Random Sampling and the Bias

Probability and Statistics for Data Science

Carlos Fernandez-Granda

COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU DATA SCIENCE

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Goal

Estimate *population parameter*

Example: Average weight of rats in New York City

Challenge: We cannot catch every rat

Simple idea: Choose a random subset of the population

Extremely effective!

Estimating a population mean

Controlled scenario: True population with $N := 4{,}082$ individuals

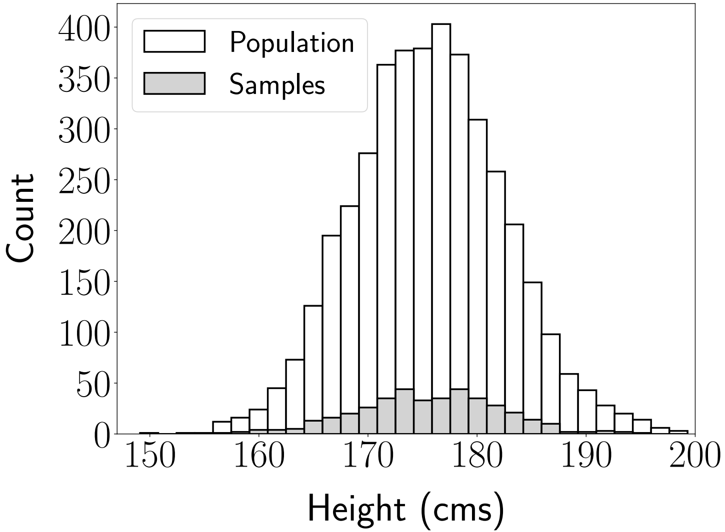
Heights: $h_1, h_2, \ldots, h_N$

Population mean:

$$\mu_{\text{pop}} := \frac{1}{N} \sum_{i=1}^{N} h_i = 175.6$$

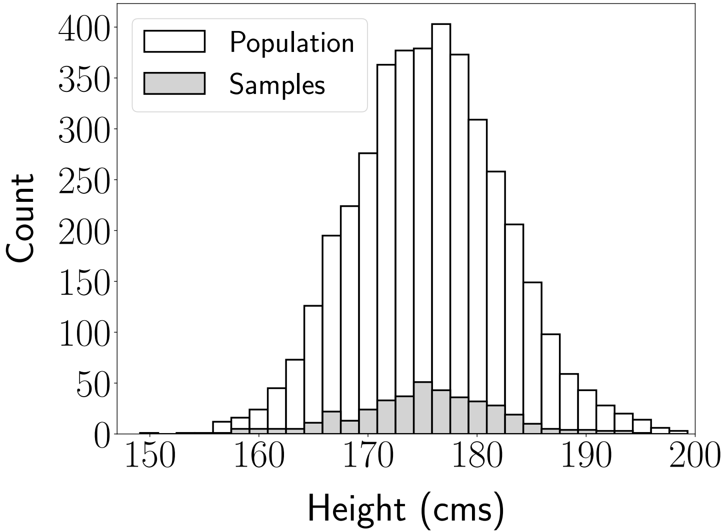
400 random samples

Sample mean = 175.5 ($\mu_{\text{pop}} = 175.6$)



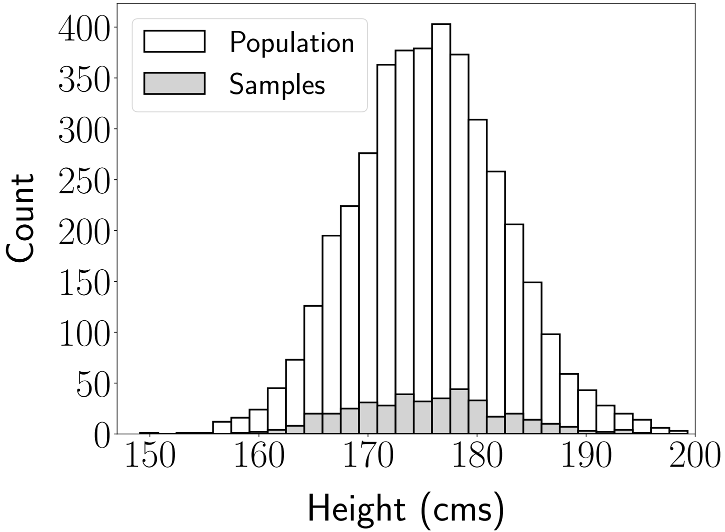
400 random samples

Sample mean = 175.2 ($\mu_{\text{pop}} = 175.6$)



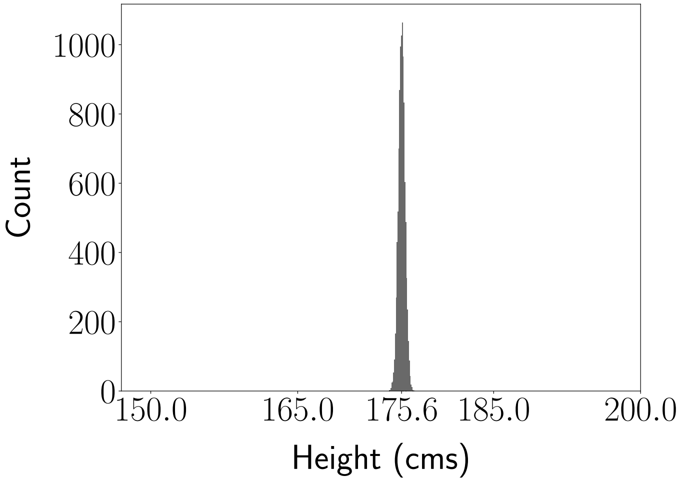
400 random samples

Sample mean = 176.1 ($\mu_{\text{pop}} = 175.6$)



Sample means of 10,000 subsets of size 400

Goal: Characterize probabilistic behavior of sample mean



Estimating a population proportion

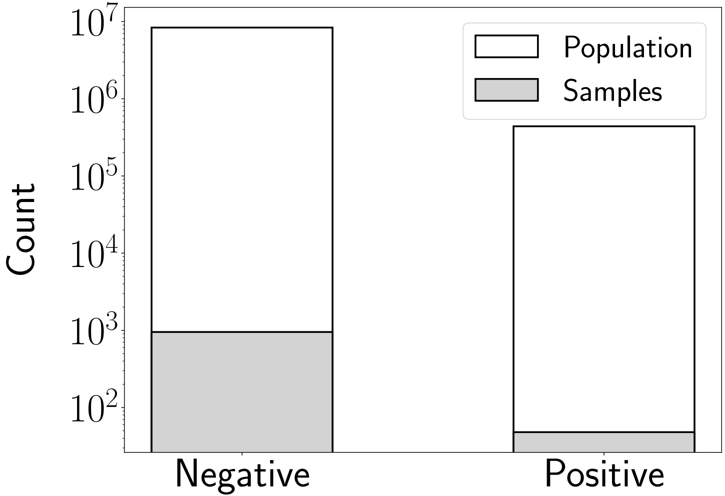
COVID-19 prevalence in New York

Population proportion:

$$\theta_{\text{pop}} = 0.05$$

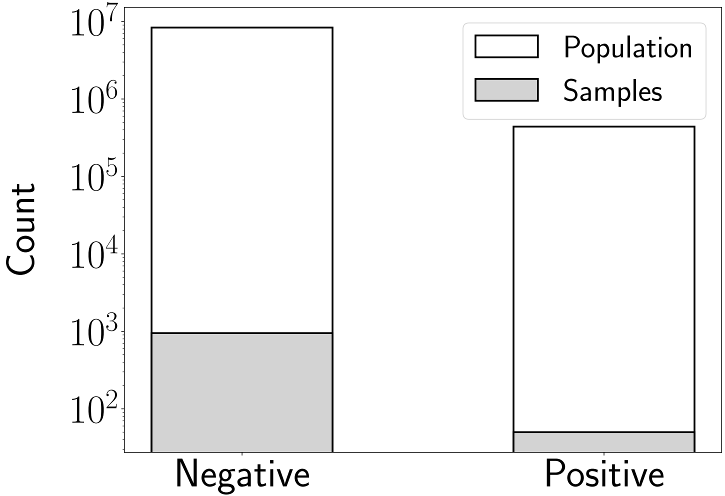
1,000 random samples out of 8.8 million

Sample proportion $= 0.055$ ($\theta_{\text{pop}} = 0.05$)



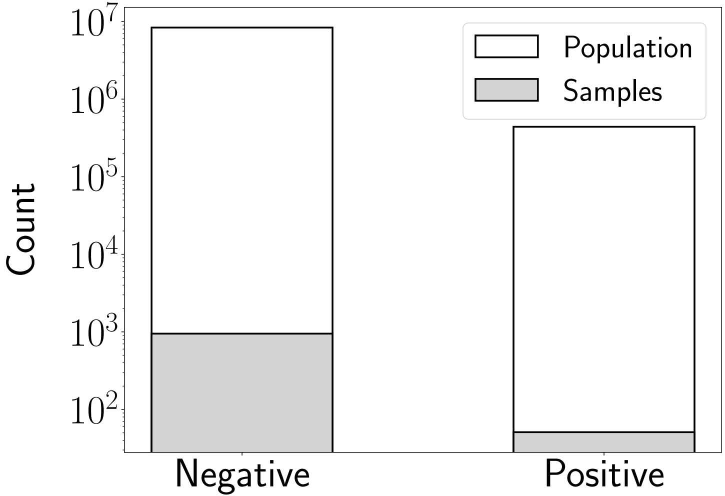
1,000 random samples out of 8.8 million

Sample proportion = 0.049 ($\theta_{\text{pop}} = 0.05$)



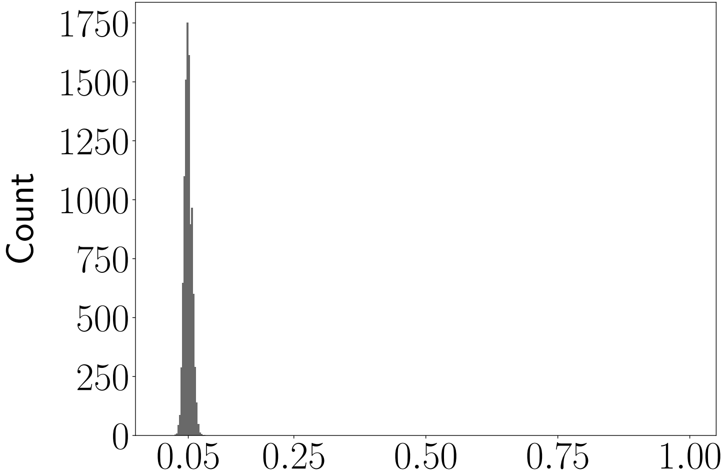
1,000 random samples out of 8.8 million

Sample proportion = 0.052 ($\theta_{\text{pop}} = 0.05$)



Sample proportions of 10,000 subsets of size 1,000

Goal: Characterize probabilistic behavior of sample proportion



Random sampling

Data: $a_1, a_2, \ldots, a_N$

Random samples: $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$

Each $\tilde{x}_i$ is selected independently and uniformly at random with replacement

Samples are independent identically distributed (i.i.d.) random variables with pmf

$$p_{\tilde{x}_j}(a_i) = \mathrm{P}(\tilde{x}_j = a_i) = \frac{1}{N}, \qquad 1 \leq i \leq N, \ 1 \leq j \leq n$$

Sample mean

Can be modeled as a random variable

$$\tilde{m} := \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i$$

Sample proportion

Data: $a_1, a_2, \ldots, a_N$

$a_i = 1$ if ith data point satisfies a certain condition
(e.g. person has COVID-19)

Random samples: $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$

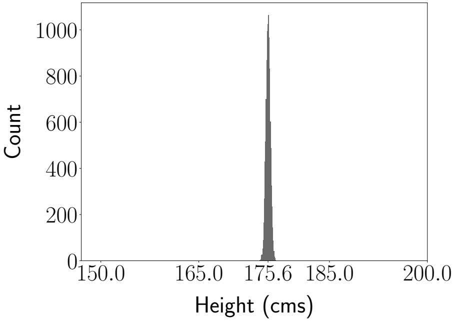
Sample proportion is just sample mean:

$$\tilde{m} := \frac{1}{n} \sum_{j=1}^{n} \tilde{x}_j$$

Estimation of population parameters

Frequentist perspective

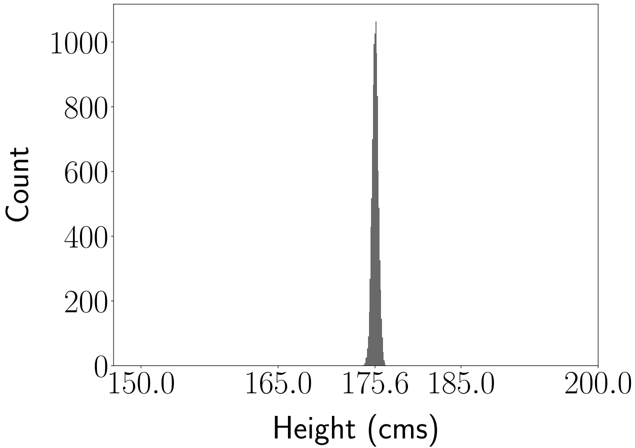
The parameter of interest is deterministic



Goal: Characterize probabilistic behavior of estimator

The bias

Is the estimator centered at the parameter?



The bias

Random measurements: $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_n$

Deterministic parameter of interest: $\gamma \in \mathbb{R}$

Estimator: $h(\tilde{x}_1, \ldots, \tilde{x}_n)$

The bias of the estimator is the mean of the error

$$\text{Bias} = \mathrm{E}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n) - \gamma\right]$$

If $\mathrm{E}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right] = \gamma$, the estimator is unbiased

Random sampling

Data: a_1, a_2, ..., a_N

Random samples: $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_n$

Samples are independent identically distributed (i.i.d.) random variables with pmf

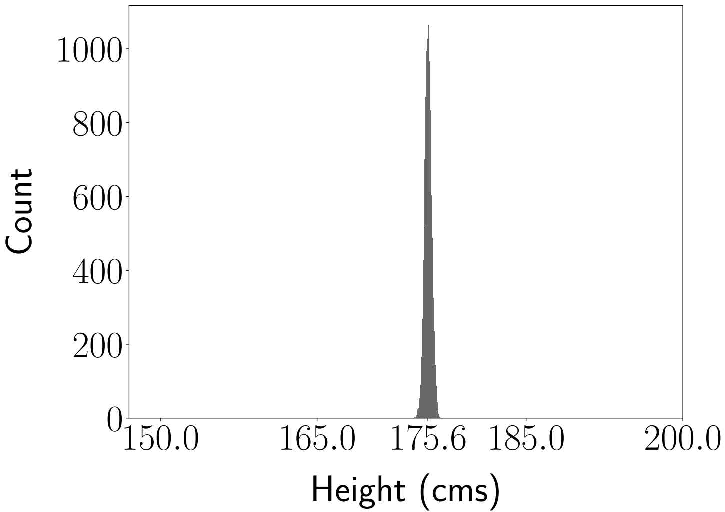
$$p_{\tilde{x}_j}(a_i) = \mathrm{P}(\tilde{x}_j = a_i) = \frac{1}{N} \qquad 1 \leq i \leq N, \ 1 \leq j \leq n$$

Sample mean is unbiased

$$
\begin{aligned}
\mathrm{E}\left[\tilde{x}_j\right] &= \sum_{i=1}^{N} a_i p_{\tilde{x}_j}(a_i) \\
&= \frac{1}{N}\sum_{i=1}^{N} a_i \\
&= \mu_{\text{pop}}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left[\widetilde{m}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j\right] \\
&= \frac{1}{n}\sum_{j=1}^{n} \mathrm{E}\left[\tilde{x}_j\right] \\
&= \mu_{\text{pop}}
\end{aligned}
$$

Sample mean is unbiased



Sample proportion is unbiased

Data: $a_1, a_2, \ldots, a_N$

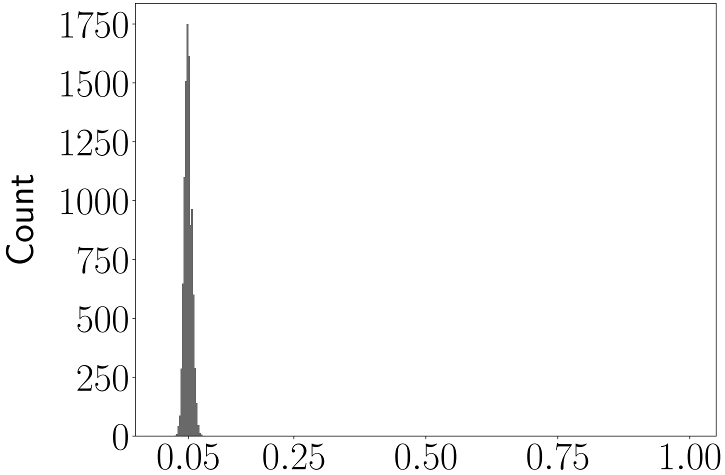
$a_i = 1$ if ith data point satisfies a certain condition

Random samples: $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$

Sample proportion is sample mean $\tilde{m} := \frac{1}{n}\sum_{j=1}^{n} \tilde{x}_j$

$$\begin{aligned} \mathrm{E}\left[\tilde{m}\right] &= \frac{1}{N}\sum_{i=1}^{N} a_i \\ &= \frac{\text{Number of COVID-19 cases}}{N} = \theta_{\text{pop}} \end{aligned}$$

Sample proportion is unbiased



Sample variance is unbiased

Data: $a_1, a_2, \ldots, a_N$ Random measurements: $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$

Population mean

$$\mu_{\text{pop}} := \frac{1}{N} \sum_{i=1}^{N} a_i$$

Population variance

$$\sigma^2_{\text{pop}} := \frac{1}{N} \sum_{i=1}^{N} (a_i - \mu_{\text{pop}})^2$$

Sample variance

$$\tilde{v} := \frac{1}{n-1} \sum_{j=1}^{n} (\tilde{x}_j - \widetilde{m})^2$$

$$\mathrm{E}[\tilde{v}] = \sigma^2_{\text{pop}}$$

What have we learned

Definition of random sampling

Definition of bias

Sample mean, proportion and variance are unbiased

Is an unbiased estimator enough?