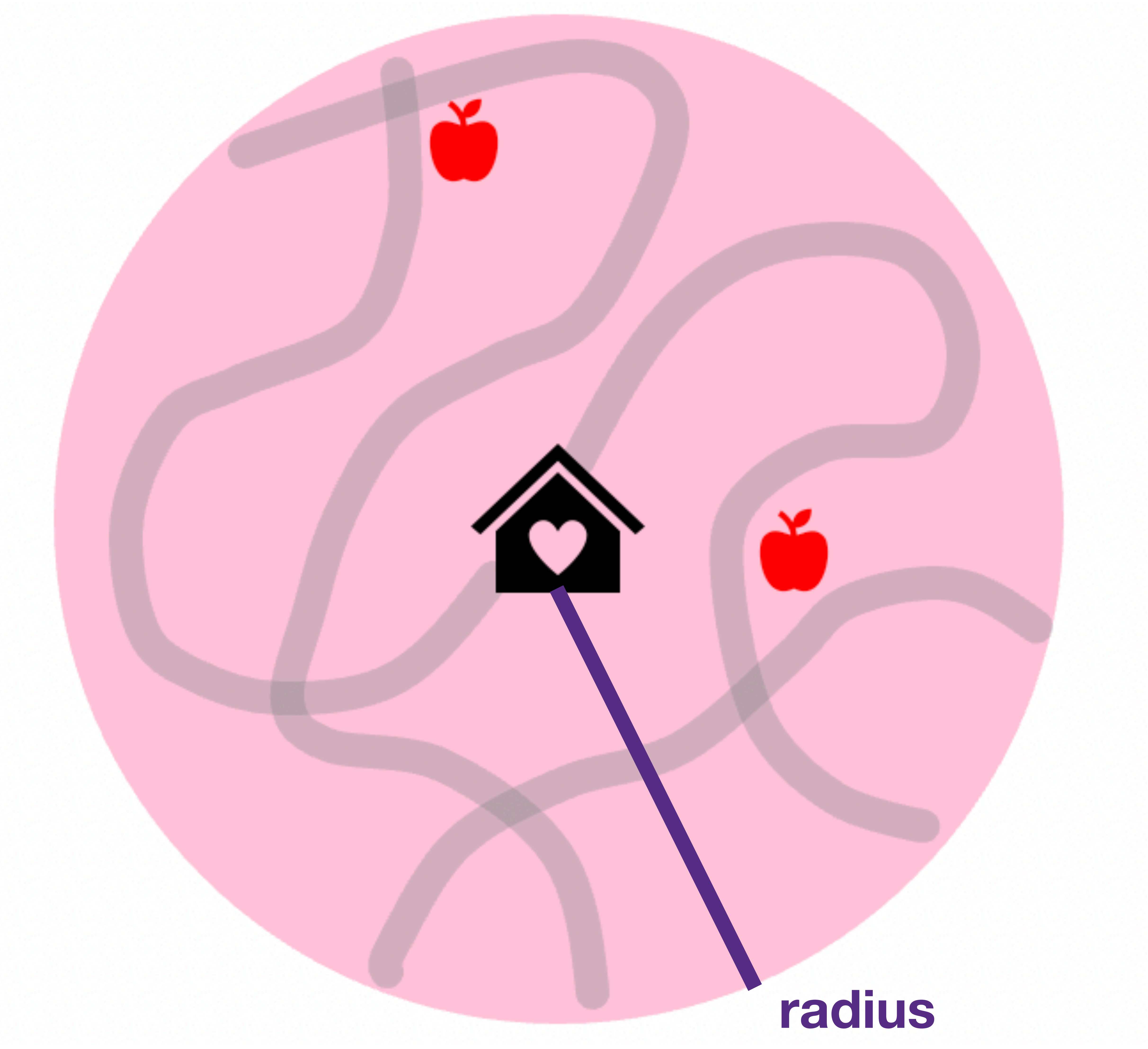
Healthy Eating ➡️ Healthy Living

- A **healthy diet** is full of fruits, vegetables, whole grains, and other high-nutrient foods.
- A healthy diet increases the likelihood of good overall health and **decreases risk of preventable illness** (World Health Organization, 2019).
- Maintaining a healthy diet requires **consistent access to healthy food**, which may be hindered by physical or social barriers like geography or income.
- Review studies found **high prevalence of diabetes** in food-insecure households (Gucciardi et al., 2014).

Measuring Food Access 📏

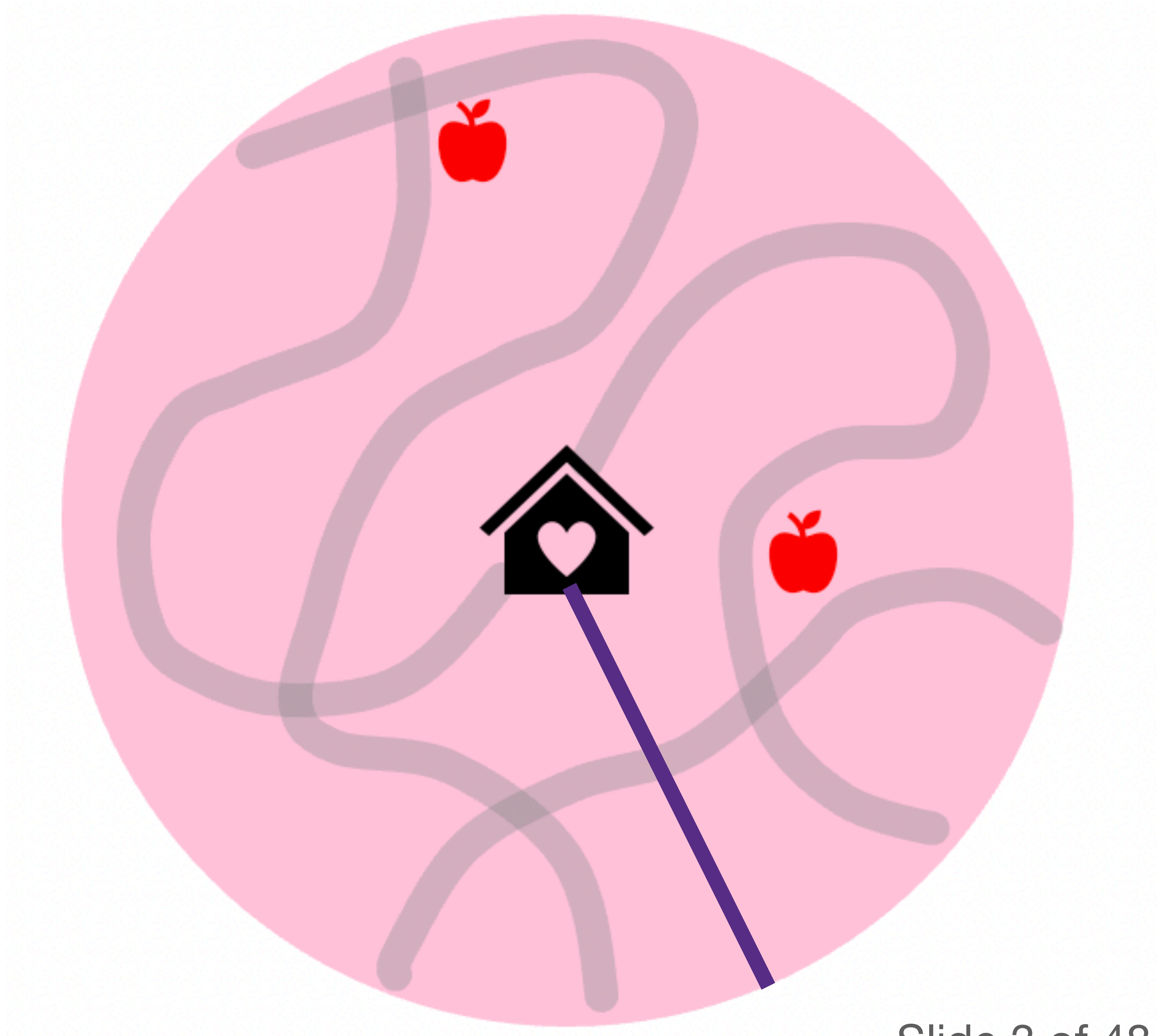


Measuring Food Access 📏



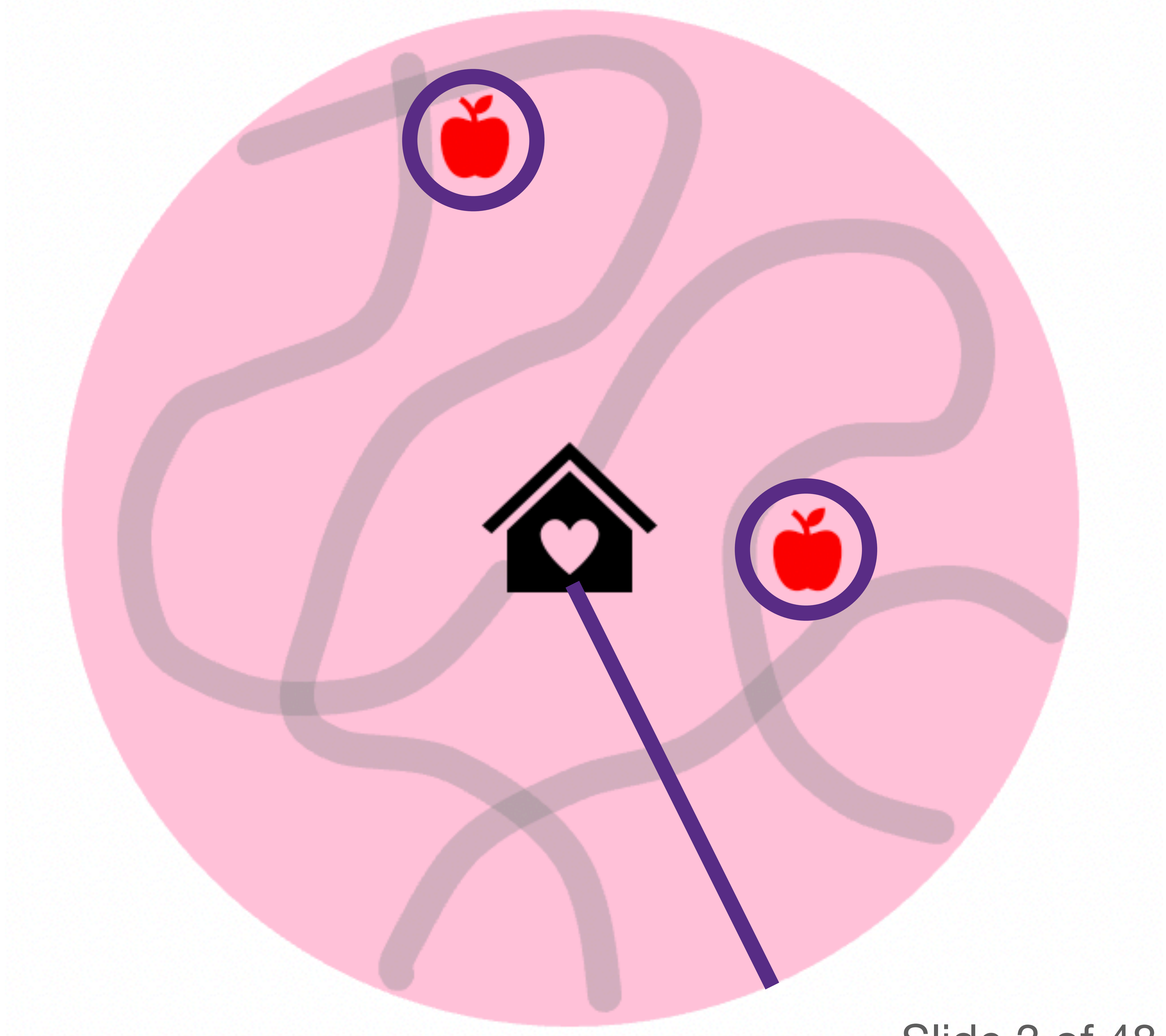
Measuring Food Access 📏

The **density** approach counts the number of healthy food retailers within a given radius.



Measuring Food Access 📏

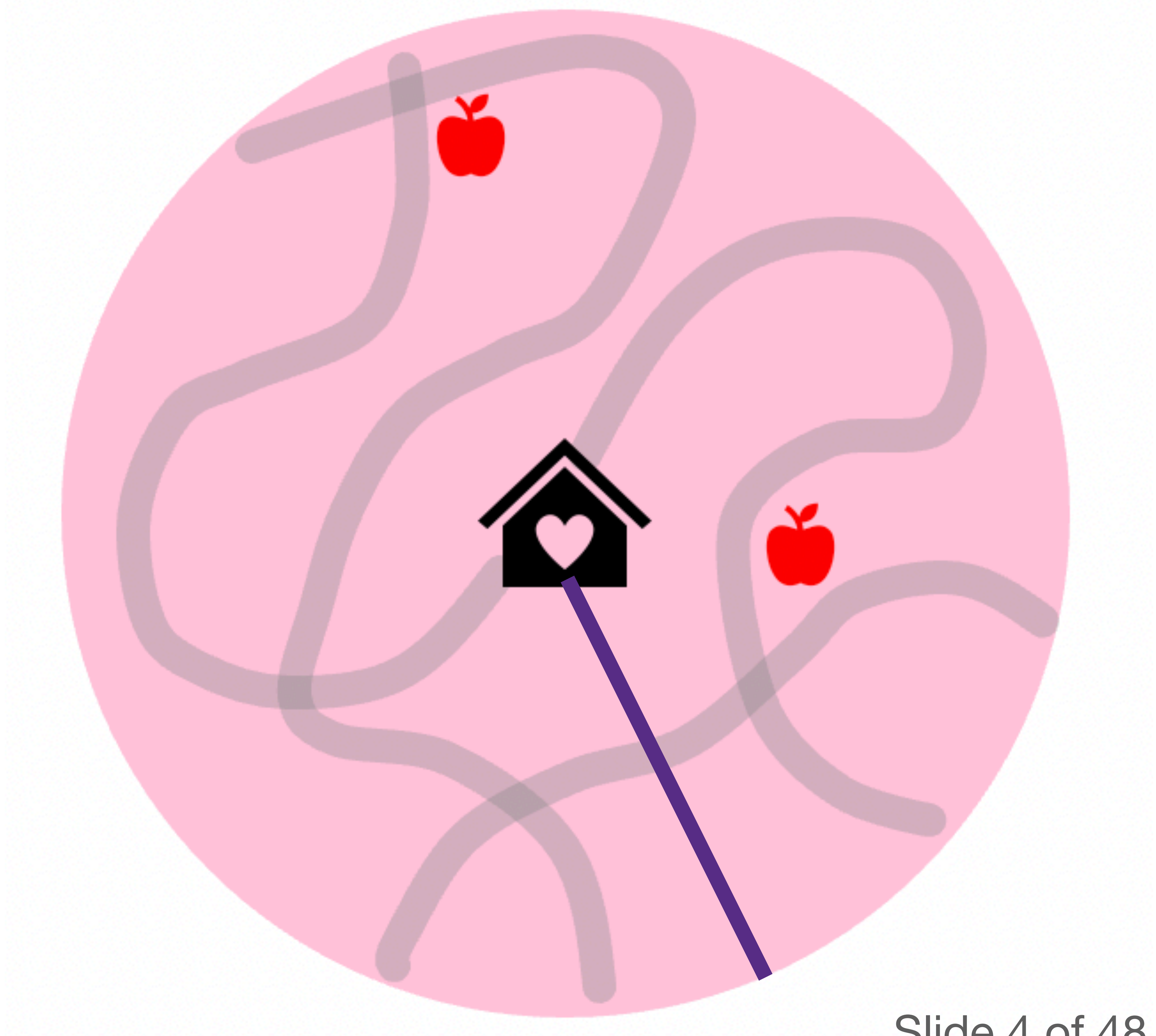
The **density** approach counts the number of healthy food retailers within a given radius.



Measuring Food Access 📏

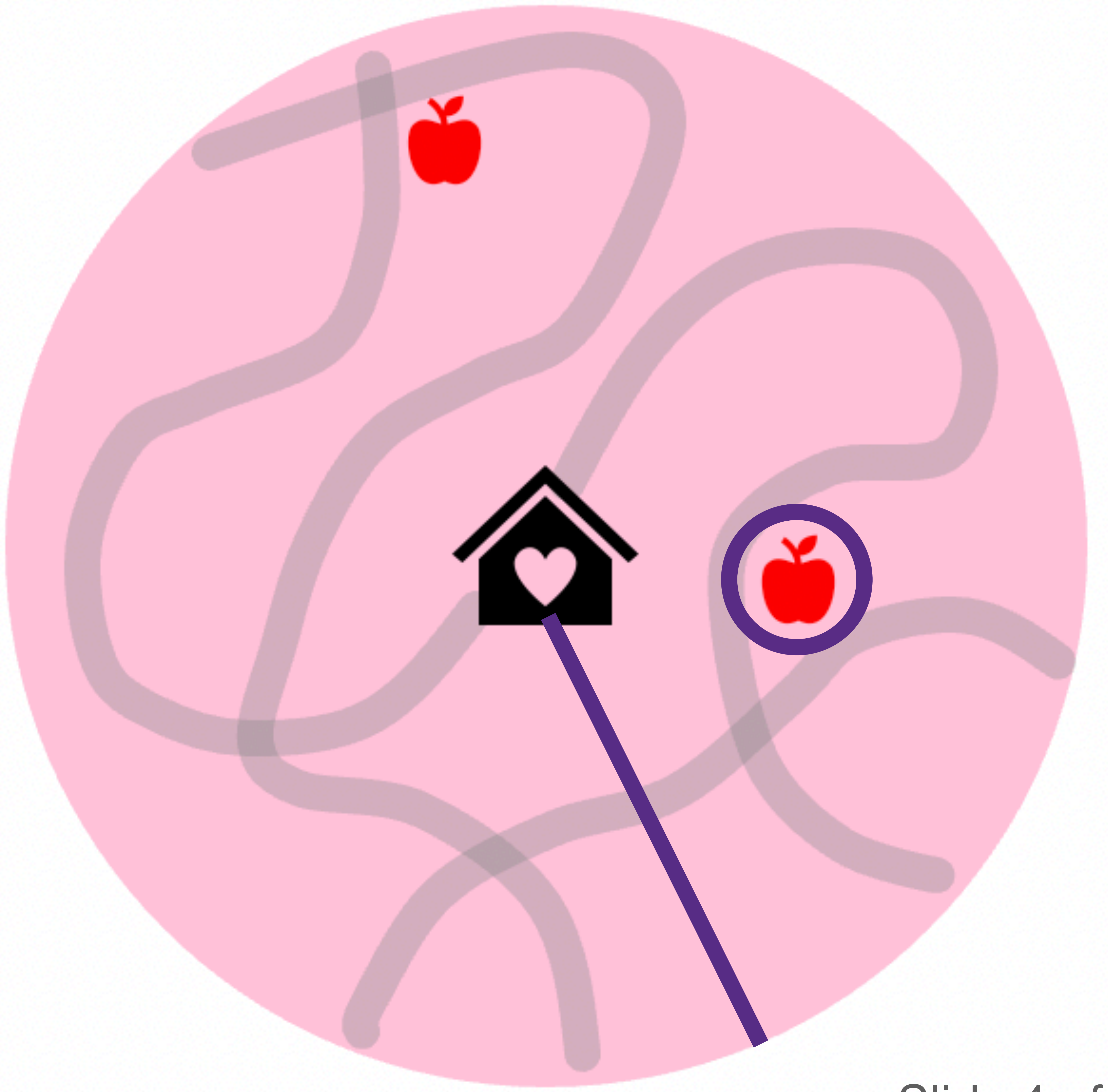
The **proximity** approach measures the distance* to the nearest healthy food retailer.

*more on that later



Measuring Food Access 📏

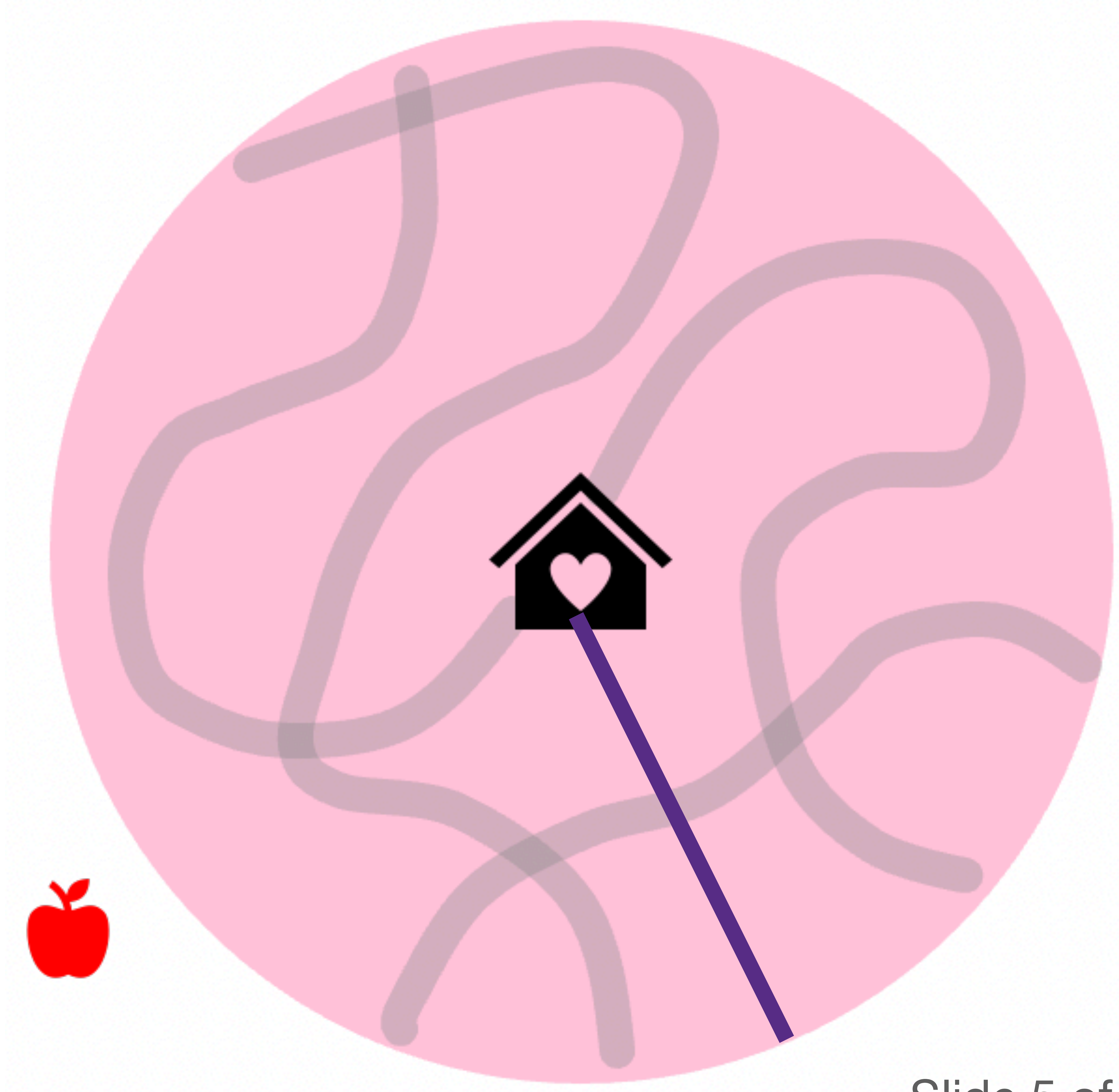
The **proximity** approach measures the distance* to the nearest healthy food retailer.



*more on that later

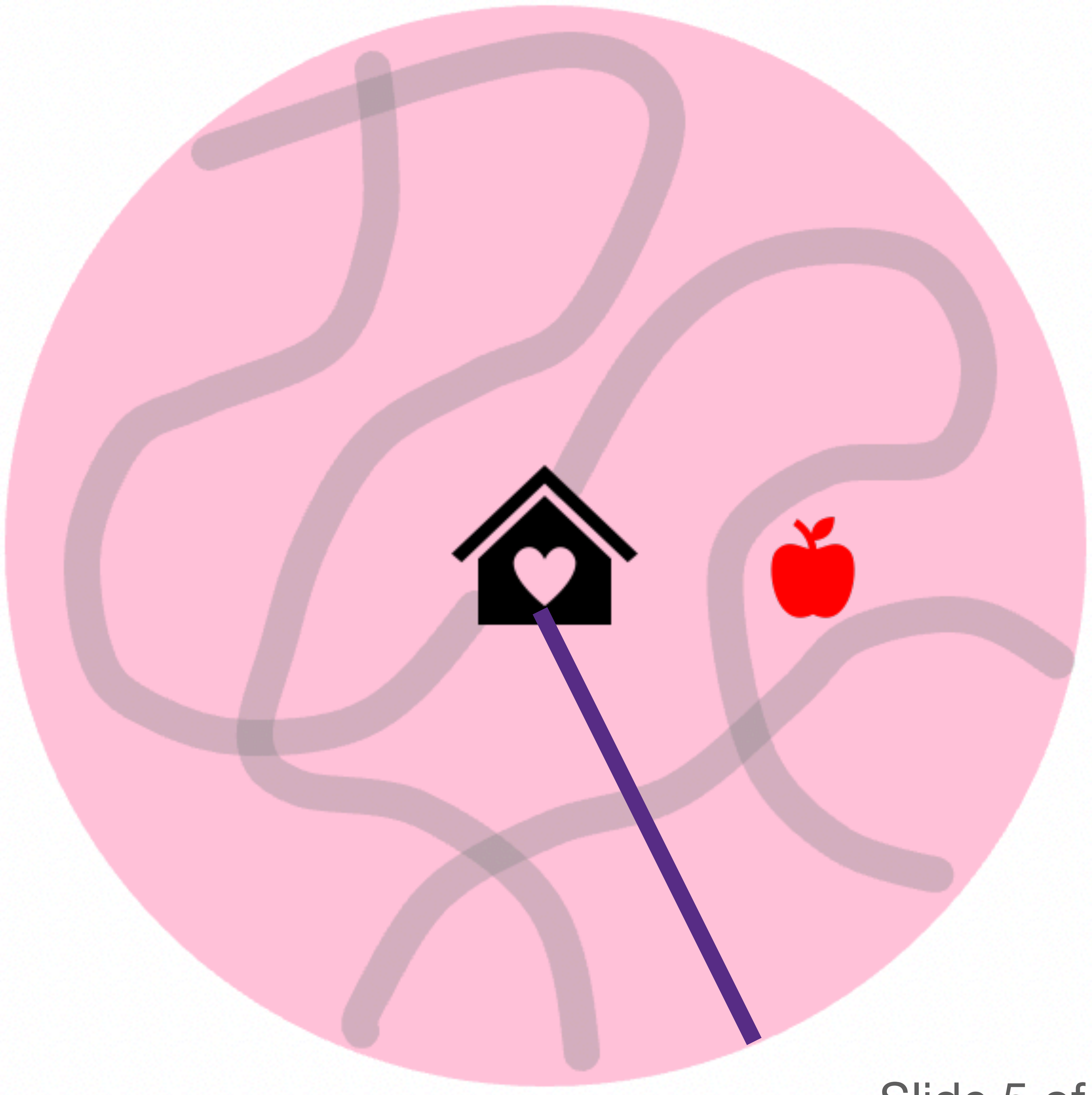
Measuring Food Access 📏

We create an **indicator** of food access that flips on if **at least one** healthy food retailer sits within our radius.



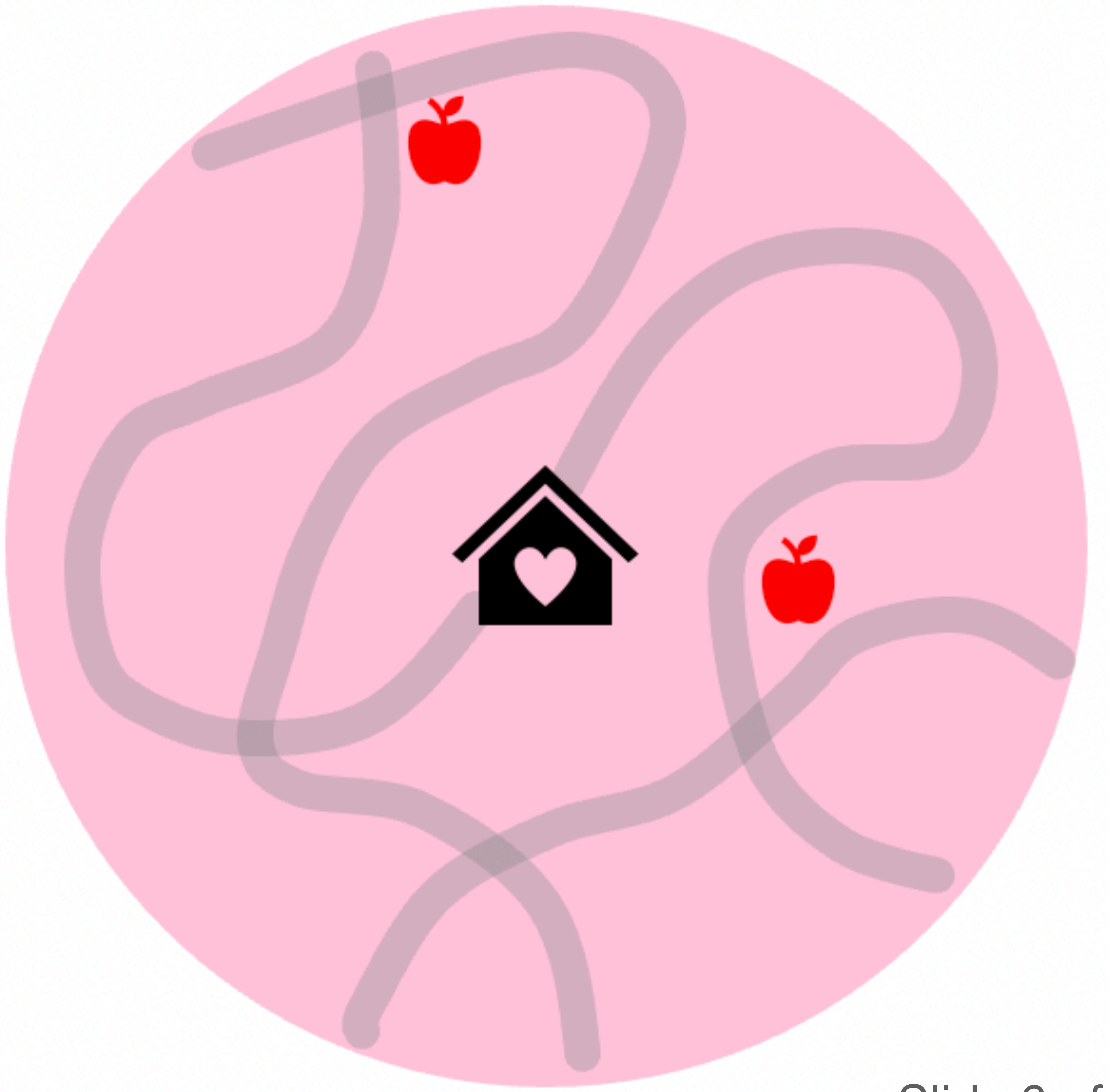
Measuring Food Access 📏

We create an **indicator** of food access that flips on if **at least one** healthy food retailer sits within our radius.



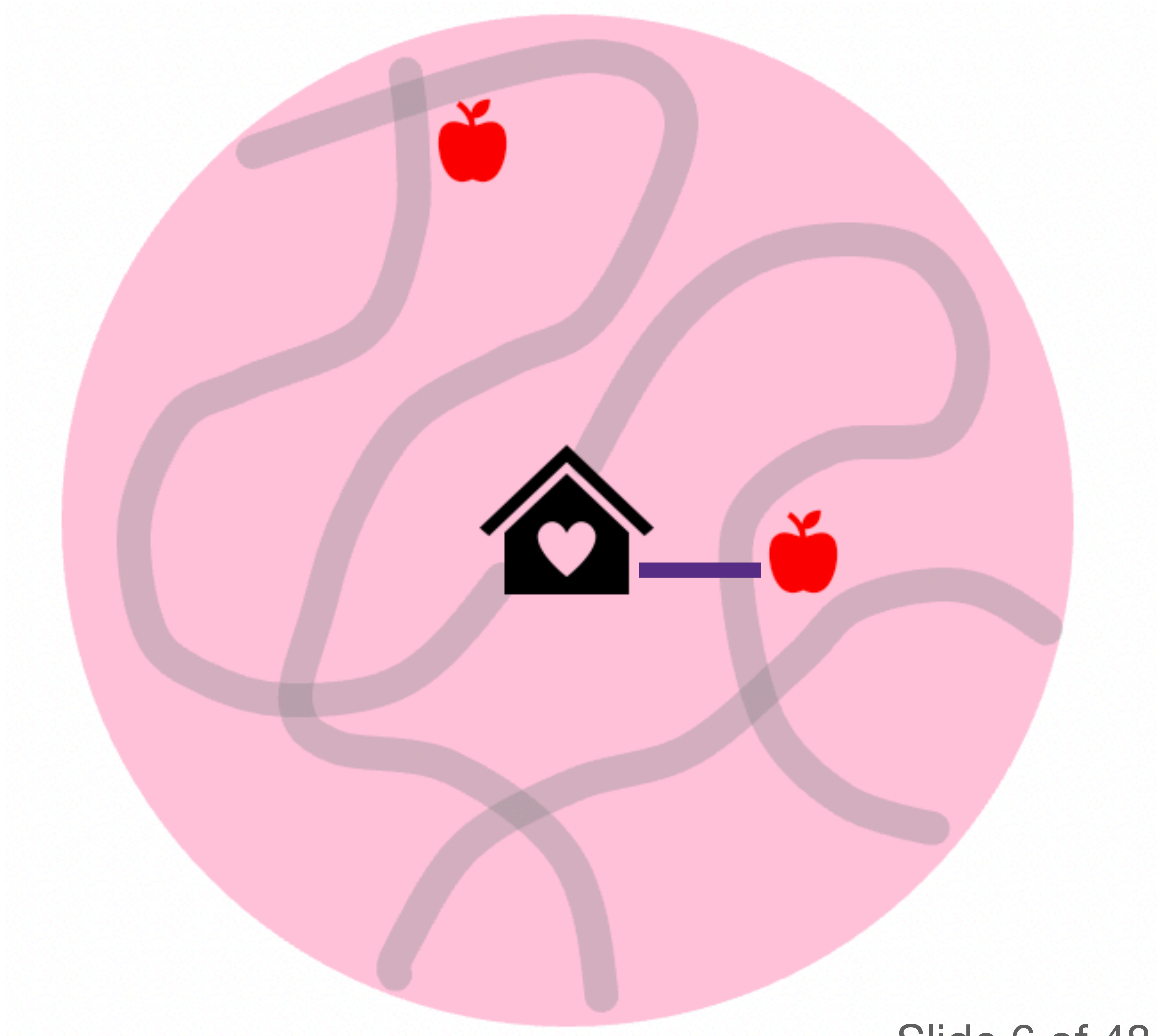
Measuring Food Access 📏

How do we measure the **distance** from our neighborhood to a healthy food retailer?



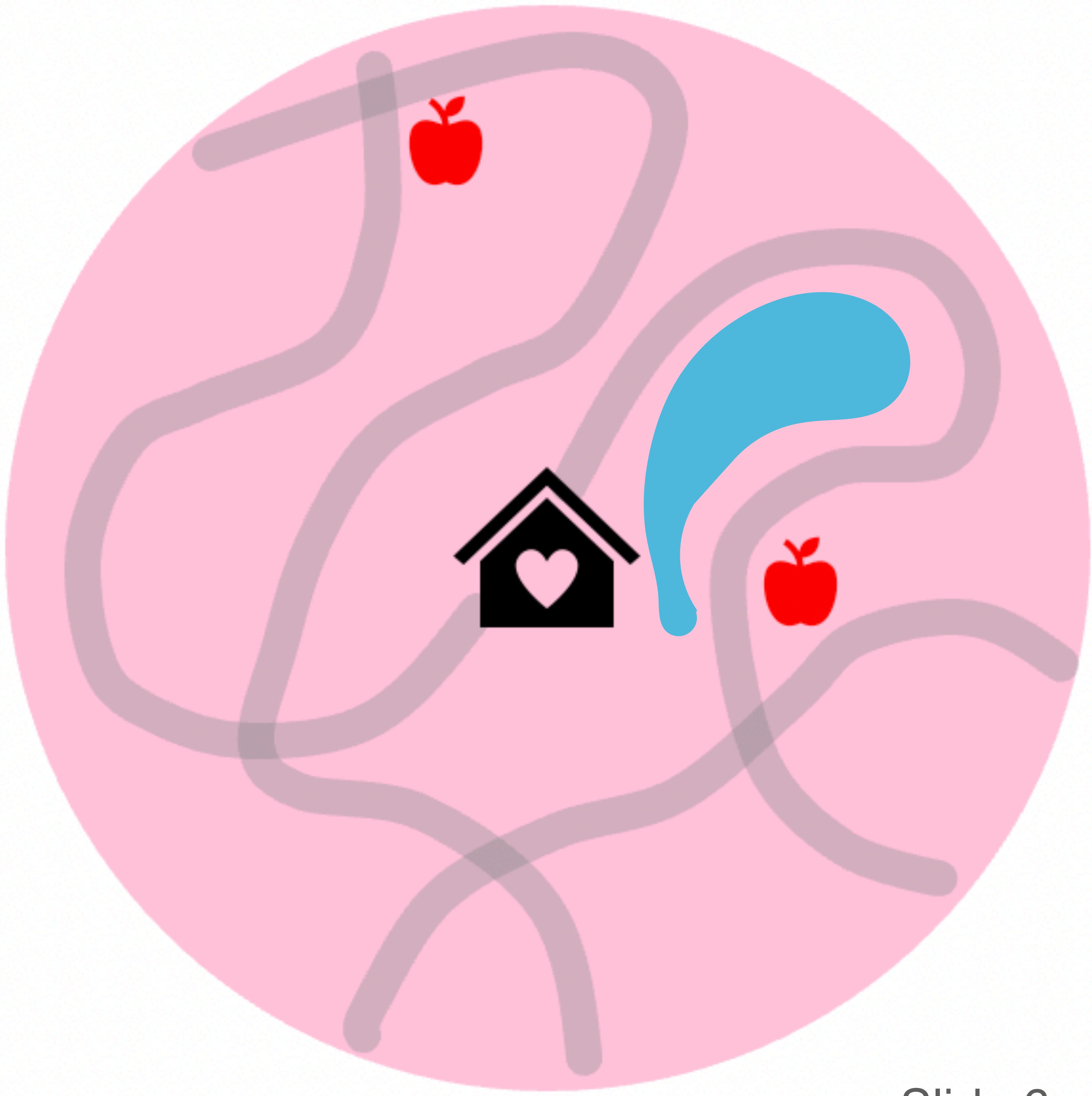
Measuring Food Access 📏

How do we measure the **distance** from our neighborhood to a healthy food retailer?



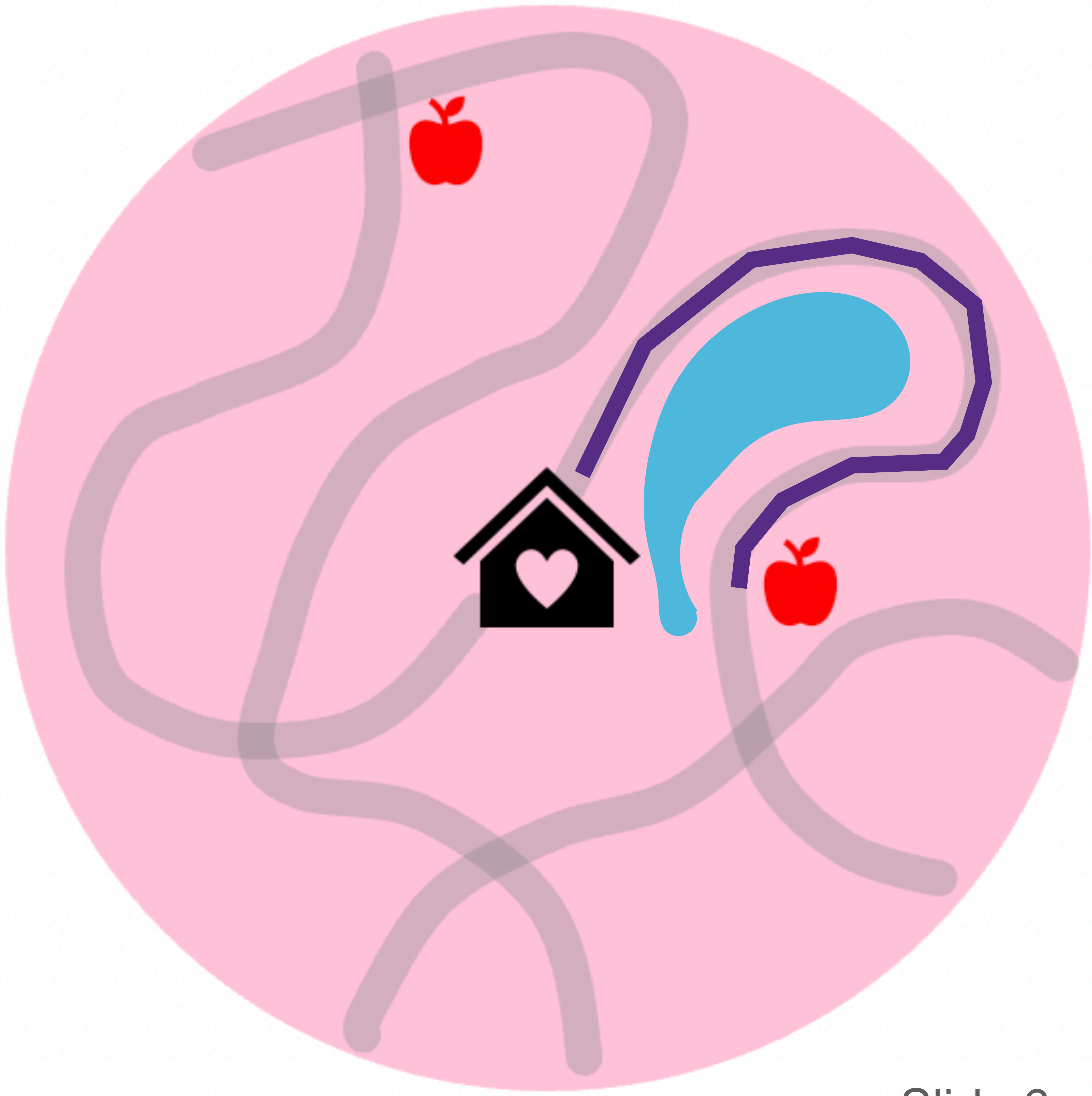
Measuring Food Access 📏

How do we measure the **distance** from our neighborhood to a healthy food retailer?



Measuring Food Access 📏

How do we measure the **distance** from our neighborhood to a healthy food retailer?



Distance Computations

- The **Haversine distance** is a trigonometric function of latitude and longitude.
- It ignores physical obstacles, so it **underestimates** the true distance between two points and is considered **error-prone**.
- The Haversine distance in the image is **impassable**, as it crosses a pond.

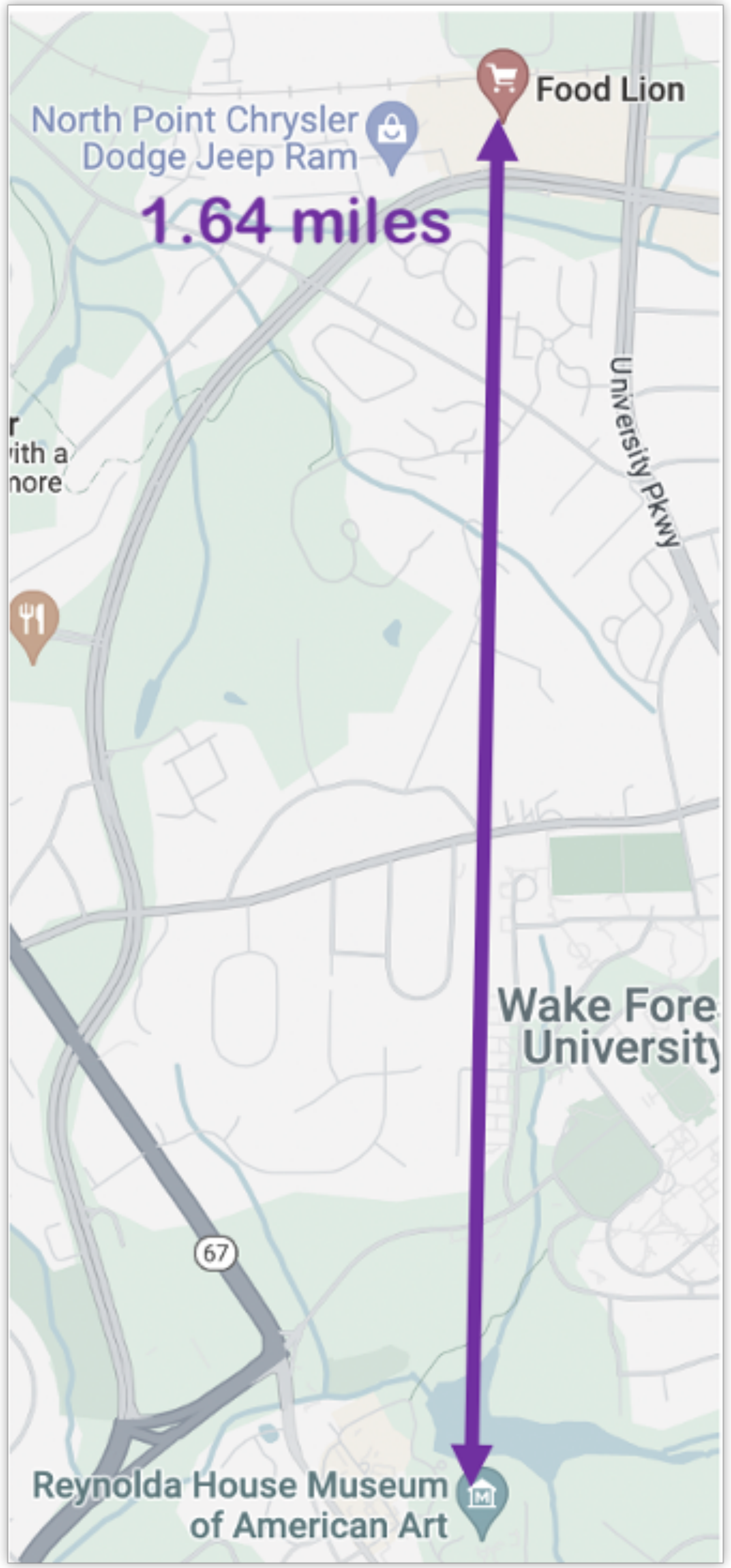


Figure: Haversine distance from Reynolda Manor House to a nearby Food Lion

Distance Computations

- The **route-based distance** works around obstacles.
- It is **more accurate** than the Haversine distance, but it is **computationally and financially expensive**.
- These distances are computed with the **ggmap** package in R, which accesses the Google Maps API.
- In our case study, these distances are **over a mile** longer than the Haversine distances for **1 in 5** neighborhoods!

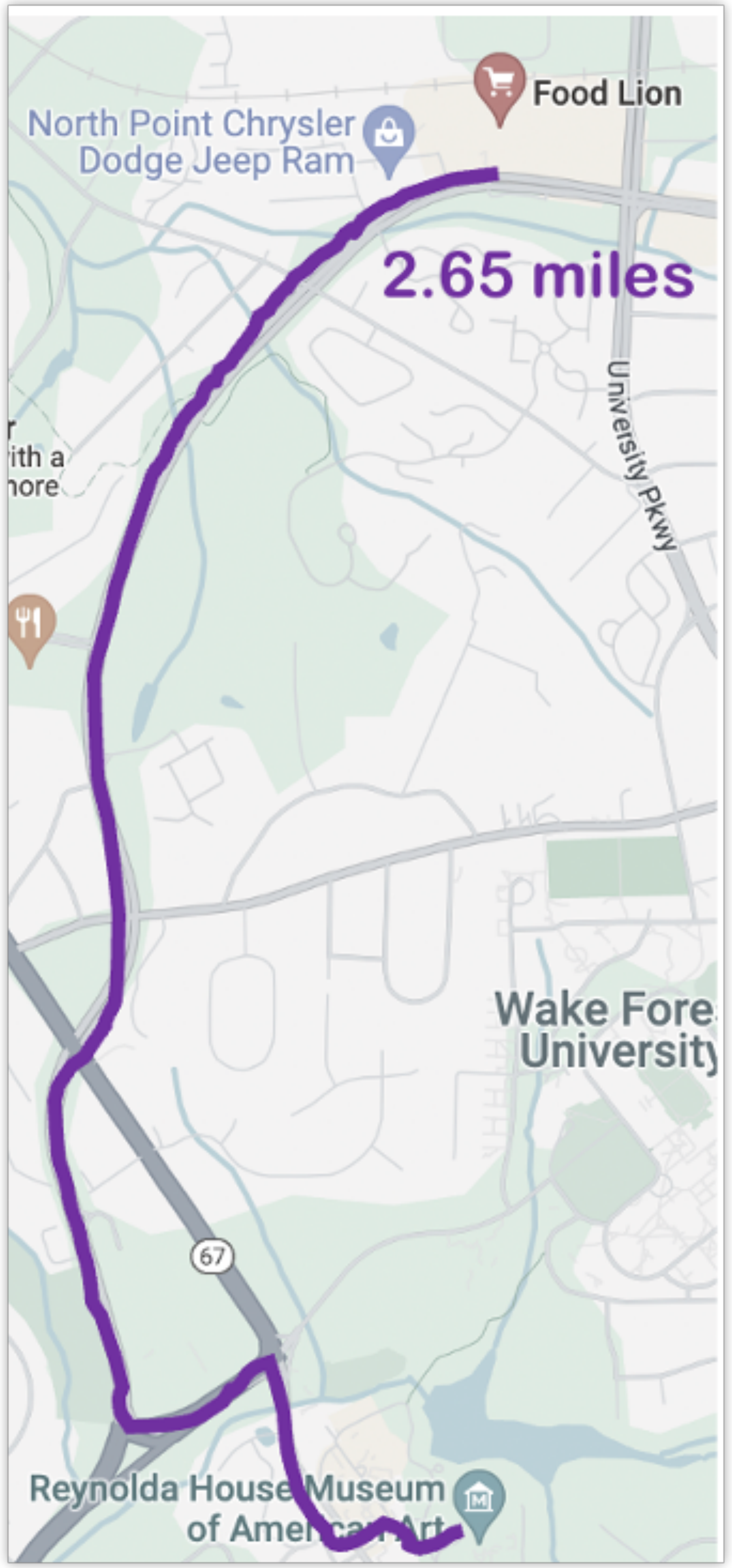
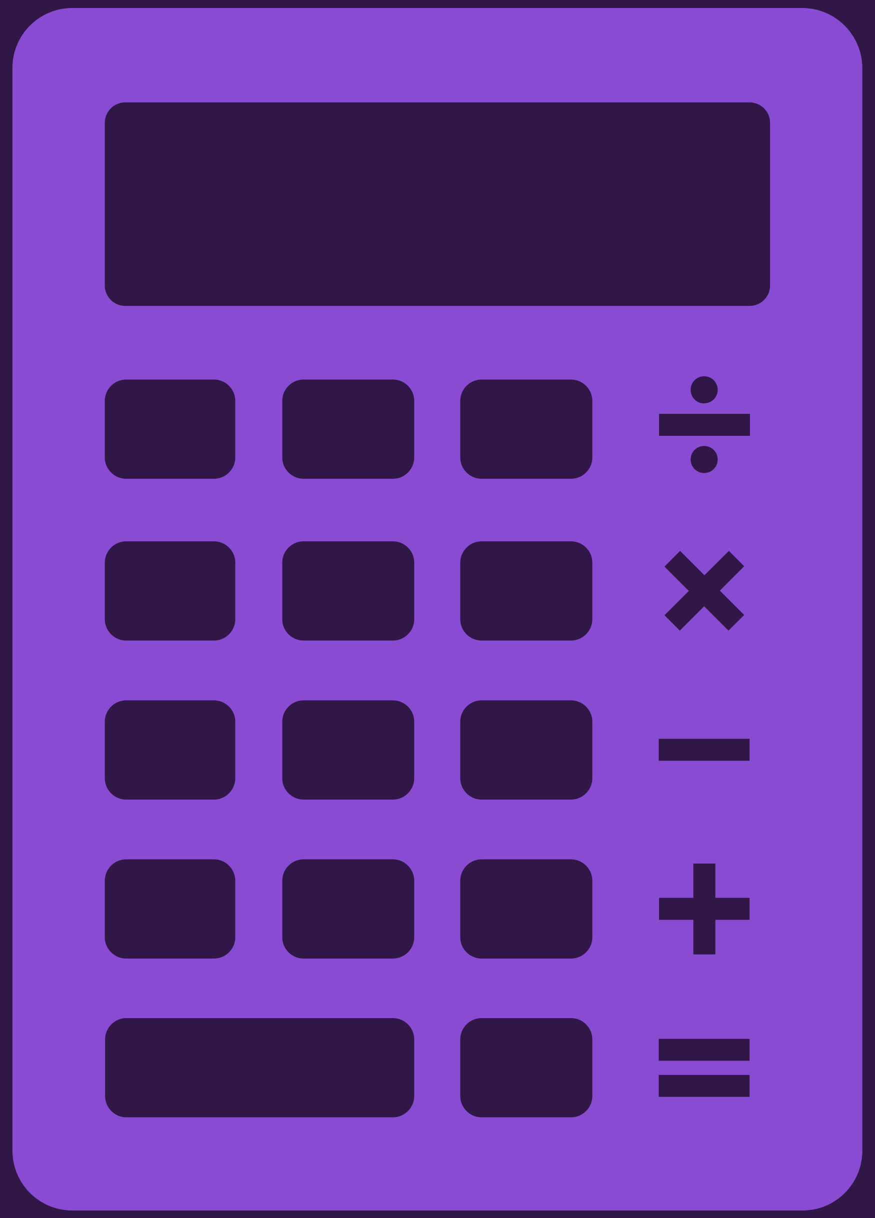


Figure: Route distance from Reynolda Manor House to a nearby Food Lion

Guiding Questions

- Can we use a function of distance to healthy food retailers to **quantify food access** in the Piedmont Triad, even if this function is **subject to misclassification**?
- Can we estimate the relationship between **food access** and **diabetes prevalence** in the presence of misclassifications and missingness?

Methods



🚨 it’s about to get math heavy 🚨

Variable Notation

Variable Notation

- X_r is an error-free binary explanatory variable for food access based on route-based distances and a radius r (e.g., $r = 1$ mile)

Variable Notation

- X_r is an error-free binary explanatory variable for food access based on route-based distances and a radius r (e.g., $r = 1$ mile)

- X^*_r is an error-prone version of X_r based on Haversine distances

Variable Notation

- X_r is an error-free binary explanatory variable for food access based on route-based distances and a radius r (e.g., $r = 1$ mile)
- X^*_r is an error-prone version of X_r based on Haversine distances
- $\mathbf{Z}$ is an error-free covariate vector

Variable Notation

- X_r is an error-free binary explanatory variable for food access based on route-based distances and a radius r (e.g., $r = 1$ mile)
- X^*_r is an error-prone version of X_r based on Haversine distances
- $\mathbf{Z}$ is an error-free covariate vector
- Y is a count of diabetes cases in the area of interest

Variable Notation

- X_r is an error-free binary explanatory variable for food access based on route-based distances and a radius r (e.g., $r = 1$ mile)
- X^*_r is an error-prone version of X_r based on Haversine distances
- $\mathbf{Z}$ is an error-free covariate vector
- Y is a count of diabetes cases in the area of interest
- Q is an indicator of whether a neighborhood has been queried

Variable Notation

- X_r is an error-free binary explanatory variable for food access based on route-based distances and a radius r (e.g., $r = 1$ mile)
- X^*_r is an error-prone version of X_r based on Haversine distances
- $\mathbf{Z}$ is an error-free covariate vector
- Y is a count of diabetes cases in the area of interest
- Q is an indicator of whether a neighborhood has been queried
- O is an offset, the population of the area

Model Notation

Model Notation

- Outcome Model

$$Y_i \mid X_{ri}, \boldsymbol{Z}_i \sim \text{Poisson}(\lambda_i)$$

$$\lambda_i = \beta_0 + \beta_1 X_{ri} + \boldsymbol{\beta_2 Z}_i$$

Model Notation

- Outcome Model

$$Y_i \mid X_{ri}, \boldsymbol{Z}_i \sim \text{Poisson}(\lambda_i)$$

$$\lambda_i = \beta_0 + \beta_1 X_{ri} + \boldsymbol{\beta_2 Z}_i$$

exponentiate to get the prevalence ratio

Model Notation

- Outcome Model

$$Y_i \mid X_{ri}, \boldsymbol{Z}_i \sim \text{Poisson}(\lambda_i)$$

$$\lambda_i = \beta_0 + \beta_1 X_{ri} + \boldsymbol{\beta_2 Z}_i$$

exponentiate to get the prevalence ratio

- Error Model

$$X_{ri} \mid X^*_{ri}, \boldsymbol{Z}_i \sim \text{Bernoulli}(\pi_i)$$

$$\pi_i = \text{expit}(\eta_0 + \eta_1 X^*_{ri} + \boldsymbol{\eta_2 Z}_i)$$

A Little More on X_r and X^*_r

- Let d be the route-based distance to the nearest healthy food retailer.
- Let h be the Haversine distance to the nearest healthy food retailer.
- Let r be the radius of interest.

$$X_r = \begin{cases} 1 \text{ if } d \leq r & \text{“Access”} \\ 0 \text{ if } d > r & \text{“No Access”} \end{cases}$$

$$X^*_r = \begin{cases} 1 \text{ if } h \leq r & \text{“Error-Prone Access”} \\ 0 \text{ if } h > r & \text{“No Access”} \end{cases}$$

Two-Phase Design

- Having **some correct** route-based distances is better than none.
- **Error-prone** Haversine distances are available for all N neighborhoods, and we can use them to create our indicator of food access X^*_r that is subject to **misclassification**.
- In addition to X^*_r, we **query** route-based distances to create our indicator X_r for n neighborhoods, where $n < N$.
- We now have a **missing data problem**, as (N - n) neighborhoods only have X^*_r.

Outcome Model Options

- **Gold Standard**
- Naive Analysis
- Complete Case Analysis
- Maximum Likelihood Estimation

👍

The model achieves optimal bias and variance.

👎

The model assumes we have all of the correct data available, but we do not.

Outcome Model Options

- Gold Standard
- **Naive Analysis**
- Complete Case Analysis
- Maximum Likelihood Estimation

👍

The model is easy to fit and utilizes information from the error-prone data for all of the neighborhoods.

👎

The model is biased by a function of the sensitivity and specificity (Shaw et al., 2020).

Outcome Model Options

- Gold Standard
- Naive Analysis
- **Complete Case Analysis**
- Maximum Likelihood Estimation

👍

The model is unbiased, as it uses the error-free measurements.

👎

The model does not take the unqueried data into account.

Outcome Model Options

- Gold Standard
- Naive Analysis
- Complete Case Analysis
- **Maximum Likelihood Estimation**

👍

The model utilizes information from both the queried and unqueried observations.

👎

The method was not yet derived or implemented in existing software.

Outcome Model Options

- Gold Standard
- Naive Analysis
- Complete Case Analysis
- **Maximum Likelihood Estimation**

👍

The model utilizes information from both the queried and unqueried observations.

👎

The method was not yet derived or implemented in existing software.

Roadmap

Putting Together the MLE

We have four cases of data quality.

1. *No misclassification or missingness* ($X_r = X^*_r$ always)
2. *Misclassification without missingness* (always have X_r and X^*_r)
3. *Misclassification and total missingness* (never have X_r but always X^*_r)
4. *Misclassification and partial missingness* (sometimes have X_r but always X^*_r)

Case 1

No misclassification or missingness

$$P_{\beta,\eta}(Y, X, Z) = P_\beta(Y \mid X, Z)P_\eta(X \mid Z)P(Z)$$

Case 1

No misclassification or missingness

$$P_{\beta,\eta}(Y, X, Z) = P_\beta(Y \mid X, Z) P_\eta(X \mid Z) P(Z)$$

outcome model

Case 1

No misclassification or missingness

$$P_{\beta,\eta}(Y, X, Z) = P_\beta(Y \mid X, Z)P_\eta(X \mid Z)P(Z)$$

Case 1

No misclassification or missingness

error model

$$P_{\beta,\eta}(Y, X, Z) = P_{\beta}(Y \mid X, Z)P_{\eta}(X \mid Z)P(Z)$$

Case 1

No misclassification or missingness

$$P_{\beta,\eta}(Y, X, Z) = P_{\beta}(Y \mid X, Z)P_{\eta}(X \mid Z)P(Z)$$

Case 1

No misclassification or missingness

$$P_{\beta,\eta}(Y, X, Z) = P_\beta(Y \mid X, Z) P_\eta(X \mid Z) P(Z)$$

drops out

Case 1

No misclassification or missingness

$$P_{\beta,\eta}(Y, X, Z) = P_\beta(Y \mid X, Z) P_\eta(X \mid Z) P(Z)$$

Case 2

Misclassification without missingness

$$P_{\beta,\eta}(Y, X, Z, X^*) = P_\beta(Y \mid X, Z)P_\eta(X \mid X^*, Z)P(X^*, Z)$$

Case 2

Misclassification without missingness

$$P_{\beta,\eta}(Y, X, Z, X^*) = P_\beta(Y \mid X, Z) P_\eta(X \mid X^*, Z) P(X^*, Z)$$

outcome model

Case 2

Misclassification without missingness

$$P_{\beta,\eta}(Y, X, Z, X^*) = P_\beta(Y \mid X, Z)P_\eta(X \mid X^*, Z)P(X^*, Z)$$

Case 2

Misclassification without missingness

error model

$$P_{\beta,\eta}(Y, X, Z, X^*) = P_\beta(Y \mid X, Z) P_\eta(X \mid X^*, Z) P(X^*, Z)$$

Case 2

Misclassification without missingness

$$P_{\beta,\eta}(Y, X, Z, X^*) = P_{\beta}(Y \mid X, Z)P_{\eta}(X \mid X^*, Z)P(X^*, Z)$$

Case 2

Misclassification without missingness

$$P_{\beta,\eta}(Y, X, Z, X^*) = P_\beta(Y \mid X, Z)P_\eta(X \mid X^*, Z)P(X^*, Z)$$

drops out

Case 2

Misclassification without missingness

$$P_{\beta,\eta}(Y, X, Z, X^*) = P_{\beta}(Y \mid X, Z)P_{\eta}(X \mid X^*, Z)P(X^*, Z)$$

Case 2

Misclassification without missingness

$$P_{\beta,\eta}(Y, X, Z, X^*) = P_\beta(Y \mid X, Z)P_\eta(X \mid X^*, Z)P(X^*, Z)$$

Case 3

Misclassification and total missingness

$$P_{\beta,\eta}(Y, X^*, Z) = \sum_{x=0}^{1} P_\beta(Y \mid X = x, Z) P_\eta(X = x \mid Z) P(X^*, Z)$$

Case 3

Misclassification and total missingness

$$P_{\beta,\eta}(Y, X^*, Z) = \sum_{x=0}^{1} P_{\beta}(Y \mid X = x, Z) P_{\eta}(X = x \mid Z) P(X^*, Z)$$

outcome model

Case 3

Misclassification and total missingness

$$P_{\beta,\eta}(Y, X^*, Z) = \sum_{x=0}^{1} P_{\beta}(Y \mid X = x, Z) P_{\eta}(X = x \mid Z) P(X^*, Z)$$

Case 3

Misclassification and total missingness

error model

$$P_{\beta,\eta}(Y, X^*, Z) = \sum_{x=0}^{1} P_\beta(Y \mid X = x, Z) P_\eta(X = x \mid Z) P(X^*, Z)$$

Case 3

Misclassification and total missingness

$$P_{\beta,\eta}(Y, X^*, Z) = \sum_{x=0}^{1} P_\beta(Y \mid X = x, Z) P_\eta(X = x \mid Z) P(X^*, Z)$$

Case 3

Misclassification and total missingness

$$P_{\beta,\eta}(Y, X^*, Z) = \sum_{x=0}^{1} P_{\beta}(Y \mid X = x, Z) P_{\eta}(X = x \mid Z) P(X^*, Z)$$

drops out

Case 3

Misclassification and total missingness

$$P_{\beta,\eta}(Y, X^*, Z) = \sum_{x=0}^{1} P_{\beta}(Y \mid X = x, Z) P_{\eta}(X = x \mid Z) P(X^*, Z)$$

Case 3

Misclassification and total missingness

$$P_{\beta,\eta}(Y, X^*, Z) = \sum_{x=0}^{1} P_\beta(Y \mid X = x, Z) P_\eta(X = x \mid Z) P(X^*, Z)$$

Case 4

Misclassification and partial missingness

$$\mathscr{L}_N(\beta, \eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, Z_i)\}^{Q_i} \{P(X_i^*, Y_i, Z_i)\}^{1-Q_i}$$

Case 4

Misclassification and partial missingness

$$\mathscr{L}_N(\beta, \eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, Z_i)\}^{Q_i} \{P(X_i^*, Y_i, Z_i)\}^{1-Q_i}$$

from case 2

Case 4

Misclassification and partial missingness

$$\mathcal{L}_N(\beta, \eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, Z_i)\}^{Q_i} \{P(X_i^*, Y_i, Z_i)\}^{1-Q_i}$$

Case 4

Misclassification and partial missingness

$$\mathcal{L}_N(\beta, \eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, Z_i)\}^{Q_i} \{P(X_i^*, Y_i, Z_i)\}^{1-Q_i}$$

from case 3

Case 4

Misclassification and partial missingness

$$\mathscr{L}_N(\beta, \eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, Z_i)\}^{Q_i}\{P(X_i^*, Y_i, Z_i)\}^{1-Q_i}$$

Case 4

Misclassification and partial missingness

query indicators

$$\mathscr{L}_N(\beta, \eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, Z_i)\}^{Q_i} \{P(X_i^*, Y_i, Z_i)\}^{1-Q_i}$$

Case 4

Misclassification and partial missingness

product over all (independent) neighborhoods

$$\mathcal{L}_N(\beta, \eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, Z_i)\}^{Q_i} \{P(X_i^*, Y_i, Z_i)\}^{1-Q_i}$$

Case 4

Misclassification and partial missingness

$$\mathscr{L}_N(\beta, \eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, Z_i)\}^{Q_i} \{P(X_i^*, Y_i, Z_i)\}^{1-Q_i}$$

Maximizing the Likelihood

- We do not have an analytical form for the MLE, so we use **numerical methods**.
- We use the optim() function in R with the BFGS routine (Bonnans et al., 2006).
- We find the **minimum** of the **negative** log likelihood, which is **convex.**
- We **initialize** with the **complete case** estimates (Little and Rubin, 2002).
- We invert the numerical estimate of the **Hessian** matrix as the **standard error estimator**.

As N goes to infinity, the MLE ($\hat{\theta}_N$) is:

1. Consistent

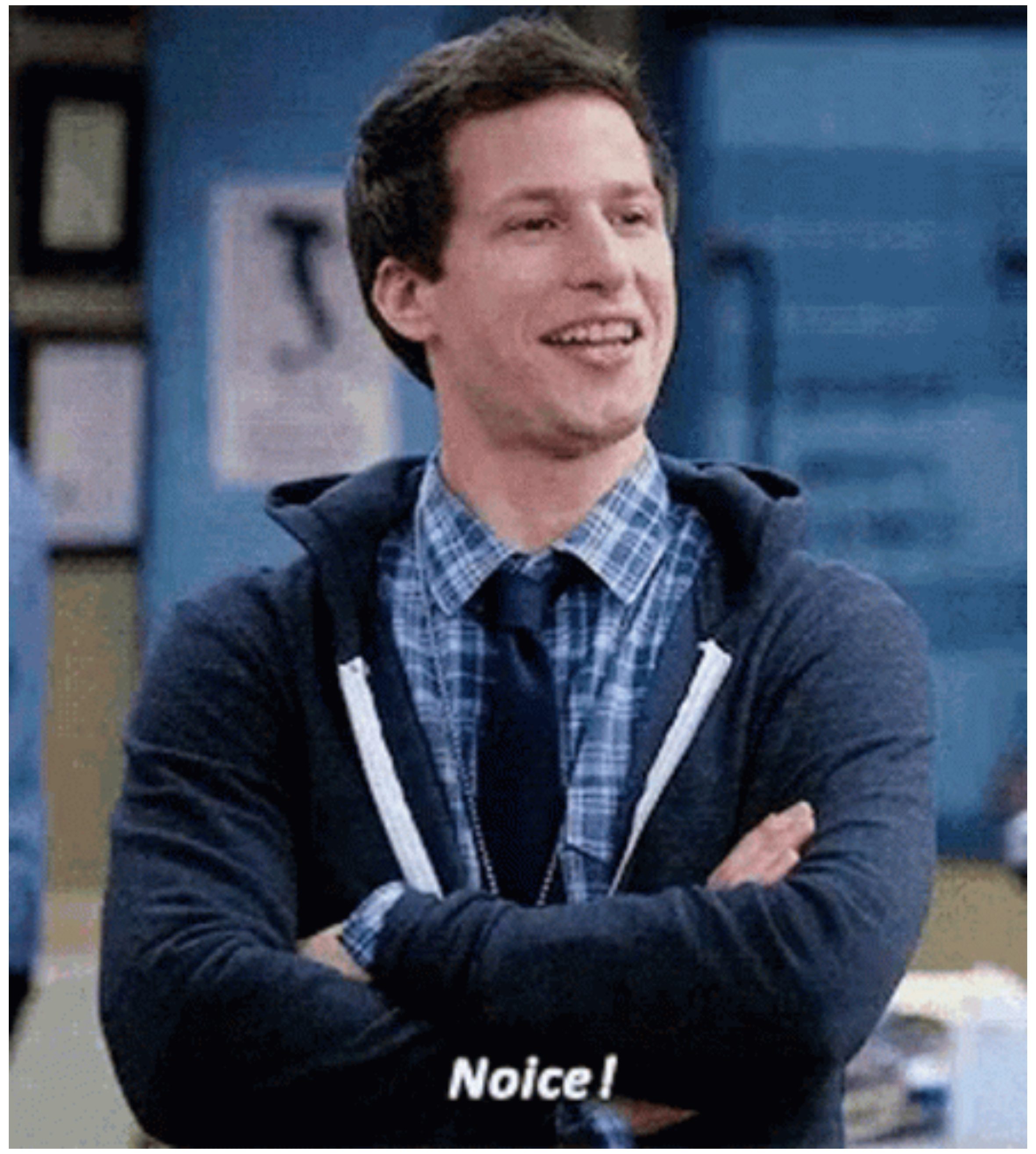
$$\hat{\boldsymbol{\theta}}_N \xrightarrow{\text{p}} \boldsymbol{\theta}$$

2. Asymptotically Normal

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}\right) \sim \text{Normal}(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}))$$

3. Asymptotically Efficient

$\mathcal{I}^{-1}(\boldsymbol{\theta})$ achieves the Cramer-Rao lower bound



As N goes to infinity, the MLE ($\hat{\theta}_N$) is:

1. Consistent

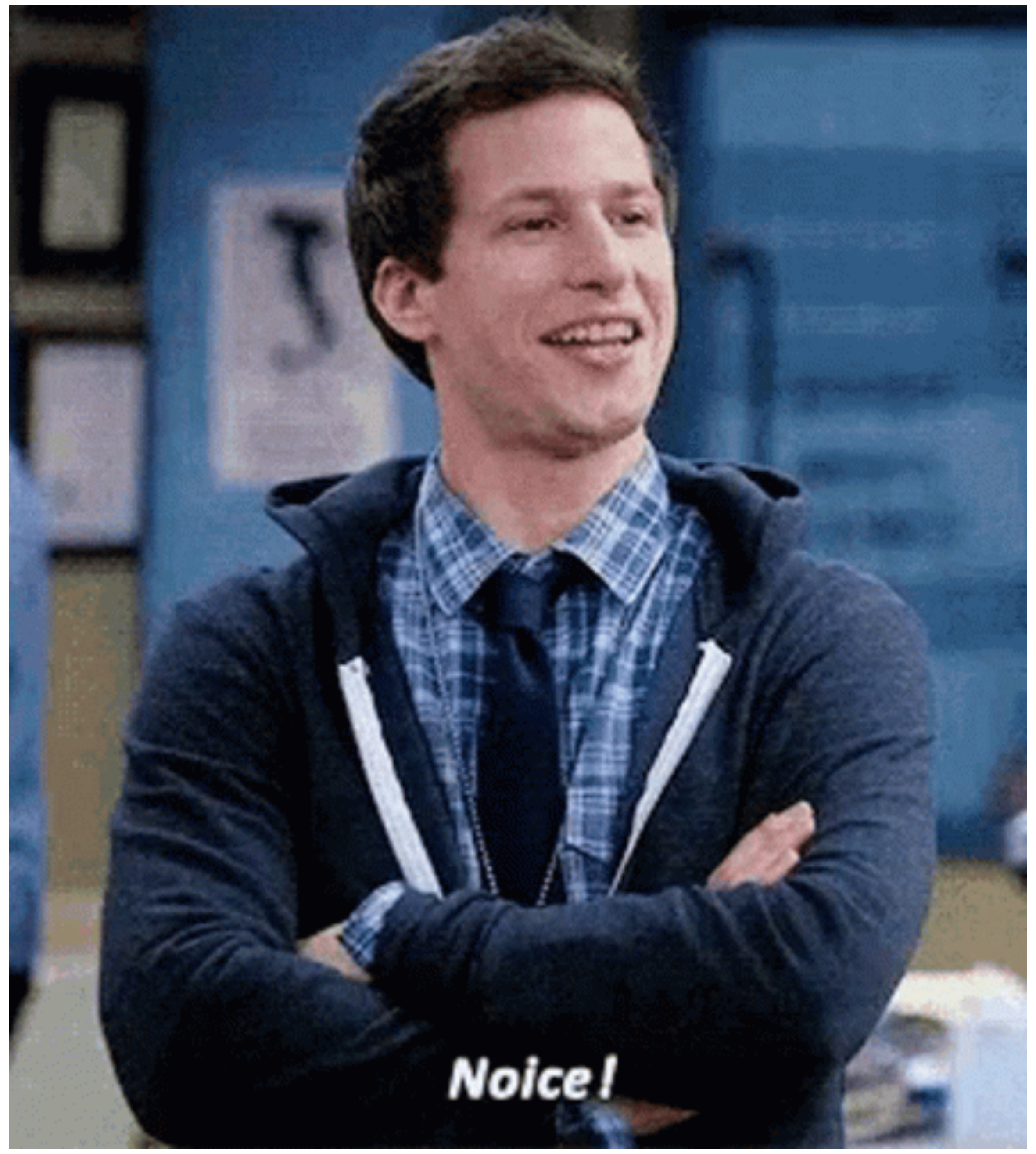
$$\hat{\boldsymbol{\theta}}_N \xrightarrow{\text{p}} \boldsymbol{\theta}$$

2. Asymptotically Normal

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}\right) \sim \text{Normal}(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}))$$

3. Asymptotically Efficient

$\mathcal{I}^{-1}(\boldsymbol{\theta})$ achieves the Cramer-Rao lower bound



POSSUM

```r
#devtools::install_github(repo = "sarahlotspeich/possum")

## Example
library(possum) #for the MLE
library(dplyr) #for data wrangling
set.seed(1031) #for reproducibility

#generate data
beta <- c(-2.2, 0.15) #governs Poisson outcome
eta <- c(-2.2, 4.4) #governs logistic error model
xstar = rbinom(n = 500, size = 1, prob = 0.5) #error-prone exposure
x = rbinom(n = 500, size = 1, #error-free exposure X|X*
           prob = 1 / (1 + exp(-(eta[1] + eta[2] * xstar))))
lambda = exp(beta[1] + beta[2] * x) #mean of Y|X
y = rpois(n = 500, lambda = lambda) #Poisson outcome with mean lambda
q = rbinom(n = 500, size = 1, prob = 0.75) #queried indicator
df <- data.frame(xstar, x, y, q) #construct complete dataset
df <- df |> mutate(x = ifelse(q == 1, x, NA)) #redact X for unqueried rows

#call MLE function
mle_output <- mlePossum(error_formula = x ~ xstar,
                        analysis_formula = y ~ x,
                        data = df)
```


```
> mle_output
$coefficients
                    Est        SE
(Intercept) -2.07087196 0.1595924
x            0.01085821 0.2378044

$convergence
[1] 0
```

Simulations

Setup

Simulation Studies

Setup

Simulation Studies

$$X^* \sim \text{Bernoulli}(0.496)$$

Setup

Simulation Studies

$$X^* \sim \text{Bernoulli}(0.496)$$

$$X \mid X^* \sim \text{Bernoulli}(\pi), \text{ where } \pi = \text{expit}(\eta_0 + \eta_1 X^*)$$

Setup

Simulation Studies

$$X^* \sim \text{Bernoulli}(0.496)$$

$$X \mid X^* \sim \text{Bernoulli}(\pi), \text{ where } \pi = \text{expit}(\eta_0 + \eta_1 X^*)$$

$$\eta_0 = -\log\left(\frac{1 - FPR}{FPR}\right) \quad \eta_1 = -\log\left(\frac{1 - TPR}{TPR}\right) - \eta_0$$

Setup

Simulation Studies

$$X^* \sim \text{Bernoulli}(0.496)$$

$$X \mid X^* \sim \text{Bernoulli}(\pi), \text{ where } \pi = \text{expit}(\eta_0 + \eta_1 X^*)$$

$$\eta_0 = -\log\left(\frac{1 - FPR}{FPR}\right) \quad \eta_1 = -\log\left(\frac{1 - TPR}{TPR}\right) - \eta_0$$

$$Y \sim \text{Poisson}(\lambda), \text{ where } \lambda = \exp(\beta_0 + \beta_1 X)$$

Setup

Simulation Studies

$$X^* \sim \text{Bernoulli}(0.496)$$

$$X \mid X^* \sim \text{Bernoulli}(\pi), \text{ where } \pi = \text{expit}(\eta_0 + \eta_1 X^*)$$

$$\eta_0 = -\log\left(\frac{1 - FPR}{FPR}\right) \quad \eta_1 = -\log\left(\frac{1 - TPR}{TPR}\right) - \eta_0$$

$$Y \sim \text{Poisson}(\lambda), \text{ where } \lambda = \exp(\beta_0 + \beta_1 X)$$

$$Q \sim \text{Bernoulli}(q)$$

Roadmap

Simulation Studies

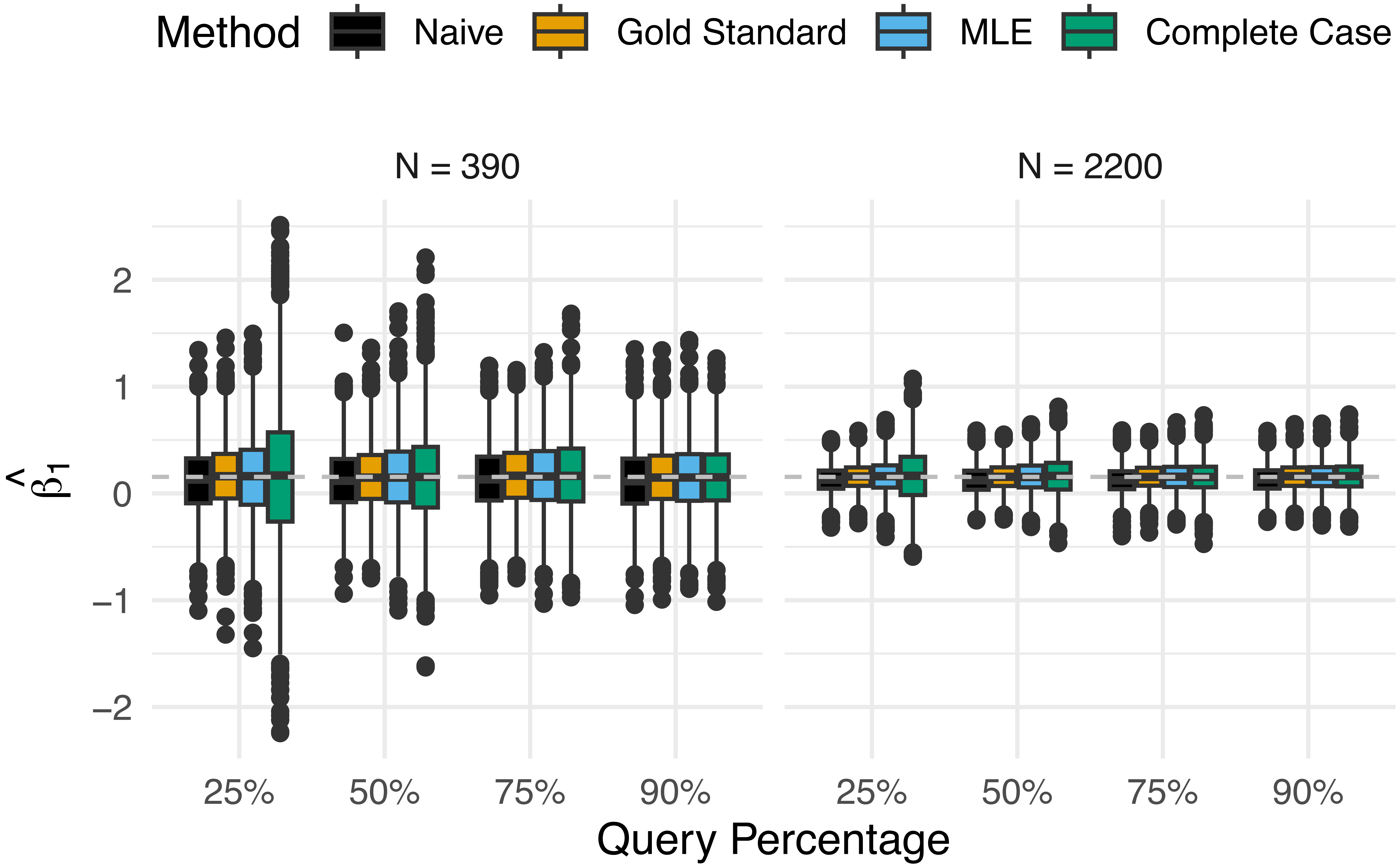
We **vary**:

- Sample size N
- Queried proportion q
- Error mechanism (FPR, TPR)
- Prevalence ratio $\exp(\beta_1)$
- Prevalence $\exp(\beta_0)$

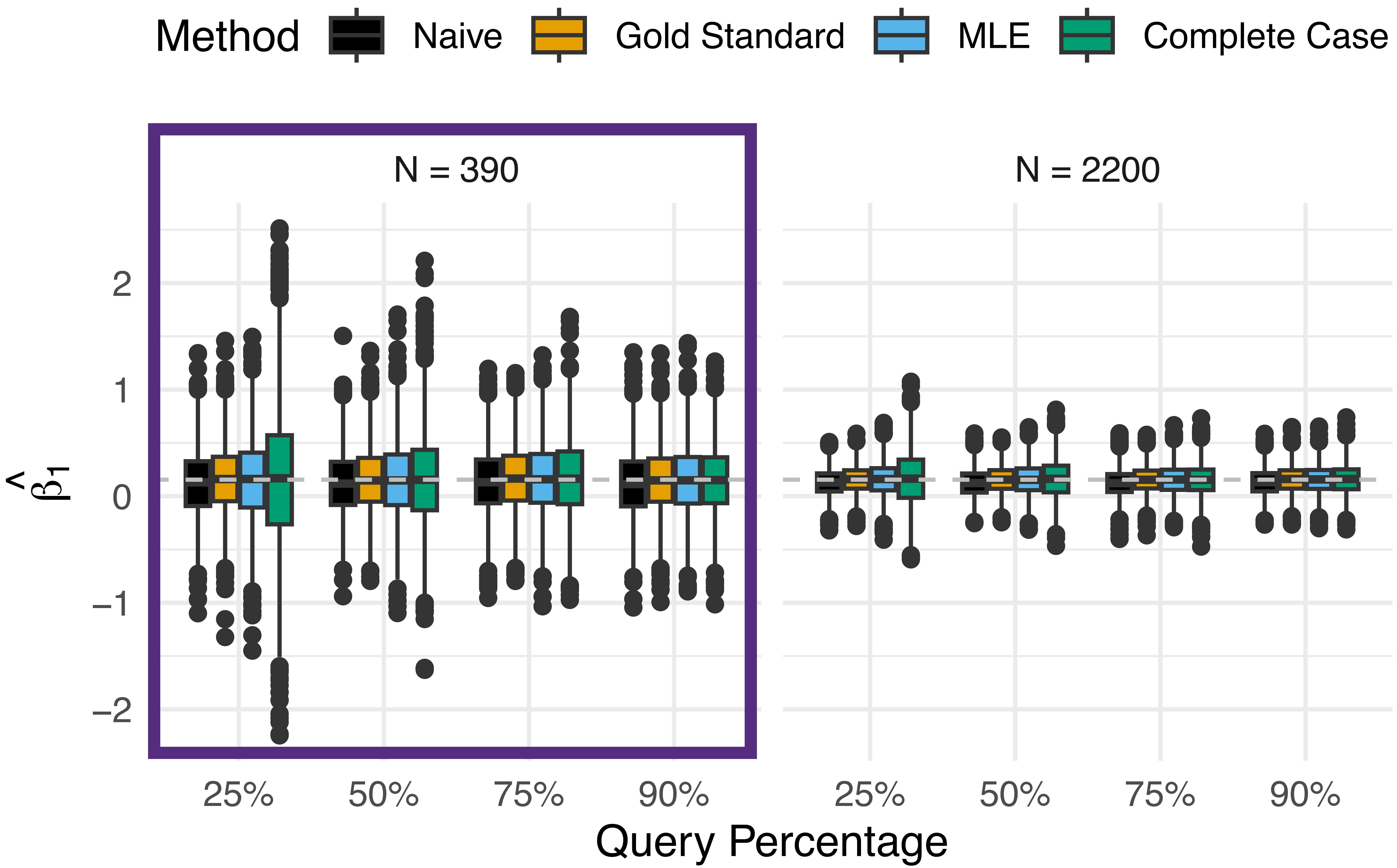
We **compare**:

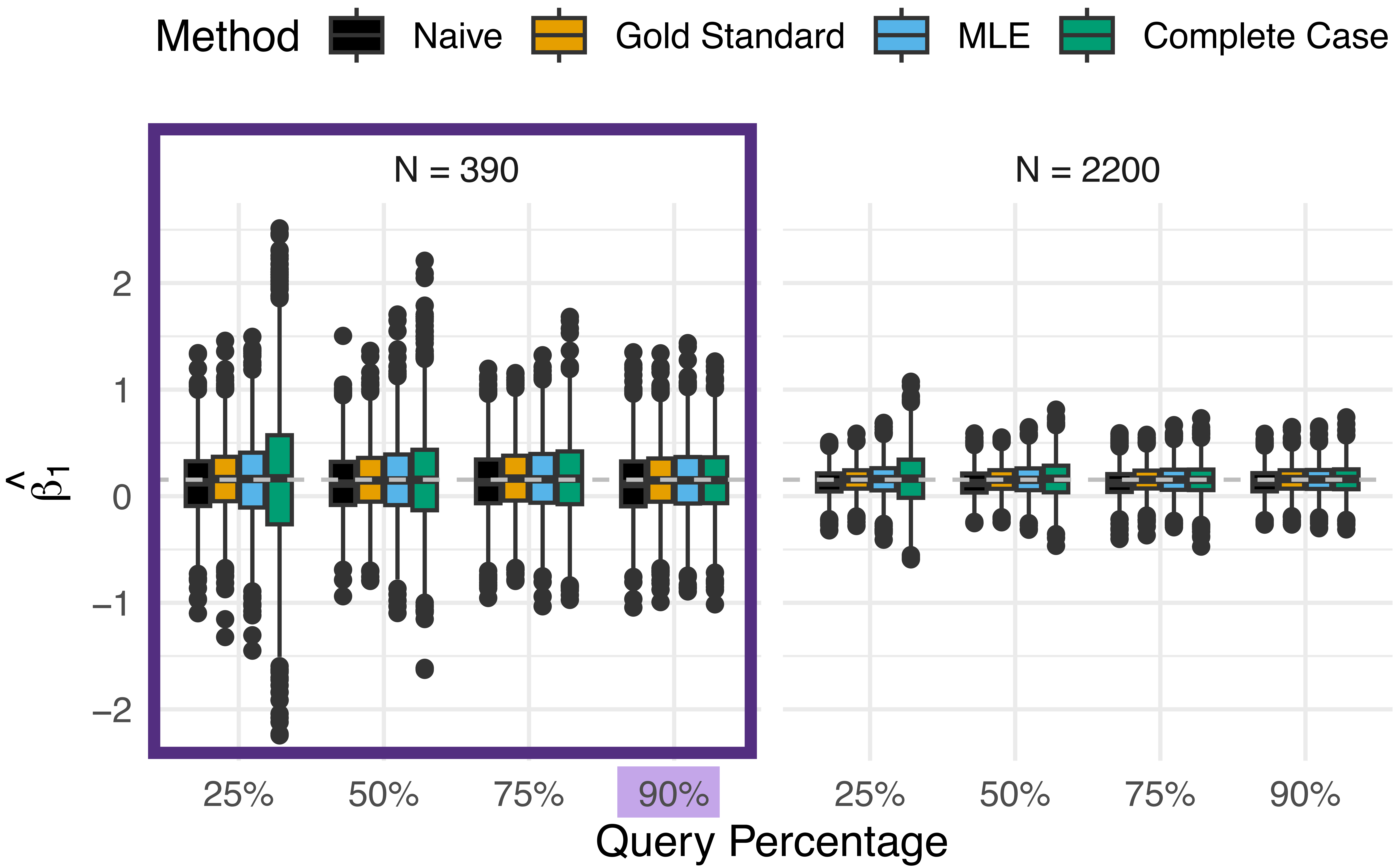
- Gold standard
- Complete case
- Naive model
- MLE

We **observe** the effect of interest $\hat{\beta}_1$ and the relative efficiency.

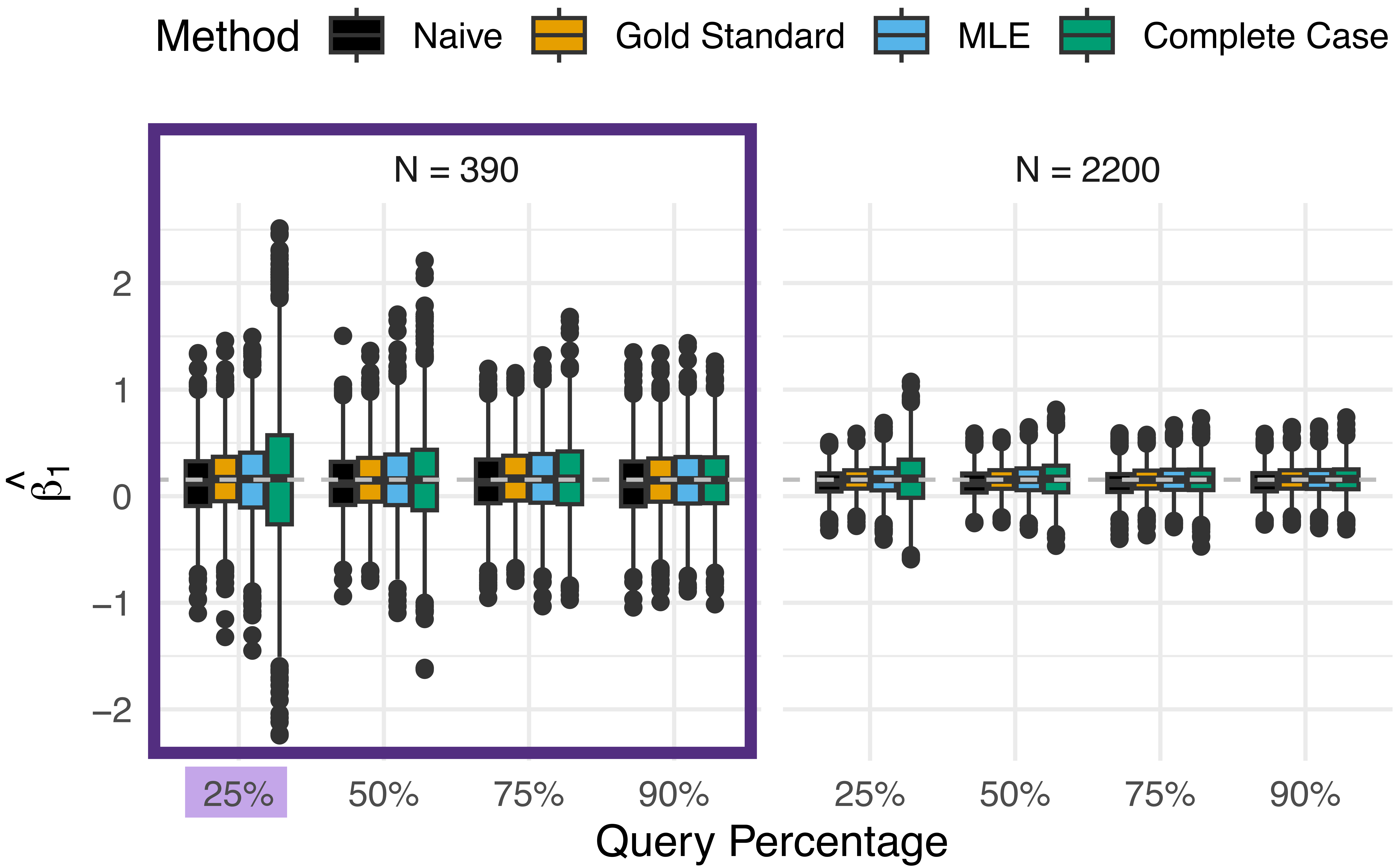


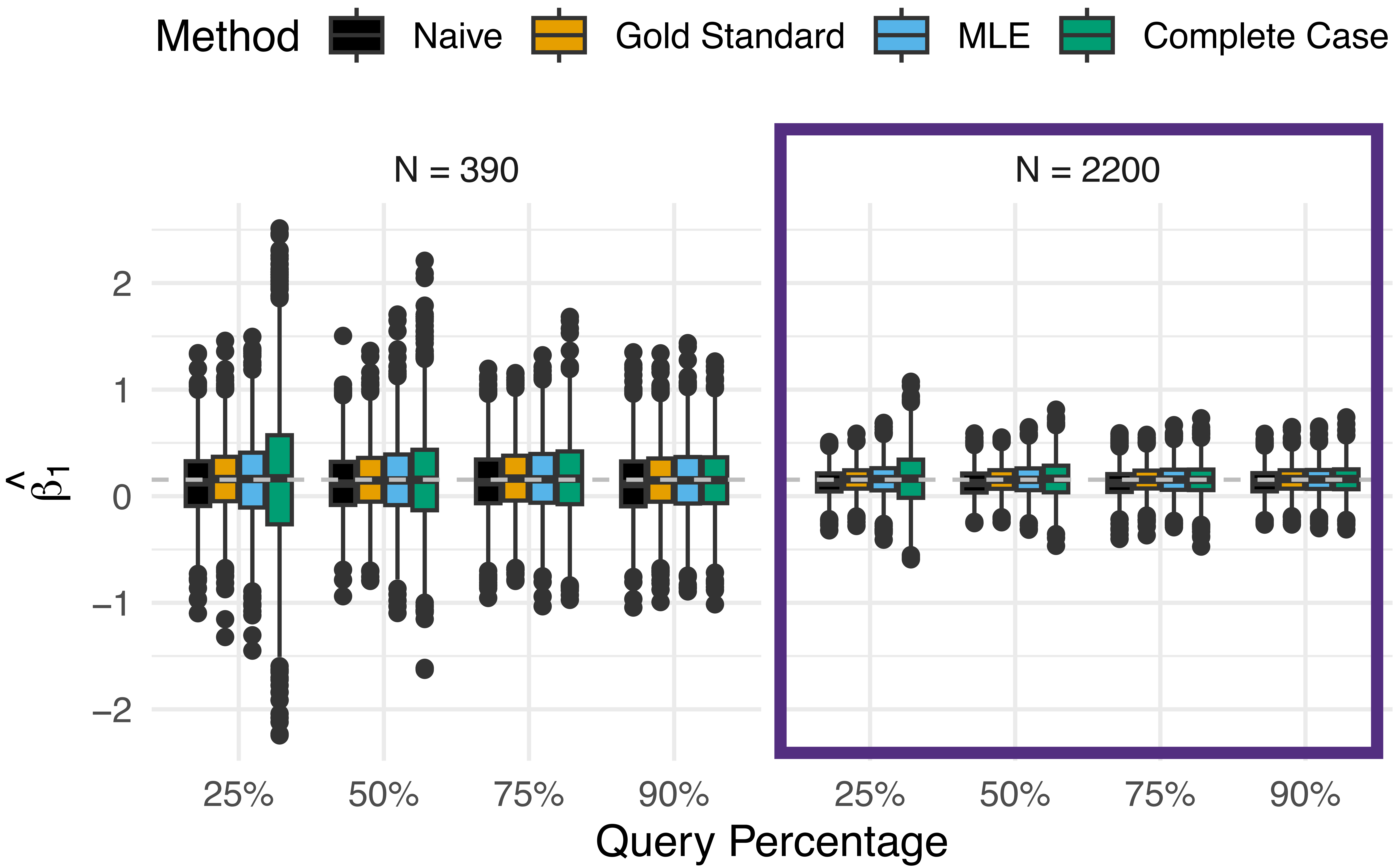
Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
2
1
0
−1
−2
$\hat{\beta}_1$
25%
50%
75%
90%
25%
50%
75%
90%
Query Percentage


Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
$\hat{\beta}_1$
2
1
0
−1
−2
25%
50%
75%
90%
25%
50%
75%
90%
Query Percentage

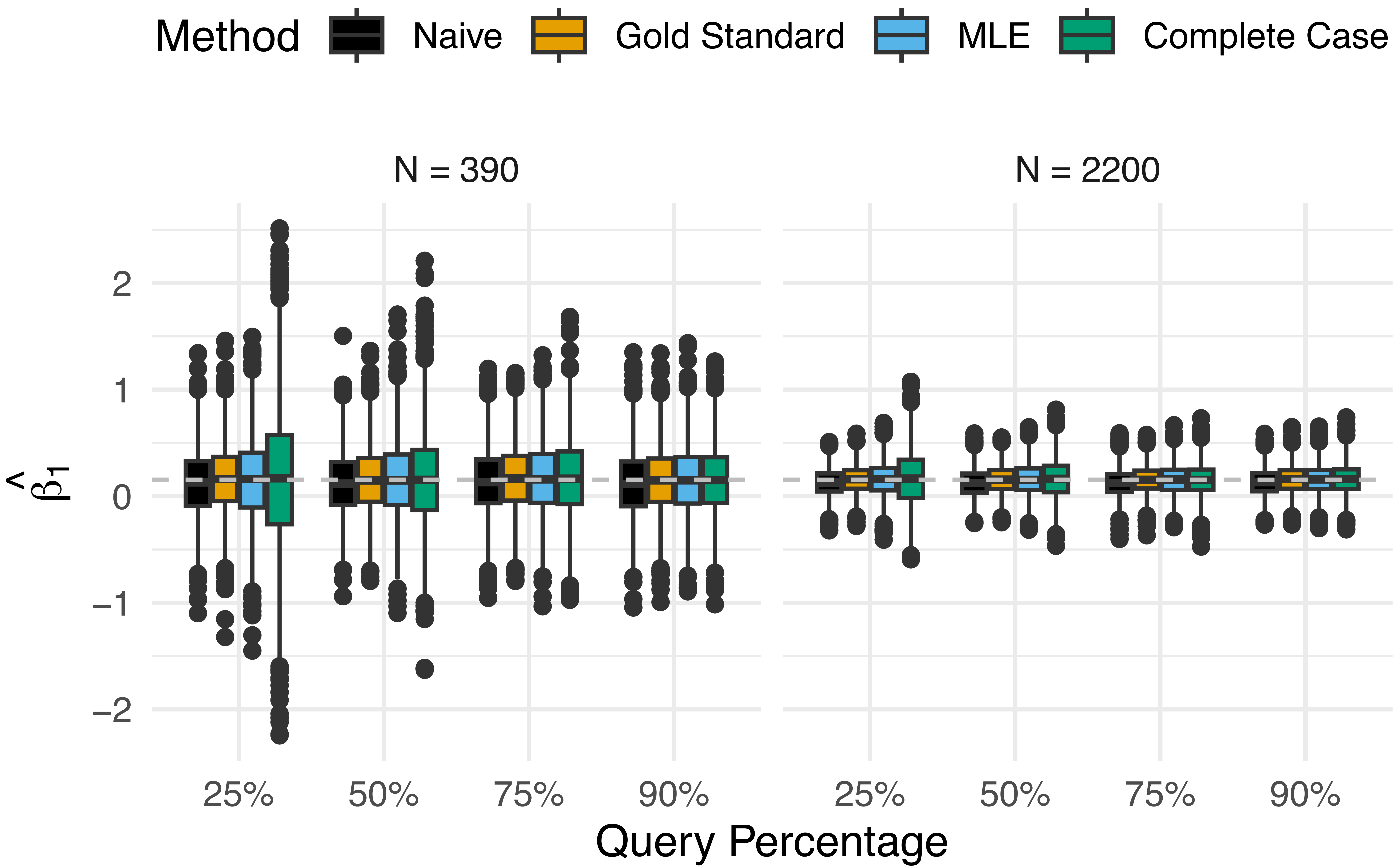


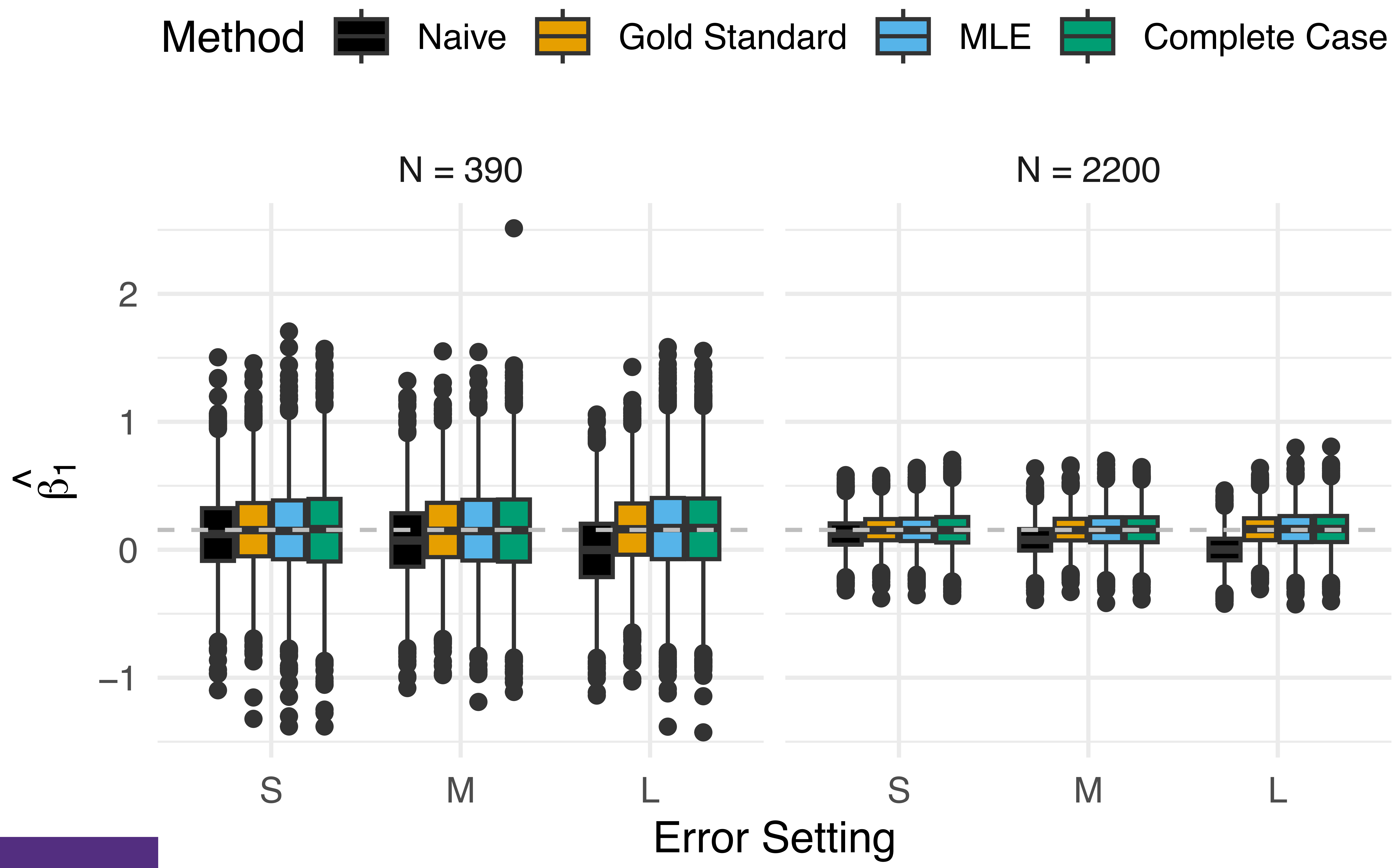
Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
$\hat{\beta}_1$
2
1
0
−1
−2
25%
50%
75%
90%
25%
50%
75%
90%
Query Percentage


Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
2
1
0
−1
−2
$\hat{\beta}_1$
25%
50%
75%
90%
25%
50%
75%
90%
Query Percentage

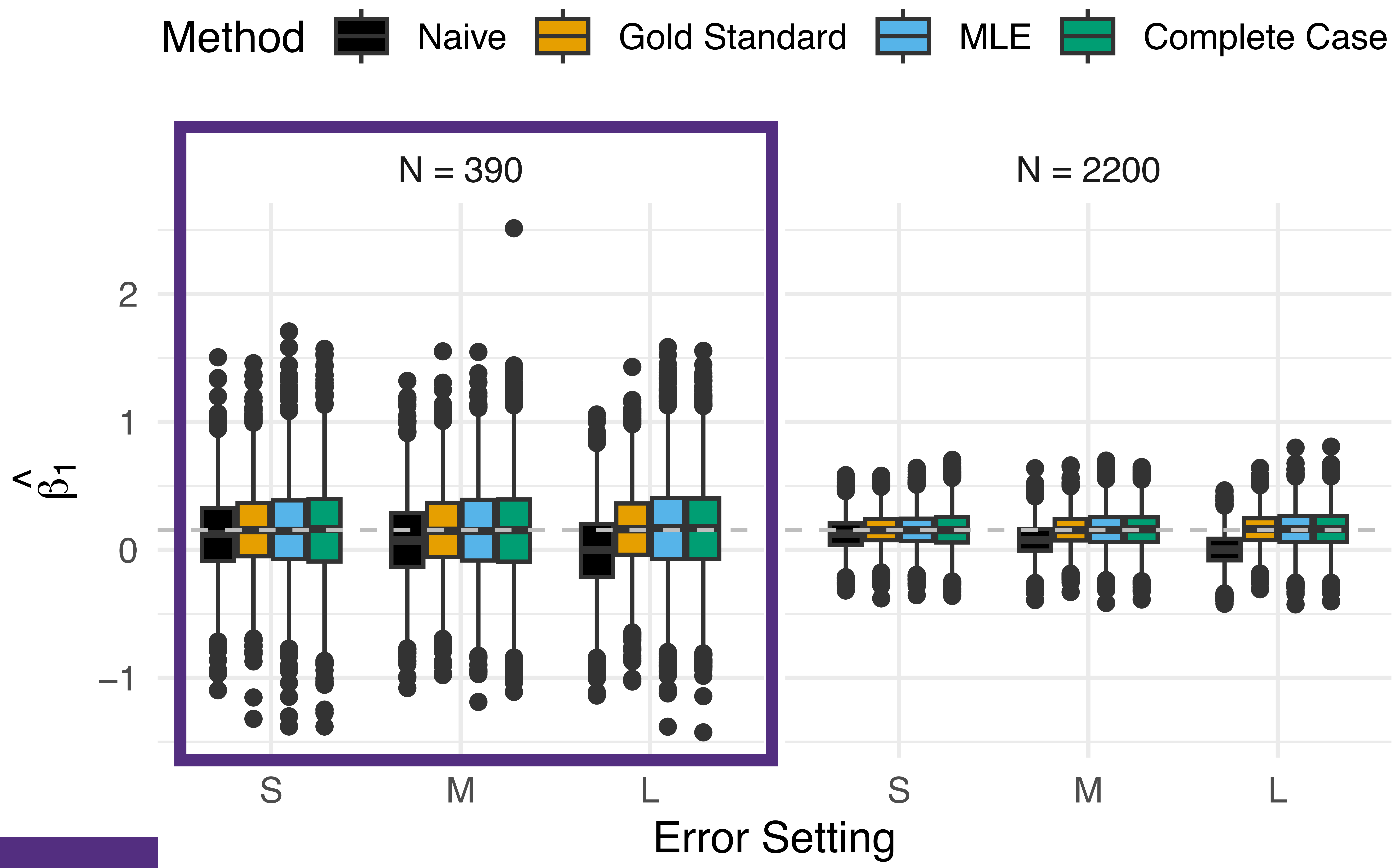


Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
2
1
0
−1
−2
$\hat{\beta}_1$
25%
50%
75%
90%
25%
50%
75%
90%
Query Percentage

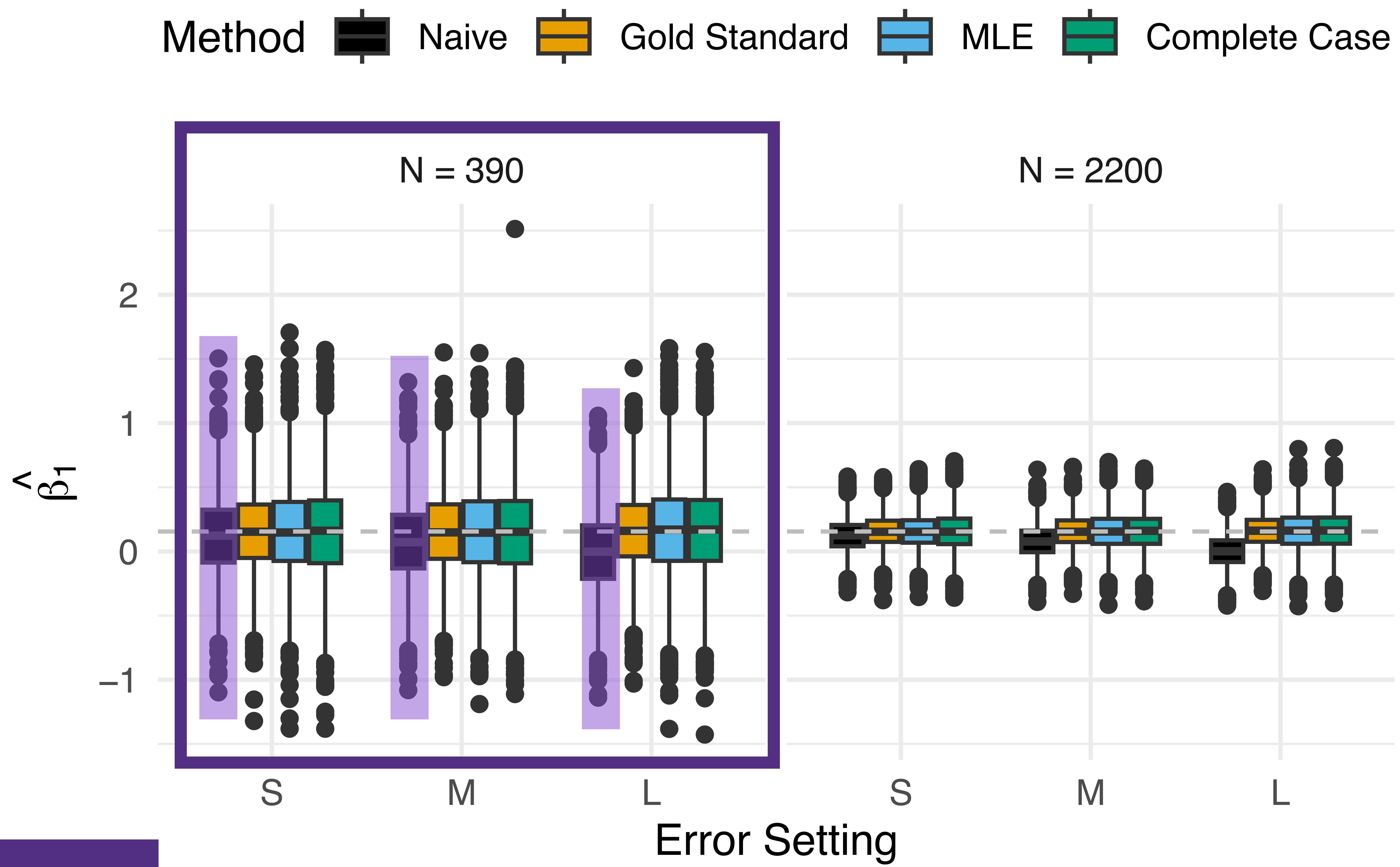

Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
2
1
0
−1
−2
$\hat{\beta}_1$
25%
50%
75%
90%
25%
50%
75%
90%
Query Percentage



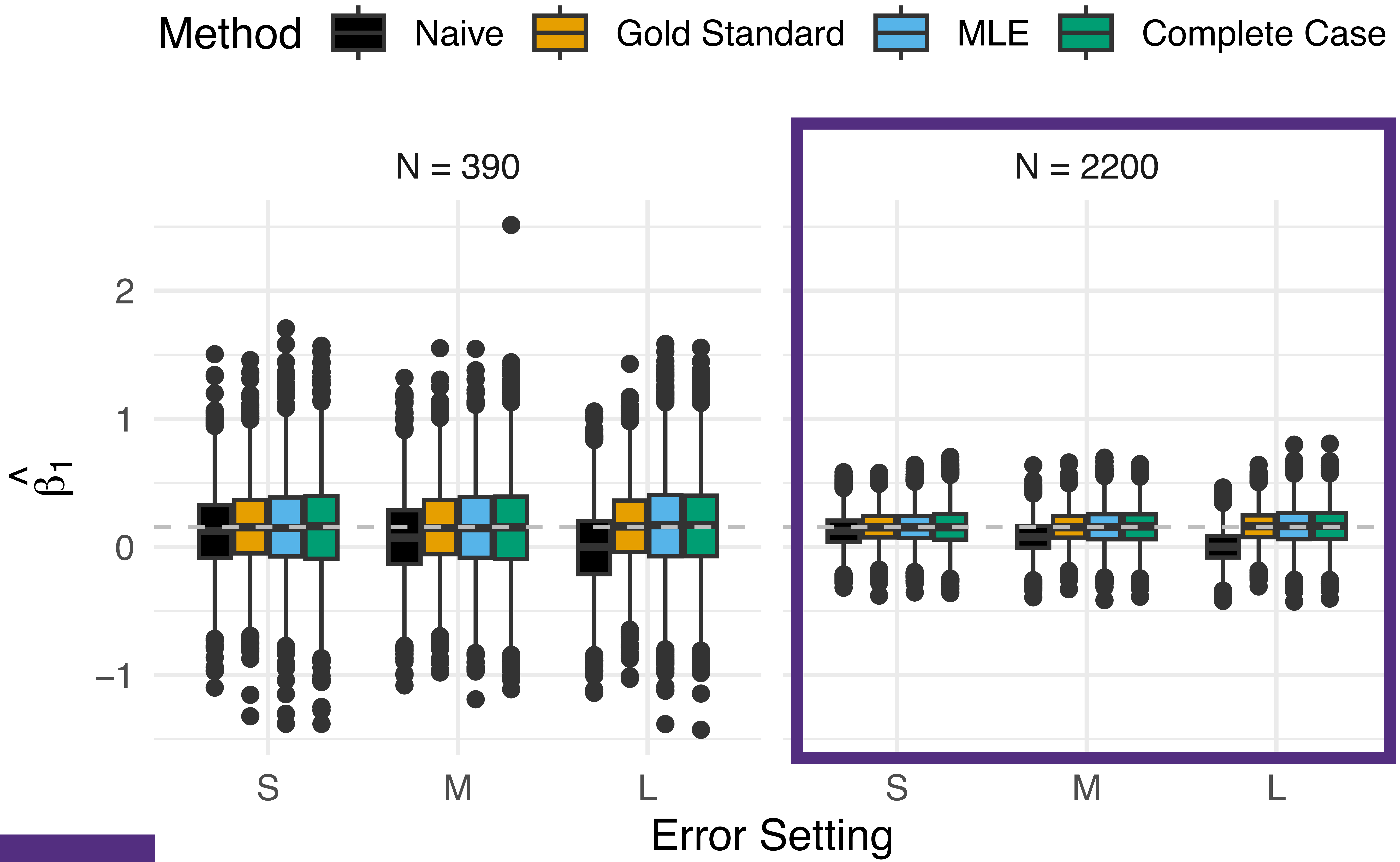
S: (10% FPR, 90% TPR)
M: (25% FPR, 75% TPR)
L: (50% FPR, 50% TPR)



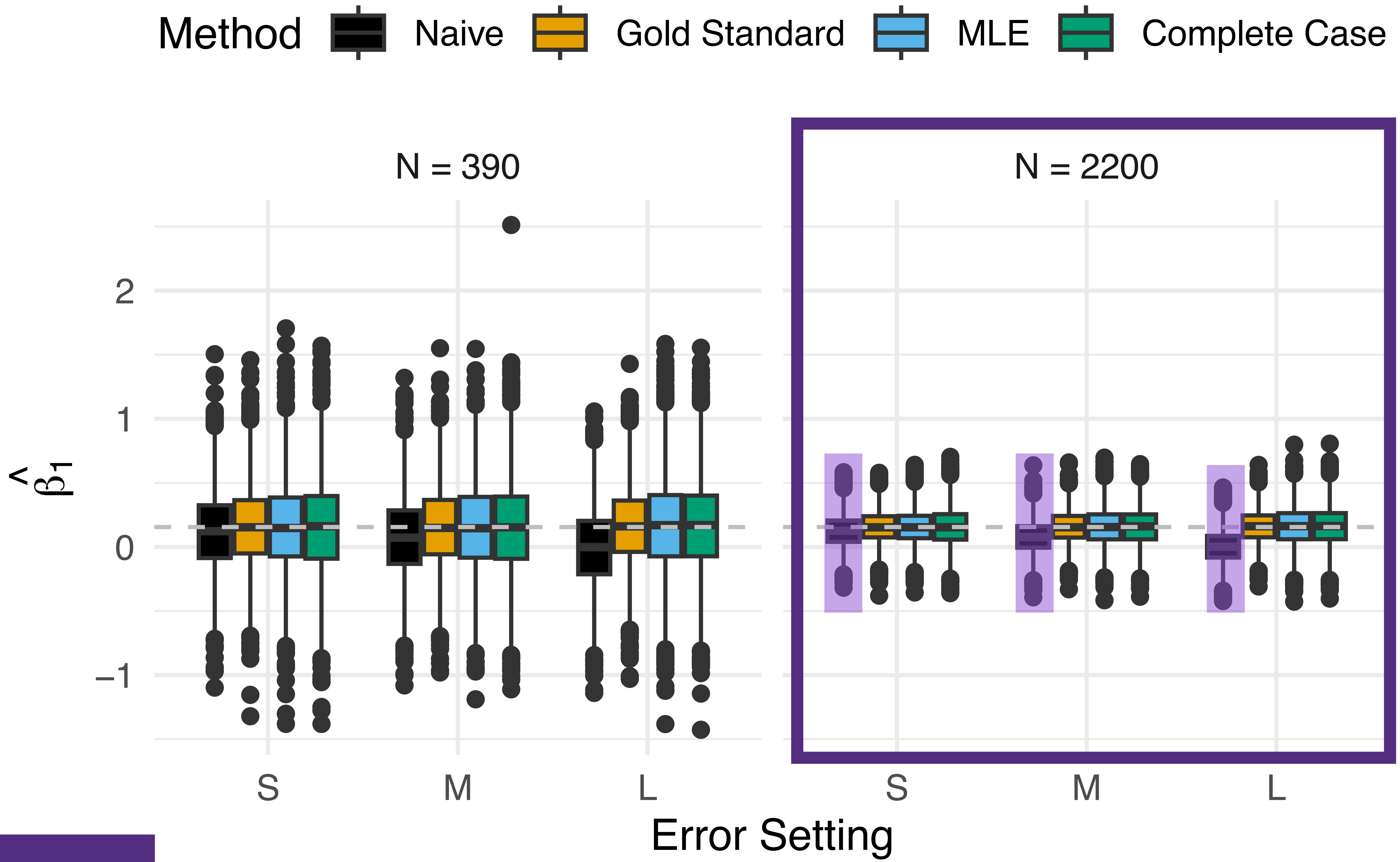
S: (10% FPR, 90% TPR)
M: (25% FPR, 75% TPR)
L: (50% FPR, 50% TPR)



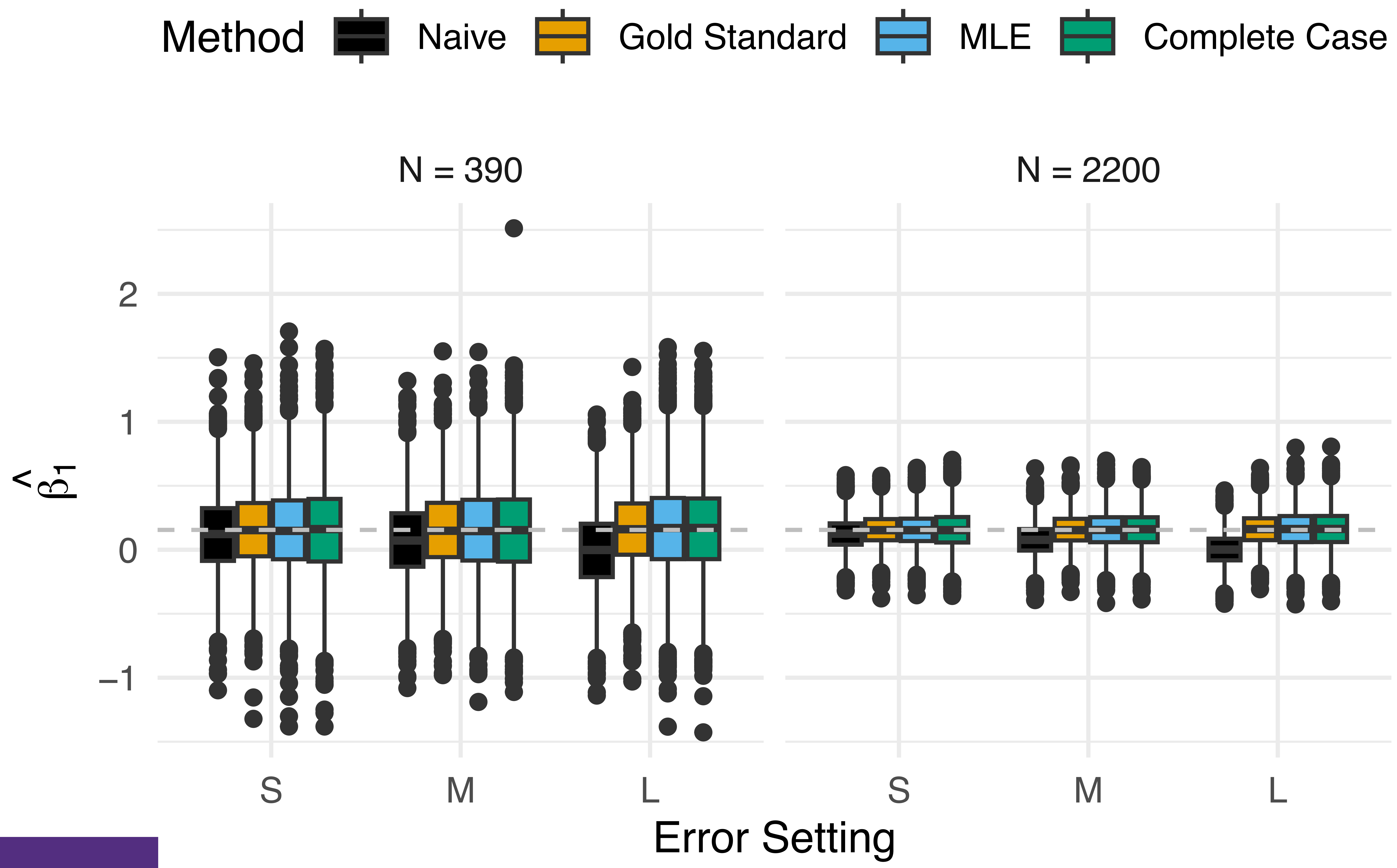
S: (10% FPR, 90% TPR)
M: (25% FPR, 75% TPR)
L: (50% FPR, 50% TPR)



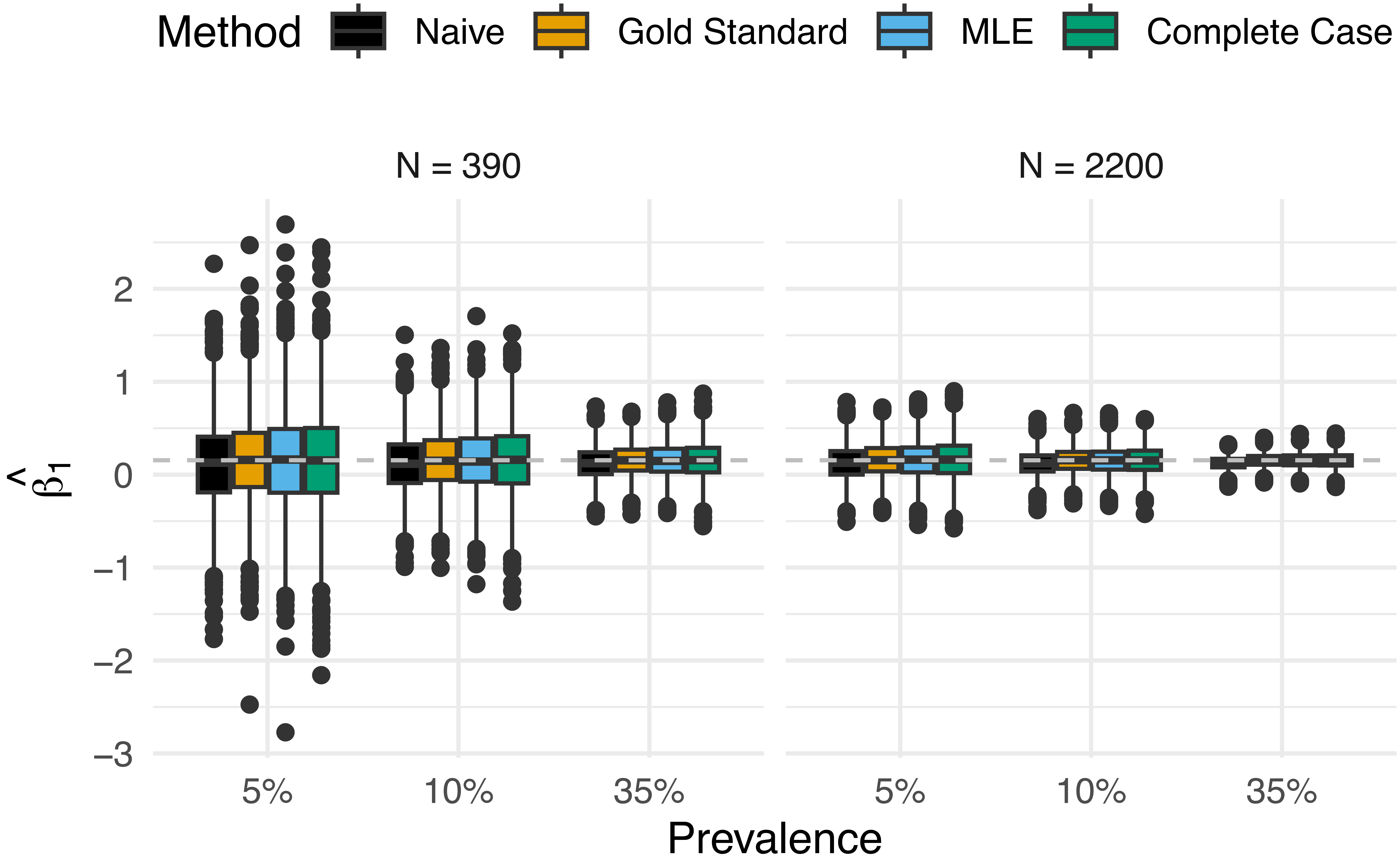
S: (10% FPR, 90% TPR)
M: (25% FPR, 75% TPR)
L: (50% FPR, 50% TPR)

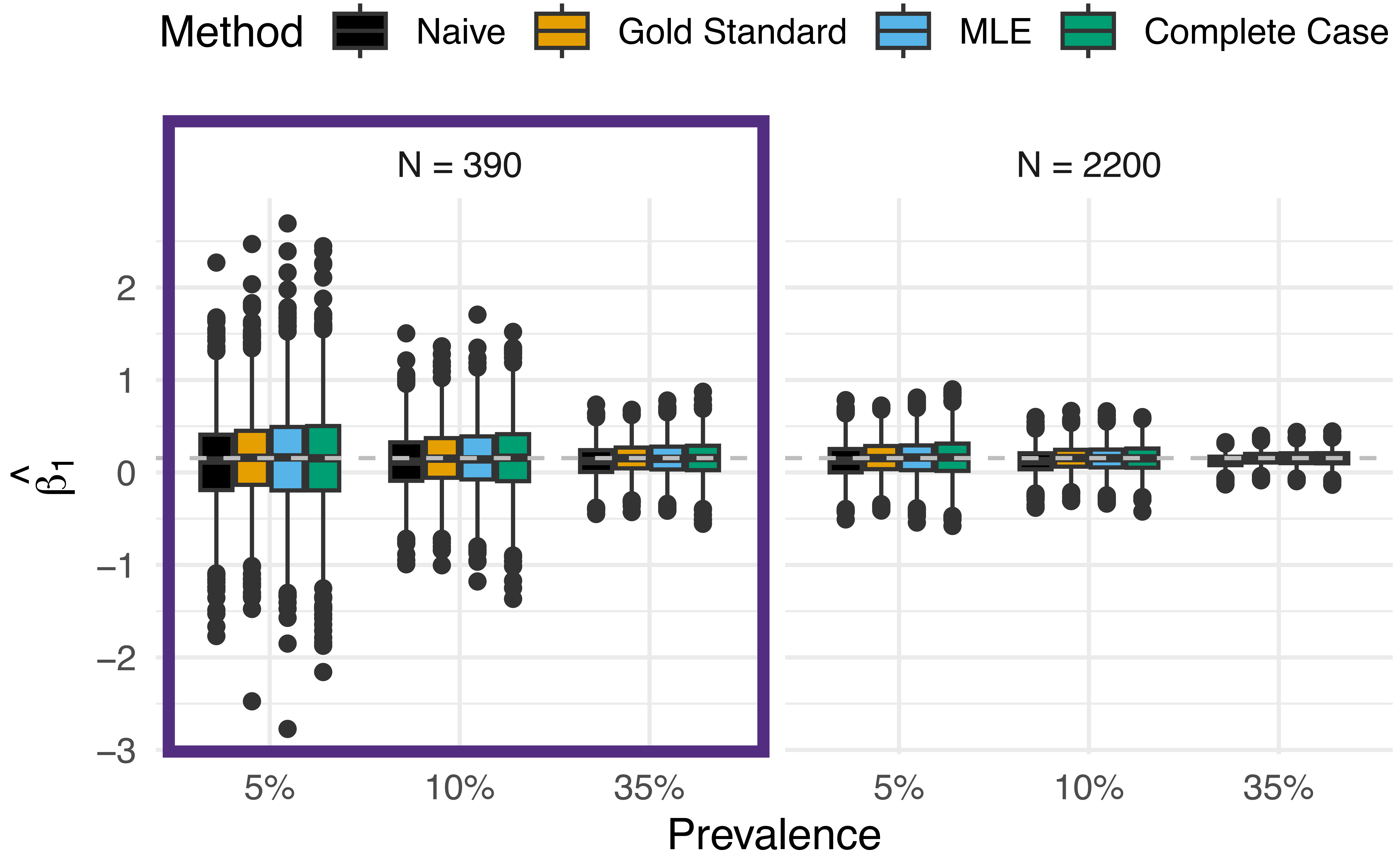


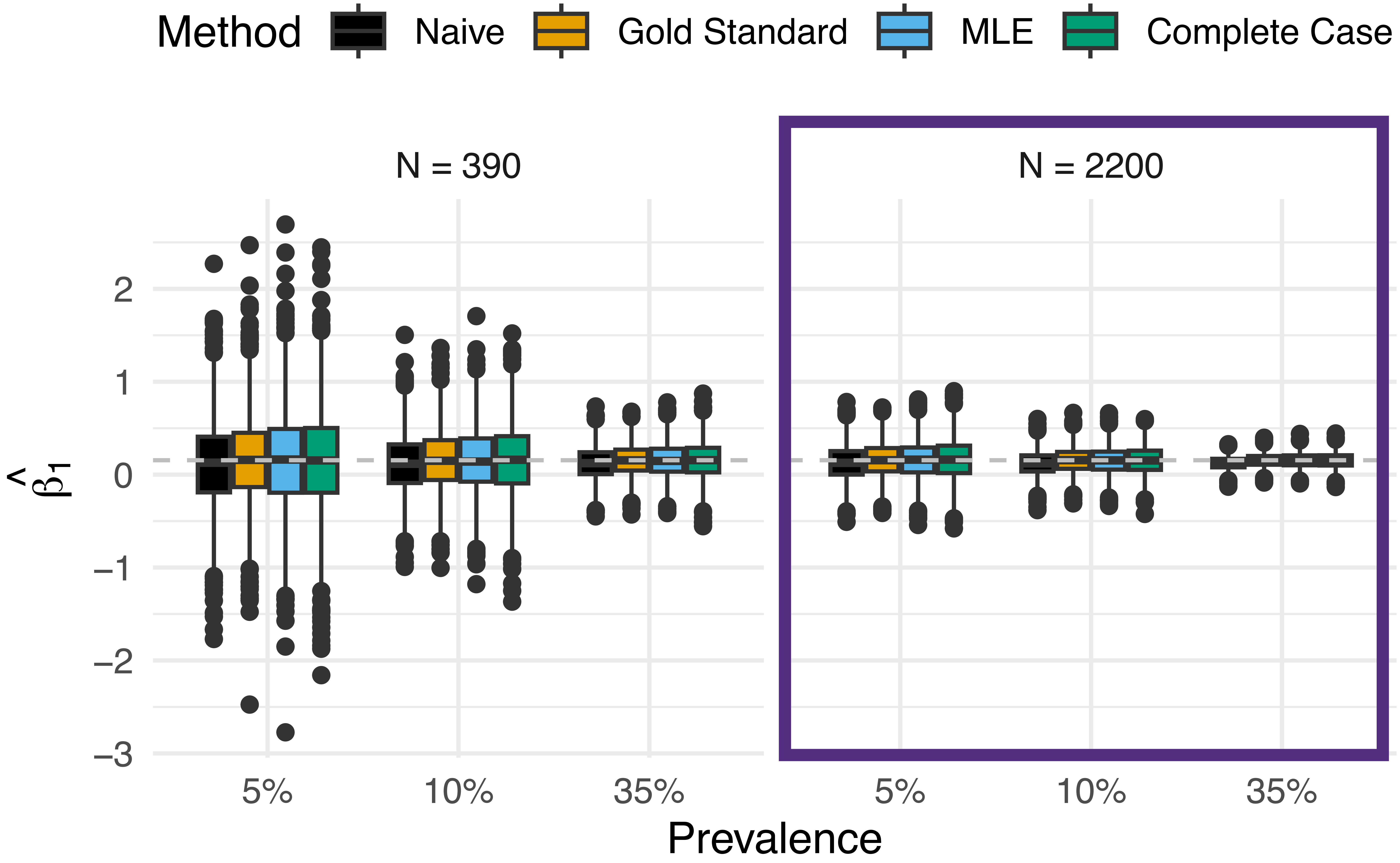
S: (10% FPR, 90% TPR)
M: (25% FPR, 75% TPR)
L: (50% FPR, 50% TPR)

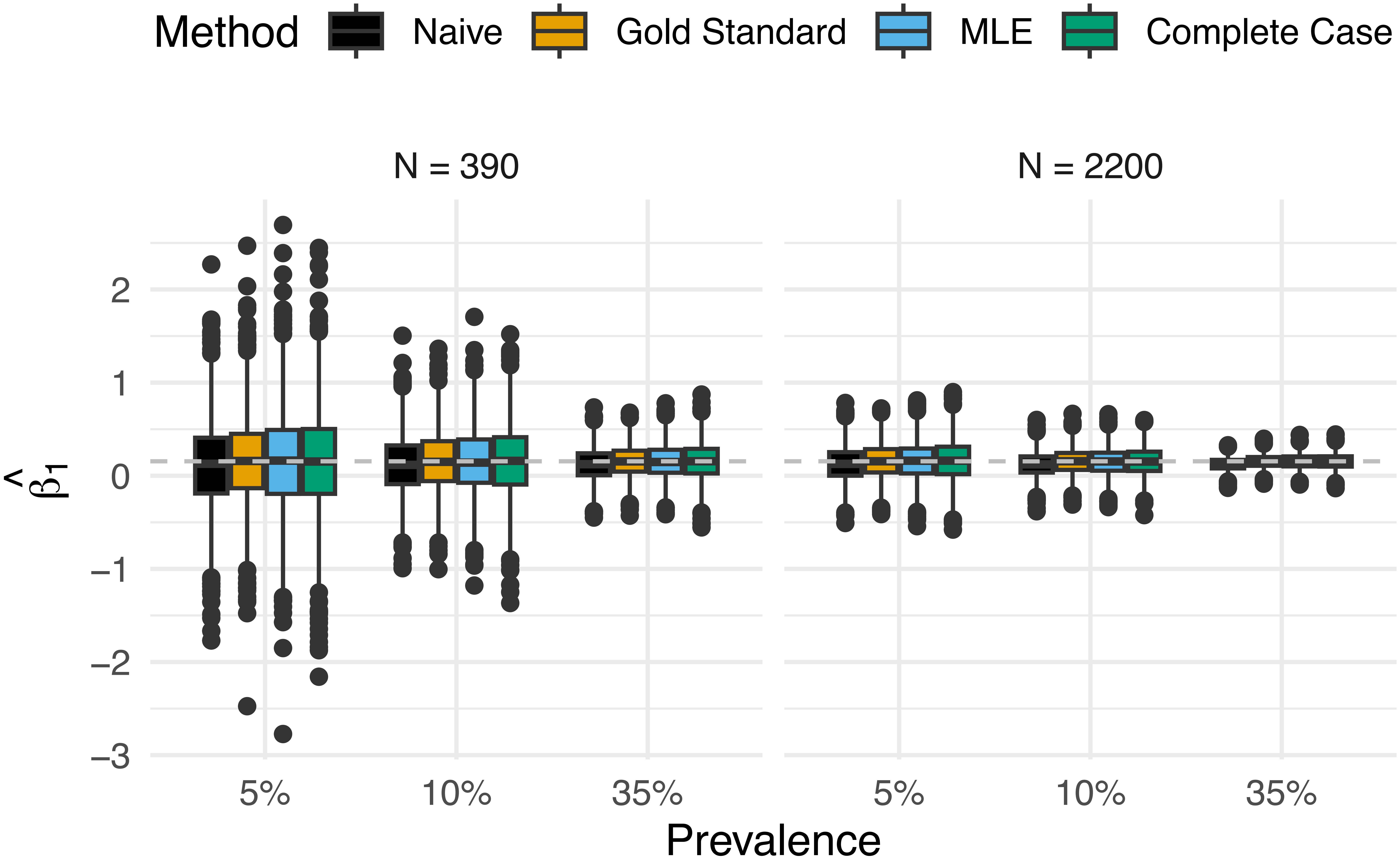


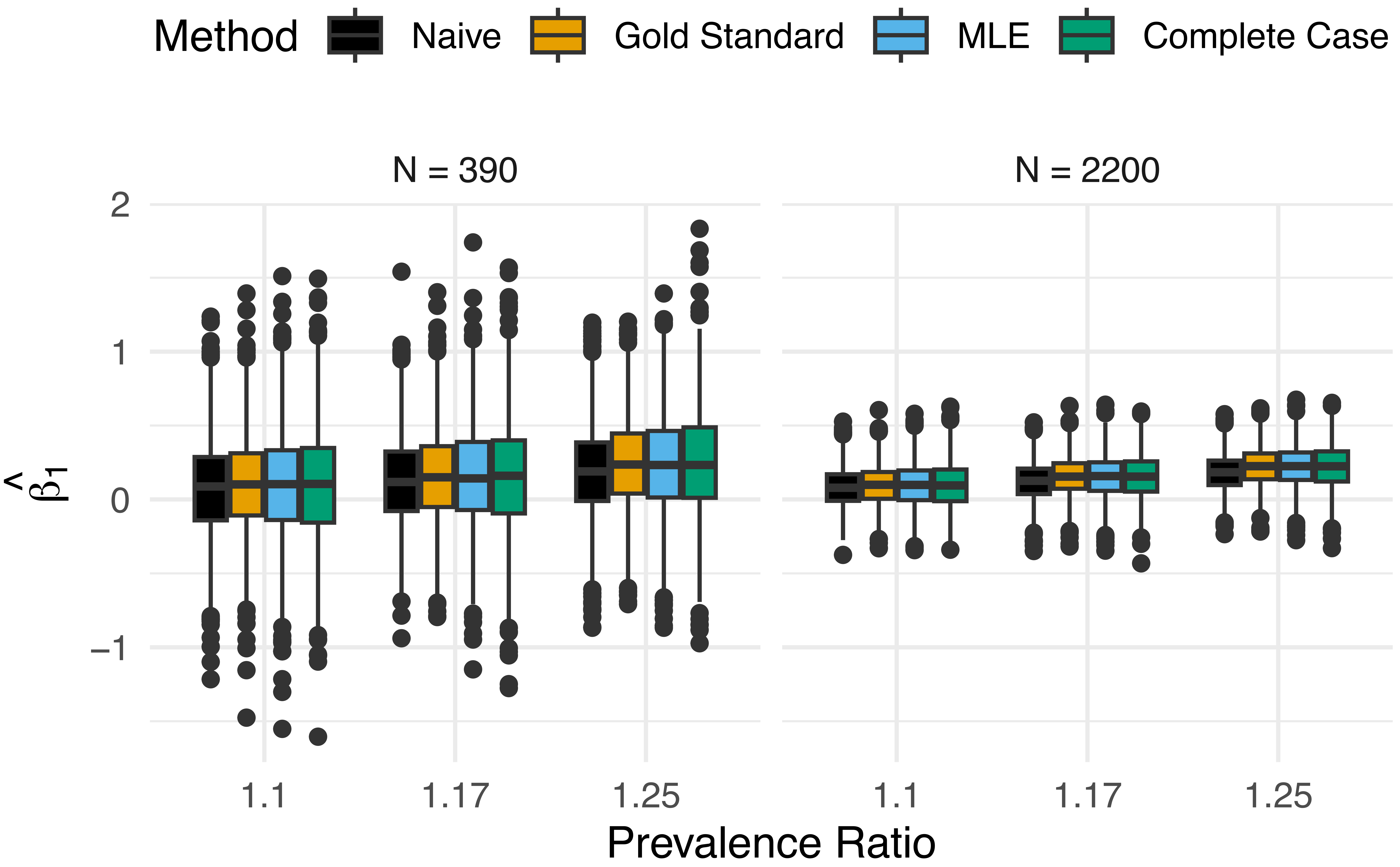
S: (10% FPR, 90% TPR)
M: (25% FPR, 75% TPR)
L: (50% FPR, 50% TPR)

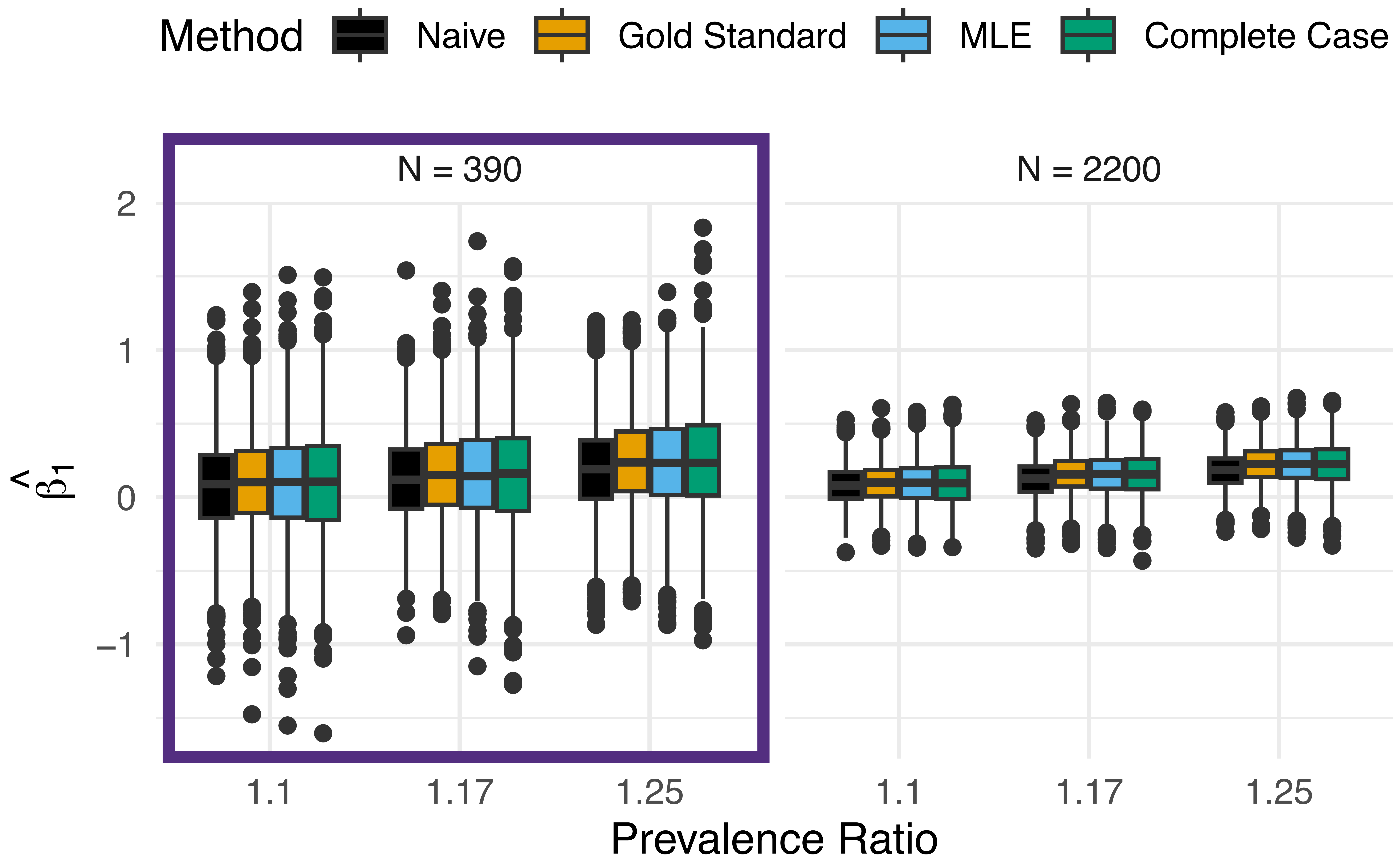

Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
$\hat{\beta}_1$
2
1
0
−1
−2
−3
5%
10%
35%
5%
10%
35%
Prevalence

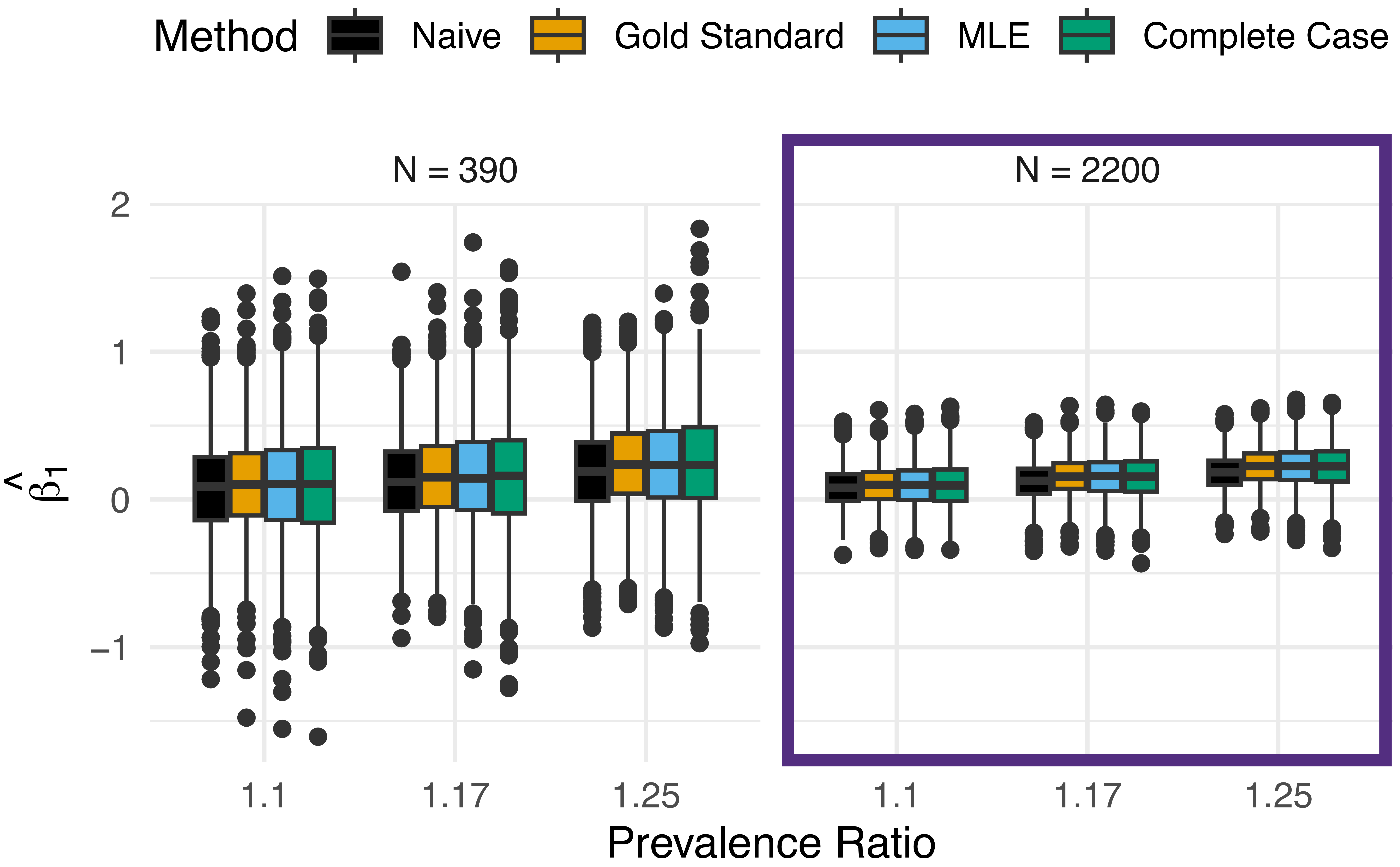

Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
2
1
0
−1
−2
−3
$\hat{\beta}_1$
5%
10%
35%
5%
10%
35%
Prevalence


Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
2
1
0
−1
−2
−3
$\hat{\beta}_1$
5%
10%
35%
5%
10%
35%
Prevalence

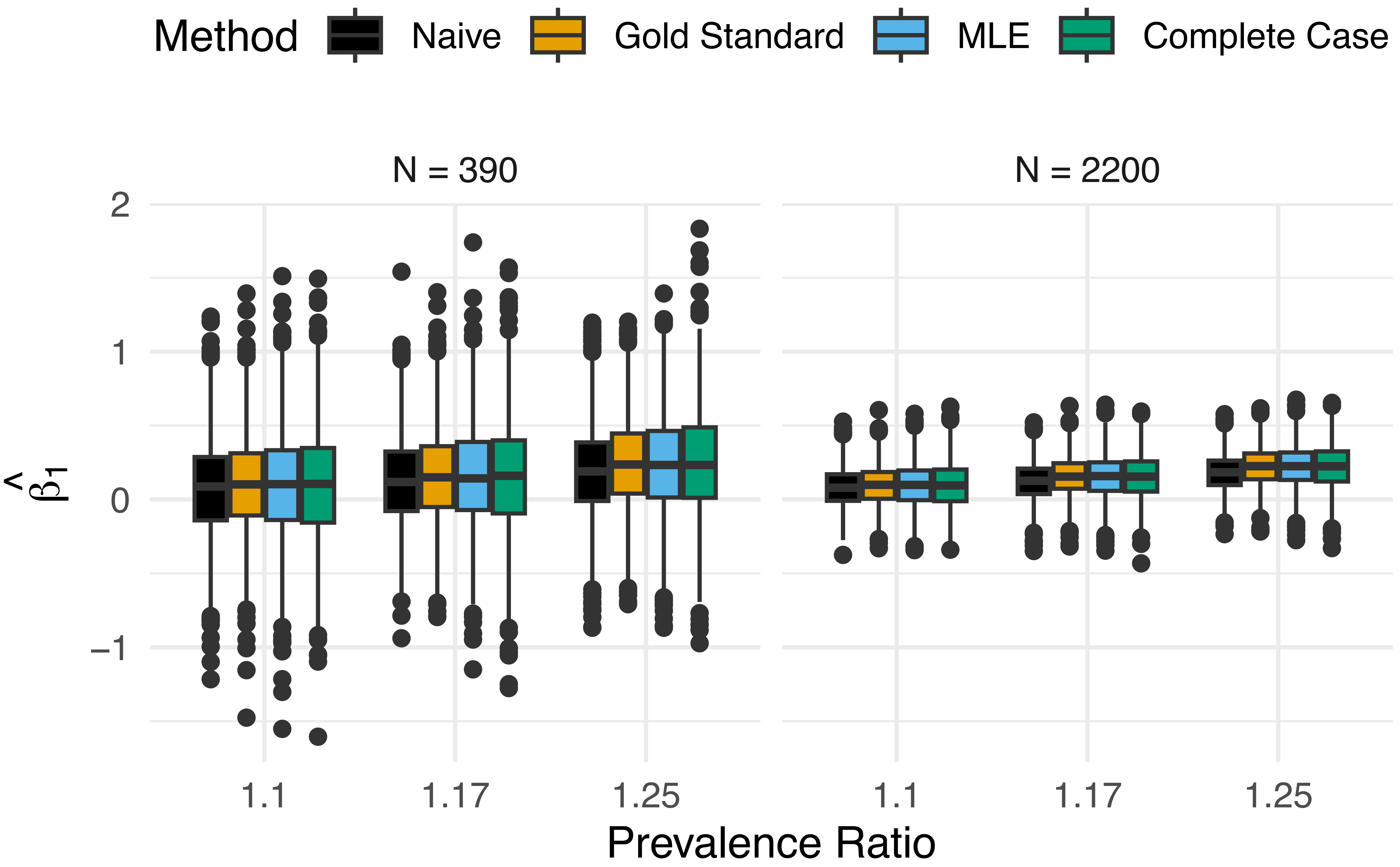

Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
$\hat{\beta}_1$
2
1
0
−1
−2
−3
5%
10%
35%
5%
10%
35%
Prevalence


Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
2
1
0
−1
$\hat{\beta}_1$
1.1
1.17
1.25
1.1
1.17
1.25
Prevalence Ratio


Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
2
1
0
−1
$\hat{\beta}_1$
1.1
1.17
1.25
1.1
1.17
1.25
Prevalence Ratio



Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
2
1
0
−1
$\hat{\beta}_1$
1.1
1.17
1.25
1.1
1.17
1.25
Prevalence Ratio



Method
Naive
Gold Standard
MLE
Complete Case
N = 390
N = 2200
2
1
0
−1
$\hat{\beta}_1$
1.1
1.17
1.25
1.1
1.17
1.25
Prevalence Ratio

Takeaways

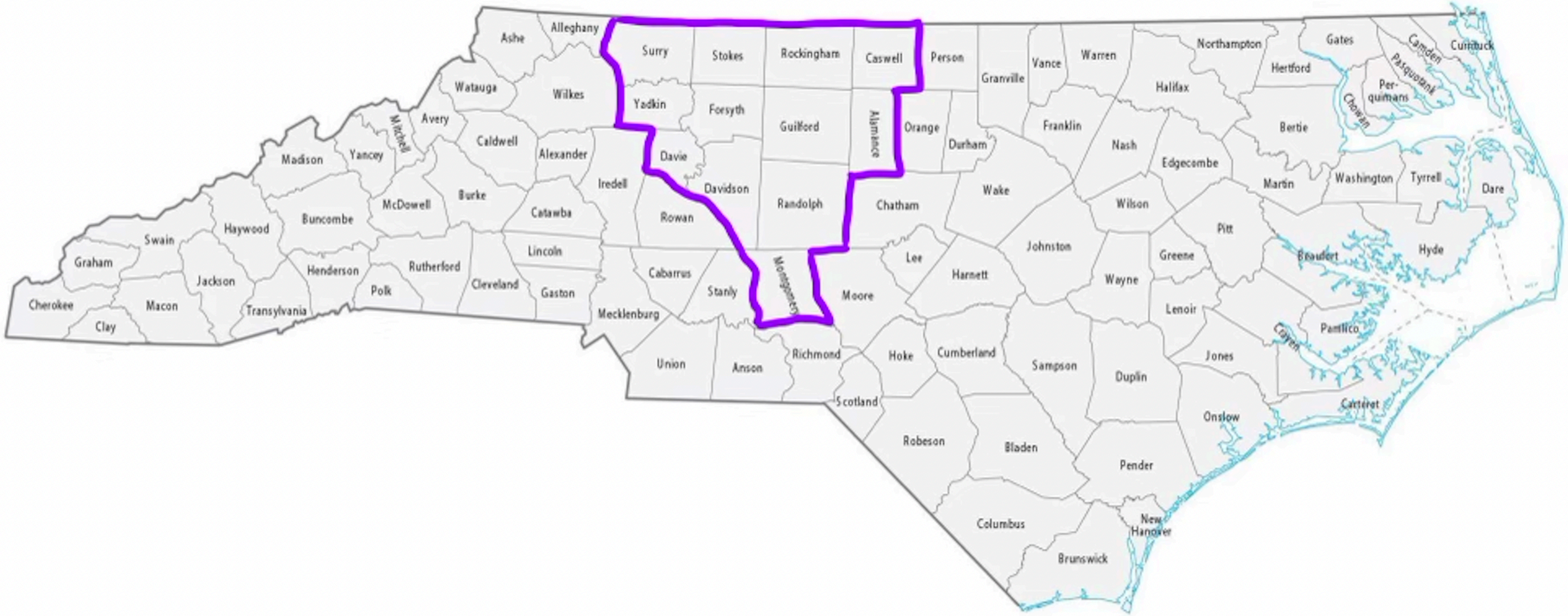
Simulation Studies

- Across all four query settings, the MLE remains **fairly unbiased**.
- As we vary the size of the queried sample, the MLE recovers up to 91% of the **efficiency** of the gold standard model and beats the complete case model in every case.
- As we introduce more error into the input data, the MLE remains **fairly unbiased**.
- As we vary the error, the MLE recovers between 70 and 83% of the **efficiency** of the gold standard model.

Case Study:
Diabetes in the Piedmont Triad

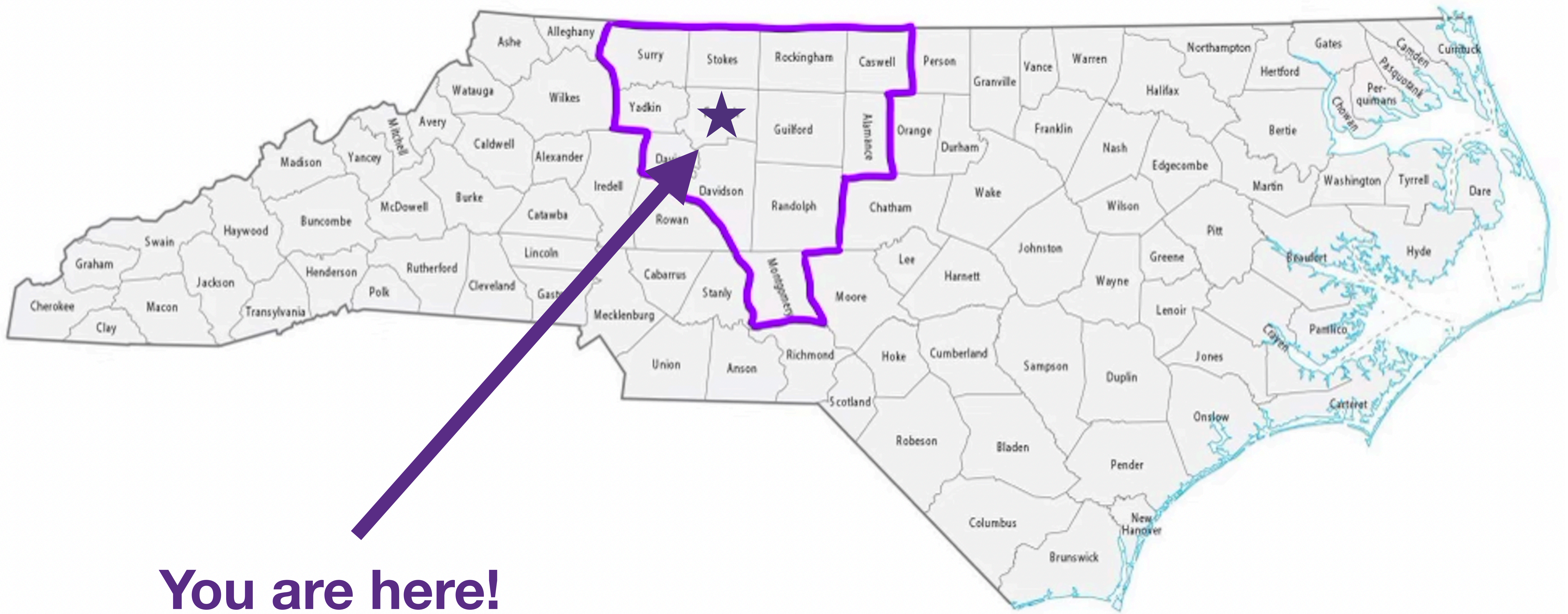
The Piedmont Triad

N = 387 Census Tracts



The Piedmont Triad

N = 387 Census Tracts



You are here!

Our “Neighborhoods”

What We Have

- **Population center** of the neighborhood
- **Haversine distance** from the nearest healthy food retailer to the center
- **Route-based distance** from the nearest healthy food retailer to the center
- **Population size** of the tract
- Count of **diabetes cases** in the tract

Our "Neighborhoods"

Where They Came From

- Neighborhood population centers (N = 387) are from the Census Bureau (census tracts, 2010 release).
- Healthy food retailers (M = 701) are from the US Department of Agriculture (historical SNAP retailer locator dataset, 2022 release).
- Diabetes prevalences are from the Centers for Disease Control and Prevention (PLACES dataset, 2022 release).
- The data were adapted from Lotspeich et al., 2023+.

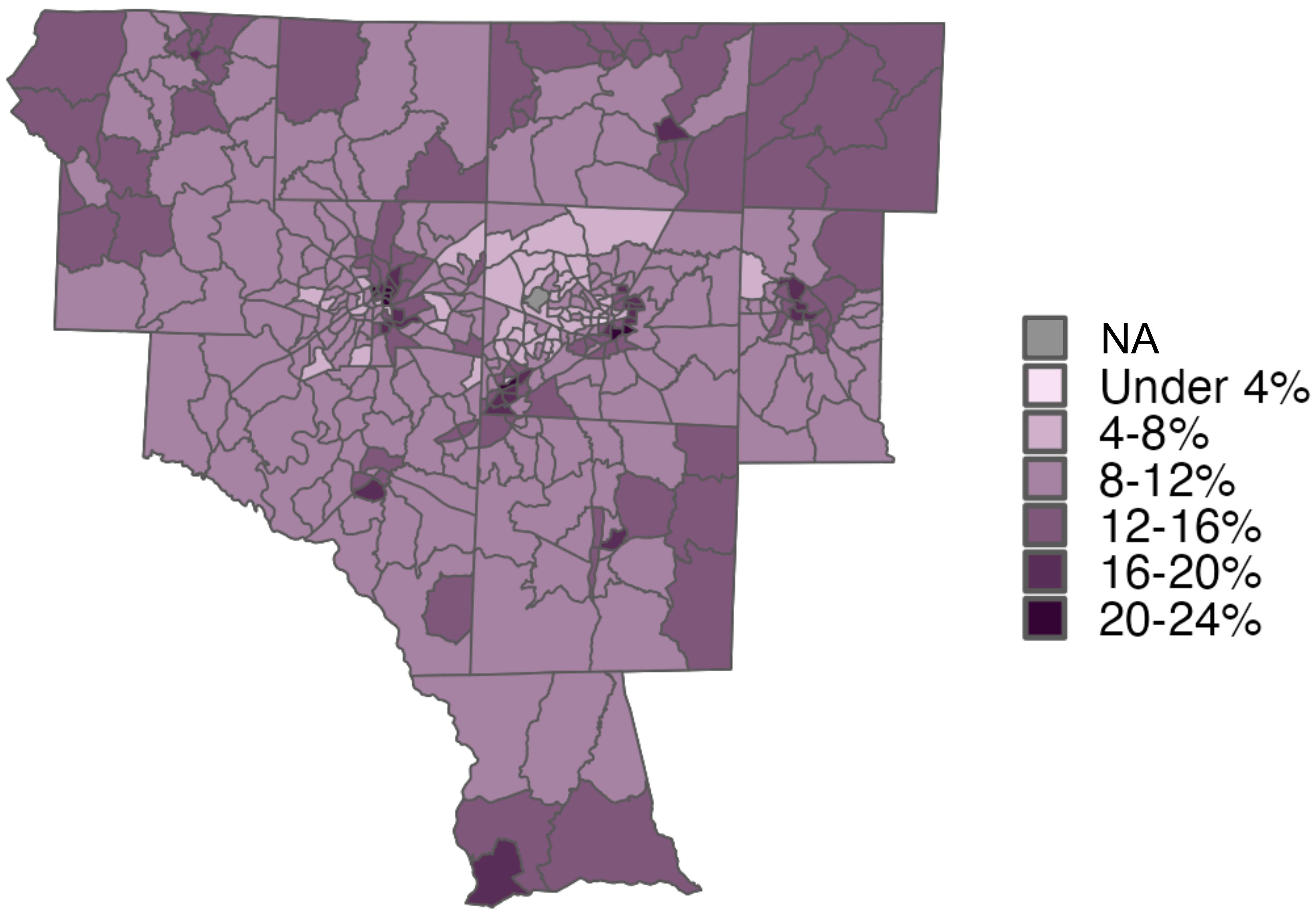
Our “Neighborhoods”

What We Did

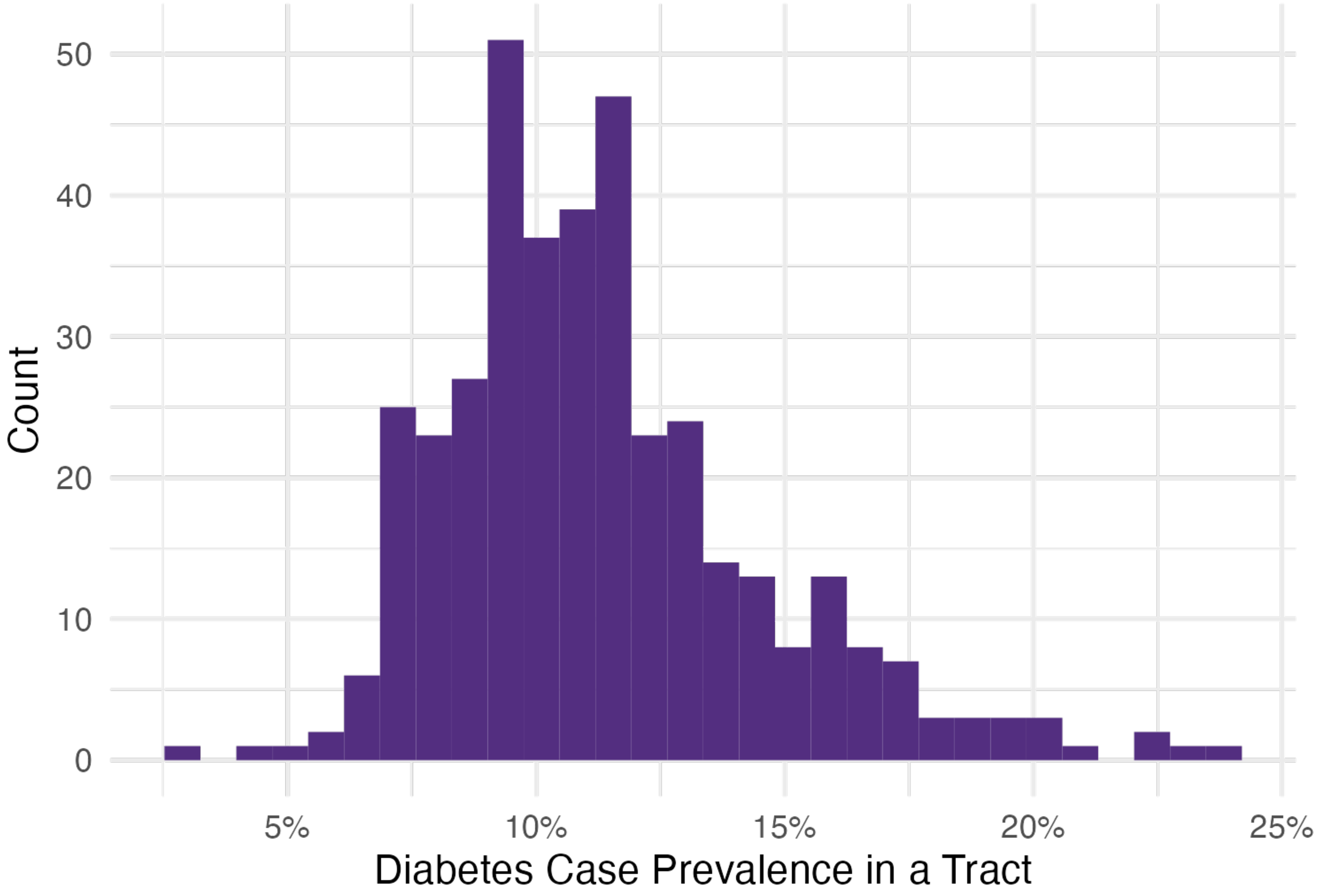
- Discretized both distance measurements to **create X_r and X^*_r**
- Used **radii** of 0.5, 1, 5, and 10 miles
- Chose 25% of the tracts randomly to **throw out X_r** (i.e., let $q = 0.75$)

Diabetes Landscape

- Statewide prevalence in 2021 was **12.4%** (American Diabetes Association)
- Most tracts have **8-12%** prevalence
- Prevalence **varies** across the Triad
- Lower prevalences coincide with smaller, urban tracts

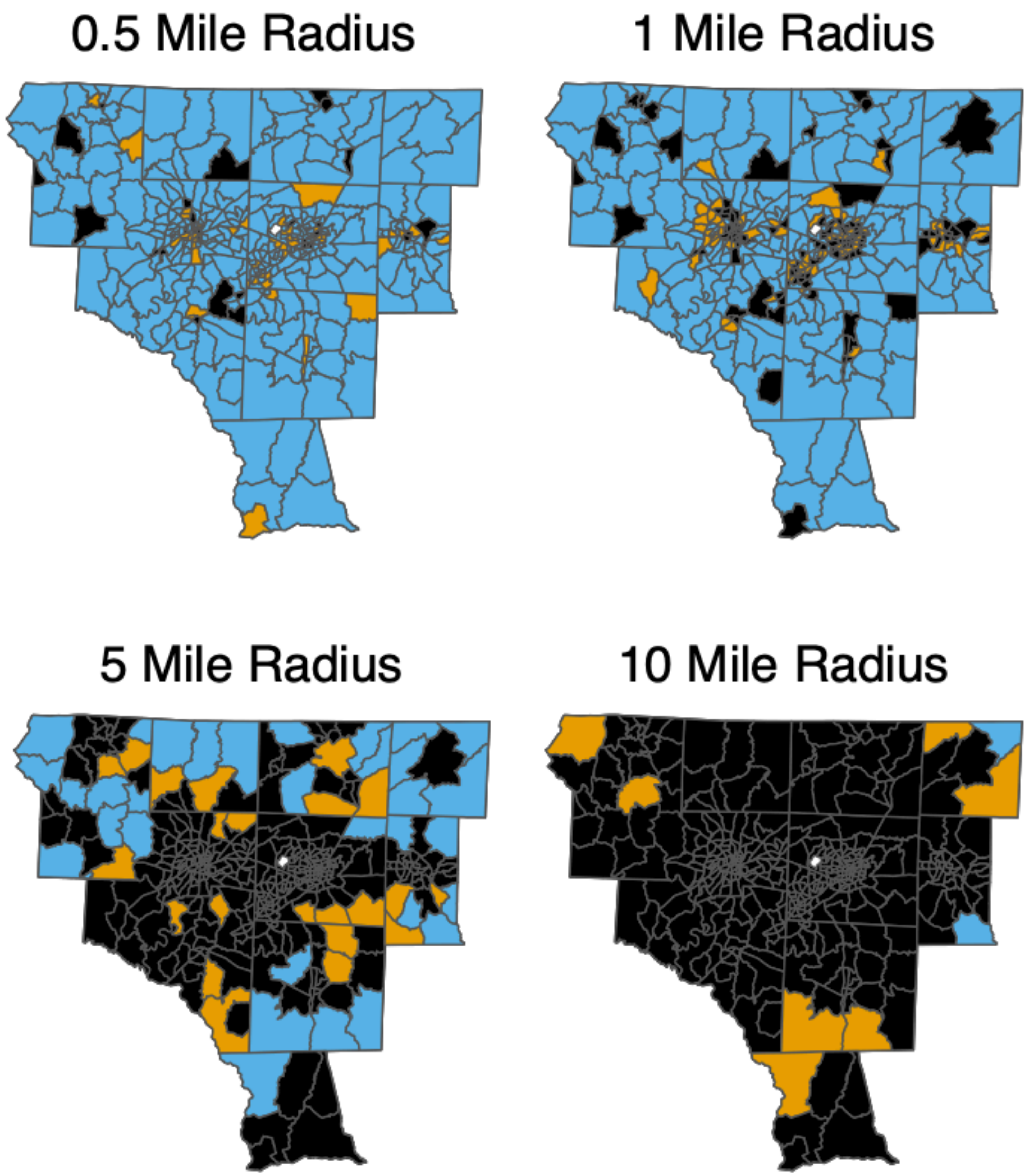


Diabetes Landscape



Food Access Landscape

- As radius **increases**, more tracts **flip** from blue to gold or black
- 22% of tracts have **over a mile difference** between their distance measures to the nearest retailer



No Access

Error-Prone Access

True Access

Error Rates

0.5 Mile Radius
Straight-Line

Route-Based	No Access	Access
No Access	302	51
Access	0	34

1 Mile Radius
Straight-Line

Route-Based	No Access	Access
No Access	195	70
Access	0	122

No Access   Error-Prone Access   True Access

Error Rates

5 Mile Radius
Straight-Line

Route-Based	No Access	Access
No Access	32	22
Access	0	333

10 Mile Radius
Straight-Line

Route-Based	No Access	Access
No Access	2	7
Access	0	378

No Access   Error-Prone Access   True Access

The Model

$$\log\{\text{E}_\beta(\text{Diabetes Cases} \mid \text{Access})\} = \beta_0 + \beta_1 \text{Access} + \log(\text{Population})$$

The Model

$$\log\{E_\beta(\text{Diabetes Cases} \mid \text{Access})\} = \beta_0 + \beta_1 \text{Access} + \log(\text{Population})$$

↓

log(outcome prevalence)

The Model

$$\log\{\mathrm{E}_\beta(\text{Diabetes Cases} \mid \text{Access})\} = \beta_0 + \beta_1\text{Access} + \log(\text{Population})$$

The Model

$$\log\{\mathrm{E}_\beta(\text{Diabetes Cases} \mid \text{Access})\} = \beta_0 + \beta_1 \text{Access} + \log(\text{Population})$$

log(prevalence ratio of exposure)

The Model

$$\log\{\text{E}_{\beta}(\text{Diabetes Cases} \mid \text{Access})\} = \beta_0 + \beta_1\text{Access} + \log(\text{Population})$$

The Model

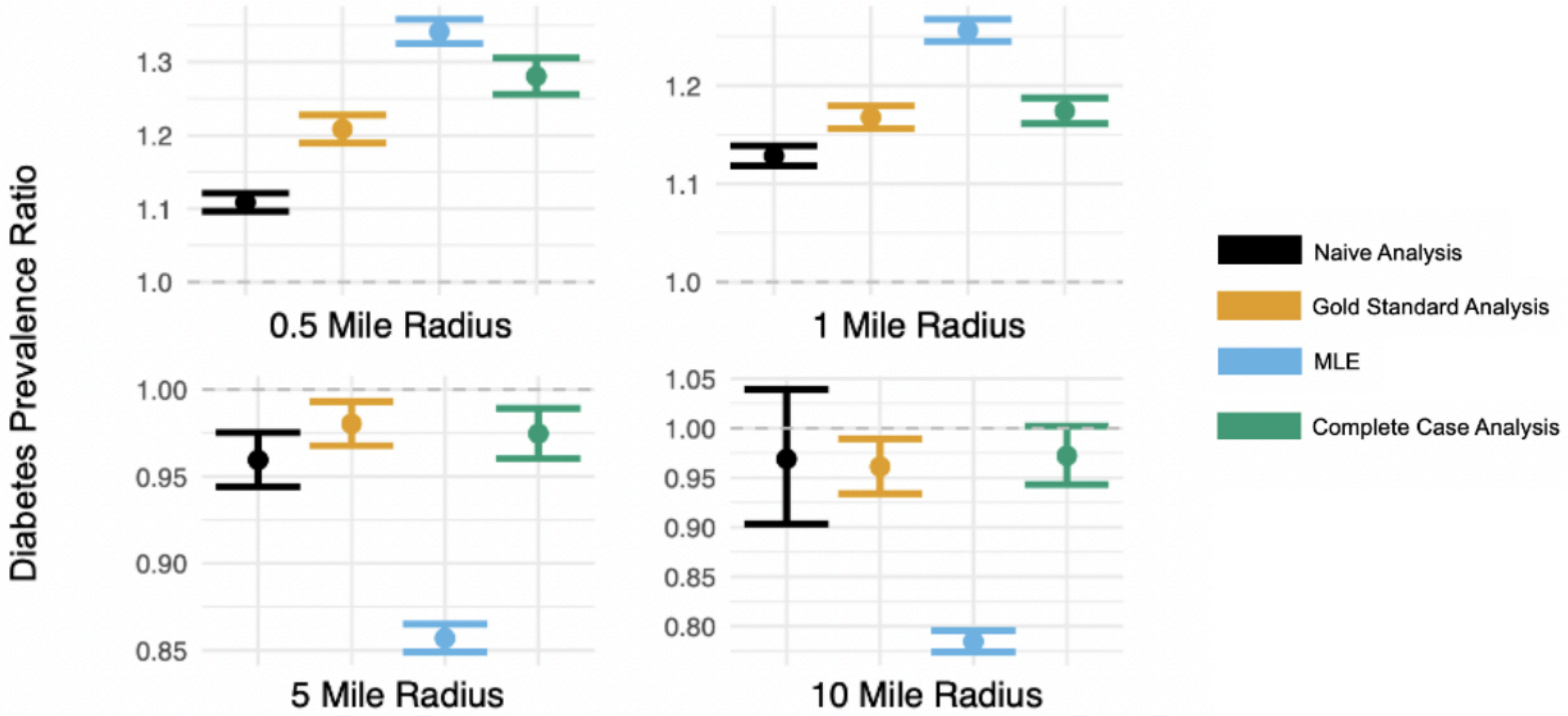
$$\log\{E_\beta(\text{Diabetes Cases} \mid \text{Access})\} = \beta_0 + \beta_1 \text{Access} + \log(\text{Population})$$

offset

The Model

$$\log\{\text{E}_\beta(\text{Diabetes Cases} \mid \text{Access})\} = \beta_0 + \beta_1 \text{Access} + \log(\text{Population})$$

Model Results



What if we missed a confounder?

Hypothetical β_2

- In the **worst case**, we need a confounder-outcome effect of **9.5%** to tip the prevalence ratio to the null.

- In the **best case**, we need a confounder-outcome effect of **54.9%** to tip the prevalence ratio.

Wrap Up 🎬

Guiding Questions

- Can we use a function of distance to healthy food retailers to **quantify food access** in the Piedmont Triad, even if this function is **subject to misclassification**?
- Can we estimate the relationship between **food access** and **diabetes prevalence** in the presence of misclassifications and missingness?

Guiding Questions

- ✓ Can we use a function of distance to healthy food retailers to **quantify food access** in the Piedmont Triad, even if this function is **subject to misclassification**?
- Can we estimate the relationship between **food access** and **diabetes prevalence** in the presence of misclassifications and missingness?

Guiding Questions

- ✓ Can we use a function of distance to healthy food retailers to **quantify food access** in the Piedmont Triad, even if this function is **subject to misclassification**?
- ✓ Can we estimate the relationship between **food access** and **diabetes prevalence** in the presence of misclassifications and missingness?

Strengths and Limitations

- ⭐Uses all available data
- ⭐Only two parametric assumptions
- ⭐Lower bias than naive analysis
- ⭐Recovers efficiency lost by the complete case analysis

- 😢Finicky numerical behavior, especially in the standard error estimators
- 😢Poisson assumptions in the case study

Recommendations

- Use the **gold standard** in a setting where there is no missingness or misclassification.
- Use the **MLE** if you have high error rates and missingness, as it **avoids the bias** of the naive analysis and **recovers more efficiency** than the complete case analysis.
- If you have very **little missingness**, you can get away with the **complete case analysis.**

Future Directions

- Incorporate a spatial model to explore relationships among adjacent tracts
- Vary the outcome model of interest
- Extend past the binary exposure case
- Improve the query design

Ashley’s Future Directions

Ashley’s Future Directions

References

1. American Diabetes Association. About diabetes, 2021. URL https://diabetes.org/about-diabetes

2. D.D. Boos and L.A. Stefanski. Essential statistical inference. Springer texts in statistics. Springer, New York, NY, 2013 edition, Feb. 2013

3. D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

4. E. Gucciardi, M. Vahabi, N. Norris, J. P. Del Monte, and C. Farnum. The intersection between food insecurity and diabetes: a review. Current nutrition reports, 3:324–332, 2014

5. McGowan LD (2022). “tipr: An R package for sensitivity analyses for unmeasured confounders.” _Journal of Open Source Software_, *7*(77), 4495. <https://doi.org/10.21105/joss.04495>.

References

6. J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. Numerical Optimization: Theoretical and Practical Aspects. Universitext. Springer, Berlin ; New York, 2nd ed edition, 2006. ISBN 978-3-540-35445-1

7. P. A. Shaw, P. Gustafson, R. J. Carroll, V. Deffner, K. W. Dodd, R. H. Keogh, V. Kipnis, J. A. Tooze, M. P. Wallace, H. Küchenhoff, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: part 2—more complex methods of adjustment and advanced topics. Statistics in medicine, 39(16):2232–2263, 2020

8. R. Little and D. Rubin. Statistical Analyses with Missing Data. New Jersey: John Wiley & Sons, Inc, 2002.

9. S. Lotspeich, A. Mullan, L. D’Agostino McGowan, and S. Hepler. Combining straight-line and map-based distances to investigate food access and health. In Preparation, 2023+

10. Walker K, Herman M (2024). _tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames_. R package version 1.6, <https://CRAN.R-project.org/package=tidycensus>.

11. World Health Organization. Healthy diet, 2019. URL https://iris.who.int/handle/10665/325828

Acknowledgements

- Dr. Sarah Lotspeich
- Dr. Staci Hepler
- Dr. Lucy D'Agostino McGowan
- Dr. Dave Kline
- Dr. Nicole Dalzell
- Dr. Lynne Yengulalp
- The Garcia-Lotspeich Lab
- The SESH Lab
- The faculty at Wake Forest
- My mentors, classmates, and collaborators at the University of Scranton, Carnegie Mellon University, and Central Washington University
- My grad student classmates at Wake Forest
- My friends and family

THE ANDREW SABIN FAMILY
CENTER for ENVIRONMENT AND SUSTAINABILITY