The Bootstrap

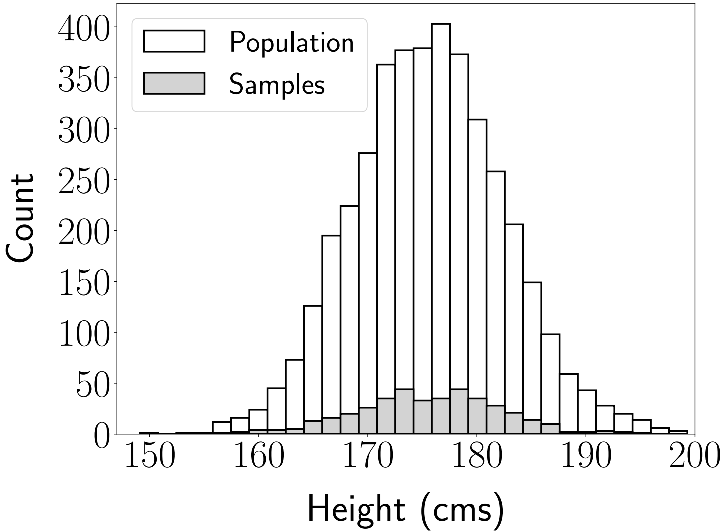
Probability and Statistics for Data Science

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

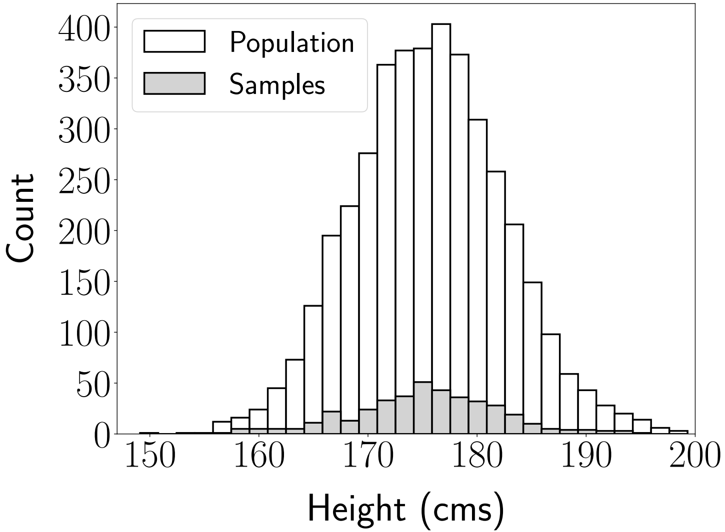
Random sampling

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



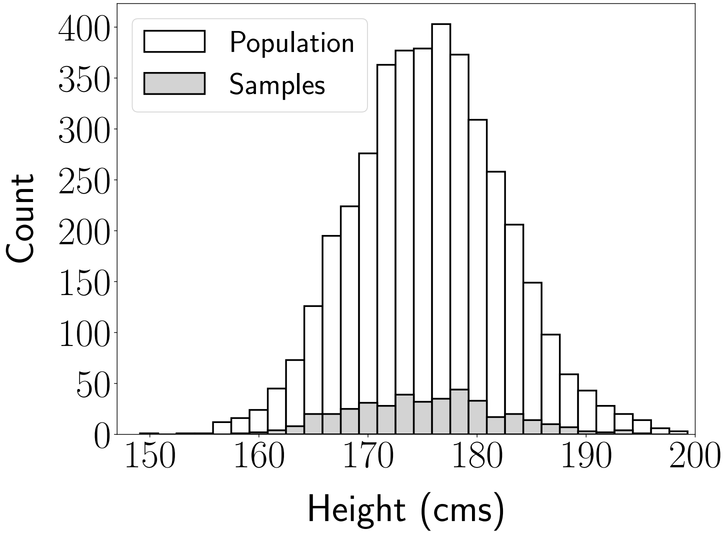
400 random samples

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



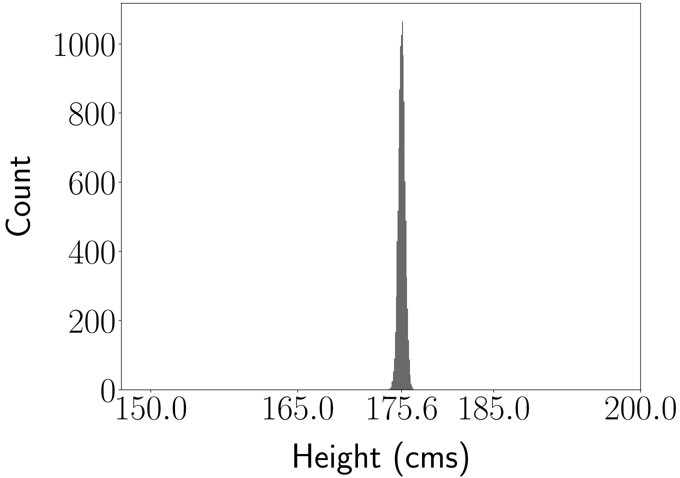
400 random samples

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



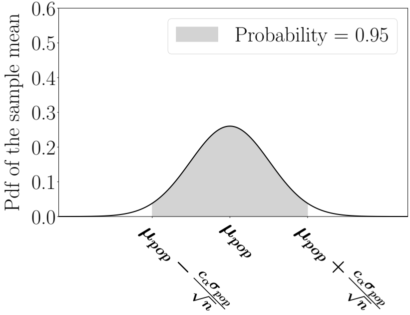
Sample means of 10,000 subsets of size 400

Goal: Quantify uncertainty from available data

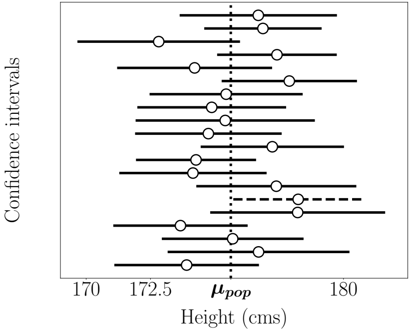


Confidence interval

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



Probability = 0.95
Pdf of the sample mean
0.0
0.1
0.2
0.3
0.4
0.5
0.6
$\mu_{pop} - \frac{c_\alpha \sigma_{pop}}{\sqrt{n}}$
μ_{pop}
$\mu_{pop} + \frac{c_\alpha \sigma_{pop}}{\sqrt{n}}$



Confidence intervals
170
172.5
μ_{pop}
180
Height (cms)

Standard error

We need to estimate standard error

For sample mean

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

We use sample standard deviation to estimate σ_{pop}

What if we don't know the formula?

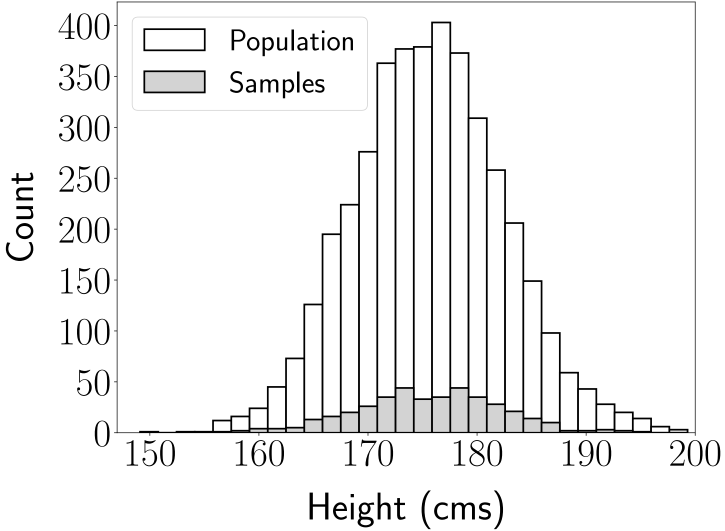
Challenge

How to estimate standard error computationally?

If we can sample more data, this is easy

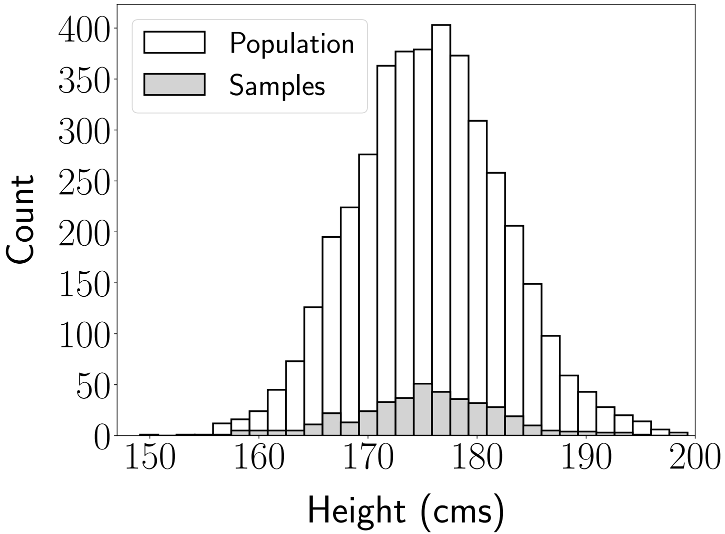
We sample n data points

Sample mean: 175.5



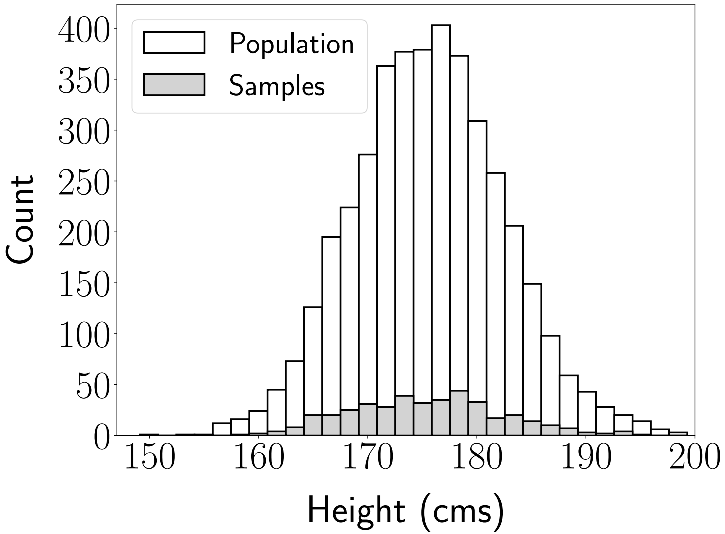
and *n* more

Sample mean: 175.2



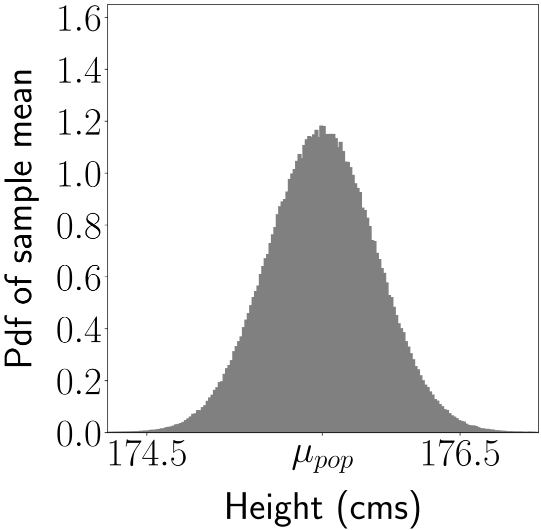
and n more

Sample mean: 176.1



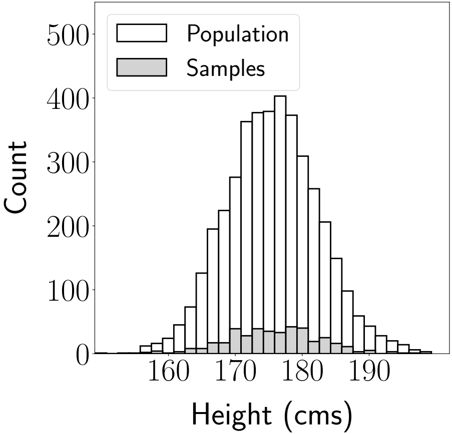
Distribution of sample means

Standard error = standard deviation = 0.343



Problem

We only have *n* data points



Idea: Sample from the *n* data as if *they were the population*

The bootstrap

Samples: $x_1, \ldots, x_n$

Bootstrap indices: $\tilde{k}_1$, $\tilde{k}_2$, ..., $\tilde{k}_n$

Sampled independently and uniformly with replacement

$$\mathrm{P}\left(\tilde{k}_j = i\right) = \frac{1}{n} \qquad 1 \leq i, j \leq n$$

Bootstrap samples: $\tilde{b}_1$, ..., $\tilde{b}_n$

$$\tilde{b}_j = x_{\tilde{k}_j} \qquad 1 \leq j \leq n$$

The bootstrap

Bootstrap standard error

Samples: $x_1, \ldots, x_n$

Estimator: $h(x_1, \ldots, x_n)$

Bootstrap samples: $\tilde{b}_1$, $\tilde{b}_2$, $\ldots$, $\tilde{b}_n$

The bootstrap standard error of h is

$$\mathsf{se}_{\mathsf{bs}} = \sqrt{\mathrm{Var}\left[h(\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_n)\right]}$$

Monte Carlo approximation

(1) Generate K batches, $b_j^{[k]}$, $1 \leq j \leq n$, $1 \leq k \leq K$

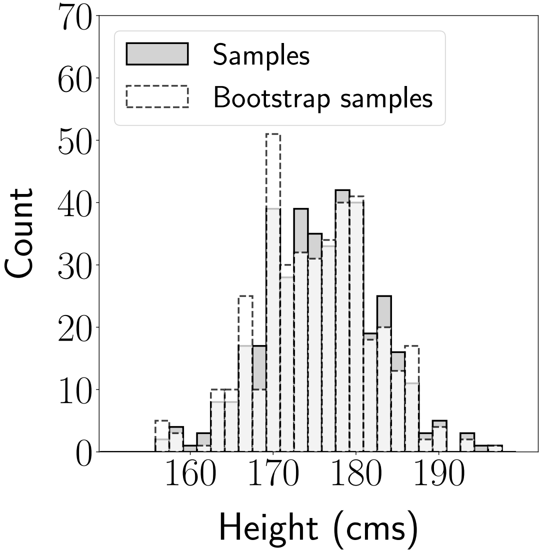
(2) Compute parameter estimates

$$W := \{w_1, w_2, \ldots, w_K\}, \qquad w_k := h(b_1^{[k]}, b_2^{[k]}, \ldots, b_n^{[k]})$$

(3) Bootstrap standard error: Sample standard deviation of W

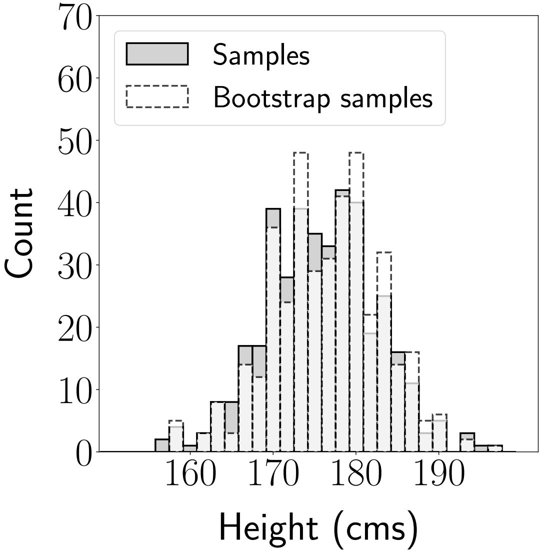
Bootstrap samples

Bootstrap sample mean: 175.3



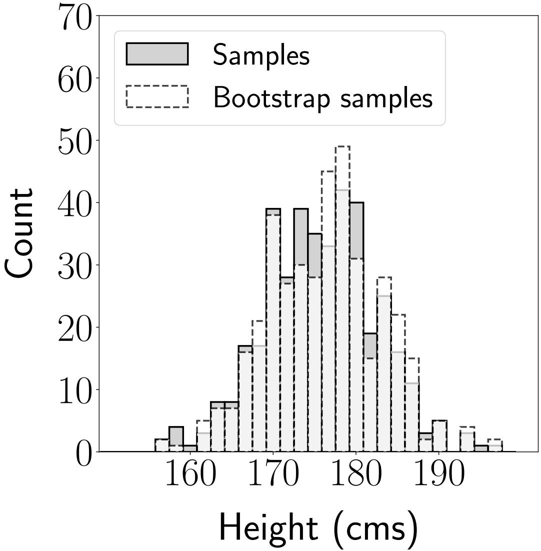
Bootstrap samples

Bootstrap sample mean: 176.6



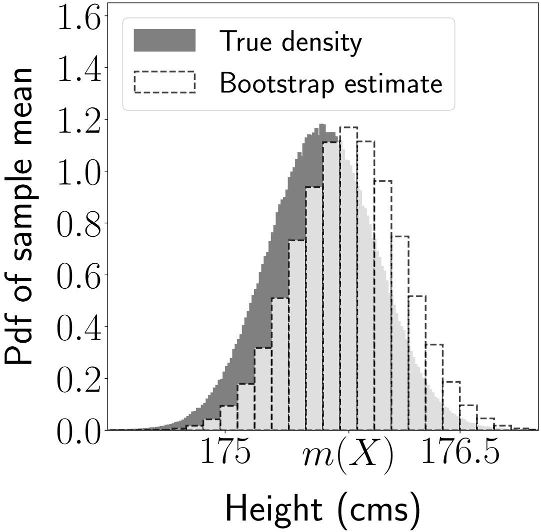
Bootstrap samples

Bootstrap sample mean: 176.2



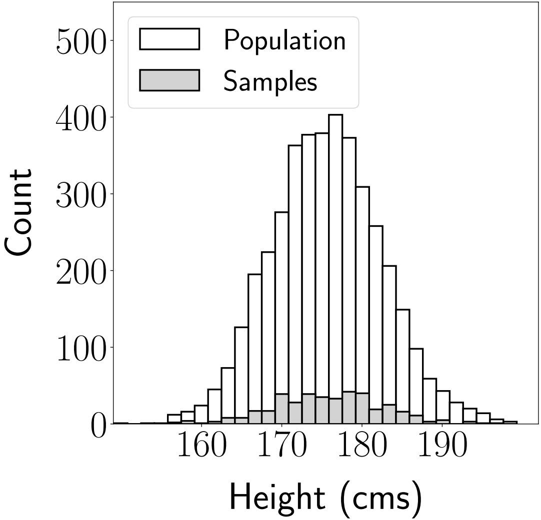
Distribution of bootstrap samples

Bootstrap standard error: 0.339 (True standard error: 0.343)



Traditional standard-error estimate

$\sqrt{\frac{v(X)}{n}} = 0.340$ (Bootstrap estimate: 0.339)



Bootstrap standard error of the sample mean

Samples $X := \{x_1, \ldots, x_n\}$ are the "population"

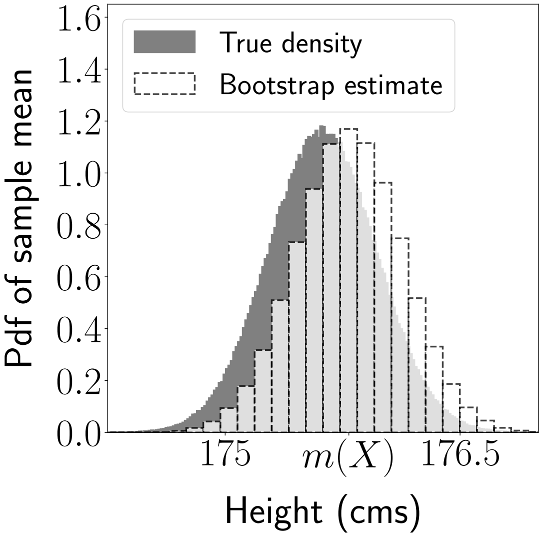
$$\widetilde{m}_{\text{bs}} := \frac{1}{n} \sum_{k=1}^{n} \tilde{b}_k$$

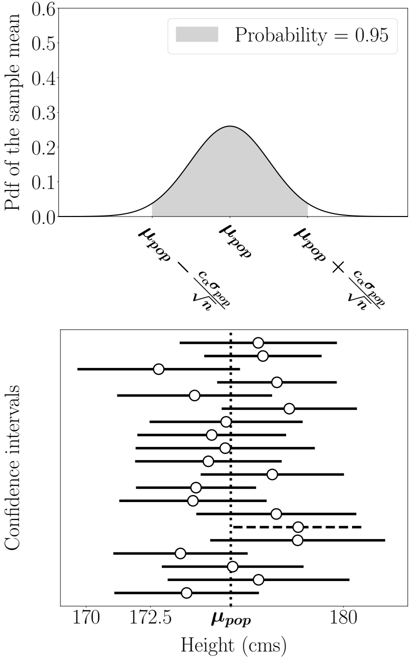
$$\begin{aligned} \mathrm{E}\left[\widetilde{m}_{\text{bs}}\right] &= \text{"Population" mean} \\ &= \frac{1}{n} \sum_{j=1}^{n} x_j = m(X) \end{aligned}$$

$$\begin{aligned} \text{se}_{\text{bs}}^2 = \mathrm{Var}\left[\widetilde{m}_{\text{bs}}\right] &= \frac{\text{"Population" variance}}{n} \\ &= \frac{\frac{1}{n} \sum_{j=1}^{n} (x_j - m(X))^2}{n} \\ &= \frac{n-1}{n^2} v(X) \end{aligned}$$

Distribution of bootstrap samples

Bootstrap standard error: 0.339 ($\sqrt{\frac{v(X)}{n}} = 0.340$)




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

Confidence interval for a Gaussian

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

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

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

Bootstrap Gaussian confidence interval

Samples: $X := \{x_1, \ldots, x_n\}$

Estimator: $h(X)$

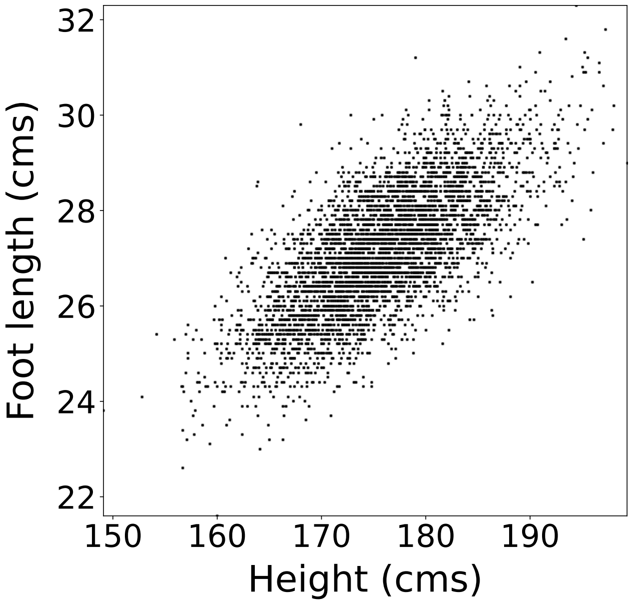
Bootstrap standard error: se_{bs}

1-α bootstrap Gaussian confidence interval

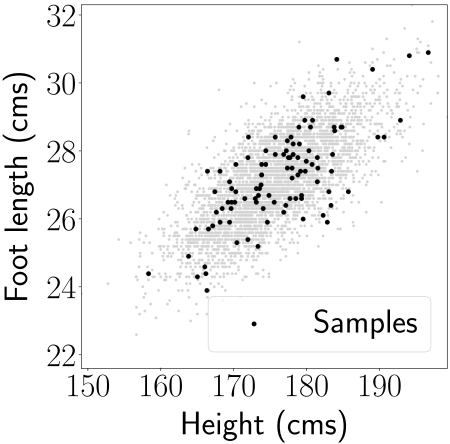
$$\mathcal{I}_{1-\alpha}^{\text{BSG}} := \left[h(X) - c_\alpha \text{se}_{\text{bs}}, h(X) + c_\alpha \text{se}_{\text{bs}}\right] \qquad c_\alpha := F_{\tilde{z}}^{-1}\left(1 - \frac{\alpha}{2}\right)$$

$$\widetilde{\mathcal{I}}_{0.95} := \left[h(X) - 1.96\, \text{se}_{\text{bs}}, h(X) + 1.96\, \text{se}_{\text{bs}}\right]$$

Population correlation coefficient: 0.718



100 samples

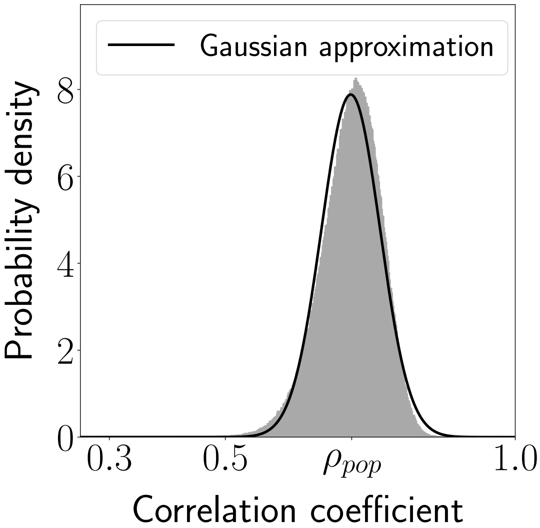


Sample correlation coefficient: $\rho_{\text{sample}} = 0.727$

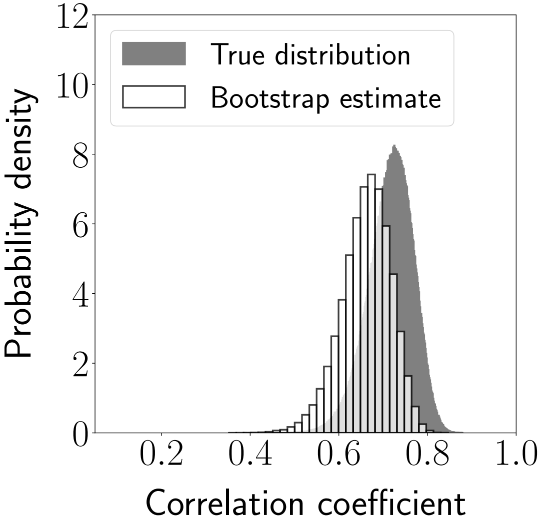
Confidence interval?

Distribution of sample correlation coefficient

True standard error: 0.051



Bootstrap standard error $\text{se}_{\text{bs}} = 0.056$



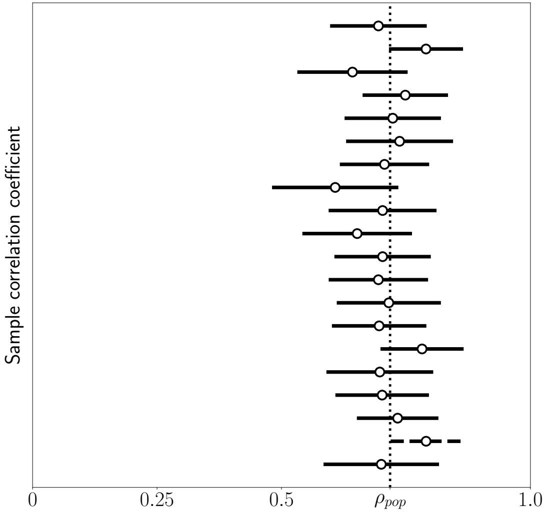
Bootstrap Gaussian confidence interval

$$\mathcal{I}_{1-\alpha}^{\mathsf{BSG}} := [\rho_{\mathsf{sample}} - c_{\alpha}\, \mathsf{se}_{\mathsf{bs}}, \rho_{\mathsf{sample}} + c_{\alpha}\, \mathsf{se}_{\mathsf{bs}}]$$

$$\begin{aligned} \mathcal{I}_{0.95}^{\mathsf{BSG}} &:= [\rho_{\mathsf{sample}} - 1.96\, \mathsf{se}_{\mathsf{bs}}, \rho_{\mathsf{sample}} + 1.96\, \mathsf{se}_{\mathsf{bs}}] \\ &= [0.617, 0.837] \end{aligned}$$

Bootstrap Gaussian confidence intervals

Coverage: 93.7% (out of 10^4)



What have we learned

Definition of the bootstrap

Bootstrap standard error

Bootstrap Gaussian confidence interval