The Standard Error

Probability and Statistics for Data Science

Carlos Fernandez-Granda

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

Estimation of population parameters

Simple idea: Choose a random subset of the population

Estimating a population mean

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

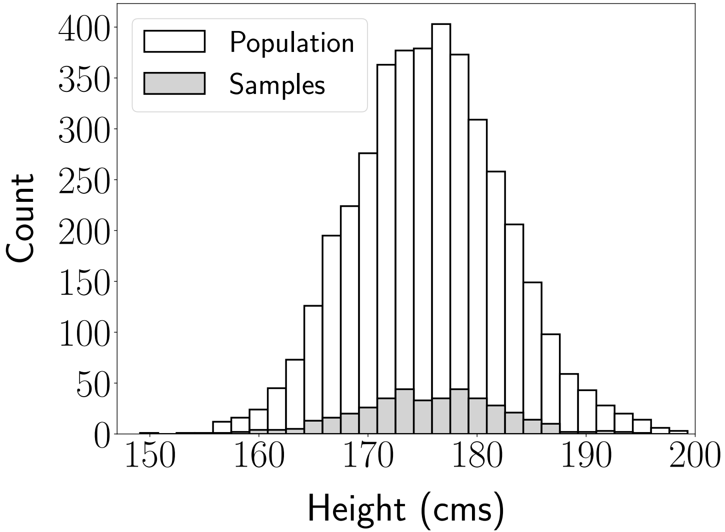
Heights: $h_1, h_2, \ldots, h_N$

Population mean:

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

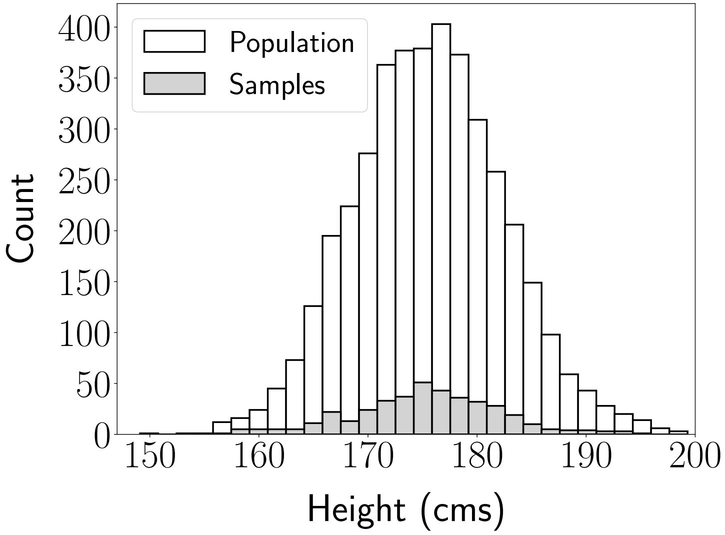
400 random samples

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



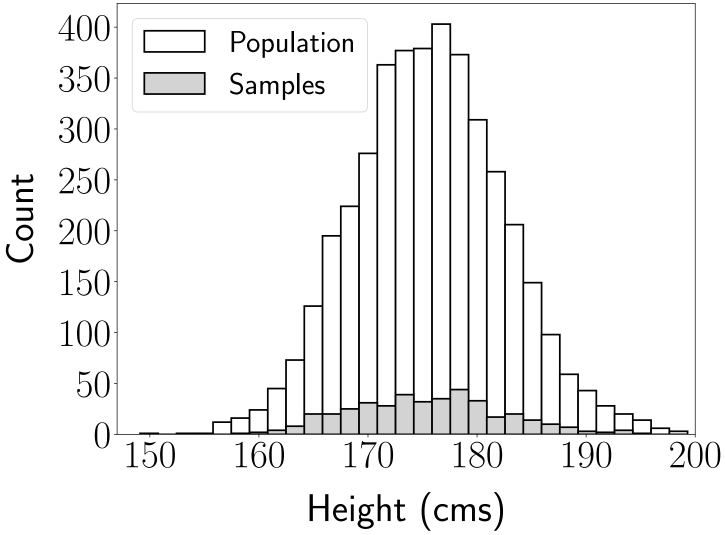
400 random samples

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



400 random samples

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



Random sampling

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

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

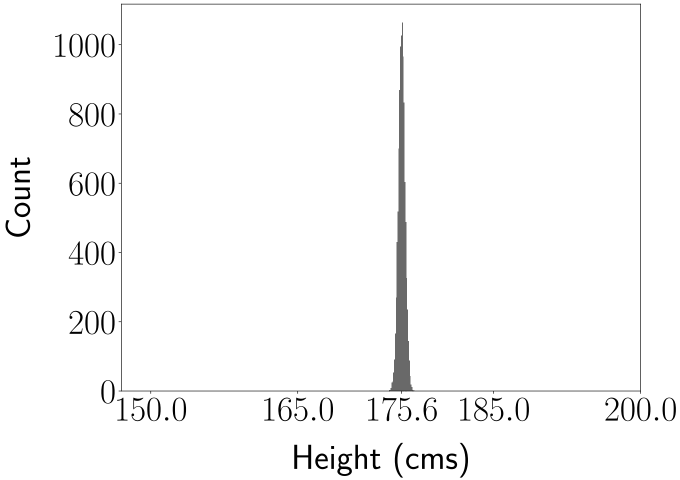
Each $\tilde{x}_i$ is selected independently and uniformly at random with replacement

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 means of 10,000 subsets of size 400

Sample mean has to be analyzed probabilistically

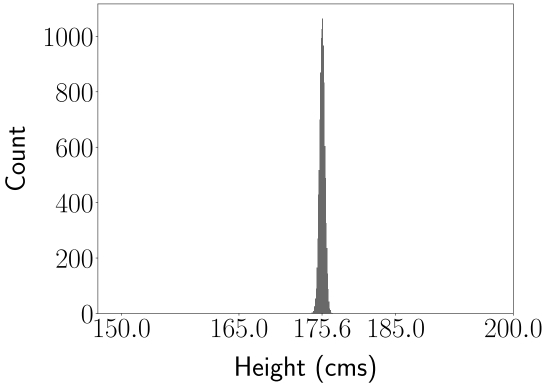


Sample mean is unbiased

Modeled as a random variable

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

$$\mathrm{E}\left[\tilde{m}\right] = \mu_{\text{pop}}$$



Standard error

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

Deterministic parameter of interest: γ

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

The standard error of the estimator is its standard deviation

$$\text{se}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right] := \sqrt{\operatorname{Var}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right]}$$

Standard error

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

$$\begin{aligned}
\mathsf{se}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right] &:= \sqrt{\mathrm{Var}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right]} \\
&= \sqrt{\mathrm{E}\left[\left(h(\tilde{x}_1, \ldots, \tilde{x}_n) - \mathrm{E}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n)\right]\right)^2\right]} \\
&= \sqrt{\mathrm{E}\left[\left(h(\tilde{x}_1, \ldots, \tilde{x}_n) - \gamma\right)^2\right]}
\end{aligned}$$

Standard error of the sample mean

$$\mathsf{se}\left[\widetilde{m}\right]^2 = \mathrm{Var}\left[\widetilde{m}\right] = \mathrm{Var}\left[\frac{1}{n}\sum_{j=1}^{n}\tilde{x}_j\right]$$

$$= \frac{1}{n^2}\mathrm{Var}\left[\sum_{j=1}^{n}\tilde{x}_j\right]$$

Uncorrelated random variables

If $\tilde{a}$ and $\tilde{b}$ are uncorrelated

$$\mathrm{Var}[\tilde{a} + \tilde{b}] = \mathrm{Var}\,[\tilde{a}] + \mathrm{Var}[\tilde{b}]$$

Sum of independent random variables

Independent random variables $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$ with finite variance

$$\begin{aligned}
\mathrm{Var}\left[\sum_{k=1}^{n} \tilde{a}_k\right] &= \mathrm{Var}\left[\tilde{a}_1\right] + \mathrm{Var}\left[\sum_{k=2}^{n} \tilde{a}_k\right] \\
&= \mathrm{Var}\left[\tilde{a}_1\right] + \mathrm{Var}\left[\tilde{a}_2\right] + \mathrm{Var}\left[\sum_{k=3}^{n} \tilde{a}_k\right] \\
&= \sum_{k=1}^{n} \mathrm{Var}\left[\tilde{a}_k\right]
\end{aligned}$$

Standard error of the sample mean

$$
\begin{aligned}
\mathsf{se}\left[\widetilde{m}\right]^2 &= \frac{1}{n^2}\mathrm{Var}\left[\sum_{j=1}^{n}\tilde{x}_j\right] \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{Var}\left[\tilde{x}_j\right] \\
&= \frac{\sigma_{\mathsf{pop}}^2}{n} \\
\mathrm{Var}\left[\tilde{x}_j\right] &:= \mathrm{E}\left[\left(\tilde{x}_j - \mathrm{E}\left[\tilde{x}_j\right]\right)^2\right] \\
&= \mathrm{E}\left[\left(\tilde{x}_j - \mu_{\mathsf{pop}}\right)^2\right] \\
&= \sum_{i=1}^{N}\left(a_i - \mu_{\mathsf{pop}}\right)^2 p_{\tilde{x}_j}(a_i) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(a_i - \mu_{\mathsf{pop}}\right)^2 = \sigma_{\mathsf{pop}}^2
\end{aligned}
$$

Standard error of the sample mean

$$\mathrm{se}\left[\widetilde{m}\right] = \frac{\sigma_{\mathrm{pop}}}{\sqrt{n}}$$

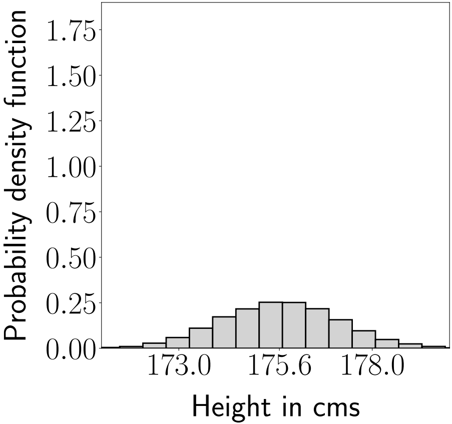
No dependence on N!

Height data: $n = 20$

$\mu_{\text{pop}} := 175.6$ cm, $\sigma_{\text{pop}} = 6.85$ cm

Total population $N := 4{,}082$

10^4 sample means

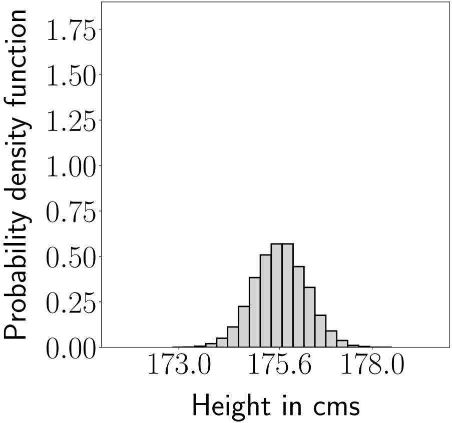


$n = 100$

$\mu_{\text{pop}} := 175.6$ cm, $\sigma_{\text{pop}} = 6.85$ cm

Total population $N := 4{,}082$

10^4 sample means

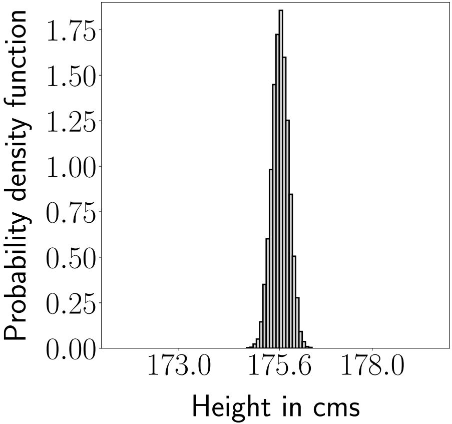


n =1,000

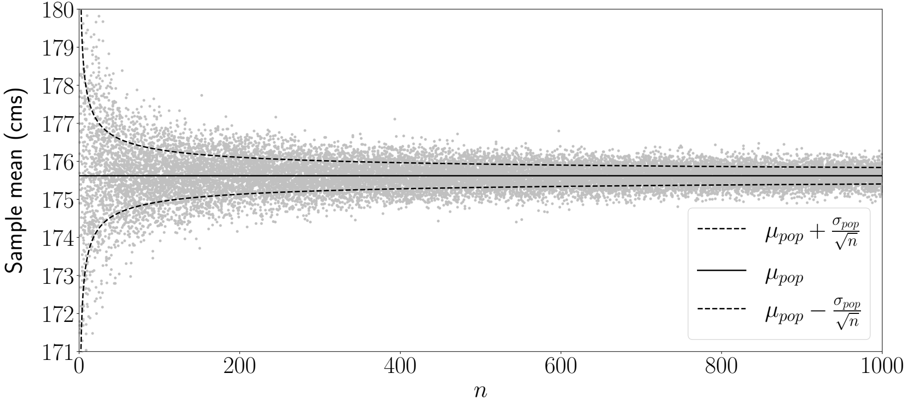
$\mu_{\text{pop}} := 175.6$ cm, $\sigma_{\text{pop}} = 6.85$ cm

Total population $N := 4{,}082$

10^4 sample means



Height data


Sample mean (cms)
180
179
178
177
176
175
174
173
172
171
0
200
400
600
800
1000
n
$\mu_{pop} + \frac{\sigma_{pop}}{\sqrt{n}}$
μ_{pop}
$\mu_{pop} - \frac{\sigma_{pop}}{\sqrt{n}}$

Estimating a population proportion

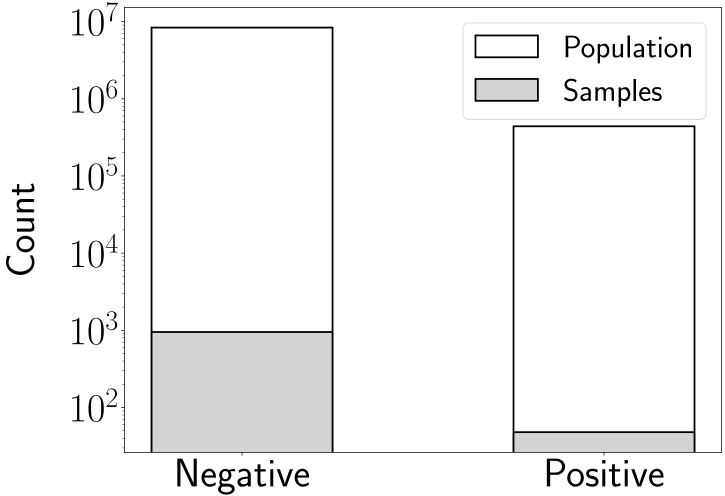
COVID-19 prevalence in New York

Population proportion:

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

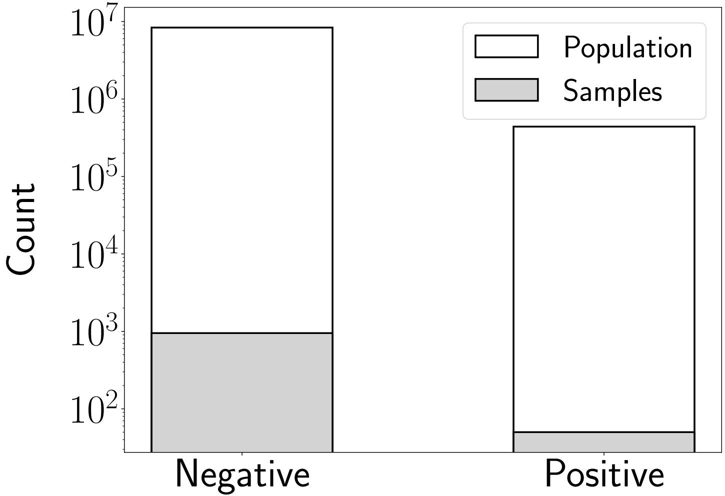
1,000 random samples out of 8.8 million

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



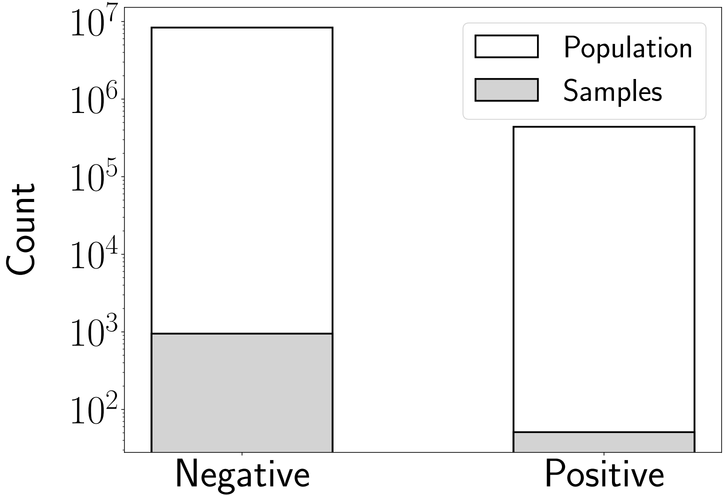
1,000 random samples out of 8.8 million

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

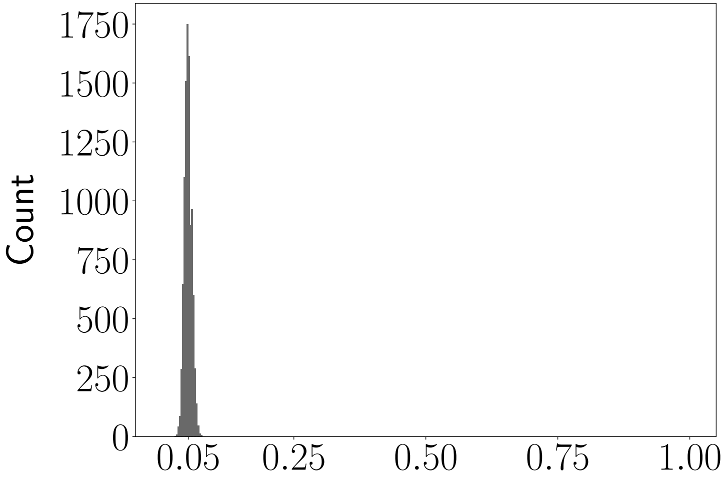


1,000 random samples out of 8.8 million

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



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



Standard error of sample proportion

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

$$\mathrm{se}\left[\tilde{m}\right] = \frac{\sigma_{\text{pop}}}{\sqrt{n}}$$

Population variance

$$\begin{aligned}
\sigma^2_{\text{pop}} &:= \frac{1}{N}\sum_{i=1}^{N}(a_i - \theta_{\text{pop}})^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} a_i^2 - \frac{2\theta_{\text{pop}}}{N}\sum_{i=1}^{N} a_i + \frac{1}{N}\sum_{i=1}^{N}\theta^2_{\text{pop}} \\
&= \theta_{\text{pop}} - 2\theta^2_{\text{pop}} + \theta^2_{\text{pop}} \\
&= \theta_{\text{pop}}(1 - \theta_{\text{pop}})
\end{aligned}$$

Standard error of sample proportion

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

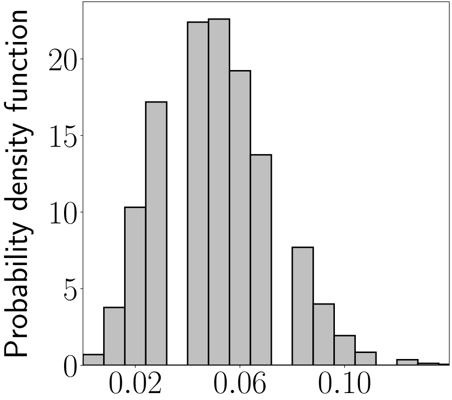
$$\begin{aligned} \text{se}\left[\tilde{m}\right] &= \frac{\sigma_{\text{pop}}}{\sqrt{n}} \\ &= \sqrt{\frac{\theta_{\text{pop}}(1-\theta_{\text{pop}})}{n}} \end{aligned}$$

COVID-19

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

Total population $N := 8$ million

Distribution of 10^4 sample means for $n = 100$

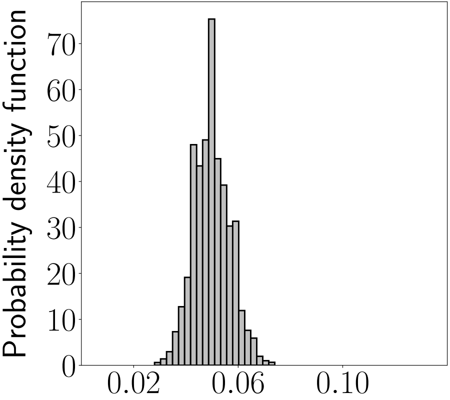


COVID-19

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

Total population $N := 8$ million

Distribution of 10^4 sample means for $n = 1{,}000$

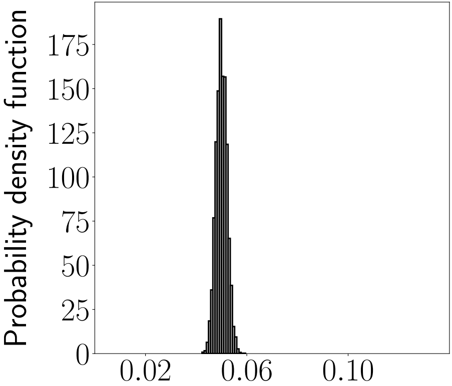


COVID-19

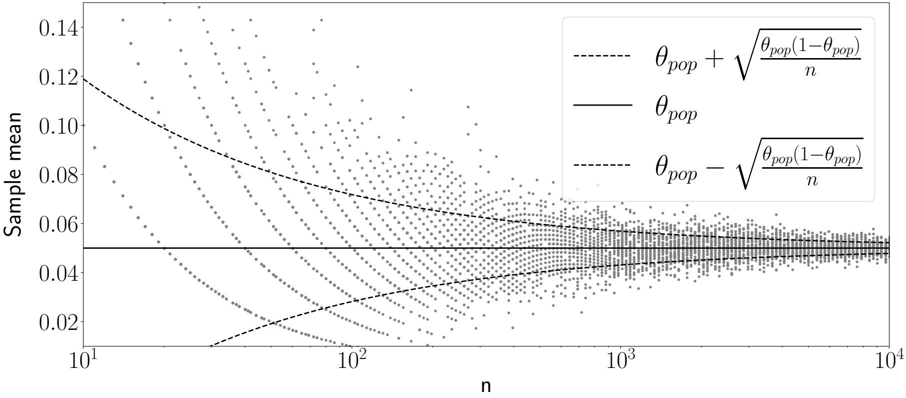
$\theta_{\text{pop}} := 0.05$

Total population $N := 8$ million

Distribution of 10^4 sample means for $n = 10{,}000$



COVID-19



$\theta_{pop} + \sqrt{\frac{\theta_{pop}(1-\theta_{pop})}{n}}$
θ_{pop}
$\theta_{pop} - \sqrt{\frac{\theta_{pop}(1-\theta_{pop})}{n}}$
Sample mean
0.14
0.12
0.10
0.08
0.06
0.04
0.02
10^1
10^2
10^3
10^4
n

What have we learned

Definition of standard error

Standard error of sample mean and sample proportion

Random sampling works!