Confidence Intervals

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

Plan

Definition of confidence intervals

How to build confidence intervals for the population mean

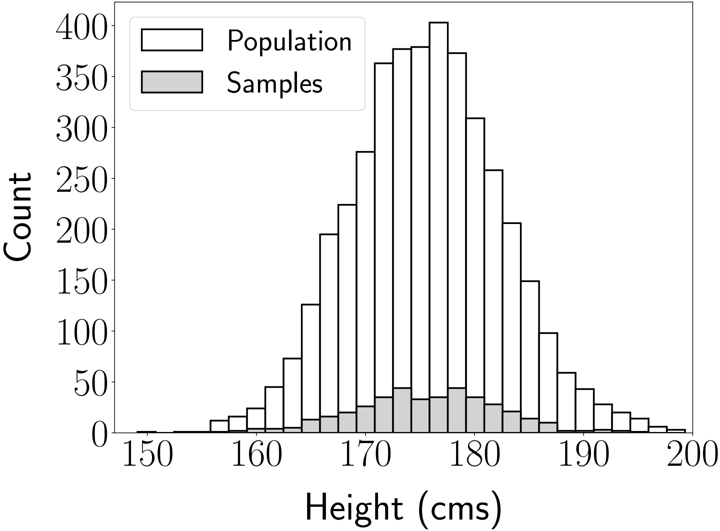
Interpretation of confidence intervals

Estimation of population parameters

Simple idea: Choose a random subset of the population

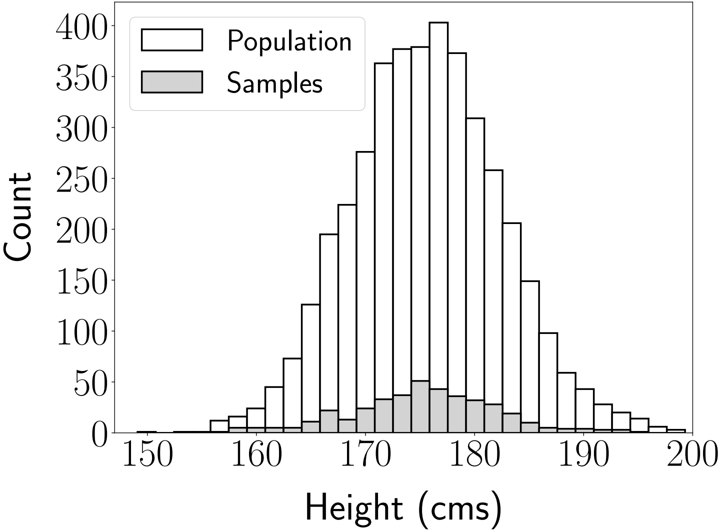
Random sampling

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



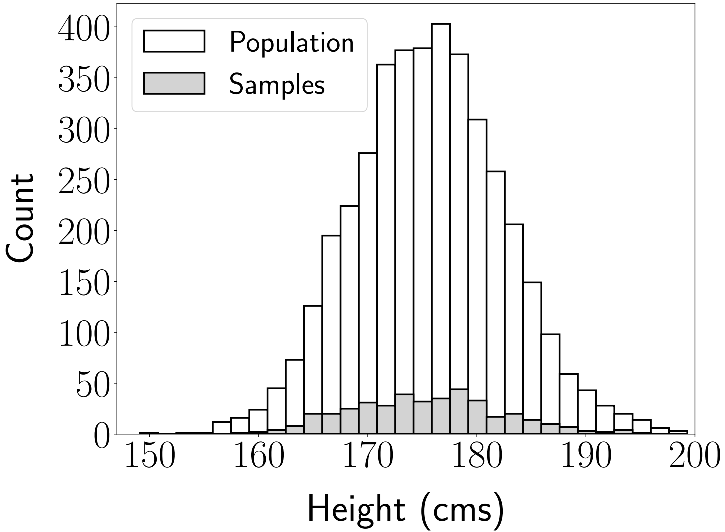
400 random samples

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



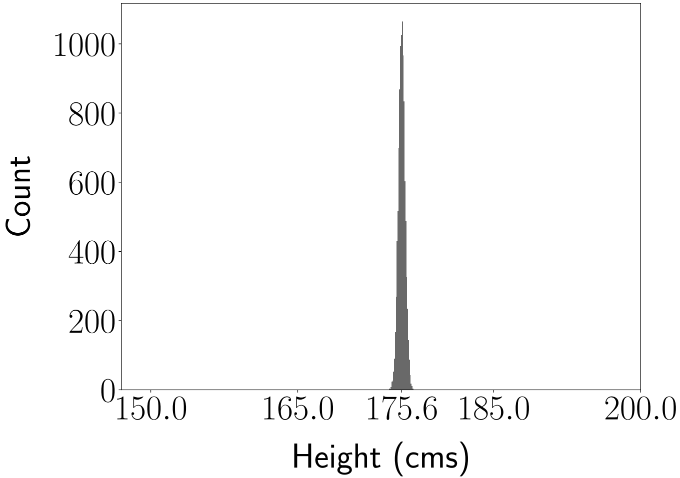
400 random samples

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



Sample means of 10,000 subsets of size 400

Goal: Quantify uncertainty from available data



Confidence interval

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

Sample mean

Population mean: μ_{pop} Population variance: σ^2_{pop}

Random samples selected independently and uniformly at random with replacement: $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$

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

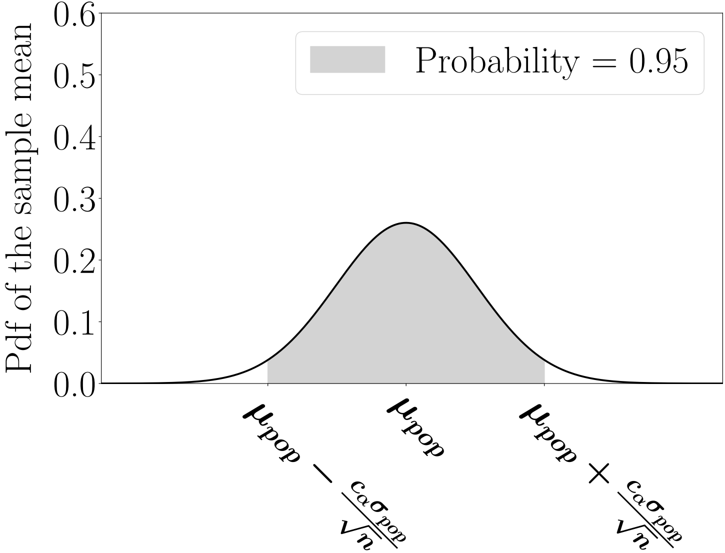
$$\mathrm{E}\left[\tilde{m}_n\right] = \mu_{\text{pop}}$$

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

As $n \to \infty$ $\tilde{m}_n$ converges in distribution to a Gaussian with mean μ_{pop} and standard deviation $\text{se}\left[\widetilde{m}\right]$

Approximate distribution of the sample mean

Can we use this interval to quantify uncertainty?

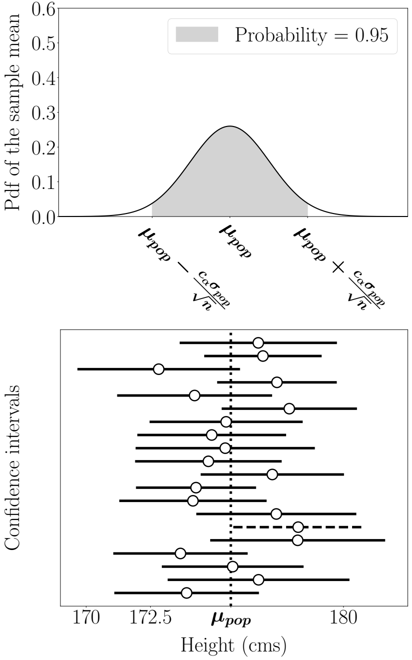


Confidence interval?

$$\widetilde{m} \in [\mu_{\text{pop}} - c, \mu_{\text{pop}} + c]$$

Problem: We don't know μ_{pop}!

$$\mu_{\text{pop}} \in [\widetilde{m} - c, \widetilde{m} + c]$$



Probability = 0.95
Pdf of the sample mean
0.0
0.1
0.2
0.3
0.4
0.5
0.6
$\boldsymbol{\mu_{pop} - \frac{c_\alpha \sigma_{pop}}{\sqrt{n}}}$
$\boldsymbol{\mu_{pop}}$
$\boldsymbol{\mu_{pop} + \frac{c_\alpha \sigma_{pop}}{\sqrt{n}}}$
Confidence intervals
170
172.5
$\boldsymbol{\mu_{pop}}$
180
Height (cms)

Reminder

If $\tilde{a}$ is a Gaussian random variable with mean μ and variance σ^2

$$\tilde{b} := \alpha\tilde{a} + \beta$$

is Gaussian with mean $\alpha\mu + \beta$ and variance $\alpha^2\sigma^2$

Confidence interval for a Gaussian

Let $\tilde{a}$ be Gaussian with mean μ and variance σ^2

$$\widetilde{\mathcal{I}}_{1-\alpha} := [\tilde{a} - c_\alpha \sigma, \tilde{a} + c_\alpha \sigma] \qquad c_\alpha := F_{\tilde{z}}^{-1}\left(1 - \frac{\alpha}{2}\right)$$
$$\widetilde{\mathcal{I}}_{0.95} := [\tilde{a} - 1.96\sigma, \tilde{a} + 1.96\sigma]$$

$$\begin{aligned}
\mathrm{P}\left(\mu \in \widetilde{\mathcal{I}}_{1-\alpha}\right) &= 1 - \mathrm{P}\left(\tilde{a} - c_\alpha \sigma > \mu\right) - \mathrm{P}\left(\tilde{a} + c_\alpha \sigma < \mu\right) \\
&= 1 - \mathrm{P}\left(\frac{\tilde{a} - \mu}{\sigma} > c_\alpha\right) - \mathrm{P}\left(\frac{\tilde{a} - \mu}{\sigma} < -c_\alpha\right) \\
&= 1 - \mathrm{P}\left(\tilde{z} > c_\alpha\right) - \mathrm{P}\left(\tilde{z} < -c_\alpha\right) \\
&= 1 - 2\mathrm{P}\left(\tilde{z} > c_\alpha\right)
\end{aligned}$$

Confidence interval for the population mean

Population mean: μ_{pop} Population variance: σ^2_{pop}

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

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

$$\mathrm{E}\left[\tilde{m}_n\right] = \mu_{\text{pop}} \qquad \text{se}\left[\widetilde{m}_n\right] = \frac{\sigma_{\text{pop}}}{\sqrt{n}}$$

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

$$\widetilde{\mathcal{I}}_{0.95} := \left[\tilde{m} - \frac{1.96\sigma_{\text{pop}}}{\sqrt{n}}, \tilde{m} + \frac{1.96\sigma_{\text{pop}}}{\sqrt{n}}\right]$$

Wait a minute

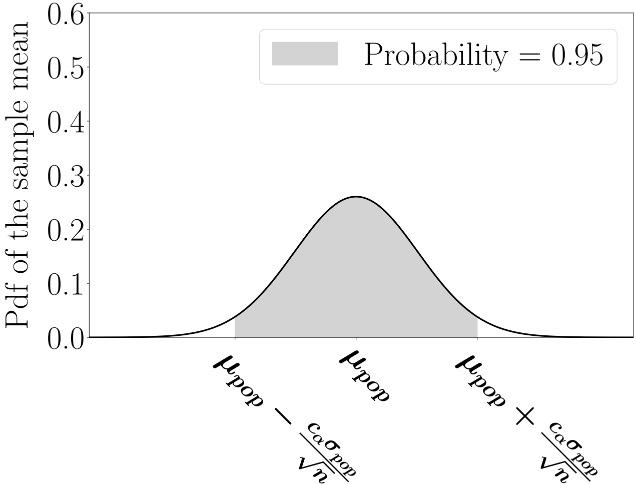
We don't know σ_{pop}!

Solution: Use sample standard deviation or an upper bound

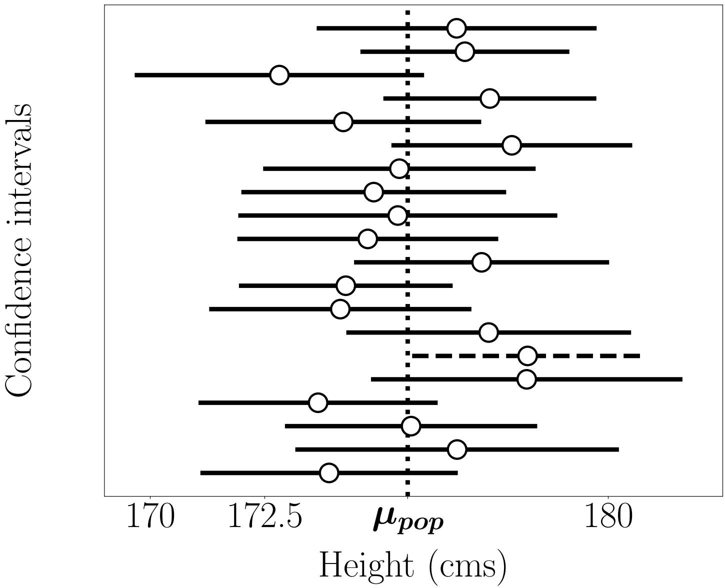
Height data: $n = 20$

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

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



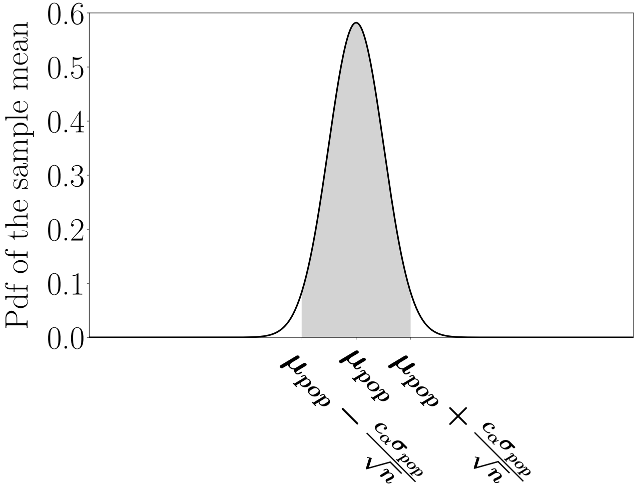
0.95 confidence intervals ($n = 20$)



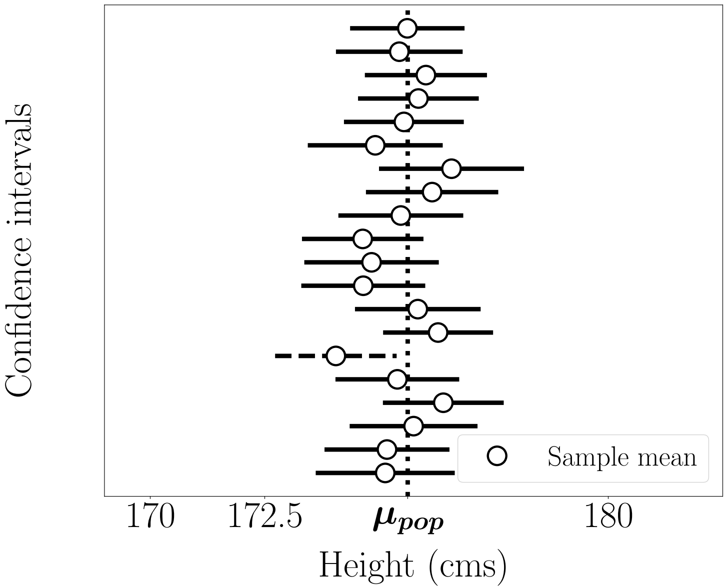
Height data: $n = 100$

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

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



0.95 confidence intervals ($n = 100$)



Interpretation of confidence intervals

Confidence interval for population mean of height data:

$$[174.6, 177.4]$$

Tempting interpretation:

The probability that the mean height is between 174.6 cms and 177.4 cms is 0.95

Problem: No random quantities, the mean height is 175.6!

Interpretation of confidence intervals

Confidence interval for population mean of height data:

$$[174.6, 177.4]$$

Tempting interpretation:

The probability that 175.6 is between 174.6 cms and 177.4 cms is 0.95

Problem: No random quantities, the mean height is 175.6!

Interpretation of confidence intervals

0.95 Confidence interval for population mean of height data

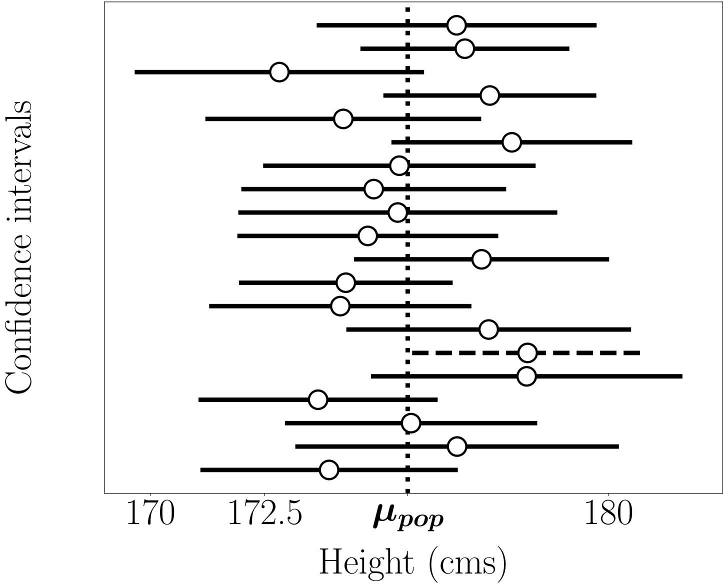
$$[174.6, 177.4]$$

Correct interpretation:

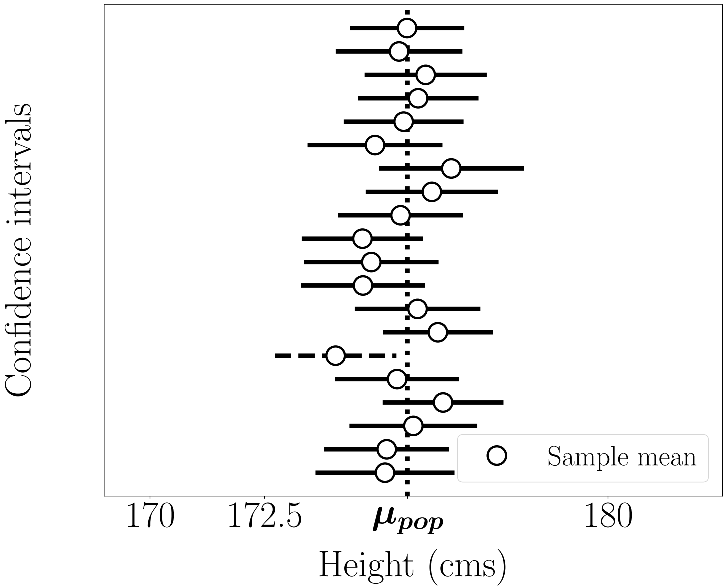
If we repeat the same procedure many times, 95% of the time the interval will contain the population mean

Equivalently, if we build many 0.95 confidence intervals, 95% of them contain the population parameter

Height data: $n = 20$



Height data: $n = 100$



What have we learned

Definition of confidence intervals

How to build confidence intervals for the population mean

Interpretation of confidence intervals