Modeling Loggerhead Nesting Patterns: How many Pseudo-Absence Points are Necessary?

Presented by Cheyenne Long
Advisor: Dr. Samantha Seals

Introduction

Do loggerhead turtles have a location preference when selecting where to nest on Pensacola Beach?

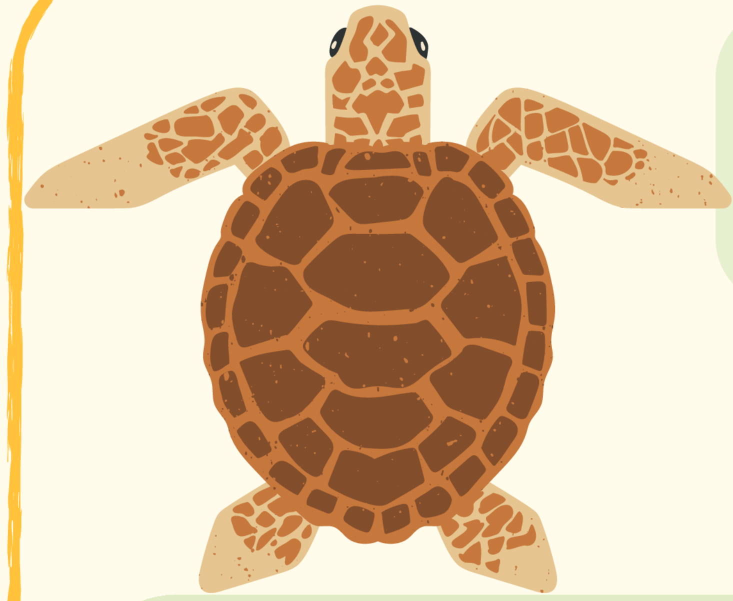
- **Can we identify preferred beach characteristics?**

Related projects from the Computational Geomorphology & Modeling Lab:

- Spring 2024: Environmental Science student examined nesting preferences of loggerhead turtles on Pensacola Beach
- Spring 2024: Mathematics & Statistics student bootstrapped different ratios of presence/psuedo-absence points
- Ongoing: Environmental Science student examining different ratios of presence/pseudo-absence points
- Current project: Simulation study to determine how analysis results are affected by increasing the number of pseudo-absence points.

What is Pseudo-Absence Data?

- Type of background data
  - A set of data points or environmental variables that represent locations in the study
  - Useful for presence only or limited data
- Not true absence points
  - Available, but uninhabited, environment in area
  - Used to model abundance data



Why use Pseudo-Absence Data for this analysis?

Creates environmentally similar, randomly generated points to use as "absence" points

10:1 Literature

1. *"Selecting Pseudo-Absences for Species Distribution Models: How, Where, and How Many?"*
    - Concluded 10:1 was necessary when looking at 10-10,000 absences

2. *"Assessing the Effects of Pseudo-Absences on Predictive Distribution Model Performance."*
    - Used a 10:1 ratio in their analysis of species distribution

3. *"Going West: Range Expansion for Loggerhead Sea Turtles in the Mediterranean Sea Under Climate Change."*
    - Used a 10:1 ratio

Relevance

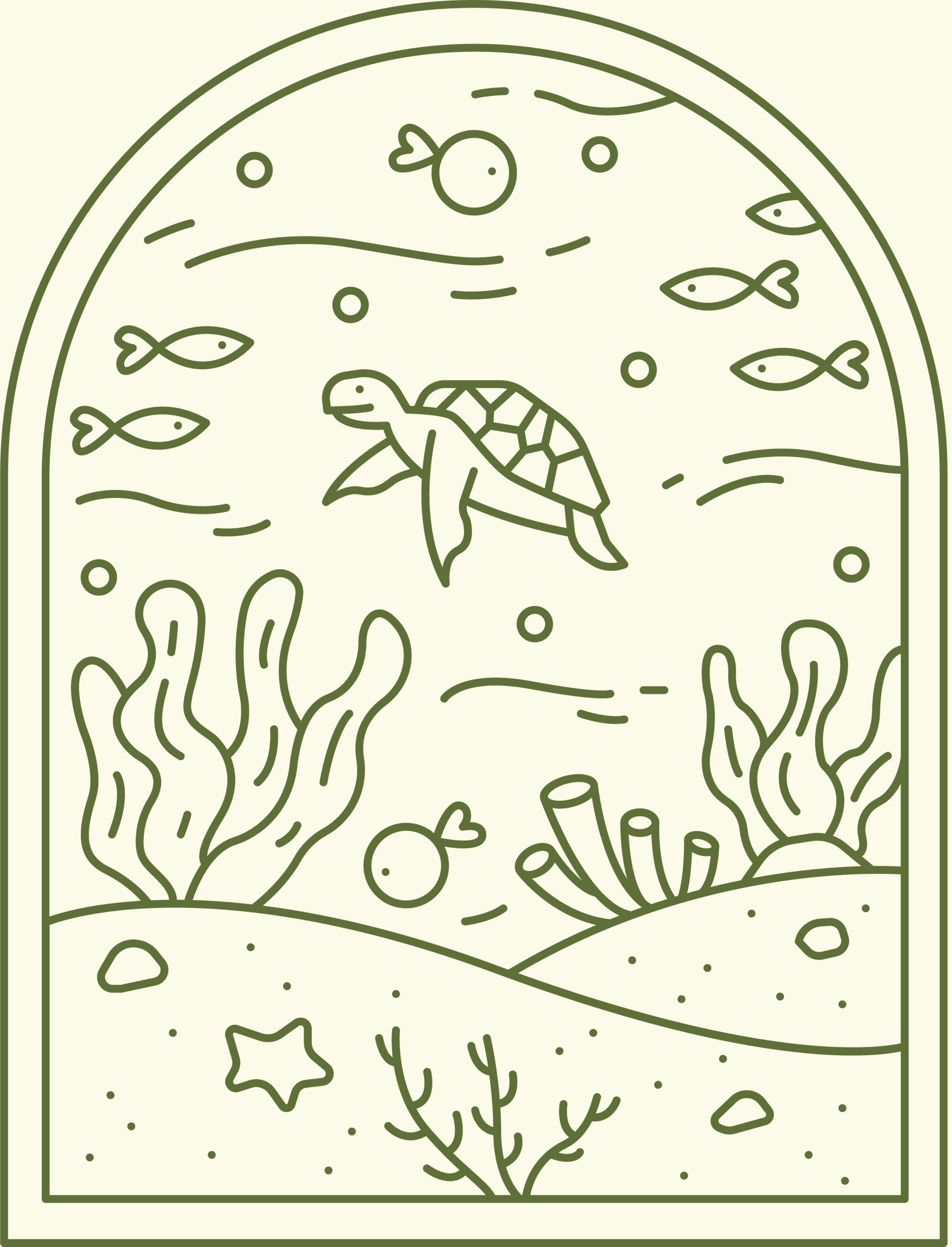
- Save time generating data points if 10:1 is not needed
  - Common ratio is 10:1
  - Is this necessary?
  - Do we get accurate results with smaller ratios?

How?

Simulate presence and pseudo-absence data using the observed characteristics from Pensacola Beach

Perform statistical analysis on each simulated dataset.

Examine the distributions of analysis results (slopes, standard errors, *p*-values)



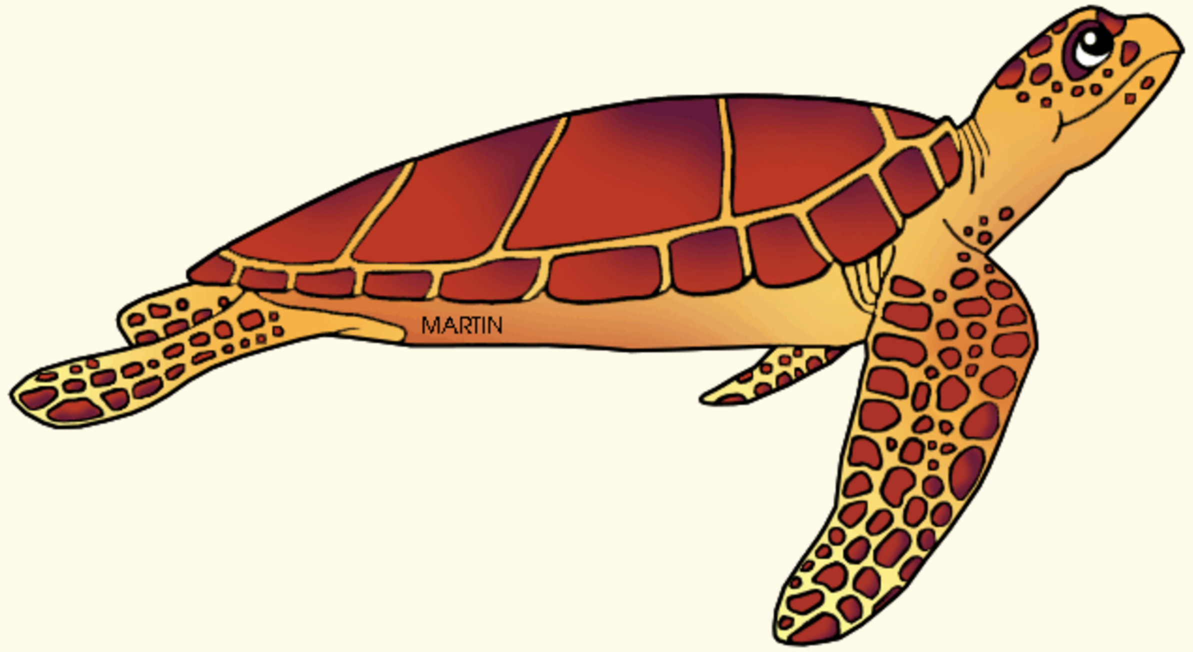
Data

My Data

y = Nested

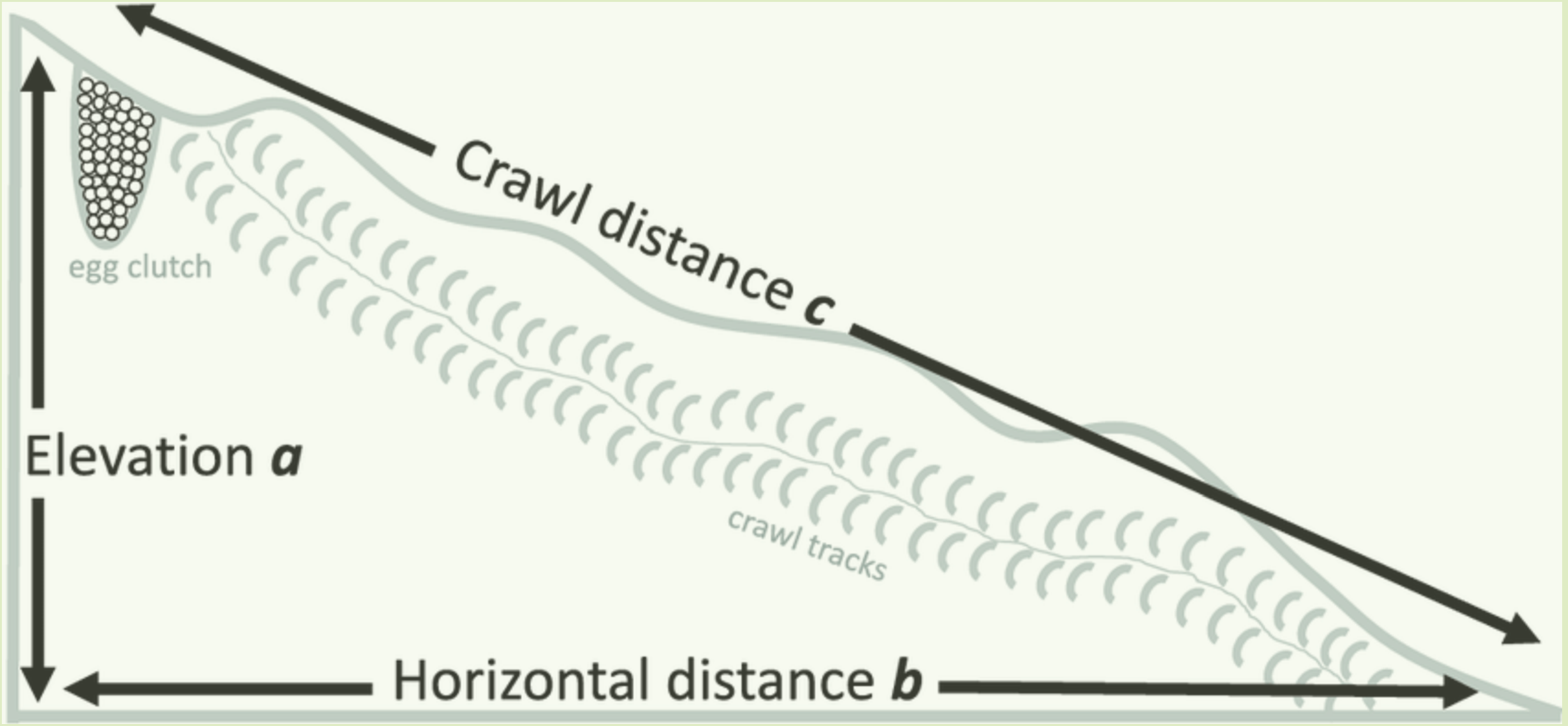
x = Nest Elevation (meters)

Nested



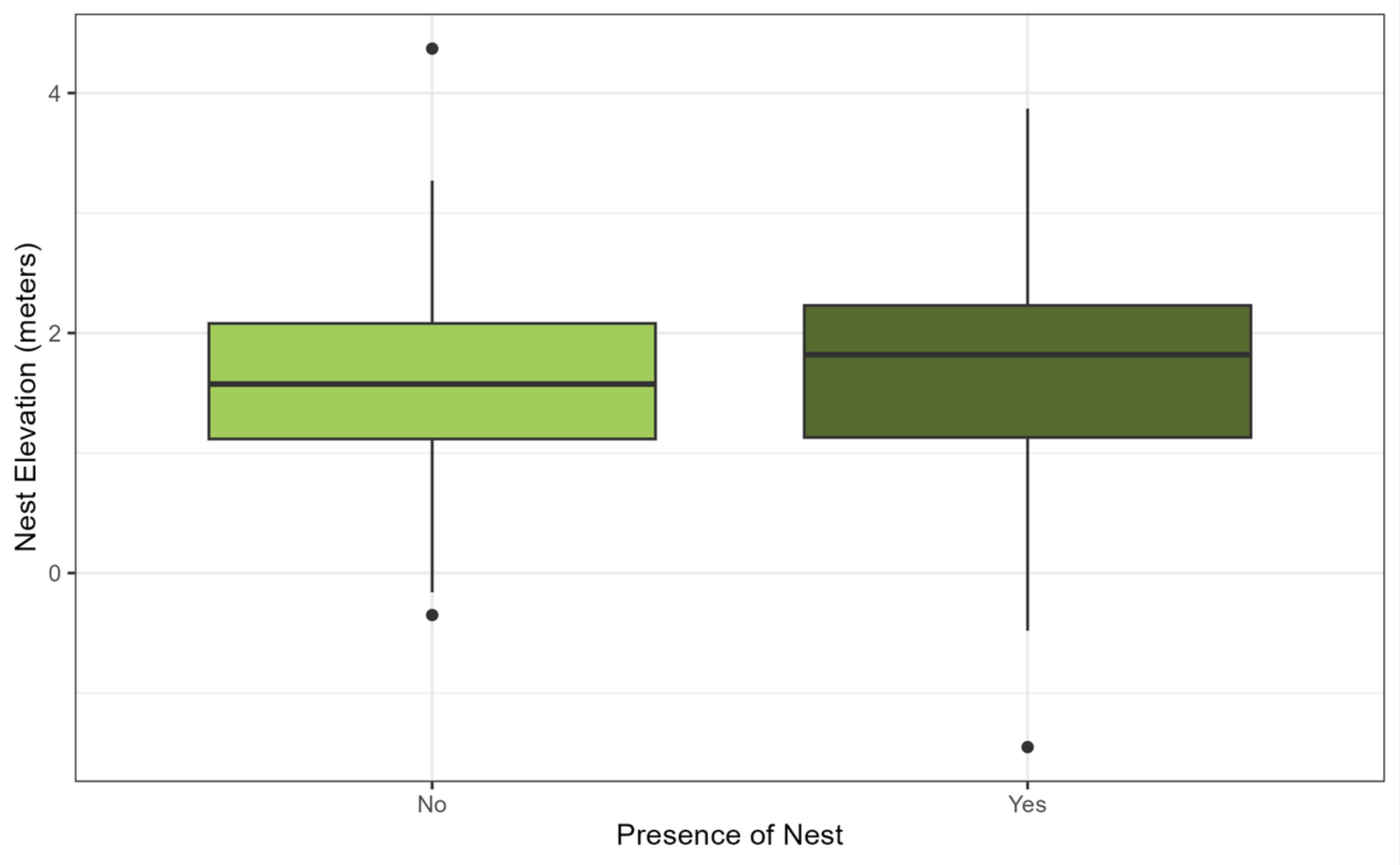
$$y = \begin{cases} 1 & \text{if a nest is present} \\ 0 & \text{if no nest is present} \end{cases}$$

Nest Elevation

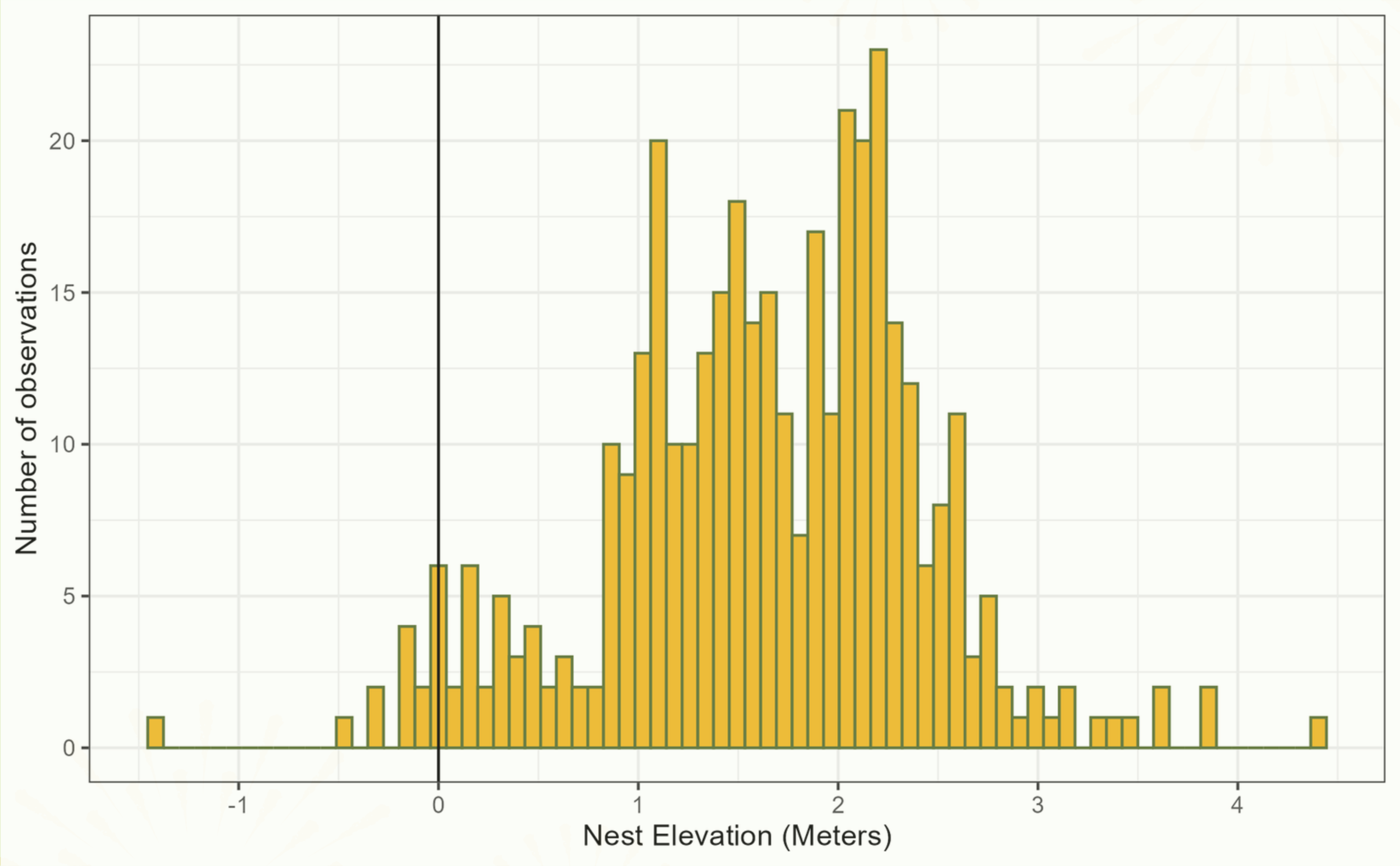


- Nest elevation for my data is defined as the elevation, in meters, of the nest above the NAVD88
- Chose because there was an observed relationship between it and nested in a previous study

Descriptive Statistics and Preliminary Analysis



Nest Elevation	Absence	Presence
Mean	1.54	1.71
Std. Dev.	0.77	0.82
Median	1.58	1.82
IQR	0.96	1.1
(Min,Max)	(-0.35,4.37)	(-1.45,3.87)

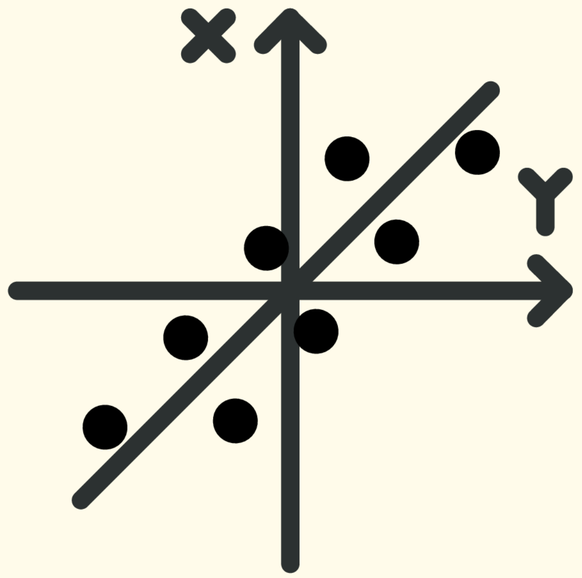


Number of observations
20
15
10
5
0
-1
0
1
2
3
4
Nest Elevation (Meters)

Methods

Linear vs. Logistic Regression

- Linear regression is used to predict continuous outcomes
- Logistic regression is used to predict categorical or qualitative dependent variables, such as binary, multinomial, or ordinal outcomes



Linear Regression

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- y is the continuous dependent variable, or the outcome variable
- x is the independent variable
- β_0 is the y-intercept of the line
- β_1 is the slope of the line
- ε is the error term

Binary Logistic Regression

- Binary logistic regression is used to model binary outcomes.
- Probability of nested in our case

$$y = \begin{cases} 1 & \text{if a nest is present} \\ 0 & \text{if no nest is present} \end{cases}$$

Binary Logistic Regression

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- π is the probability of a certain outcome
- β_0 is the y intercept
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients for the predictor variables
- $x_1, x_2, \ldots, x_k$ are the predictor variables

Simulation

What is Simulation

- Monte Carlo simulation: process of generating random data
- Parameters $\beta_0, \beta_1, \mu, \sigma$ are specified in the simulation
- Helps us understand how analysis results are affected under different scenarios
    - We know the true parameter values

Simulation Process

- Created a function in RStudio that would iteratively construct datasets
- Created data sets under different scenarios
  - Five size samples:
    - n=25, 50, 100, 150, 200
  - Four ratios of absence to presence:
    - 1:1, 2:1, 5:1, 10:1
  - 10,000 iterations under each scenario.

Simulation Process

- Simulated x (nest elevation) based on observed data from Pensacola Beach x ~ N(1.6, 0.8)
- Simulated y ~ Bin(n, 1, ratio)
- Set linear predictor to have intercept and slope of observed data
- Construct model y ~ x and save results $(\hat{\beta}_i, SE_{\hat{\beta}_i}, p-\text{value})$

Simulation Process

- Resulted in 10,000 datasets under each scenario
    - Created 200,000 individual models
- Computed bias, MSE, and rejection rate under each scenario

Evaluation of Simulation

Bias

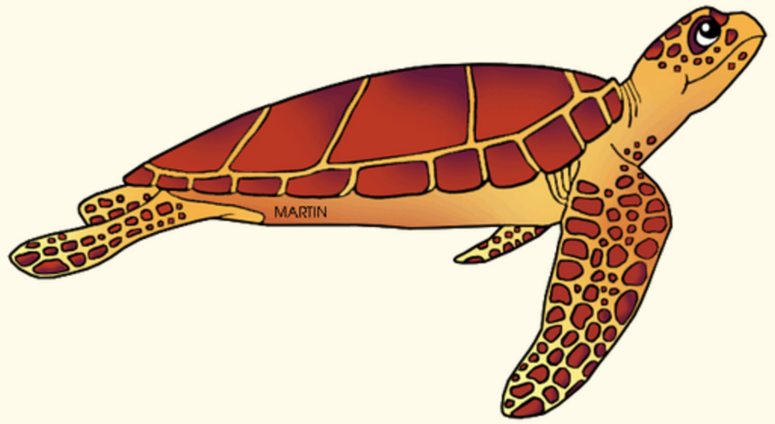
Difference between the expected value and estimated value

$$\text{Bias}(\hat{\beta}) = \mathbb{E}[\hat{\beta}] - \beta$$

$$\mathbb{E}[\hat{\beta}] = \text{the expected value of the estimator } \hat{\beta},$$

$$\beta = \text{the true value of the parameter}$$

MSE

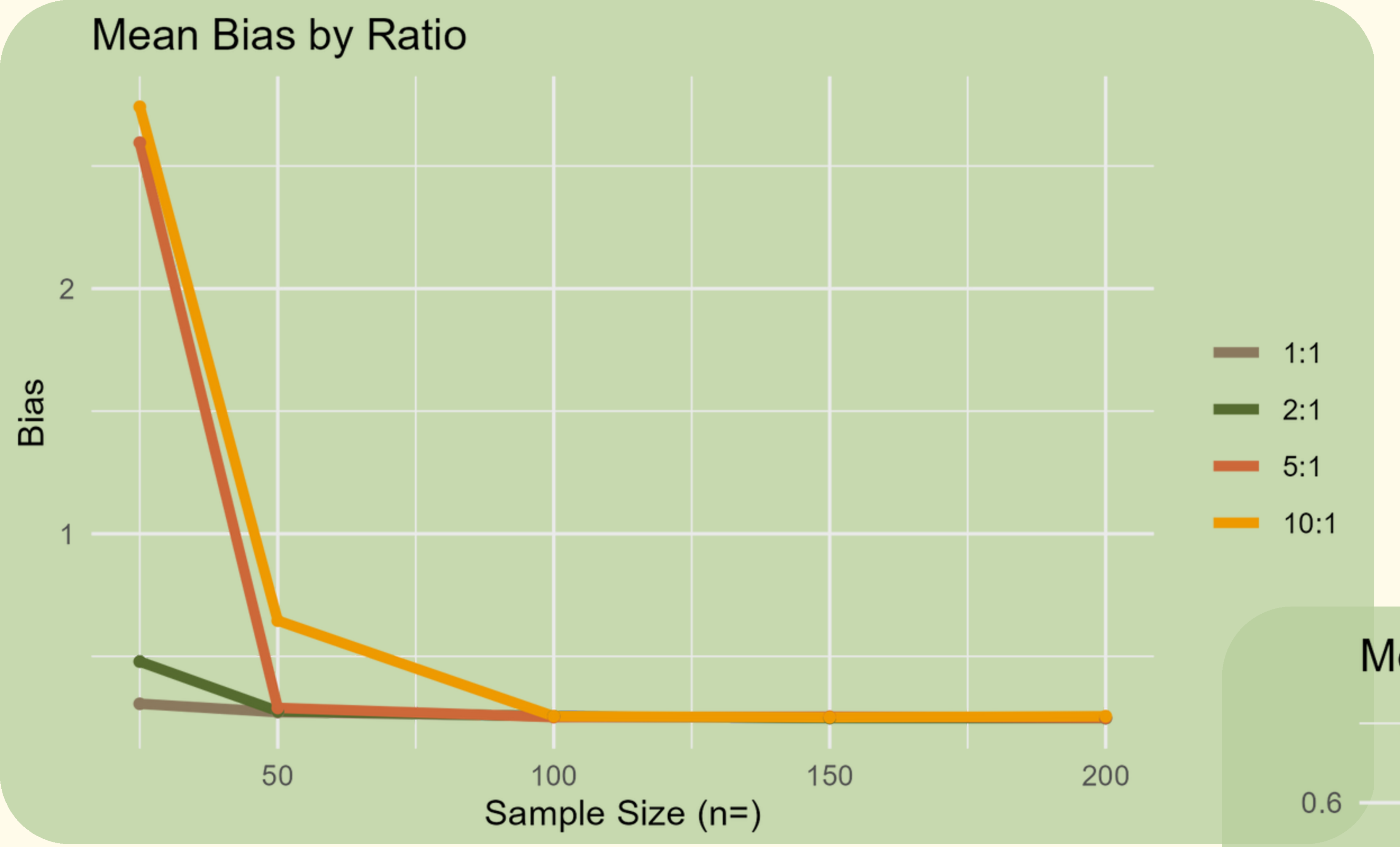


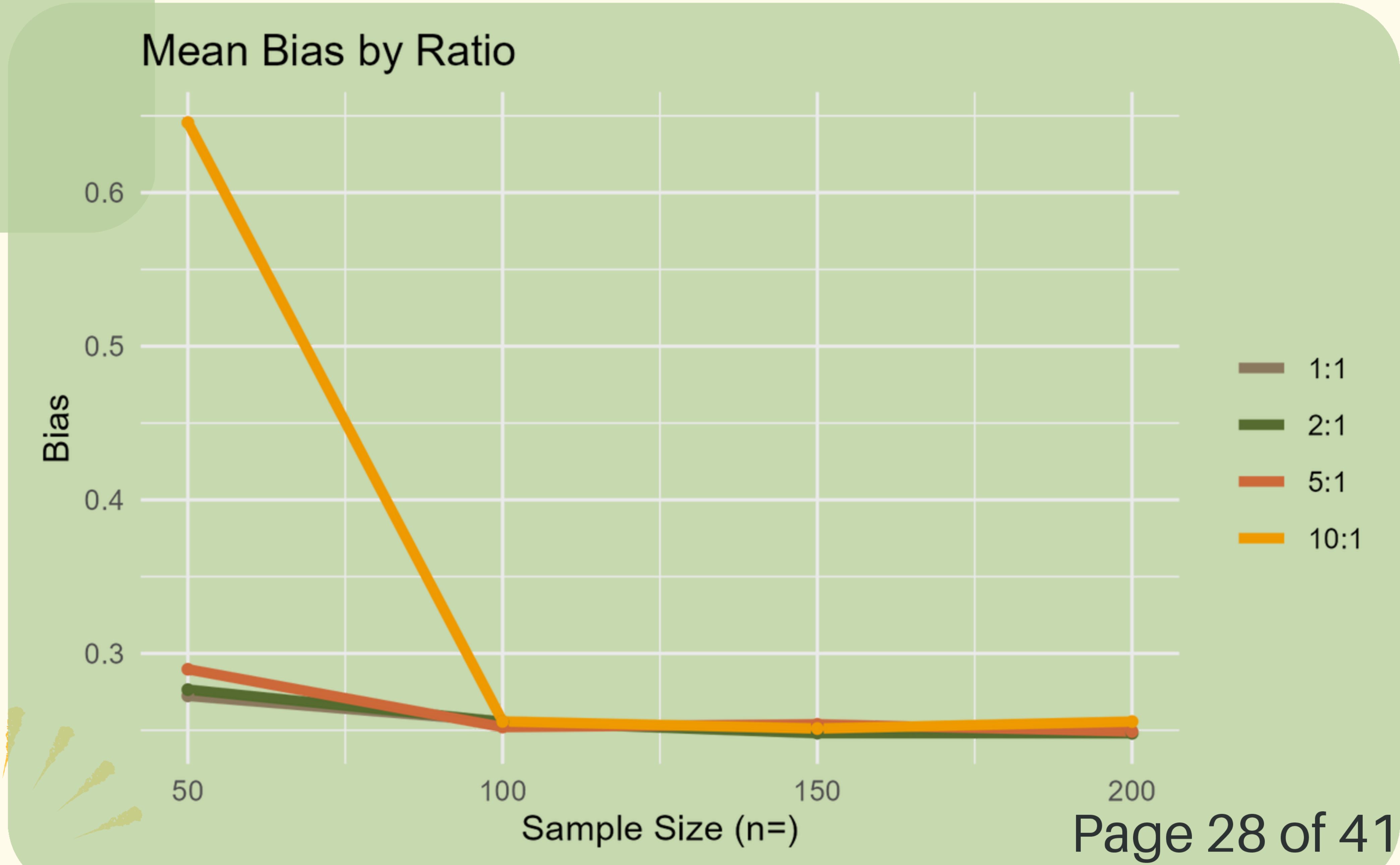
- MSE-Mean Squared Error
- Measures the average squared difference between actual and predicted values

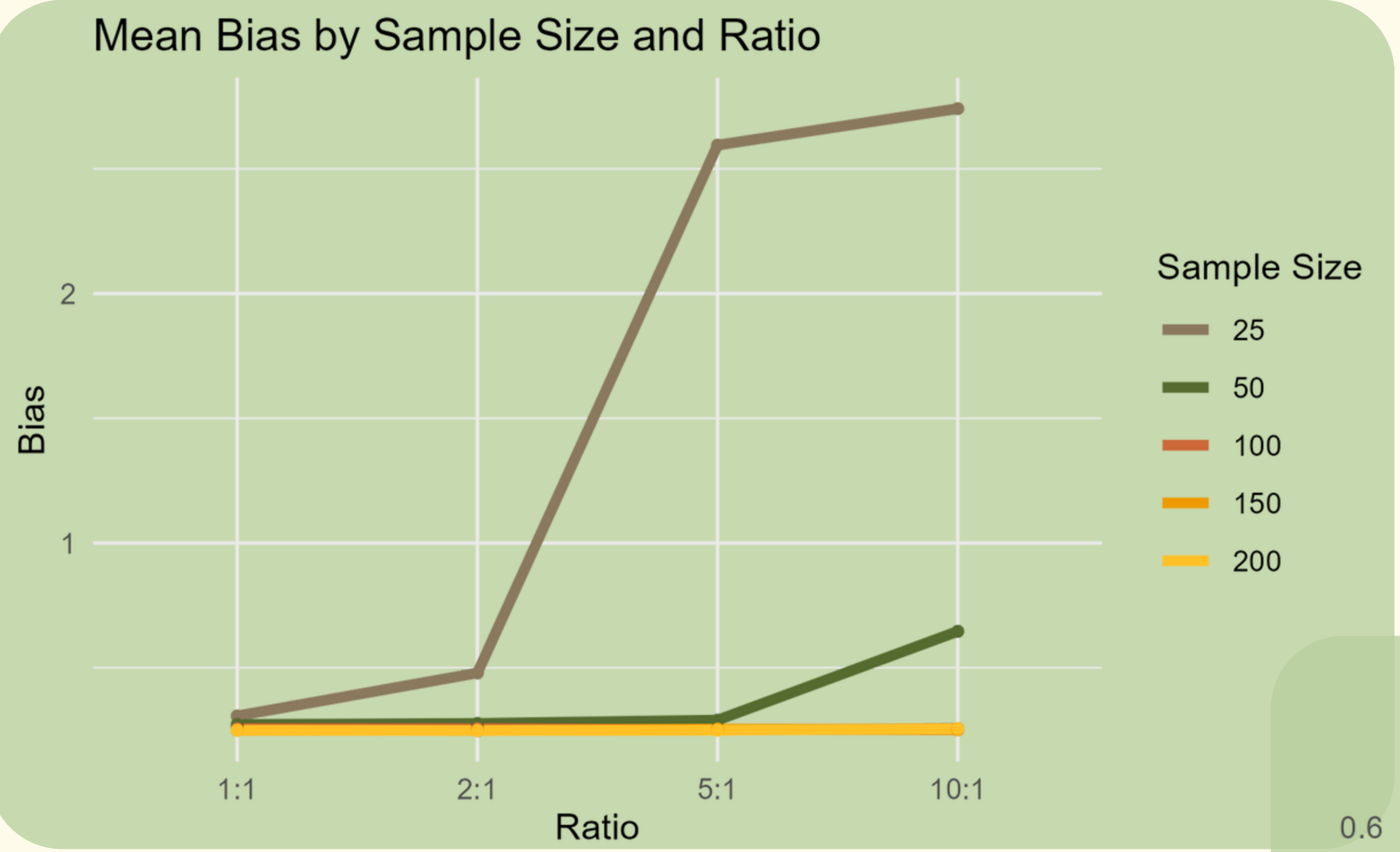
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

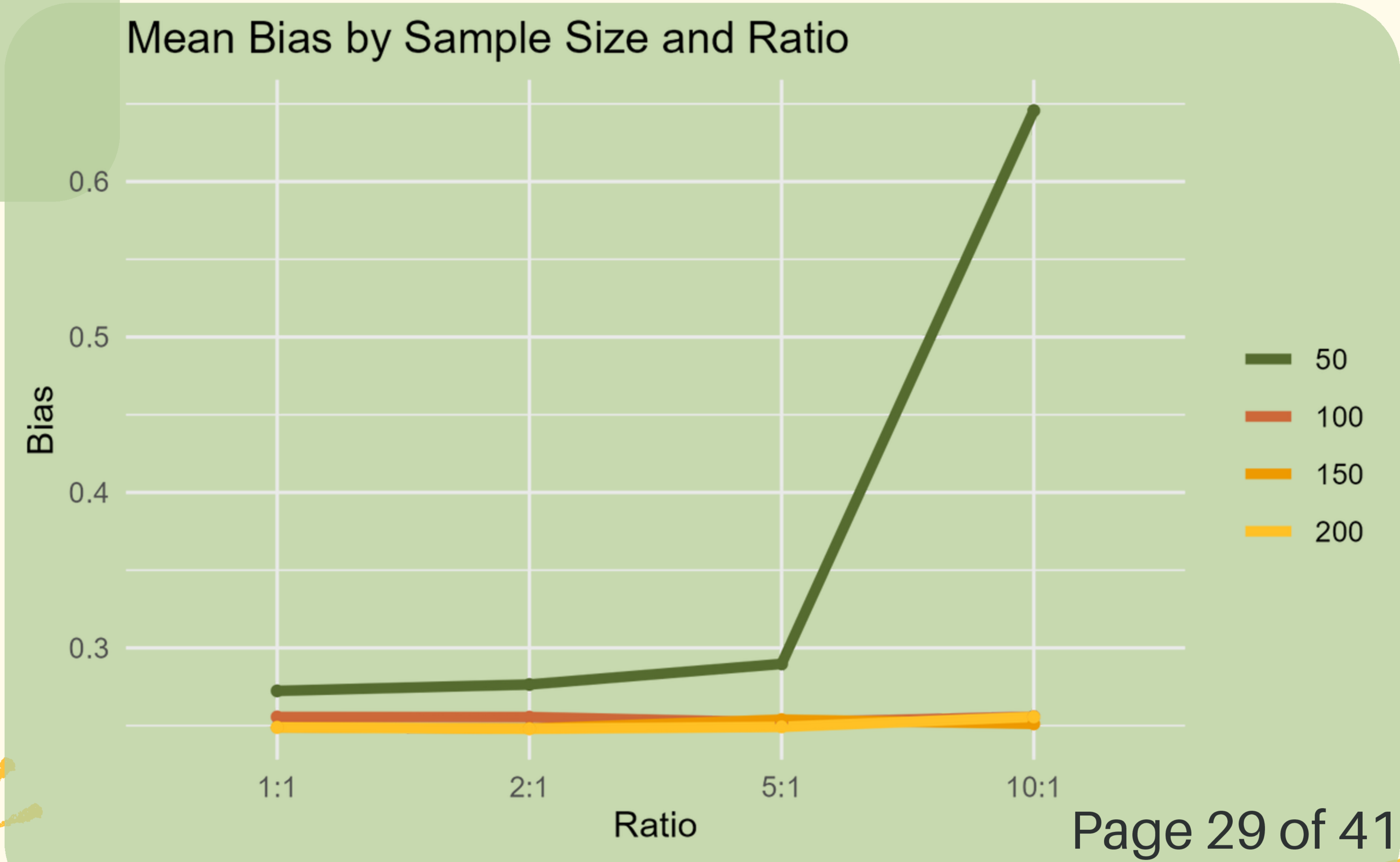
$$\mathrm{MSE}(\hat{\beta}) = \mathrm{Bias}^2(\hat{\beta}) + \mathrm{Var}(\hat{\beta})$$

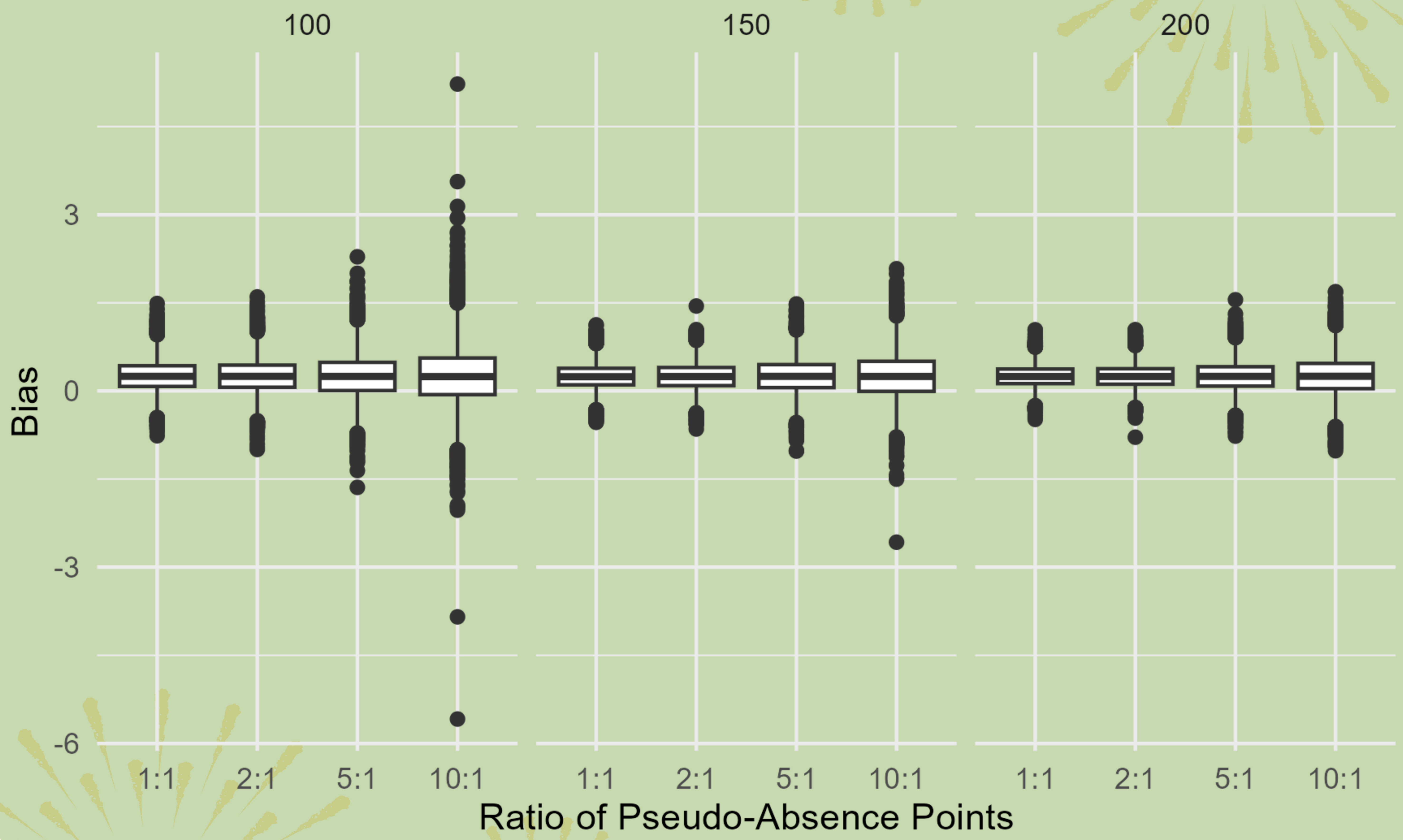
$$\text{where, } \mathrm{Var}(\hat{\theta}) = \mathbb{E}[\hat{\theta}^2] - (\mathbb{E}[\hat{\theta}])^2$$

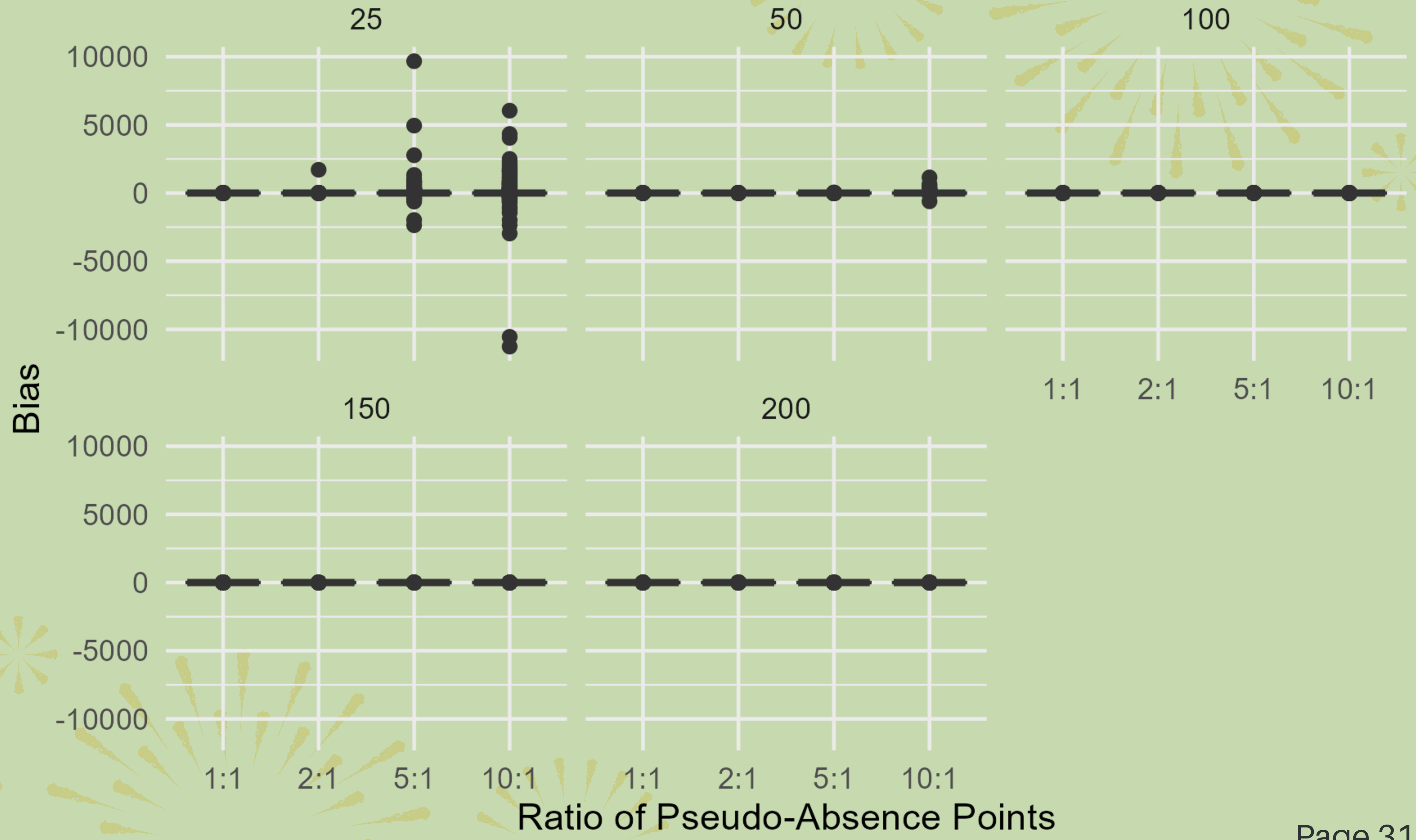

Mean Bias by Ratio
Bias
2
1
50
100
150
200
Sample Size (n=)
1:1
2:1
5:1
10:1

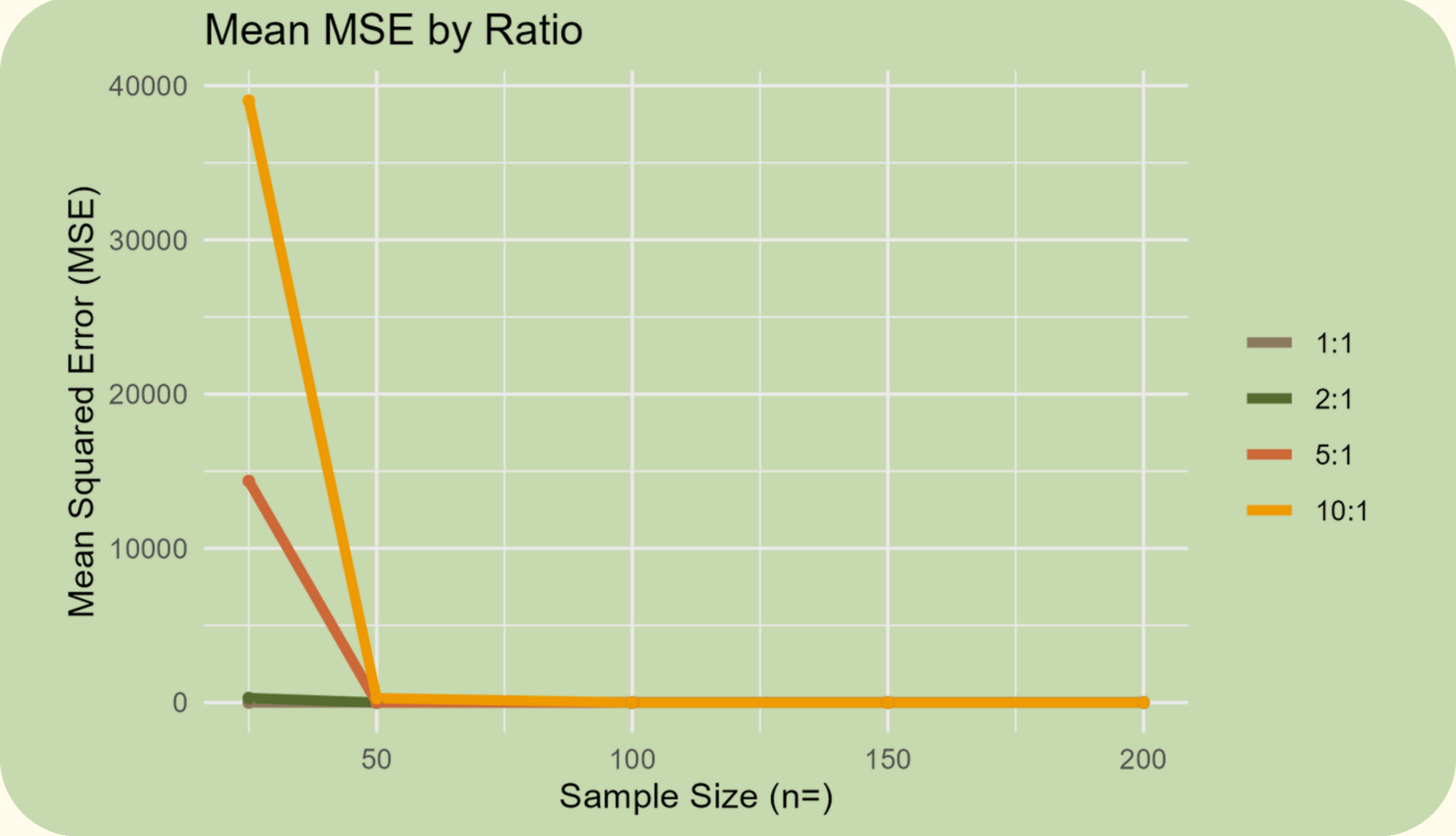

Mean Bias by Ratio
Bias
0.6
0.5
0.4
0.3
50
100
150
200
Sample Size (n=)
1:1
2:1
5:1
10:1

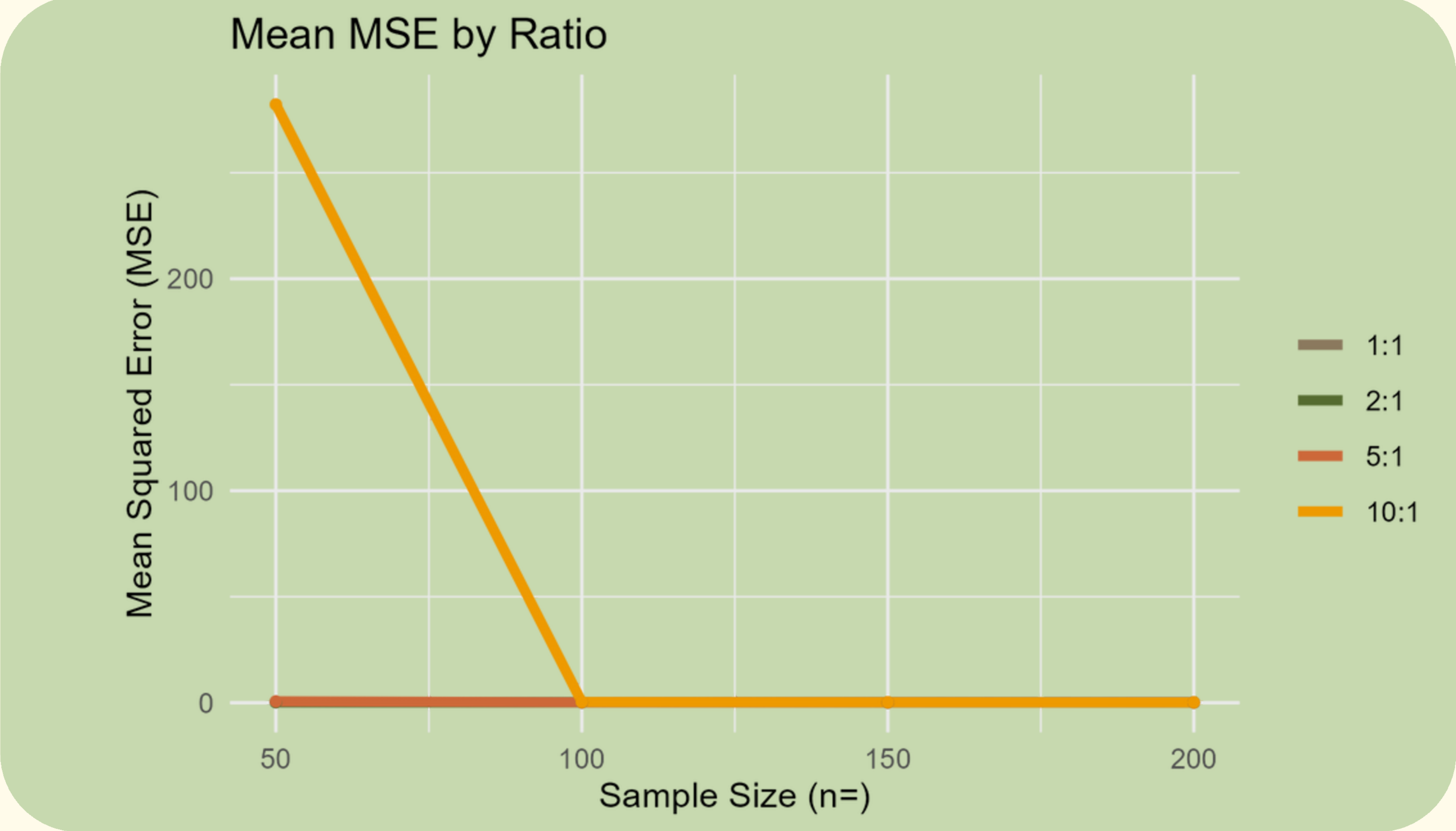

Mean Bias by Sample Size and Ratio
Bias
2
1
1:1
2:1
5:1
10:1
Ratio
Sample Size
25
50
100
150
200

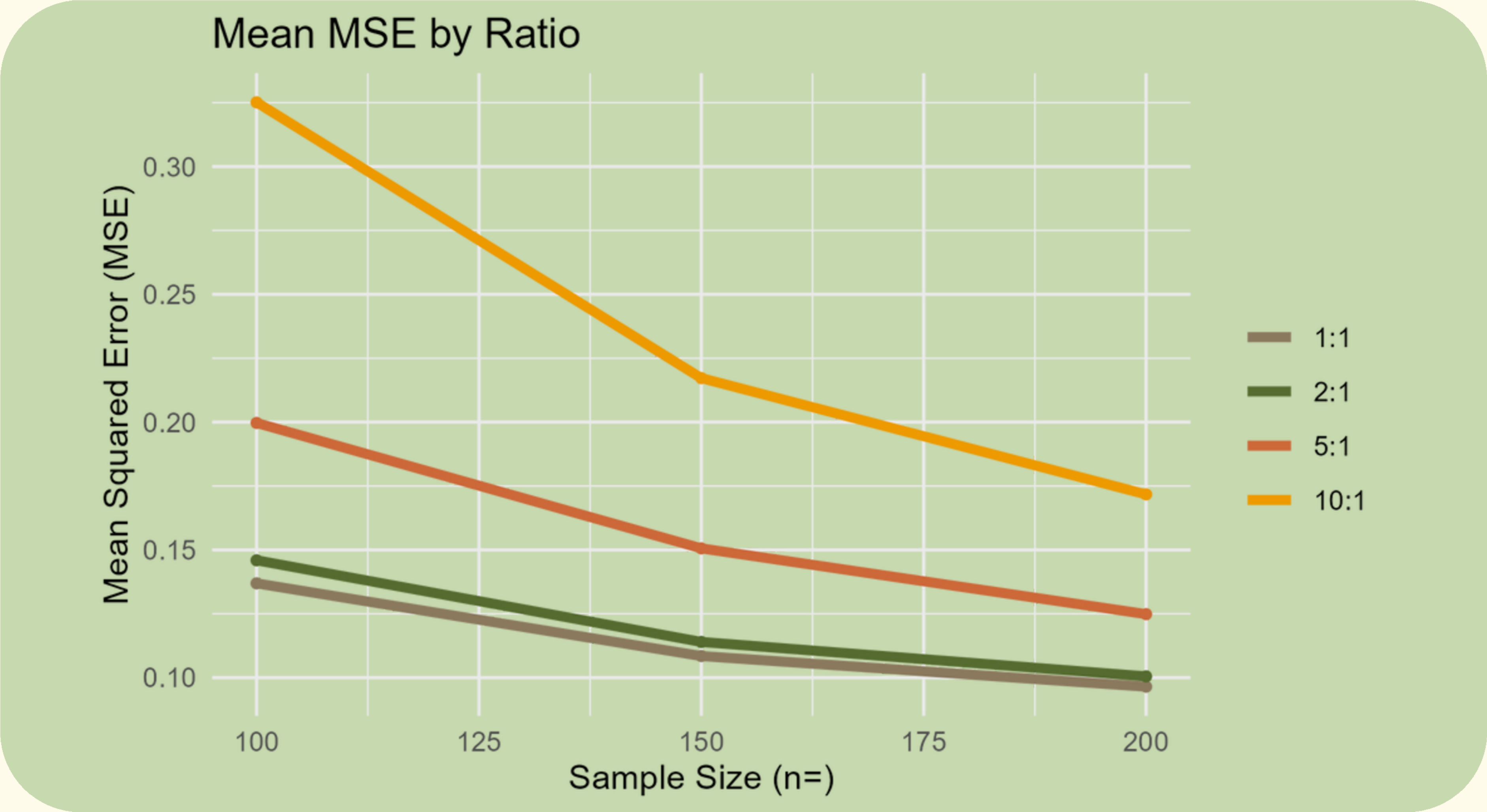

Mean Bias by Sample Size and Ratio
Bias
0.6
0.5
0.4
0.3
1:1
2:1
5:1
10:1
Ratio
50
100
150
200

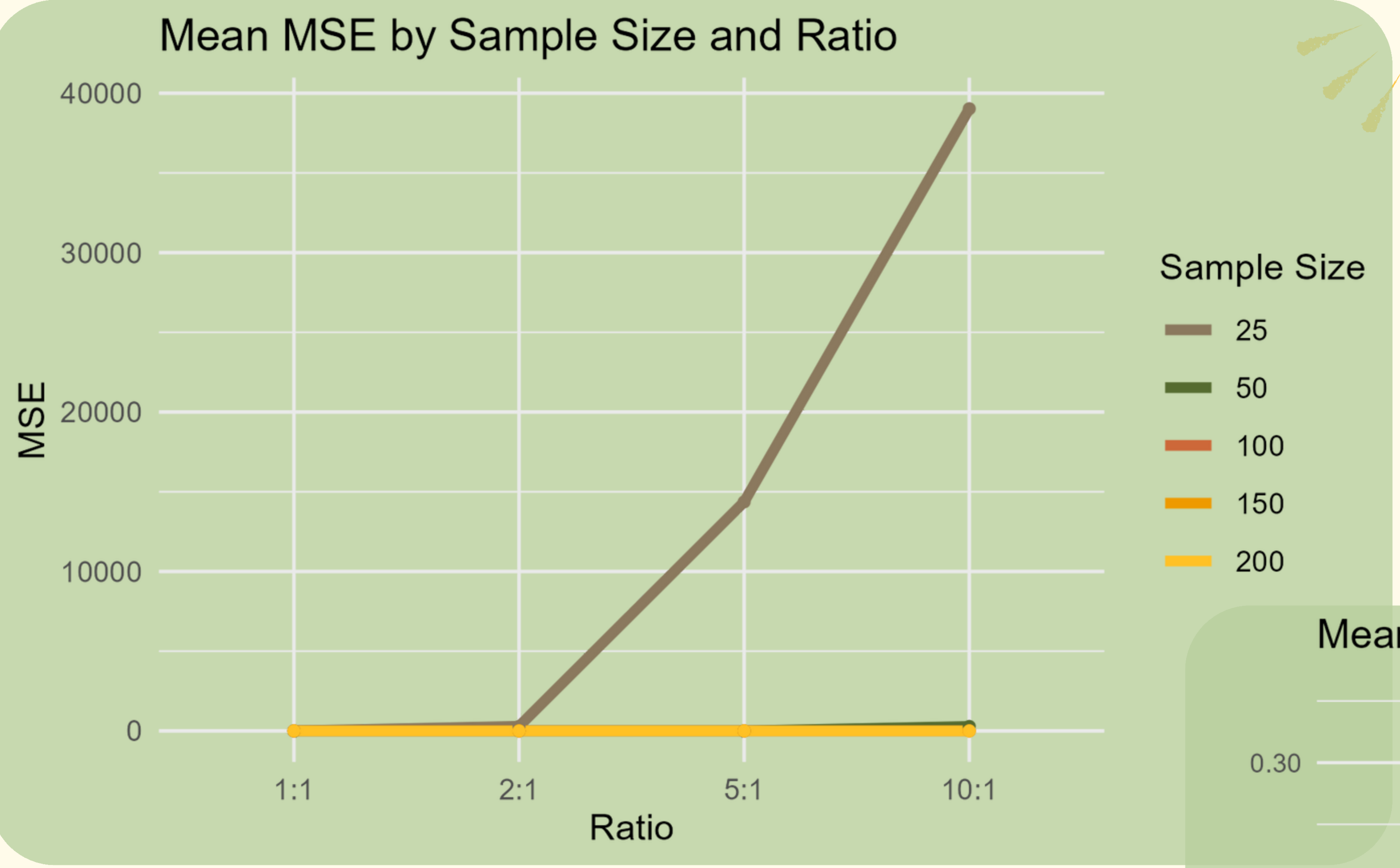

100
150
200
3
0
-3
-6
Bias
1:1
2:1
5:1
10:1
1:1
2:1
5:1
10:1
1:1
2:1
5:1
10:1
Ratio of Pseudo-Absence Points

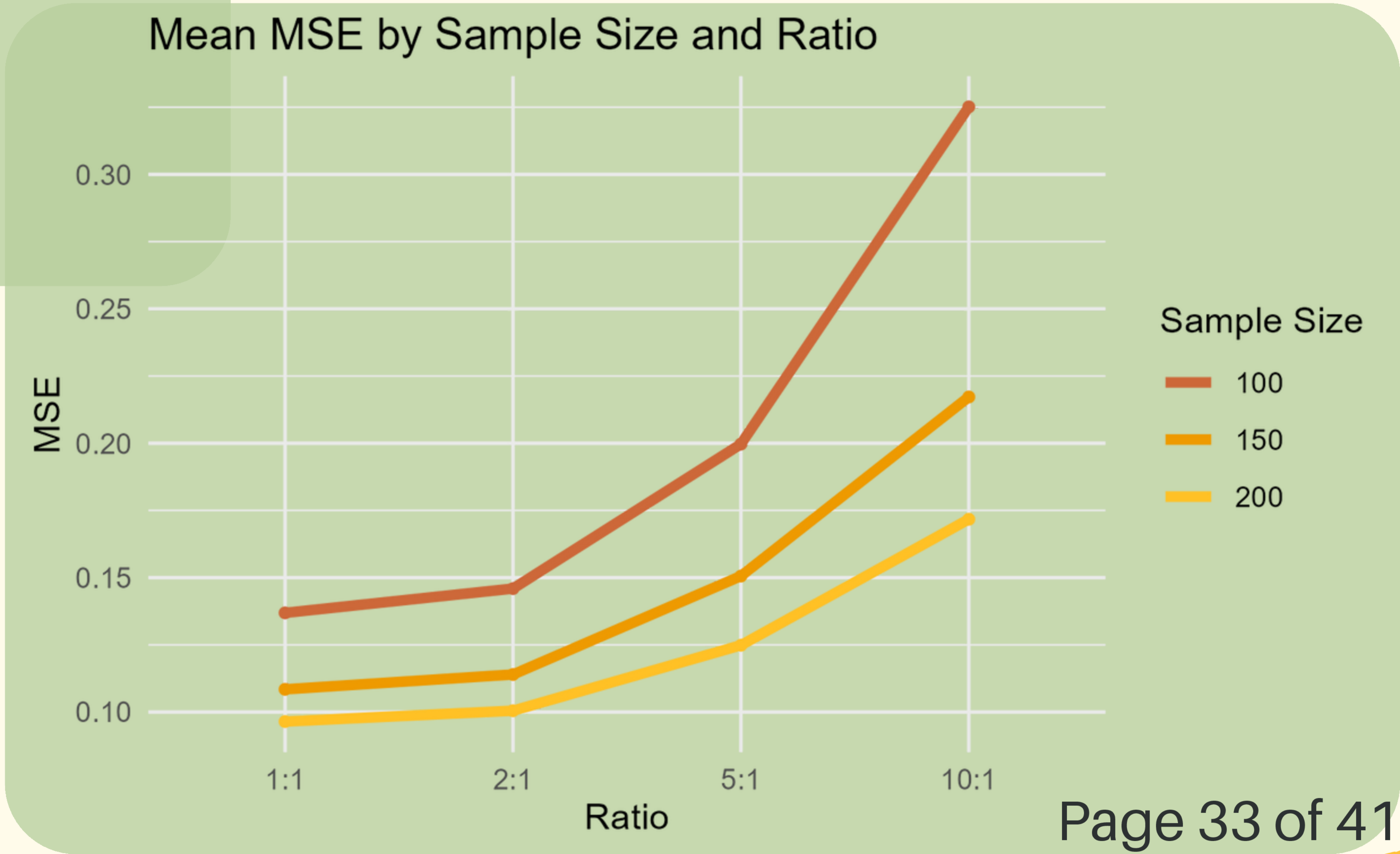

25
50
100
150
200
Bias
10000
5000
0
-5000
-10000
1:1
2:1
5:1
10:1
Ratio of Pseudo-Absence Points

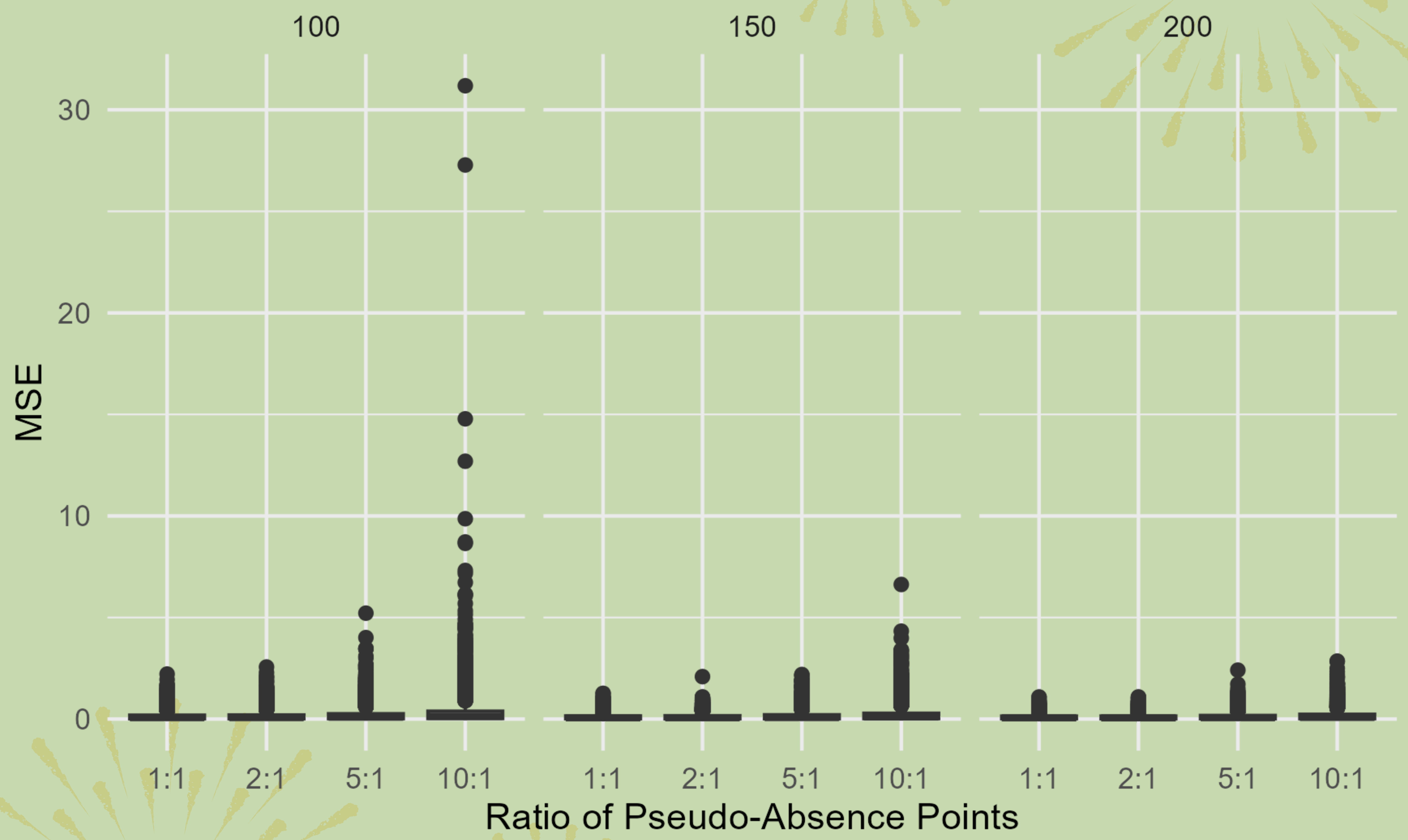

Mean MSE by Ratio
40000
30000
20000
10000
0
Mean Squared Error (MSE)
50
100
150
200
Sample Size (n=)
1:1
2:1
5:1
10:1

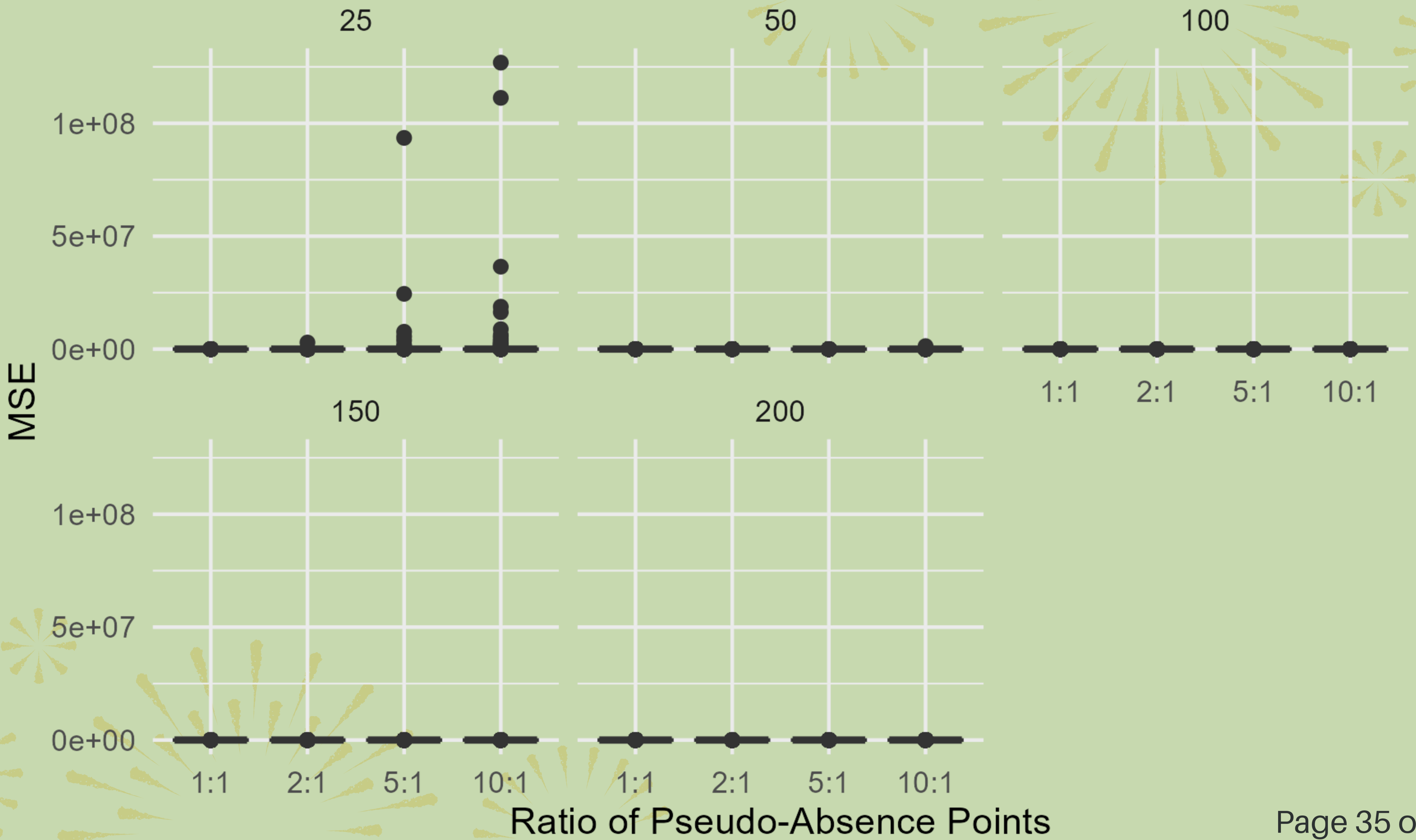

Mean MSE by Ratio
200
100
0
Mean Squared Error (MSE)
50
100
150
200
Sample Size (n=)
1:1
2:1
5:1
10:1

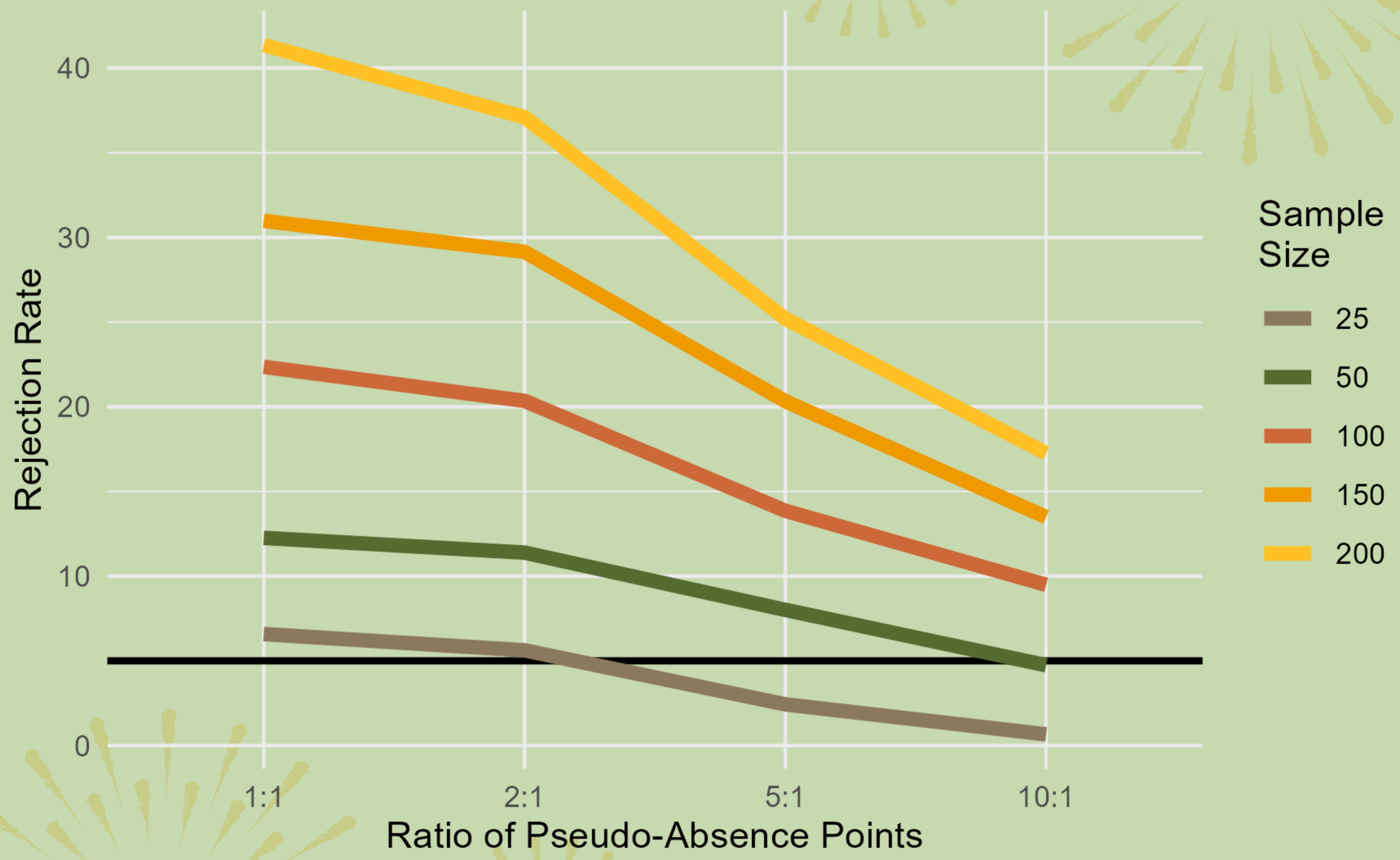

Mean MSE by Ratio
0.30
0.25
0.20
0.15
0.10
Mean Squared Error (MSE)
100
125
150
175
200
Sample Size (n=)
1:1
2:1
5:1
10:1


Mean MSE by Sample Size and Ratio
40000
30000
20000
10000
0
MSE
1:1
2:1
5:1
10:1
Ratio
Sample Size
25
50
100
150
200


Mean MSE by Sample Size and Ratio
0.30
0.25
0.20
0.15
0.10
MSE
1:1
2:1
5:1
10:1
Ratio
Sample Size
100
150
200


100
150
200
MSE
30
20
10
0
1:1
2:1
5:1
10:1
1:1
2:1
5:1
10:1
1:1
2:1
5:1
10:1
Ratio of Pseudo-Absence Points


25
50
100
150
200
MSE
1e+08
5e+07
0e+00
1:1
2:1
5:1
10:1
Ratio of Pseudo-Absence Points

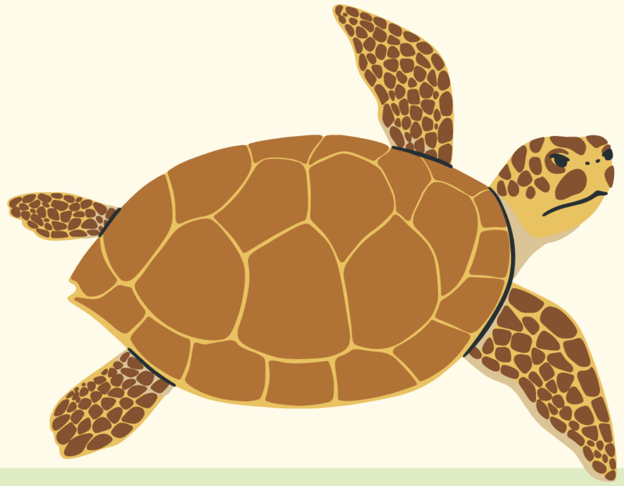

Rejection Rate
40
30
20
10
0
1:1
2:1
5:1
10:1
Ratio of Pseudo-Absence Points
Sample Size
25
50
100
150
200

Conclusions

- Ratio matters in smaller sizes
  - Hypothesis: Unbalanced data
- Saw a noticeable difference in bias, MSE, and rejection rates across different ratios.
- As the sample size increases, the bias and MSE decrease
- and the rejection rate increases
- A 10:1 ratio is not necessary and may be harmful in smaller datasets.

Future Research

- Manuscript plans:
  - Further examine imbalance in nesting outcome
    - Logistic regression for rare events - Firth correction
  - Include additional predictors with multiple logistic regression
  - Confusion matrix to examine classification



Bibliography

1. Barbet-Massin, Morgane, et al. “Selecting Pseudo-Absences for Species Distribution Models: How, Where, and How Many?” Methods in Ecology and Evolution, vol. 3, no. 2, 19 Jan. 2012, dx.doi.org/10.1111/j.2041-210X.2011.00172.x.
2. Chefaoui, Rosa M., and Jorge M. Lobo. “Assessing the Effects of Pseudo-Absences on Predictive Distribution Model Performance.” Ecological Modelling, vol. 210, no. 4, Feb. 2008, pp. 478–486, https://doi.org/10.1016/j.ecolmodel.2007.08.010. Accessed 24 Feb. 2020.
3. Chiara Mancino, Daniele Canestrelli, Luigi Maiorano, Going west: Range expansion for loggerhead sea turtles in the Mediterranean Sea under climate change, Global Ecology and Conservation, Volume 38, 2022, e02264, ISSN 2351-9894, https://doi.org/10.1016/j.gecco.2022.e02264. (https://www.sciencedirect.com/science/article/pii/S2351989422002669)

Thank You

Let’s discuss. Any questions?

106