Confidence Intervals For Proportions And Probabilities

Probability and Statistics for Data Science

Carlos Fernandez-Granda

NYU 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

Plan

How to build confidence intervals for proportions and probabilities

Confidence intervals for Monte Carlo simulations

Limitations of the confidence-interval framework

Estimating a population proportion

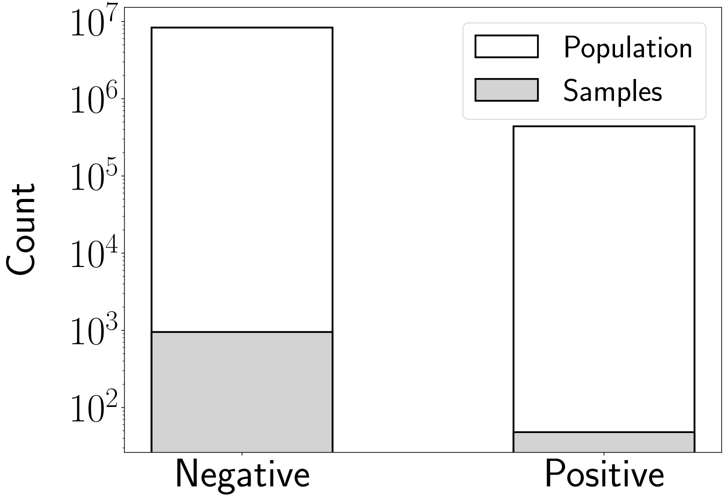
COVID-19 prevalence in New York

Population proportion:

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

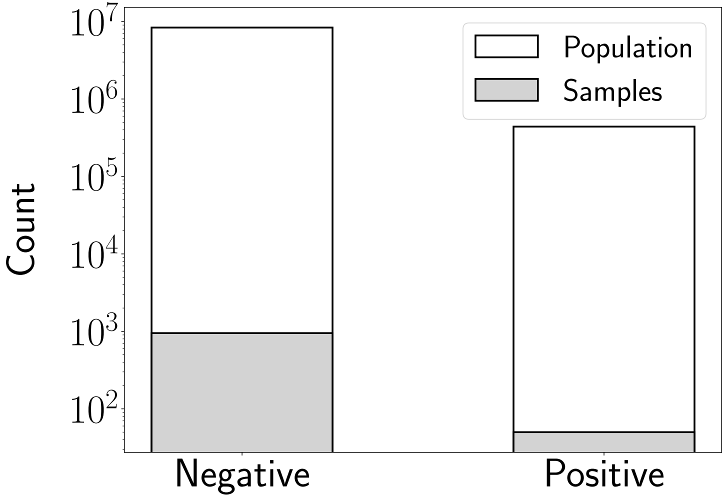
1,000 random samples out of 8.8 million

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



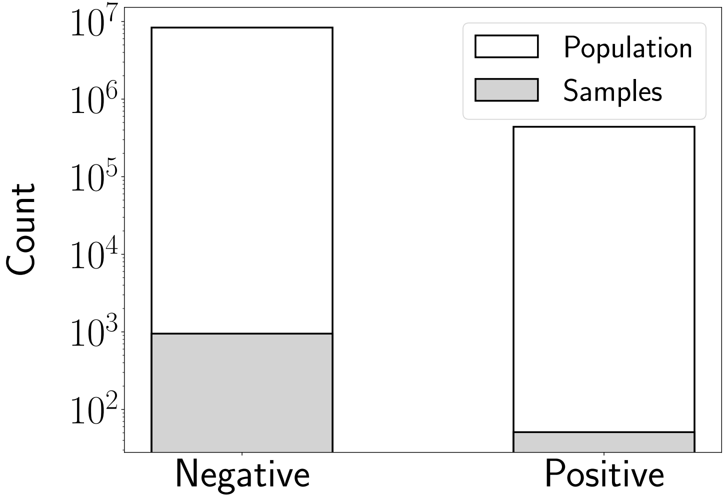
1,000 random samples out of 8.8 million

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



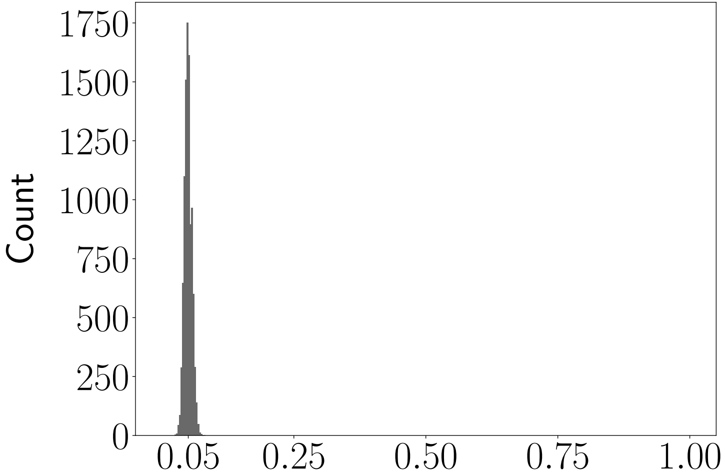
1,000 random samples out of 8.8 million

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



Confidence interval

Main idea: Report a range of values that contain parameter with high probability (e.g. 95%)

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:

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

Confidence interval for the mean

If $\tilde{x}_1$, $\tilde{x}_2$, $\tilde{x}_3$, ... are independent random variables with mean μ and variance σ^2

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

$$\mathrm{E}\left[\tilde{m}\right] = \mu$$

$$\mathrm{Var}\left[\tilde{m}\right] = \frac{\sigma^2}{n}$$

$$\widetilde{\mathcal{I}}_{1-\alpha} := \left[\tilde{m} - \frac{c_\alpha \sigma}{\sqrt{n}}, \tilde{m} + \frac{c_\alpha \sigma}{\sqrt{n}}\right] \qquad c_\alpha := F_{\tilde{z}}^{-1}\left(1 - \frac{\alpha}{2}\right)$$

$$\widetilde{\mathcal{I}}_{0.95} := \left[\tilde{a} - \frac{1.96\sigma}{\sqrt{n}}, \tilde{a} + \frac{1.96\sigma}{\sqrt{n}}\right]$$

Confidence interval for a probability

If $\tilde{b}_1$, $\tilde{b}_2$, $\tilde{b}_3$, ... are Bernoulli random variables with parameter θ

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

$$\mathrm{E}\left[\tilde{m}\right] = \theta$$

$$\mathrm{Var}\left[\tilde{m}\right] = \frac{\theta(1-\theta)}{n}$$

$$\widetilde{\mathcal{I}}_{1-\alpha} := \left[\tilde{m} - c_{\alpha}\sqrt{\frac{\theta(1-\theta)}{n}}, \tilde{m} + c_{\alpha}\sqrt{\frac{\theta(1-\theta)}{n}}\right]$$

Confidence interval for a probability

$$\widetilde{\mathcal{I}}_{1-\alpha} := \left[\tilde{m} - c_\alpha \sqrt{\frac{\theta(1-\theta)}{n}}, \tilde{m} + c_\alpha \sqrt{\frac{\theta(1-\theta)}{n}}\right]$$

$$h(\theta) := \theta(1-\theta) \leq 0.25$$

$$\frac{\mathrm{d}h(\theta)}{\mathrm{d}\theta} = 1 - 2\theta \qquad \frac{\mathrm{d}^2 h(\theta)}{\mathrm{d}\theta^2} = -2$$

$$\widetilde{\mathcal{I}}_{1-\alpha} \subset \left[\tilde{m} - \frac{0.5 c_\alpha}{\sqrt{n}}, \tilde{m} + \frac{0.5 c_\alpha}{\sqrt{n}}\right]$$

$$\widetilde{\mathcal{I}}_{0.95} \subset \left[\tilde{m} - \frac{0.98}{\sqrt{n}}, \tilde{m} + \frac{0.98}{\sqrt{n}}\right]$$

Confidence interval for population proportion θ_{pop}

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

Bernoulli random variables with parameter θ_{pop}

$$\widetilde{\mathcal{I}}_{1-\alpha} \subset \left[\tilde{m} - \frac{0.5 c_\alpha}{\sqrt{n}}, \tilde{m} + \frac{0.5 c_\alpha}{\sqrt{n}}\right]$$

$$\widetilde{\mathcal{I}}_{0.95} \subset \left[\tilde{m} - \frac{0.98}{\sqrt{n}}, \tilde{m} + \frac{0.98}{\sqrt{n}}\right]$$

Prevalence of COVID-19

Goal: Estimate prevalence θ_{pop} of COVID-19 in New York City

How many tests so error $\leq 1\%$ with probability at least 0.95?

$$\widetilde{\mathcal{I}}_{0.95} \subset \left[\tilde{m} - \frac{0.98}{\sqrt{n}}, \tilde{m} + \frac{0.98}{\sqrt{n}}\right]$$

$$\frac{0.98}{\sqrt{n}} < 0.01 \implies n \geq 9604$$

The Monte Carlo method

Idea: Estimate $P(A)$ by simulating outcomes and checking how many are in A

Key question: Have we done enough simulations?

Use confidence intervals!

2021 Tokyo Olympics

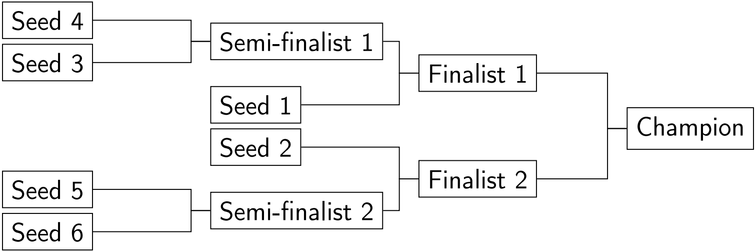
3x3 basketball tournament

Participants: Belgium, China, Japan, Latvia, the Netherlands, Poland, the Russian Olympic Committee (ROC), and Serbia

Goal: Estimate probability that each team wins

Tournament

Group stage followed by bracket



Monte Carlo method

To estimate probability θ that a team wins:

1. We simulate the tournament n times independently
2. In each simulation, $\mathrm{P}(\text{team wins}) = \theta$
3. Compute the fraction of simulations $\widetilde{\mathrm{P}}_{\mathsf{MC}}$ in which team wins

Sample mean of n Bernoulli random variables with parameter θ

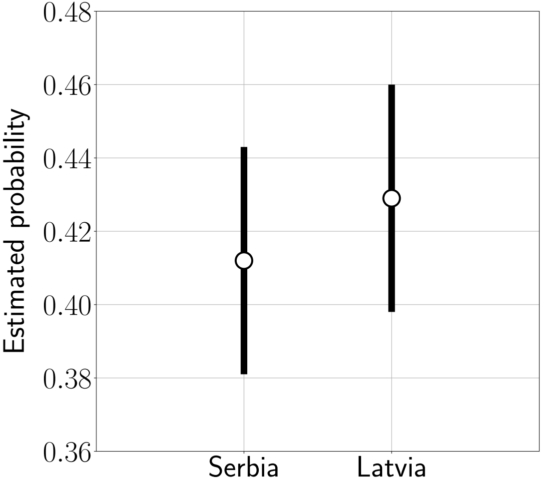
$$\widetilde{\mathcal{I}}_{1-\alpha} \subset \left[\widetilde{\mathrm{P}}_{\mathsf{MC}} - \frac{0.5 c_\alpha}{\sqrt{n}}, \widetilde{\mathrm{P}}_{\mathsf{MC}} + \frac{0.5 c_\alpha}{\sqrt{n}}\right]$$

$$\widetilde{\mathcal{I}}_{0.95} \subset \left[\widetilde{\mathrm{P}}_{\mathsf{MC}} - \frac{0.98}{\sqrt{n}}, \widetilde{\mathrm{P}}_{\mathsf{MC}} + \frac{0.98}{\sqrt{n}}\right]$$

Results

1,000 simulations: Latvia wins more often

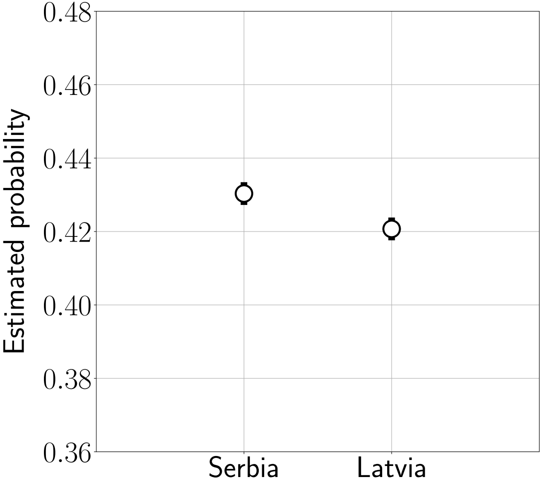
Have we done enough simulations? No



2021 Tokyo Olympics

100,000 simulations: Serbia wins more often

Have we done enough simulations? Yes



Real poll (Pennsylvania)

Data: 281 people intend to vote for Trump, 300 for Biden

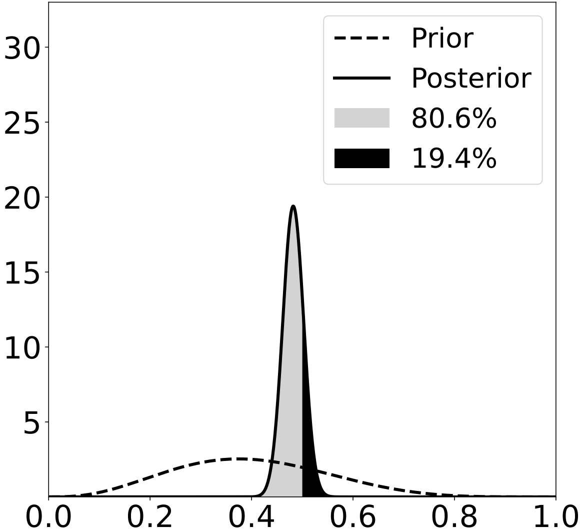
Parameter: Fraction of Trump voters in population θ

$$
\begin{aligned}
\widetilde{\mathcal{I}}_{0.95} &\subset \left[\tilde{m} - \frac{0.98}{\sqrt{n}}, \tilde{m} + \frac{0.98}{\sqrt{n}}\right] \\
&= \left[0.484 - \frac{0.98}{\sqrt{581}}, 0.484 + \frac{0.98}{\sqrt{581}}\right] = [0.444, 0.524]
\end{aligned}
$$

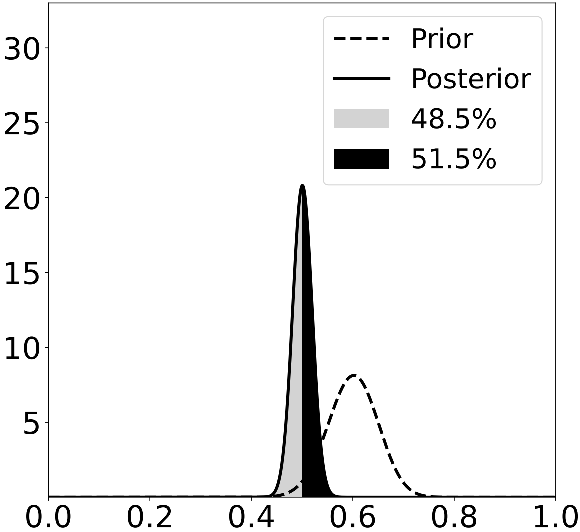
Probability that Trump wins, $\mathrm{P}\left(\theta \geq 0.5\right)$?

¯_(ツ)_/¯

Bayesian model



Bayesian model



Precipitation

Goal: Estimate fraction of time that it rains in Coos Bay

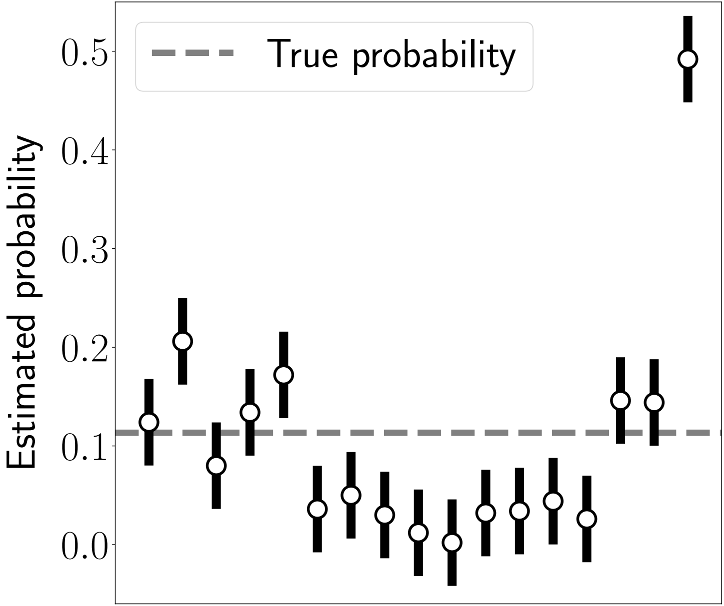
Ground truth: 11.3%

Data: 500 hourly measurements

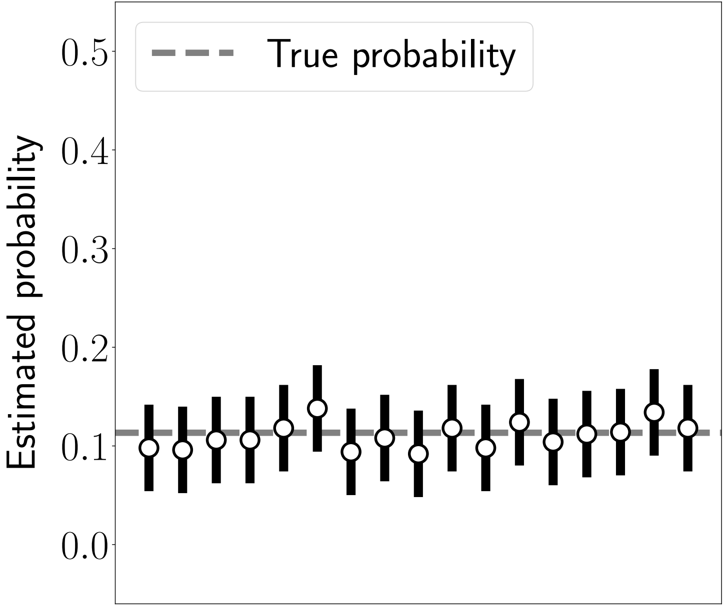
0.95 confidence interval

$$\widetilde{\mathcal{I}}_{0.95} \subset \left[\tilde{m} - \frac{0.98}{\sqrt{n}}, \tilde{m} + \frac{0.98}{\sqrt{n}}\right]$$

Sequential measurements



Randomized measurements



What have we learned

How to build confidence intervals for proportions and probabilities

Confidence intervals for Monte Carlo simulations

Limitations of the confidence-interval framework