# Big Picture Science and Modeling

### What's the point of statistics?

- a process of learning through empirical observations

. . .

- together, science philosophy and statistical modeling is the backbone to empirical learning

### Today

**1.** Study Objectives, Hypotheses, and Predictions

**2.** Big Data and Sampling

**3.** Inference and Prediction

**4.** Model-Based vs Design-Based Sampling/Inference

. . .

*Lab*: Simulation and Markdown

## Study Objective

What you want to accomplish; can have multiple related objectives in a single manuscript.

Our objective is to understand the space-use of urban living coyotes.


Framing the importance of the objective(s) provides the *justification* and depends on the audience.

## Hypothesis

1. A story that explains how the world works
2. An explanation for an observed phenomenon

Coyotes have small home ranges in urban areas

## Research/Scientific Hypothesis

"A statement about a phenomenon that also includes the potential mechanism or cause of that phenomenon". (Betts et al. 2021)

Coyotes have small home ranges in urban areas because resource density is high, leading to reduced ranging.

### Statistical Model/Hypothesis

- An explicit mathematical and stochastic representation of the observational & mechanistic process of the empirical observations.

### Statistical Model/Hypothesis

Example:

$$
\begin{gathered}
\mathbf{y} = \beta_0 + \beta_1 \times \mathbf{x} + \\
\sim \text{Normal}(0, \sigma^2)
\end{gathered}
$$

where…

$\mathbf{y}$ = vector of home range sizes of coyotes $\beta_0$ = intercept $\beta_1$ = effect diff. of HR size for urban coyotes $\mathbf{x}$ = indicator of HR in urban (1) or not in urban (0) $\sigma^2$ = uncertainty / unknown variability

### Statistical Model/Hypothesis

Example:

$$
\begin{gathered}
\mathbf{y} = \beta_0 + \beta_1 \times \mathbf{x} + \\
\sim \text{Normal}(0, \sigma^2)
\end{gathered}
$$

#### Evidence of hypothesis support

$\beta_1$ is negative and **statistically clearly different**[1] than zero

### Prediction

The expected outcome from a hypothesis. If agrees with data, it would support the hypothesis or at least not reject it.

1. Okay: Coyote home ranges are smaller in urban areas compared to non-urban areas

2. Better: Coyote home ranges in urban areas with high available food resources are smaller than coyote home ranges in urban areas with low available food resources and smaller than coyotes living in non-urban areas

## Types of Studies

- Descriptive/Naturalist (not hypothetico-deductive)
- Hypothetico-Deductive Observational
- Hypothetico-Deductive Experimental

## Take-Aways?

**1.** Study Objectives, Hypotheses, and Predictions

## Big Data Problems

"*The hidden Biases of Big Data*" by Kate Crawford (2013)

"with enough data, the numbers speak for themselves"- Wired Magazine Editor

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0,33 | 202,45 | 4862,46 | 1,21 | 1,45 | 0,85 | 8,51 | 6,35 | 300,48 | 1,15 | 3,38 | 0,30 | 2,19 | 0,78 | 1357,77 | 0,83 | 169,07 | 1,26 | 0,48 | 7,61 | 14,17 |
| 0,84 | 846,55 | 3484,47 | 1,64 | 3,52 | 0,52 | 6,76 | 9,41 | 1445,78 | 0,17 | 3,18 | 0,26 | 6,21 | 0,47 | 140,43 | 0,45 | 509,51 | 3,11 | 0,48 | 8,29 | 13,26 |
| 0,04 | 676,31 | 1427,13 | 2,87 | 3,28 | 0,62 | 1,97 | 9,46 | 1929,65 | 2,86 | 0,49 | 0,44 | 6,27 | 0,09 | 825,70 | 0,74 | 849,80 | 1,99 | 0,46 | 1,13 | 10,81 |
| 0,96 | 43,69 | 1611,10 | 0,77 | 3,69 | 0,26 | 6,47 | 3,11 | 893,24 | 0,58 | 1,07 | 0,61 | 4,08 | 0,65 | 1108,70 | 0,08 | 692,38 | 1,53 | 0,49 | 0,09 | 5,43 |
| 0,00 | 468,37 | 1310,22 | 2,71 | 0,70 | 0,47 | 6,93 | 9,02 | 1729,76 | 3,15 | 0,13 | 0,96 | 1,27 | 0,89 | 1101,45 | 0,49 | 198,52 | 0,20 | 0,73 | 4,08 | 5,49 |
| 0,71 | 762,70 | 4356,19 | 0,85 | 2,96 | 0,33 | 0,73 | 8,46 | 2463,40 | 2,51 | 1,43 | 0,95 | 6,26 | 0,83 | 1197,00 | 0,63 | 1338,64 | 3,22 | 0,45 | 4,91 | 14,47 |
| 0,95 | 84,85 | 457,84 | 2,98 | 3,55 | 0,29 | 4,14 | 6,52 | 416,30 | 2,30 | 2,07 | 0,62 | 6,94 | 0,76 | 675,53 | 0,45 | 1172,56 | 0,34 | 0,19 | 8,28 | 3,40 |
| 0,86 | 292,41 | 3760,23 | 0,77 | 0,59 | 0,25 | 4,86 | 0,06 | 999,49 | 1,79 | 2,81 | 0,63 | 6,39 | 0,64 | 1041,13 | 0,94 | 640,03 | 2,99 | 0,68 | 3,64 | 4,16 |
| 0,35 | 822,09 | 910,04 | 1,35 | 1,87 | 0,64 | 7,83 | 1,02 | 653,83 | 1,50 | 4,00 | 0,21 | 5,11 | 0,64 | 880,67 | 0,92 | 319,24 | 2,42 | 0,86 | 6,15 | 13,39 |
| 0,92 | 1389,45 | 4532,60 | 1,16 | 2,04 | 0,31 | 8,12 | 13,82 | 2281,72 | 2,02 | 3,17 | 0,45 | 3,48 | 0,89 | 1492,56 | 0,91 | 808,82 | 2,13 | 0,88 | 7,44 | 14,87 |
| 0,20 | 92,57 | 1702,20 | 3,71 | 1,66 | 0,04 | 0,96 | 13,73 | 2695,60 | 3,95 | 0,04 | 0,94 | 9,50 | 0,69 | 130,65 | 0,34 | 1200,34 | 1,39 | 0,64 | 8,65 | 9,02 |
| 0,12 | 794,10 | 4055,57 | 3,31 | 3,90 | 0,60 | 8,27 | 1,11 | 2765,63 | 3,85 | 3,52 | 0,48 | 1,10 | 0,07 | 960,19 | 0,17 | 1022,54 | 0,26 | 0,27 | 0,83 | 3,16 |
| 0,26 | 515,07 | 4431,19 | 3,36 | 0,10 | 0,44 | 7,40 | 8,96 | 581,78 | 0,59 | 3,04 | 0,12 | 9,75 | 0,89 | 675,84 | 0,14 | 106,58 | 3,39 | 0,47 | 6,61 | 5,54 |
| 0,30 | 181,71 | 4448,61 | 3,71 | 0,87 | 0,14 | 6,89 | 14,14 | 240,59 | 3,74 | 0,27 | 0,59 | 5,33 | 0,22 | 469,72 | 0,28 | 131,10 | 2,60 | 0,34 | 7,00 | 8,19 |
| 0,67 | 1021,48 | 559,76 | 1,88 | 1,42 | 0,28 | 5,23 | 0,83 | 890,83 | 1,05 | 1,87 | 0,21 | 8,26 | 0,38 | 1056,64 | 0,43 | 139,13 | 0,70 | 0,23 | 3,10 | 13,44 |
| 0,25 | 875,30 | 2191,24 | 3,22 | 3,10 | 0,10 | 9,60 | 2,21 | 1446,21 | 2,09 | 0,89 | 0,76 | 8,62 | 0,87 | 375,25 | 0,95 | 953,54 | 0,65 | 0,50 | 2,68 | 2,09 |
| 0,32 | 537,19 | 3975,58 | 2,19 | 0,88 | 0,99 | 3,28 | 1,79 | 2226,90 | 2,19 | 1,22 | 0,13 | 8,65 | 0,96 | 79,53 | 0,08 | 540,44 | 1,12 | 0,72 | 1,66 | 11,85 |
| 0,75 | 78,39 | 5775,90 | 1,76 | 2,19 | 0,40 | 8,39 | 14,70 | 497,83 | 2,88 | 1,62 | 0,58 | 8,71 | 0,87 | 230,22 | 0,39 | 1094,07 | 3,10 | 0,40 | 1,29 | 8,73 |
| 0,09 | 1220,90 | 1320,91 | 0,01 | 0,88 | 0,10 | 5,21 | 7,38 | 1757,83 | 0,57 | 0,57 | 0,13 | 5,40 | 0,62 | 1460,17 | 0,18 | 310,98 | 1,82 | 0,63 | 9,16 | 11,20 |
| 0,49 | 938,05 | 252,56 | 2,09 | 1,83 | 0,47 | 0,96 | 8,67 | 1892,22 | 3,64 | 2,13 | 0,93 | 7,78 | 0,73 | 861,45 | 0,11 | 562,53 | 1,69 | 0,67 | 1,71 | 0,49 |
| 0,05 | 1079,24 | 4665,57 | 0,18 | 0,30 | 0,34 | 7,75 | 8,75 | 1758,69 | 3,13 | 2,99 | 0,76 | 3,66 | 0,84 | 656,11 | 0,98 | 1267,72 | 3,61 | 0,40 | 0,01 | 6,05 |
| 0,84 | 1333,38 | 4464,09 | 0,14 | 2,81 | 0,21 | 0,94 | 0,99 | 1734,68 | 0,57 | 1,23 | 0,65 | 9,99 | 0,36 | 922,84 | 0,35 | 640,88 | 1,62 | 0,08 | 9,62 | 2,66 |
| 0,01 | 651,62 | 4129,64 | 2,23 | 1,94 | 0,02 | 5,38 | 8,18 | 348,67 | 3,64 | 1,23 | 0,15 | 7,13 | 0,77 | 140,68 | 0,73 | 342,64 | 2,66 | 0,04 | 8,82 | 13,42 |
| 0,82 | 1204,43 | 5247,37 | 1,44 | 2,70 | 0,20 | 7,50 | 0,55 | 2087,03 | 3,56 | 3,59 | 0,79 | 4,62 | 0,23 | 361,32 | 0,59 | 1418,51 | 3,75 | 0,17 | 5,84 | 2,08 |
| 0,74 | 1354,16 | 1633,83 | 1,56 | 1,07 | 0,82 | 3,90 | 10,01 | 307,46 | 3,09 | 3,33 | 0,52 | 7,11 | 0,54 | 478,56 | 0,96 | 438,50 | 3,50 | 0,03 | 1,29 | 5,35 |
| 0,13 | 741,44 | 485,14 | 1,25 | 0,84 | 0,84 | 7,59 | 0,30 | 271,90 | 3,98 | 2,94 | 0,27 | 8,50 | 0,66 | 123,32 | 0,96 | 462,81 | 1,02 | 0,93 | 6,54 | 0,49 |
| 0,42 | 1072,16 | 4361,62 | 0,12 | 1,44 | 0,26 | 7,68 | 9,73 | 228,51 | 0,85 | 2,01 | 0,26 | 1,76 | 0,40 | 1375,86 | 0,69 | 461,35 | 0,48 | 0,57 | 0,30 | 9,27 |
| 0,04 | 1385,13 | 2222,11 | 0,20 | 0,36 | 0,06 | 1,55 | 1,59 | 2402,97 | 0,52 | 3,31 | 0,21 | 6,95 | 0,29 | 325,73 | 0,17 | 814,54 | 0,27 | 0,18 | 0,24 | 7,78 |
| 0,44 | 1234,49 | 5302,87 | 3,09 | 1,51 | 0,76 | 6,82 | 7,28 | 530,72 | 1,78 | 1,57 | 0,57 | 4,13 | 0,15 | 126,16 | 0,33 | 460,85 | 1,62 | 0,07 | 2,04 | 9,10 |
| 0,94 | 1011,60 | 243,58 | 1,82 | 0,56 | 0,64 | 8,03 | 12,96 | 819,19 | 1,09 | 3,70 | 0,68 | 6,64 | 0,74 | 1350,17 | 0,54 | 830,75 | 2,45 | 0,31 | 2,73 | 6,43 |
| 0,76 | 137,92 | 4465,07 | 2,16 | 2,40 | 0,60 | 1,73 | 11,90 | 887,07 | 1,28 | 2,47 | 0,29 | 2,79 | 0,66 | 177,63 | 0,88 | 1339,96 | 1,95 | 0,00 | 5,16 | 9,69 |
| 0,47 | 23,55 | 1498,98 | 2,88 | 0,39 | 0,16 | 5,57 | 11,38 | 390,01 | 1,58 | 2,90 | 0,78 | 3,87 | 0,71 | 1334,75 | 0,51 | 408,46 | 3,02 | 0,21 | 7,18 | 11,83 |
| 0,19 | 848,16 | 4611,45 | 2,82 | 2,89 | 0,33 | 8,28 | 4,65 | 2046,08 | 1,64 | 0,82 | 0,91 | 7,42 | 0,93 | 1298,08 | 0,65 | 355,08 | 0,18 | 0,44 | 4,56 | 11,40 |
| 0,74 | 1089,34 | 5220,37 | 0,14 | 1,69 | 0,51 | 8,99 | 5,03 | 2443,55 | 0,84 | 1,18 | 0,53 | 7,29 | 0,30 | 1324,57 | 0,38 | 86,08 | 2,10 | 0,83 | 7,54 | 0,37 |
| 0,37 | 1407,80 | 3176,49 | 3,81 | 2,47 | 0,36 | 1,83 | 3,89 | 2491,19 | 2,13 | 3,83 | 0,00 | 6,97 | 0,82 | 1064,45 | 0,80 | 989,09 | 1,66 | 0,86 | 7,65 | 10,32 |
| 0,65 | 693,17 | 3245,95 | 3,05 | 1,86 | 0,40 | 2,56 | 8,90 | 352,00 | 3,25 | 0,60 | 0,10 | 2,68 | 0,13 | 1300,58 | 0,57 | 277,47 | 2,99 | 0,70 | 5,83 | 9,82 |
| 0,36 | 538,38 | 3348,08 | 1,25 | 0,17 | 0,85 | 5,89 | 1,36 | 1139,67 | 1,29 | 1,72 | 0,36 | 4,04 | 0,60 | 1102,77 | 0.47 | 599,78 | 2,01 | 0,82 | 1,05 | 4,71 |
| 0,19 | 1462,37 | 4621,41 | 1,66 | 2,26 | 0,64 | 0,07 | 2,19 | 2428,62 | 1,25 | 1,35 | 0,15 | 7,45 | 0,55 | 801,78 | 0,56 | 789,26 | 0,16 | 0,87 | 4,62 | 11,12 |
| 0,69 | 1052,61 | 1428,54 | 0,46 | 3,39 | 0,20 | 5,04 | 1,67 | 921,29 | 2,14 | 2,65 | 0,99 | 5,42 | 0,94 | 837,13 | 0,42 | 1149,36 | 3,20 | 0,08 | 5,03 | 3,39 |
| 0,25 | 680,56 | 5268,96 | 0,04 | 3,90 | 0,80 | 6,04 | 14,78 | 253,83 | 1,82 | 2,96 | 0,55 | 5,47 | 0,92 | 1181,12 | 0,43 | 319,59 | 0,55 | 0,41 | 8,73 | 5,36 |
| 0,40 | 627,42 | 5719,07 | 0,32 | 2,41 | 0,64 | 2,94 | 13,76 | 77,46 | 2,50 | 2,11 | 0,64 | 4,52 | 0,59 | 1445,81 | 0,57 | 1141,31 | 3,50 | 0,14 | 5,54 | 8,03 |
| 0,84 | 1328,00 | 5822,49 | 1,07 | 1,46 | 0,79 | 4,03 | 0,66 | 1282,37 | 3,60 | 3,58 | 0,33 | 4,88 | 0,88 | 941,87 | 0,46 | 955,18 | 2,48 | 0,58 | 3,18 | 7,20 |
| 0,15 | 180,07 | 3312,23 | 3,30 | 3,63 | 0,79 | 0,65 | 10,16 | 264,08 | 3,32 | 2,00 | 0,34 | 9,79 | 0,98 | 212,20 | 0,12 | 1439,36 | 0,97 | 0,34 | 5,69 | 8,44 |
| 0,52 | 253,05 | 5288,15 | 0,47 | 3,35 | 0,05 | 8,28 | 7,81 | 141,28 | 0,39 | 2,68 | 0,54 | 6,93 | 0,49 | 1355,70 | 0,48 | 1416,98 | 1,66 | 0,02 | 3,06 | 1,67 |
| 0,43 | 482,62 | 2222,80 | 0,10 | 0,75 | 0,65 | 7,42 | 8,76 | 879,10 | 3,49 | 1,55 | 0,25 | 8,52 | 0,62 | 1449,49 | 0,95 | 527,27 | 2,75 | 0,19 | 2,00 | 11,06 |
| 0,26 | 777,78 | 1632,70 | 2,88 | 1,89 | 0,56 | 7,35 | 8,57 | 716,36 | 3,39 | 1,27 | 0,44 | 2,70 | 0,81 | 549,92 | 0,88 | 457,23 | 3,70 | 0,96 | 7,01 | 6,87 |
| 0,90 | 1052,10 | 3363,54 | 1,55 | 1,38 | 0,39 | 1,18 | 9,84 | 2633,26 | 1,08 | 1,24 | 0,66 | 2,21 | 0,21 | 454,82 | 0,33 | 1475,64 | 1,60 | 0,64 | 8,21 | 2,49 |
| 0,67 | 894,53 | 147,80 | 1,48 | 2,40 | 0,03 | 7,09 | 7,72 | 1356,17 | 1,29 | 0,24 | 0,04 | 7,06 | 0,63 | 62,90 | 0,05 | 1185,04 | 3,52 | 0,59 | 0,46 | 1,15 |

## Using assumptions to move from correl

Data never speak by themselves. To derive meaningful c
on well-defined hypotheses, statistical models grounde
testable and untestable assumptions[57,58,59]. The import
statistical models, and statistical assumptions is well kr

- Correia et al. 2026. Best practices for moving from correlation to causation in ecological research. Nature Communications.

#### Big Data Problems

"*The hidden Biases of Big Data*" by Kate Crawford (2013)

. . .

#### Big Data Problems

**The Annals of Applied Statistics (2018); Xiao Li Meng,**

. . .

. . .

Sampling Processes == Human Design

#### Big Data Problems

**Bradley et al. 2021 (Nature)**

### Big Data Problems

Using eBird data w/o accounting for sampling biases.

. . .

**PNAS**

## Global abundance estimates for 9,700 bird species

Corey T. Callaghan, Shinichi Nakagawa, and William K. Cornwell Authors Info & Affiliations

**May 17, 2021** | 118 (21) e2023170118 | https://doi.org/10.1073/pnas.2023170118

## Extreme uncertainty and unquantifiable bias do not inform population sizes

Orin J. Robinson, Jacob B. Socolar, Erica F. Stuber, +30, and Alison Johnston Authors Info & Affiliations

**March 1, 2022** | 119 (10) e2113862119 | https://doi.org/10.1073/pnas.2113862119

Link1. Link2.

### The Questioning Scientist

In regard to *data* and *statistical models*, 21st century scientists should be pragmatic, excited, and questioning.

- How and why did these data come to be?
  - *understand how the data came to be*
  - *ask this even when you design the study, after data collection*

## The Questioning Scientist

In regard to *data* and *statistical models*, 21$^{st}$ century scientists should be pragmatic, excited, and questioning.

- What do these data look like?
  - *visualize the data in many dimensions*
  - *keep in mind - not all outcomes are visible in data – example?*

## The Questioning Scientist

In regard to *data* and *statistical models*, 21$^{st}$ century scientists should be pragmatic, excited, and questioning.

- How does this statistical model work?
  - *statistical notation, explicit and implicit assumptions*, *optimization*

## The Questioning Scientist

In regard to *data* and *statistical models*, 21$^{st}$ century scientists should be pragmatic, excited, and questioning.

- How does this statistical model fail in theory and in practice?
  - *statistical robustness and identifiability*

## Data vs. Information

. . .

- the question being asked of the data
- how the data came to be

## The question and the data

Surveillance monitoring data will generally have *lower quality information* to answer post-hoc hypotheses when compared to a designed study with a priori hypotheses.

## The goal of the question

- learn about the data (data summary)
- apply learning outside of the data to a 'population' (inference)
- learn about conditions relevant to but not observed in the data (prediction)

. . .

. . .

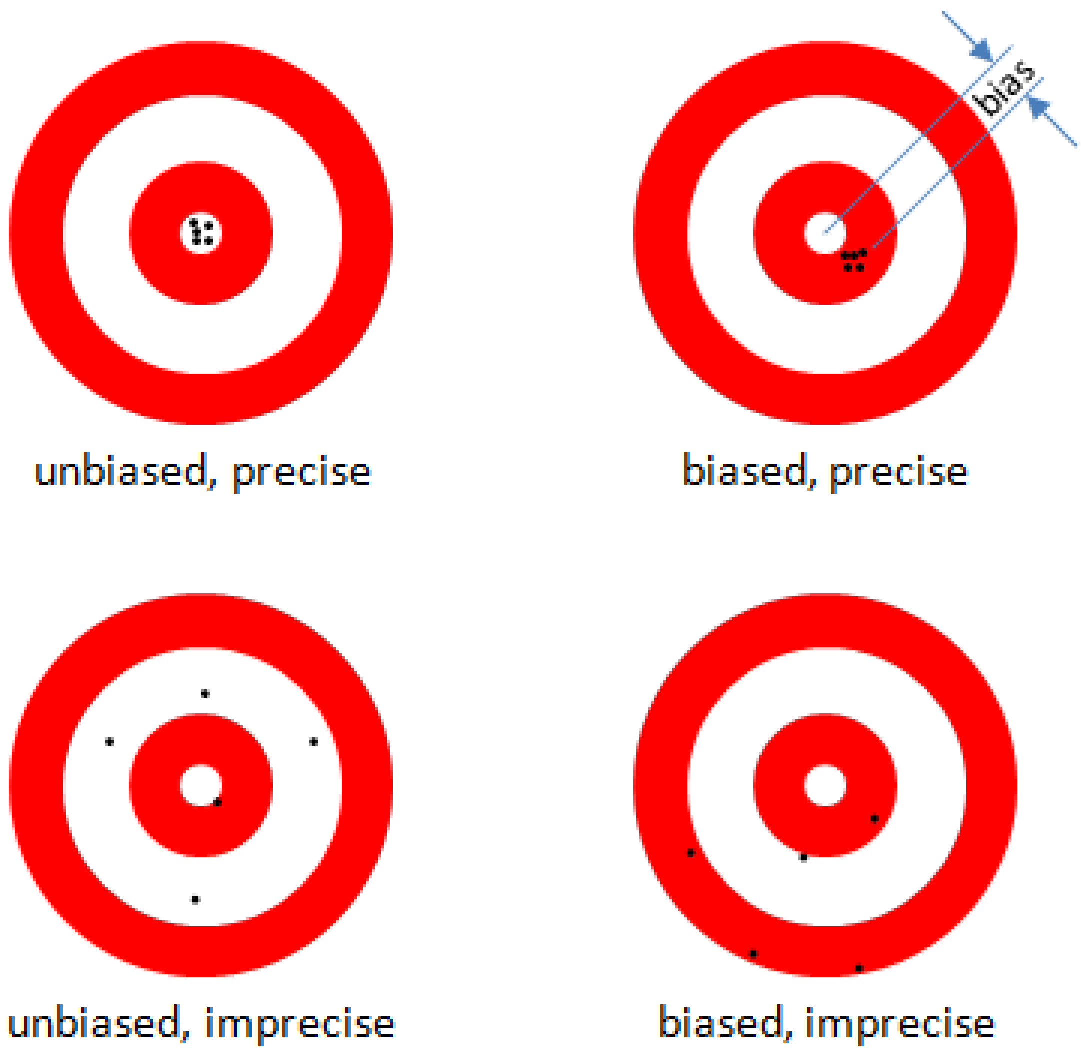
## Take-Aways

**1.** Study Objectives, Hypotheses, and Predictions

**2.** Big Data and Sampling

**Which is worse?**

- unbiased imprecise result
- biased precise result



### Inference and Prediction

Explanatory modeling focuses on minimizing (statistical) bias to obtain the most accurate representation of the underlying theory.

. . .

Predictive modeling focuses on minimizing both bias and estimation variance; this may sacrifice theoretical accuracy for improved empirical precision.

### Inference and Prediction

. . .

BUT ...

### Inference and Prediction

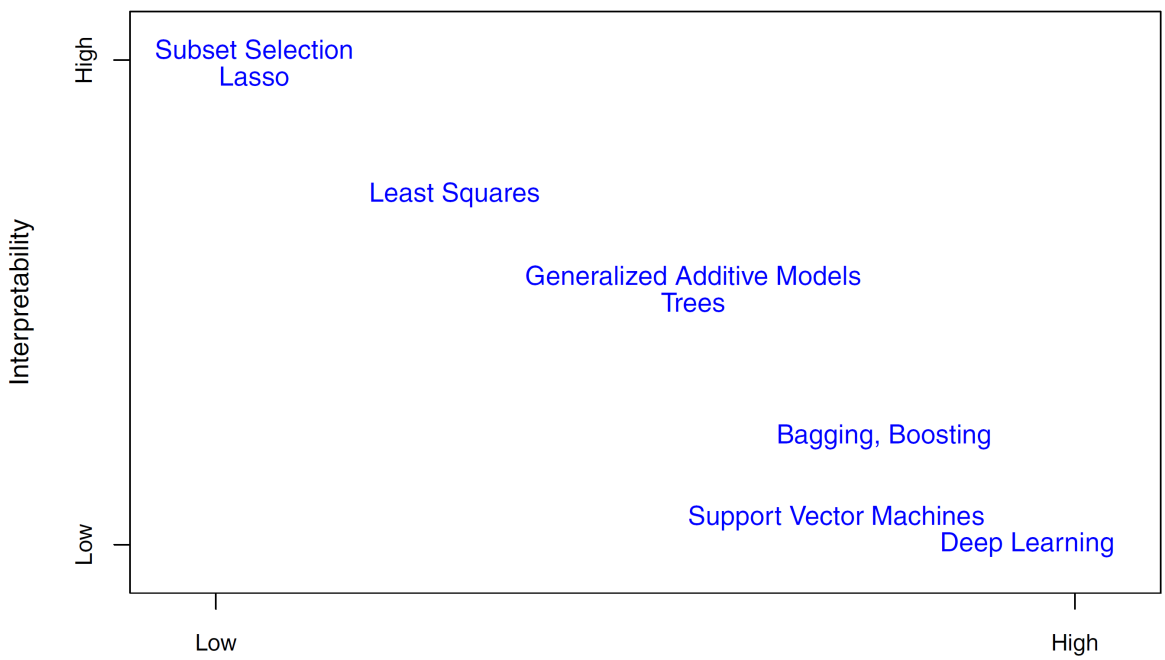



**FIGURE 2.7.** *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

## Take-Aways

**1.** Study Objectives, Hypotheses, and Predictions

**2.** Big Data and Sampling

**3.** Inference and Prediction

## Design-Based

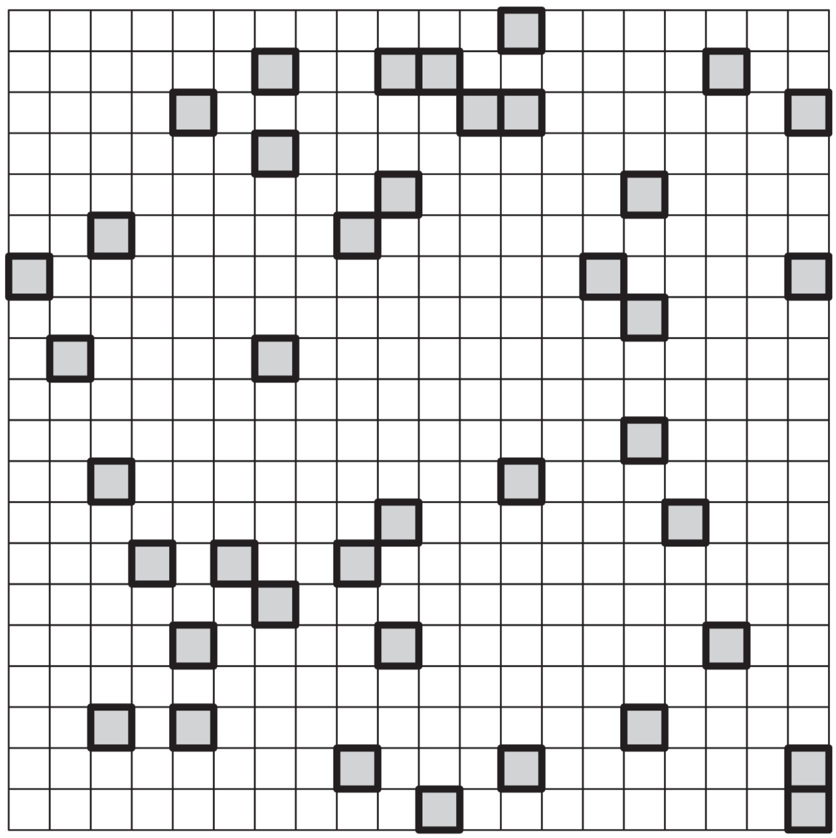


**Figure 2.2.** Another simple random sample of 40 units.

Thompson, 2012. Sampling.

The sample and population are what??

## Design-Based

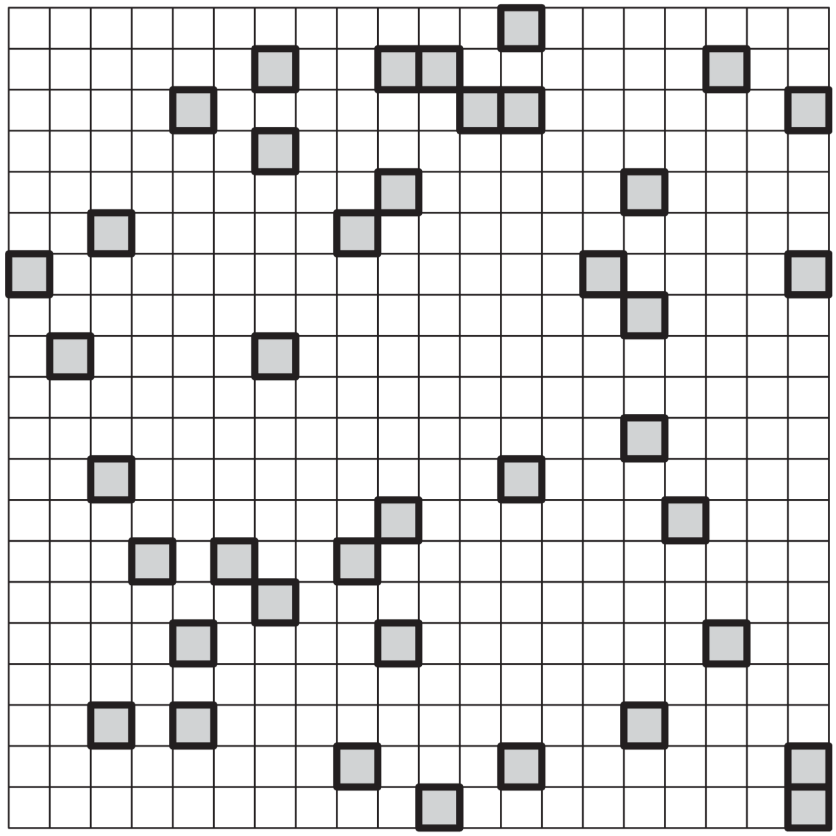


**Figure 2.2.** Another simple random sample of 40 units.

- inference relies on probabilistic assigning some units to be in the sample (e.g., random sampling).

## Design-Based

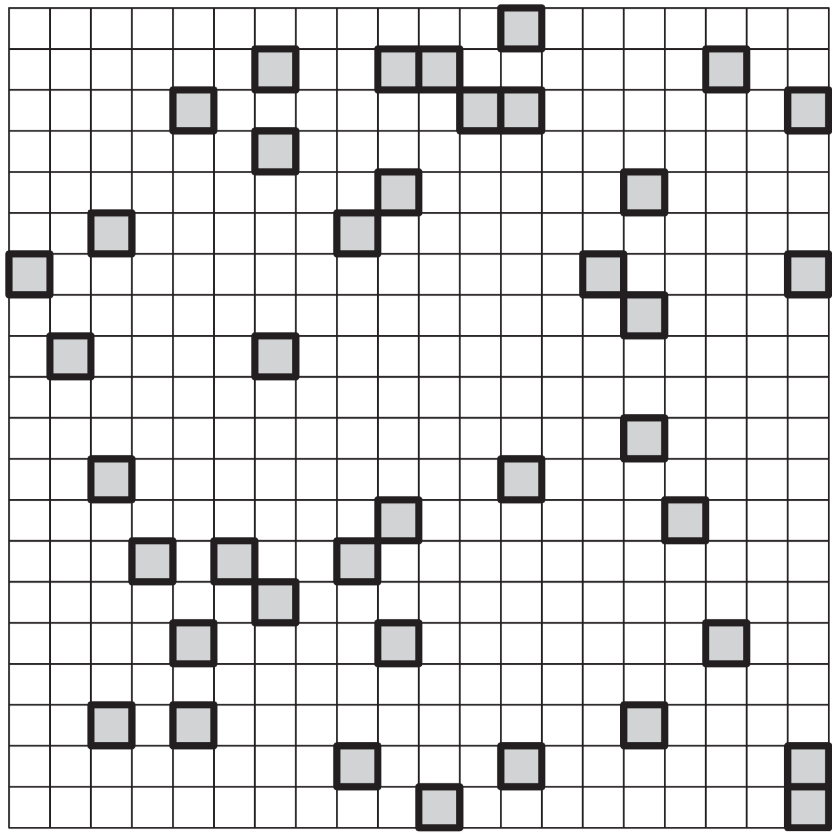
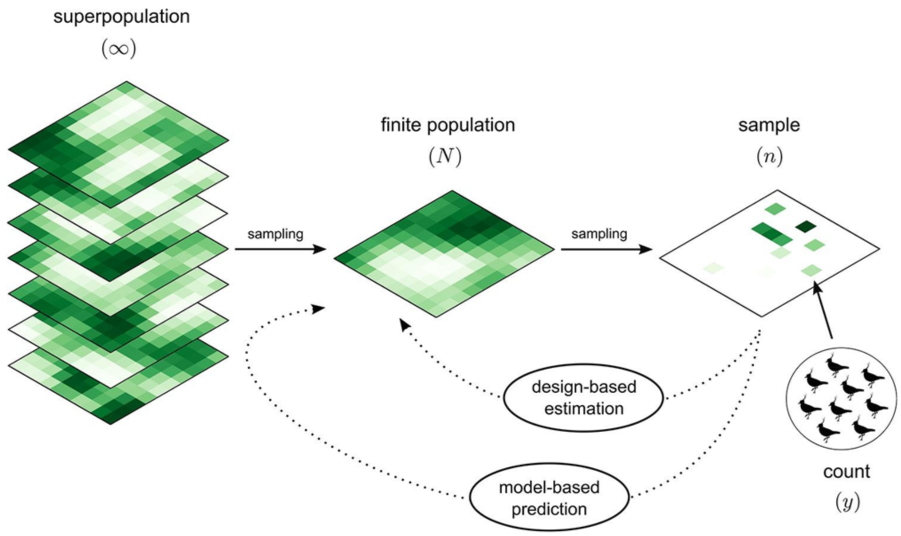


**Figure 2.2.** Another simple random sample of 40 units.

- the values themselves are held to be **fixed**, whereas the sampling process is random.

## Two types of Inference




Aubry, P., & Francesiaz, C. (2022). On comparing design-based estimation versus model-based prediction to assess the abundance of biological populations. Ecological Indicators, 144, 109394.

## Design-Based

- Key Strengths: the population of interest is often defined (e.g., grid area); does not relying on stochastic models representing the structure of the data for reliable inference

. . .

- Key Weaknesses: limited in application; still requires models to accommodate observational processes, such as detection probability

## Design-Based

- $\mathbf{Y} = [y_1,...,y_N]$

. . .

- $\mathbf{y} = [y_1,...,y_n]$

### Design-Based

- $\mathbf{Y} = [y_1,...,y_N]$
- $\mathbf{y} = [y_1,...,y_n]$
- The population mean is $\bar{Y} = \sum_{i=1}^{N} Y_i/N$ and the sample mean is $\hat{\bar{y}} = \sum_{i=1}^{n} y_i/n$

. . .

- The population mean describes ....?

. . .

- $\mathbf{y}$ is a random vector that has $n$ random values, e.g., one sample of 4 cells.

. . .

$y = [y_1 \quad y_2 \quad y_3 \quad y_4]$

. . .

$$y' = y^T = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

### Random Variable

Wikipedia: A random variable (also called 'random quantity' or 'stochastic variable') is a mathematical formalization of a quantity or object which depends on random events.

. . .

We observe samples from the domain or population or sampling frame.

. . .

**Samples** are observed with some probability.

### Statistic

- $\hat{\bar{y}}$ is a 'statistic' (# computed from a sample) and is also a random variable

. . .

- statistics have a sampling distribution, describing the probability associated to observing different values of the statistic

## Design-Based Code

### TRUTH

```
# Random discrete uniform sampler
rdu <- function(n, lower, upper){
                              sample(lower:upper,
                                     n,
                                     replace=TRUE
                                     )
                             }
N = 25

# This is our truth, which we will sample from
set.seed(4343)
mat = matrix(rdu(N,
                 lower = 0,
                 upper = 400
                 ),
             nrow = 5,
             ncol = 5
             )
```

## Design-Based Code

### TRUTH

```
     [,1] [,2] [,3] [,4] [,5]
[1,]   63  390  209  376  331
[2,]  318  368   51  222  105
[3,]  296  288   82  200  217
[4,]  267   78   82  139  236
[5,]  360  139  369   97  339
```

What is the population mean for this truth?

```
mean(mat)
```

```
[1] 224.88
```

## Design-Based Code

What is the theoretical population mean that generated these values?

. . .

$$E[Y] = \sum_{i=1}^{N} y \times P(Y = y) = \sum_{i=1}^{N} y \times 1/N$$

```r
sum(0:400 * 1/401)
```

```
[1] 200
```

$$E[Y] = \frac{0 + 400}{2} = 200$$

Wikipedia

## Design-Based Code

### Random Sampling

```r
n = 10
y = sample(
          c(mat),
          n,
          replace = TRUE
          )
y
```

```
 [1] 139 209 139 200 390 368  97 288 105 267
```

### Estimator for this truth's population mean

```r
mean(y)
```

```
[1] 220.2
```

## Design-Based Code

### Sampling Distribution

```
# How many ways can we uniquely sample 10 things from 25
combs = function(n, x) {
  factorial(n) / factorial(n-x) / factorial(x)
}

combs(N, n)
```

```
[1] 3268760
```

## Design-Based Code

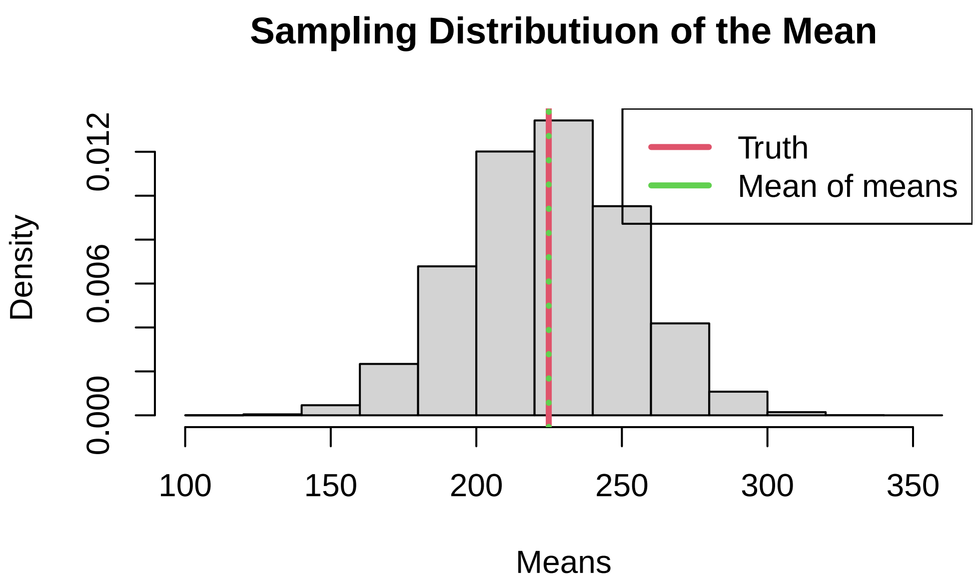
Get every combination and then calculate the mean for each sample of 10

```
  set.seed(5411)
  all.combs = utils::combn(c(mat),
                           n,
                           simplify = TRUE
                           )
  dim(all.combs)
```

```
[1]      10 3268760
```

```
  mean.all.combs = apply(all.combs,
                         2,
                         mean
                         )
```

**Design-Based Code**




**Design-Based Code**

OR, we can sample enough times to approximate it

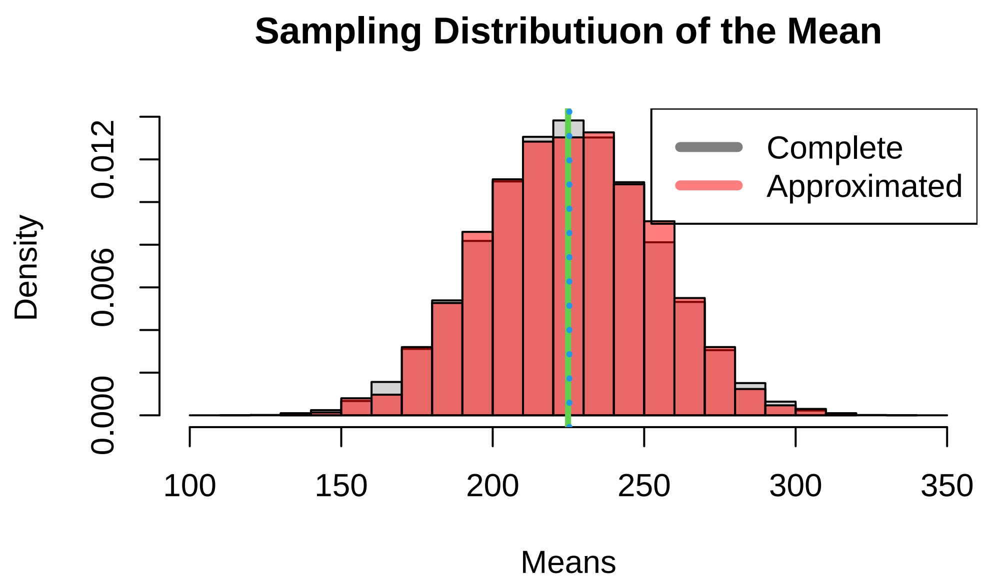
```
set.seed(54352)
sim.sampling.dist=replicate(3000,
                            sample(c(mat),n)
                            )

dim(sim.sampling.dist)
```

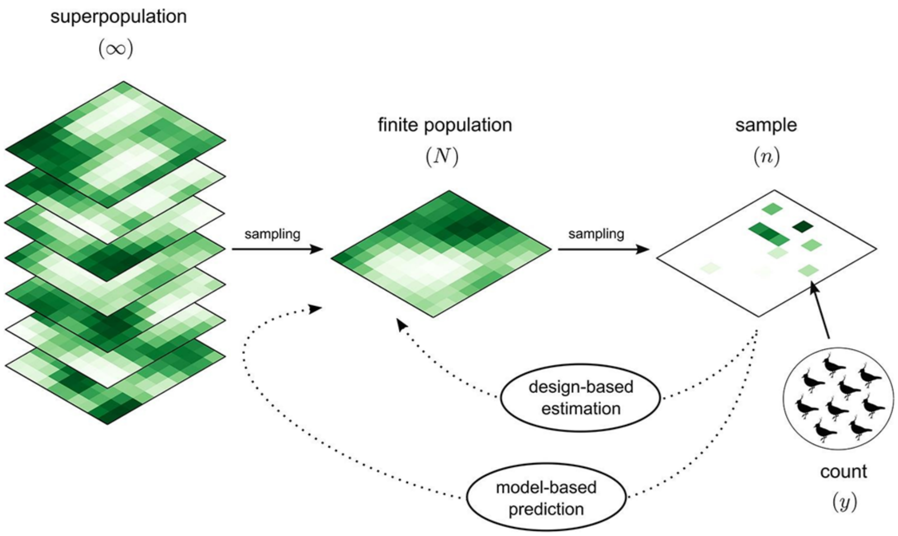
```
[1]   10 3000
```

```
mean.samples = apply(sim.sampling.dist,2,mean)
```

**Design-Based Code**



## Two types of Inference




Aubry, P., & Francesiaz, C. (2022). On comparing design-based estimation versus model-based prediction to assess the abundance of biological populations. Ecological Indicators, 144, 109394.

## Model-Based

### Inference relies on ...

"a statistical model describing how observations on population units are thought to have been generated from a super-population with potentially infinitely many observations for each unit;" Williams and Brown, 2019

. . .

"The analysis need not account for sampling randomization, because the sample is considered fixed. However, the unit values are considered random." Williams and Brown, 2019

## Model-Based

BUT....

when linking 'unit values' in a model, we need to account for their dependence.

. . .

Randomization allows us to make conditional independence claims among data in our sample, thus the model is simpler.

. . .

$P(y_2|y_1) = P(y_2)$

## Model-Based

- Key Strengths: Very flexible. Modeling is *magic*.

. . .

- Key Weaknesses:
  - Can be difficult to assess assumptions
  - Sampling frame is not always clear and thus the population you are infering to is not entirely clear

## Model-Based

Wikipedia link for Poisson

- $\mathbf{y} \sim \text{Poisson}(\lambda)$
- $y_i \sim \text{Poisson}(\lambda)$

. . .

- $\lambda$ is the population mean and variance

. . .

- Population mean estimator $\lambda = \sum_{i=1}^{N} Y_i/N$
- Sample mean estimator $\hat{\lambda} = \sum_{i=1}^{n} y_i/n$

. . .

- Maximum-Likelihood Estimate (MLE)

## Model-Based Code

```
# Create a function, to be replicated
  lambda = 200

  mat.fn = function(lambda){matrix(
                                   rpois(N, lambda = lambda),
                                   nrow = 5,
                                   ncol = 5
                                  )
                            }
```

### Model-Based Code

```
# Repeat the function n.sim times
  n.sim = 1000
  list.mat = replicate(n.sim,
                       mat.fn(lambda),
                       simplify = FALSE
                       )
  length(list.mat)
```

```
[1] 1000
```

. . .

```
# One realization from the super-population
  list.mat[[1]]
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]  170  196  179  214  197
[2,]  214  201  217  216  189
[3,]  176  195  176  212  190
[4,]  201  194  191  210  212
[5,]  195  205  203  208  192
```

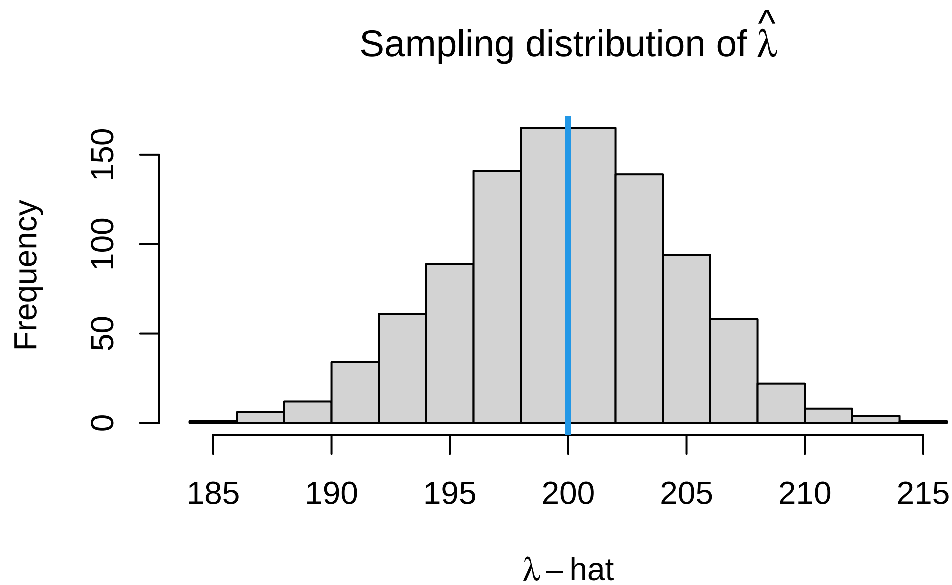
```
# Population mean for the first realization
  mean(list.mat[[1]])
```

```
[1] 198.12
```

### Model-Based Code

#### Sampling Distribution

```
samples.list = lapply(list.mat,FUN=function(x){sample(x,size=n)})
lambda.hat = unlist(lapply(samples.list,FUN=mean))
```




### Statistical Bias

the difference b/w the true value and the mean of the sampling distribution of all possible values; applies to design- and model-based sampling

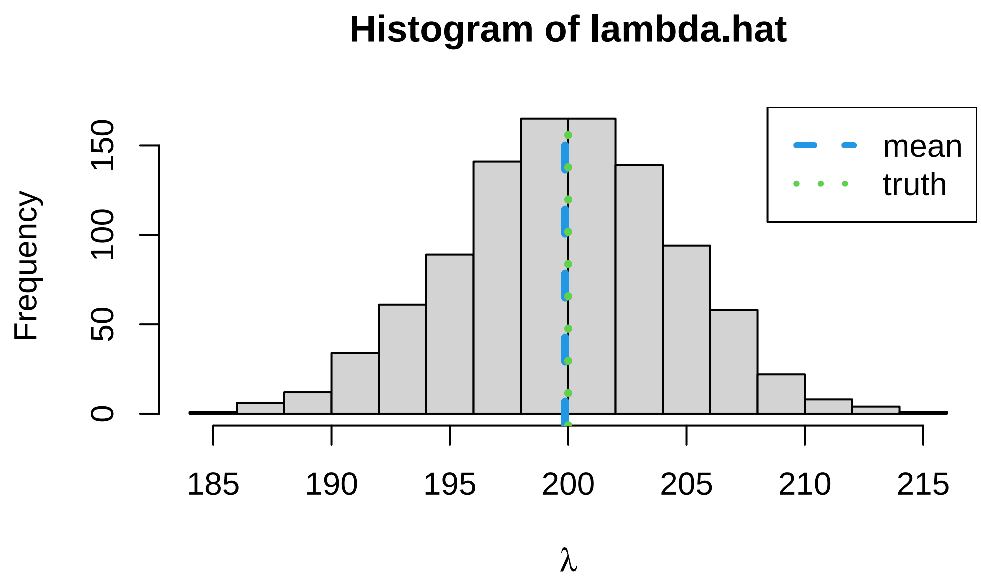
### Statistical Bias (Code)

```
# Bias (in this case, Monte Carlo error)
  mean(lambda.hat) - lambda
```

```
[1] -0.1286
```

```
# relative bias
  (mean(lambda.hat) - lambda)/lambda
```

```
[1] -0.000643
```




**Precision of the mean (Code)**

What is the probability that we will observe a mean within 5% of the truth?

. . .

We can calculate this using *Monte Carlo integration*

. . .

```
  diff = 0.025*lambda
  diff
```

```
[1] 5
```

```
  lower = lambda-diff
  upper = lambda+diff

  index = which(lambda.hat>=lower & lambda.hat <= upper)

# Probability of getting a mean within 5% of the truth
  length(index)/length(lambda.hat)
```

```
[1] 0.715
```

### Take-Aways

**1.** Study Objectives, Hypotheses, and Predictions

**2.** Big Data and Sampling

**3.** Inference and Prediction

**4.** Model-Based vs Design-Based Inference

### Lab

#### Objectives

- Introduce R Markdown
- Use simulation and design-based sampling to investigate bias and precision

### Lab Setup

Let's add some more *reality* in our work while using design-based sampling in R.

. . .

**Objective**: Evaluate sample size trade-offs for estimating white-tailed deer abundance throughout Rhode Island.

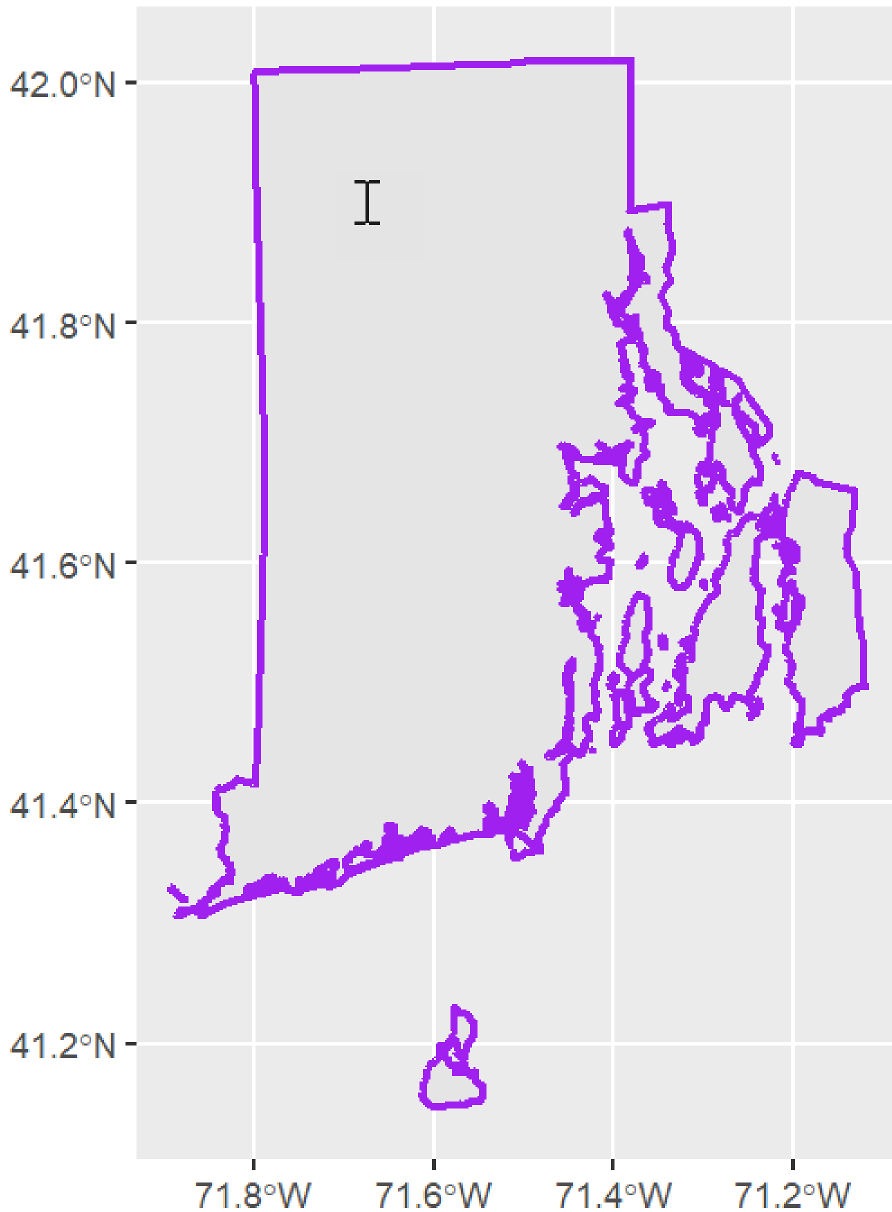
. . .

**Methodology**: Count deer in 1 sq. mile cells using FLIR technology attached to a helicopter.

## Lab Setup

Steps to consider

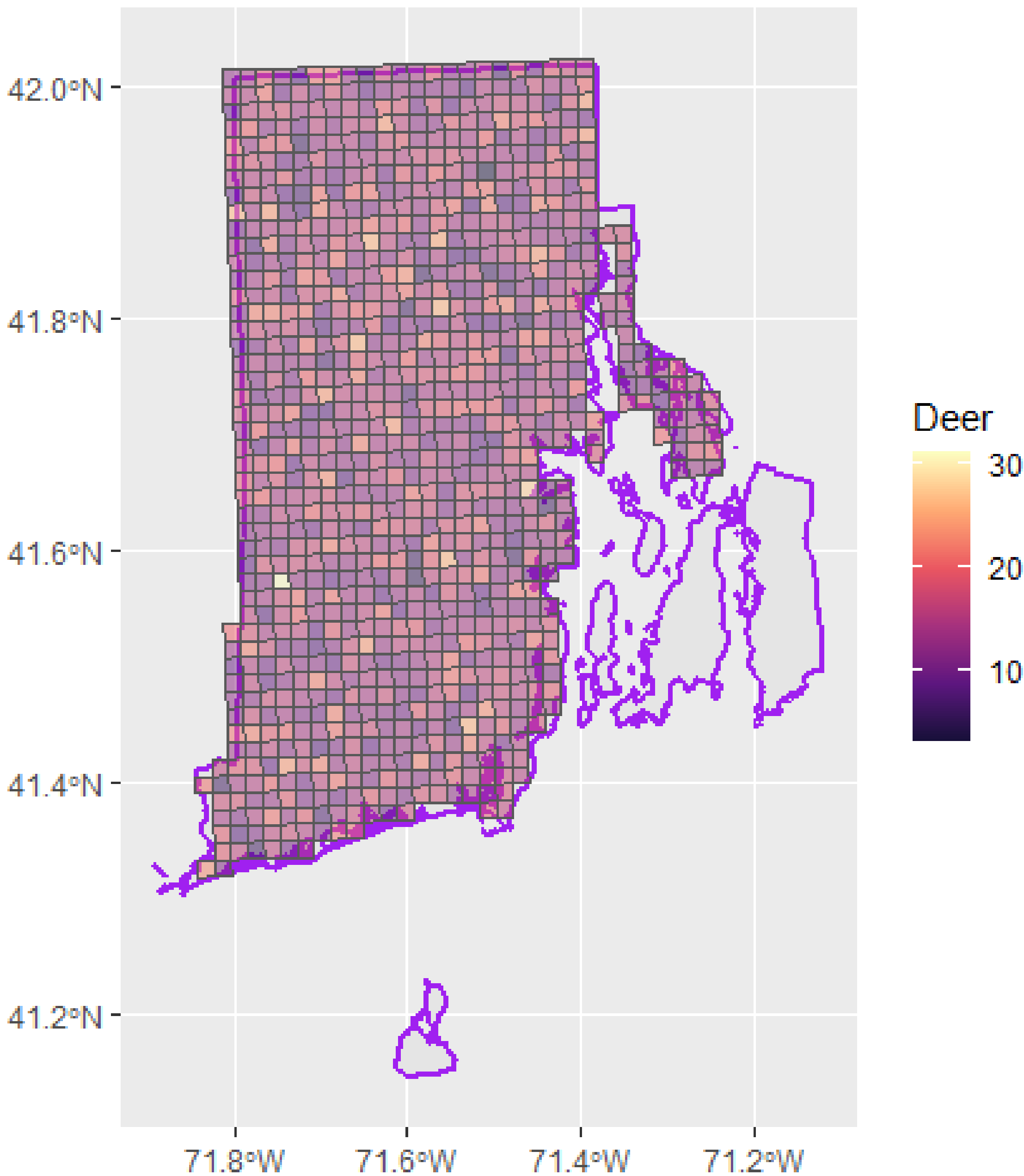
- **Sampling Frame**
  - all of RI or some subset




### Lab Setup

Steps to consider

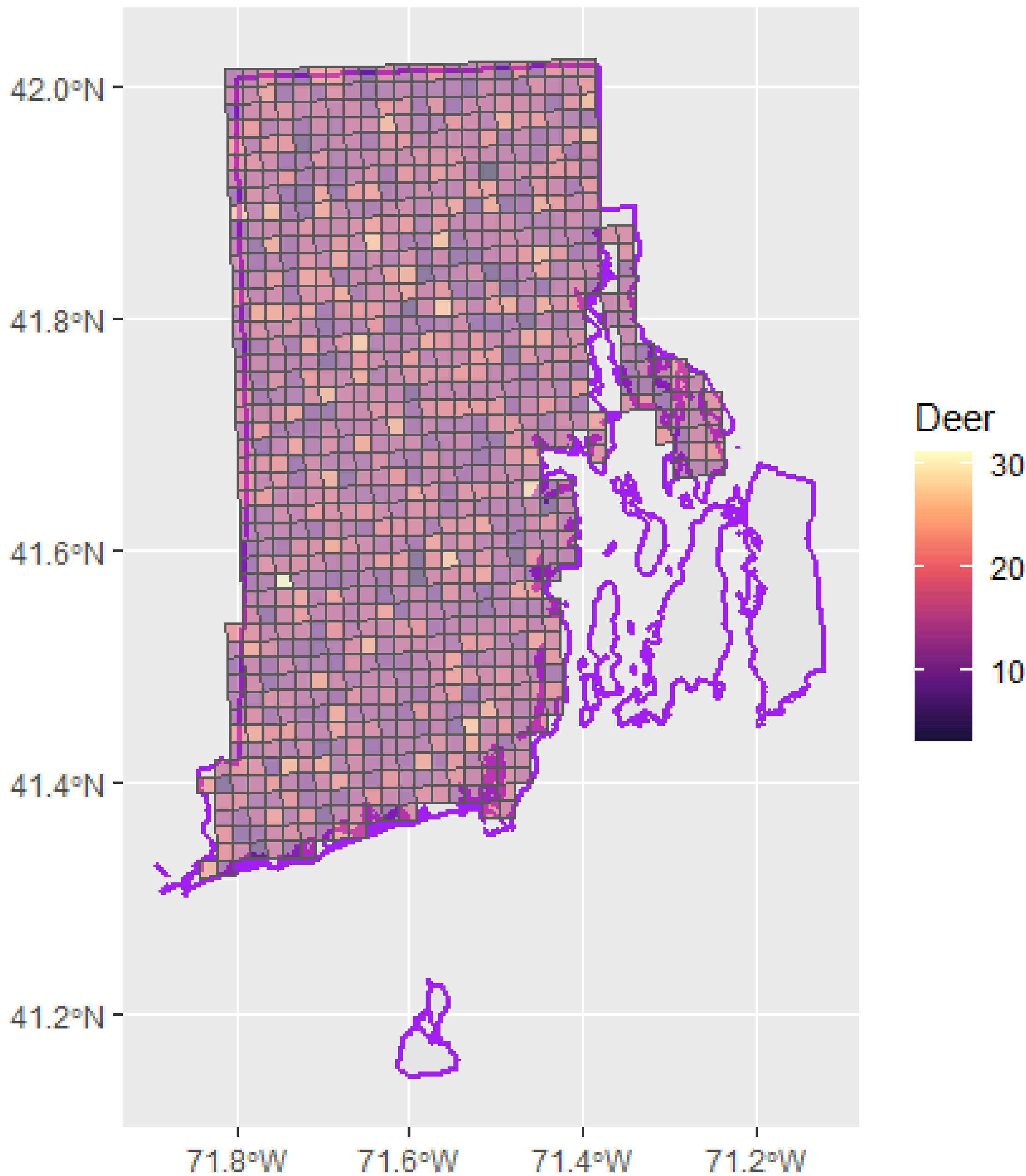
- **"Truth"**
  - how many deer per cell; how variable




**Lab Setup**

Steps to consider

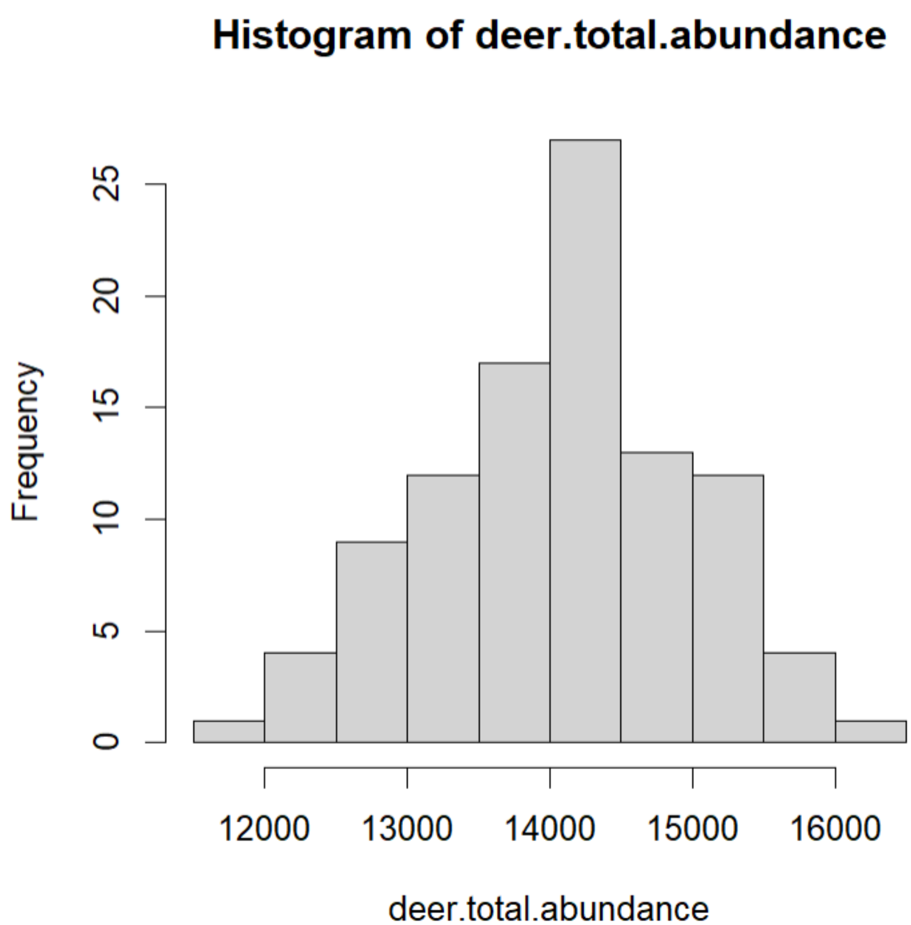
- **Sampling Process**
  - how to pick each cell



### Lab Setup

Steps to consider

- **Estimation Process**
  - estimate total deer population from the sample
- **Criteria to Evaluate**
  - use sampling distribution of deer abundance estimate or some other statistic




### Go to code