Estimation of Population Parameters

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

1. Random sampling
2. The bias
3. The standard error
4. The law of large numbers
5. The central limit theorem
6. Confidence intervals
7. The bootstrap

Random sampling

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

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

Goal: Estimate population mean

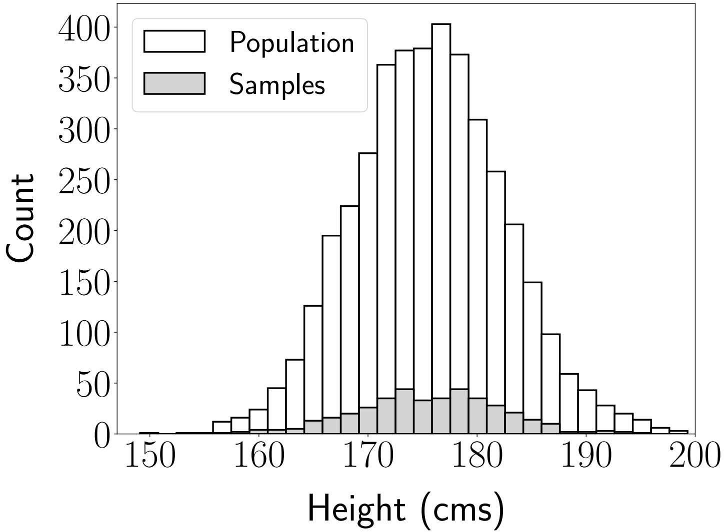
$$\mu_{\text{pop}} := \frac{1}{N} \sum_{i=1}^{N} h_i = 175.6$$

Challenge: We cannot measure everyone

Solution: Choose a random subset

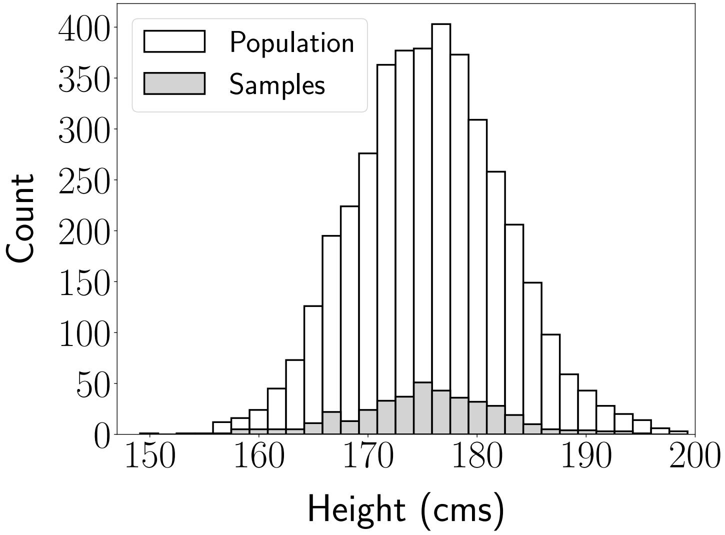
400 random samples

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



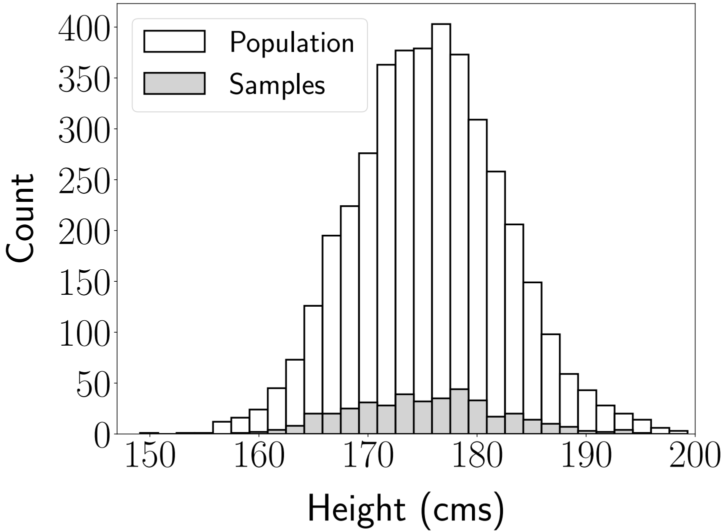
400 random samples

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



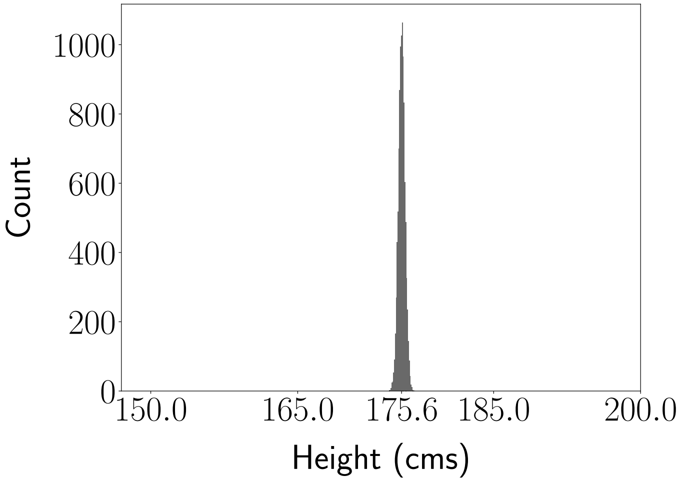
400 random samples

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



Random sampling

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

Random indices: $\tilde{k}_1, \tilde{k}_2, \ldots, \tilde{k}_n$

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

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

$$\tilde{x}_j = a_{\tilde{k}_j} \qquad 1 \leq j \leq n$$

Sample mean

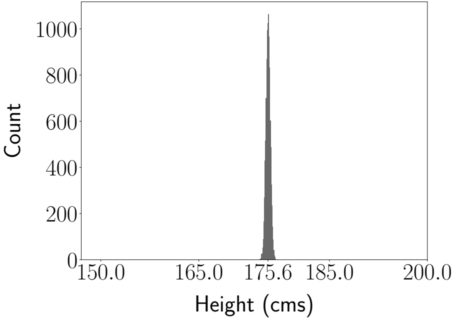
Modeled as a random variable

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

Estimation of population parameters

Frequentist perspective

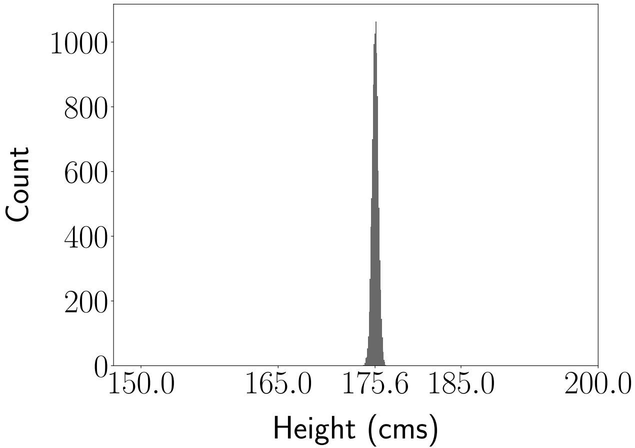
The parameter of interest is deterministic



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: $\gamma \in \mathbb{R}$

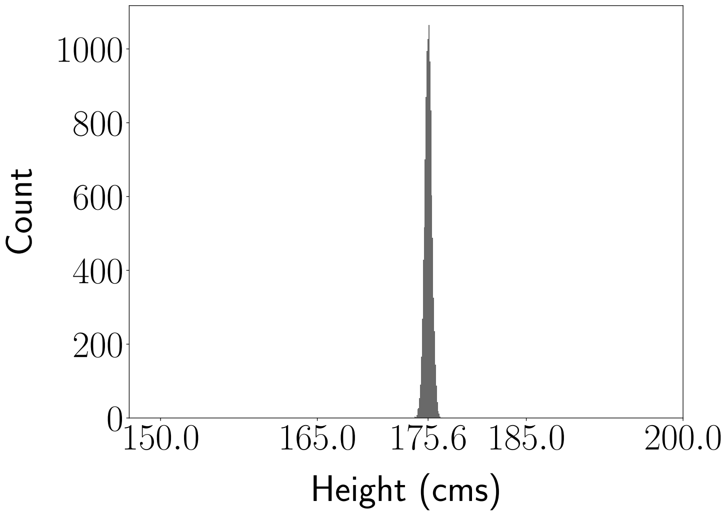
Estimator: $h(\tilde{x}_1, \ldots, \tilde{x}_n)$

The bias of the estimator is the mean of the error

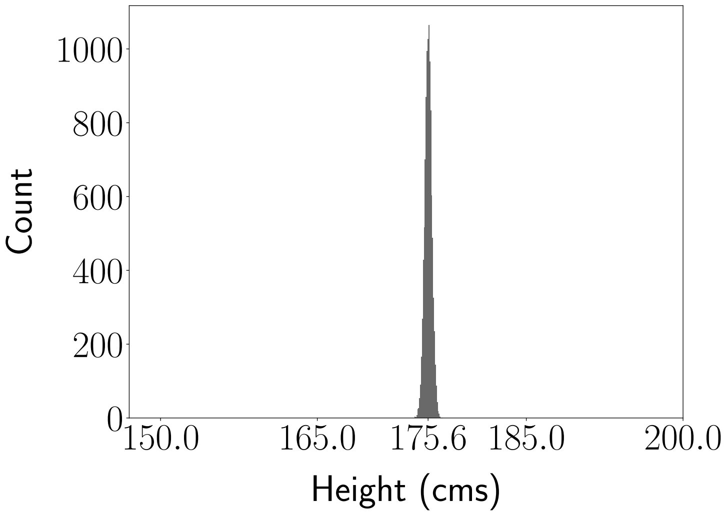
$$\text{Bias} = \mathrm{E}\left[h(\tilde{x}_1, \ldots, \tilde{x}_n) - \gamma\right]$$

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

The sample mean is unbiased



Is an unbiased estimator enough?



Standard error

The standard error of the estimator is its standard deviation

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

Standard error of the sample mean

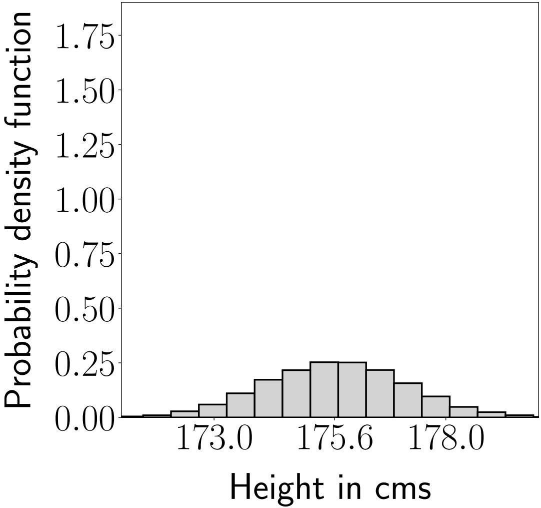
$$\mathrm{se}\left[\widetilde{m}\right] = \frac{\sigma_{\mathrm{pop}}}{\sqrt{n}}$$

No dependence on N!

Height data: $n = 20$

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

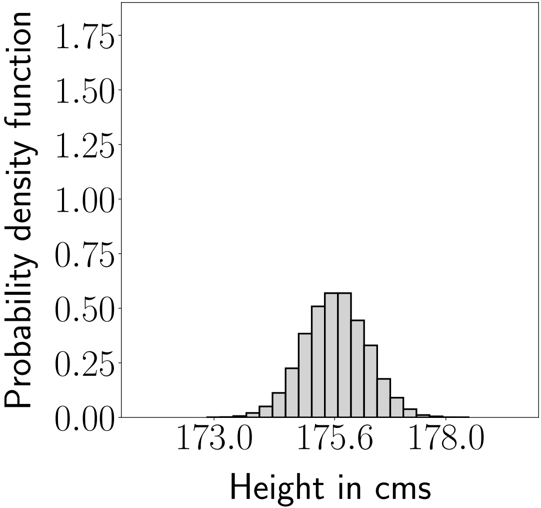
10^4 sample means



$n = 100$

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

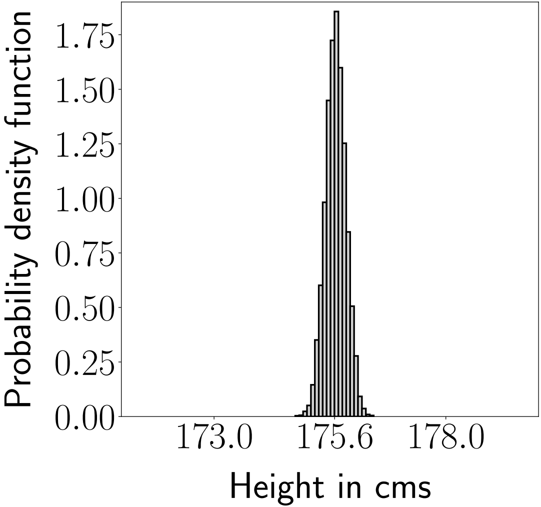
10^4 sample means



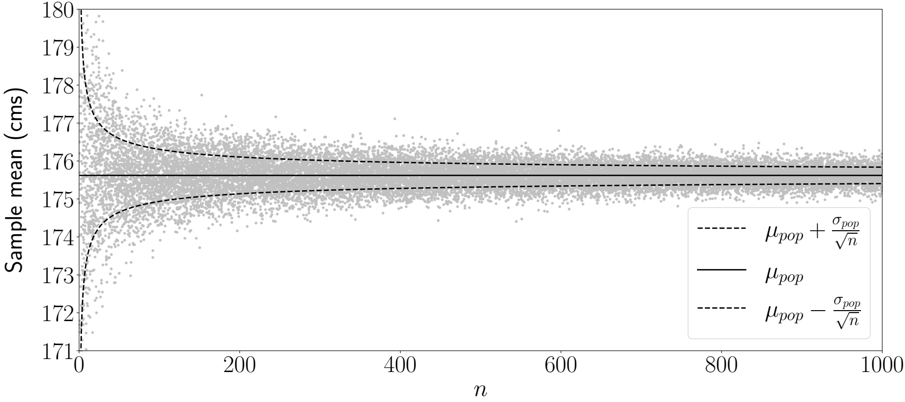
n =1,000

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

10^4 sample means



Height data



$$\lim_{n \to \infty} \text{MSE}_n = \lim_{n \to \infty} \mathrm{E}\left[\left(\widetilde{m}_n - \mu_{\text{pop}} \right)^2 \right] = 0$$

Convergence in probability

Probability of deviating by ϵ

$$p_n := \mathrm{P}\left(|\widetilde{m}_n - \mu_{\text{pop}}| > \epsilon\right)$$

$$p_1,\ p_2,\ p_3,\ p_4,\ \ldots$$

Chebyshev's inequality

A random variable with small variance cannot be far from its mean μ with high probability

Law of large numbers

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

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

$$\mathrm{P}\left(|\tilde{m}_n - \mu| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

Converges to zero for any ϵ!

Consistency

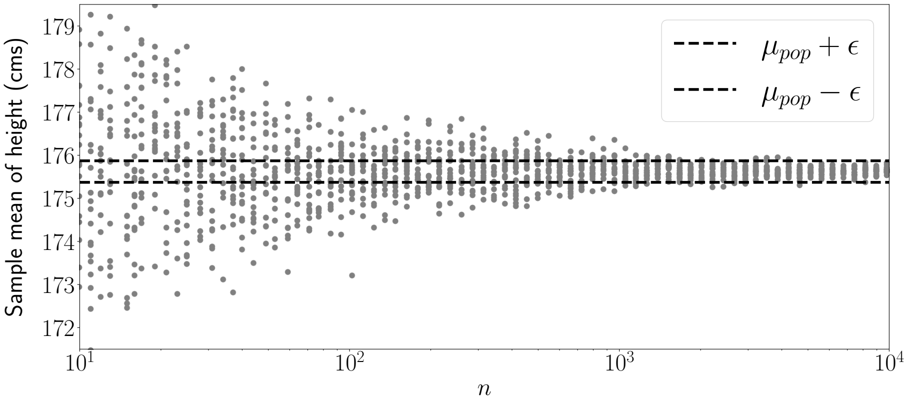
Random measurements: $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_n$

Deterministic parameter of interest: γ

An estimator $h(\tilde{x}_1, \ldots, \tilde{x}_n)$ is consistent if for any $\epsilon > 0$

$$\lim_{n \to \infty} \mathrm{P}\left(|h(\tilde{x}_1, \ldots, \tilde{x}_n) - \gamma| > \epsilon\right) = 0$$

The sample mean is consistent

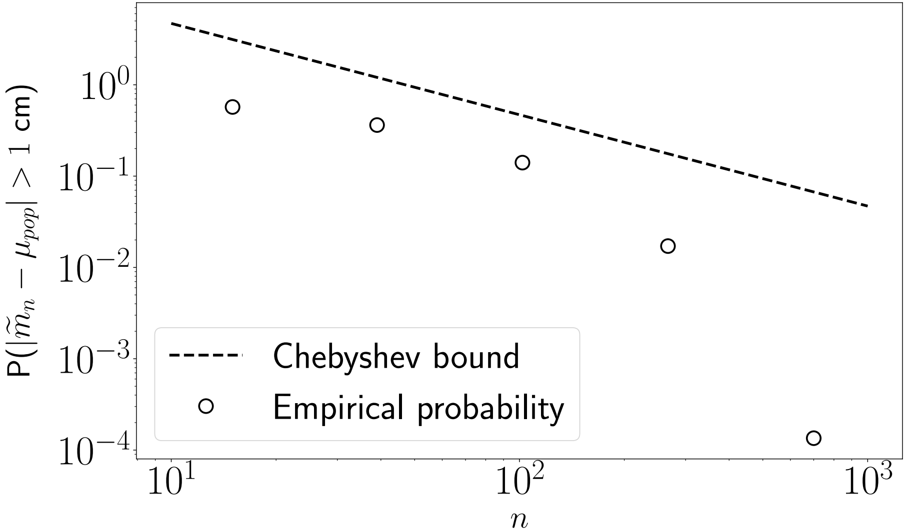


Chebyshev bound

$$\mathrm{P}\left(|\tilde{m}_n - \mu_{\mathsf{pop}}| > \epsilon\right) \leq \frac{\sigma_{\mathsf{pop}}^2}{n\epsilon^2}$$

Is this a good approximation?

No!



$P(|\widetilde{m}_n - \mu_{pop}| > 1 \text{ cm})$
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^1
10^2
10^3
n
Chebyshev bound
Empirical probability

Goal

Approximate the distribution of the sample mean

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

Sum of independent discrete random variables

Independent discrete random variables $\tilde{a}$ and $\tilde{b}$ with integer values

The pmf of $\tilde{s} = \tilde{a} + \tilde{b}$ is

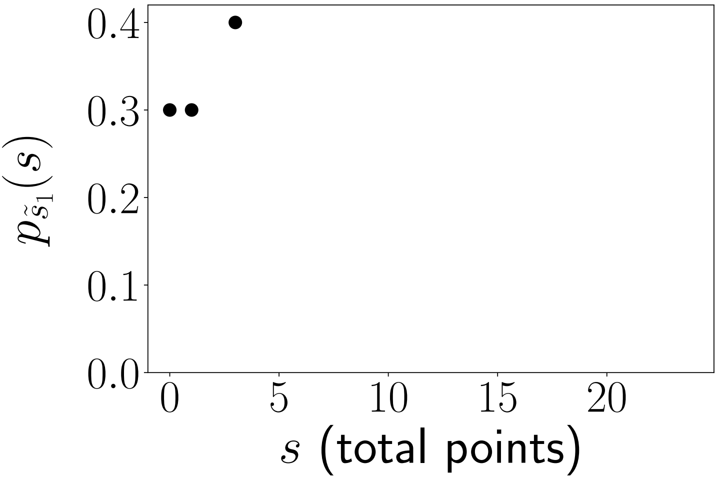
$$p_{\tilde{s}}(s) = \sum_{a=-\infty}^{\infty} p_{\tilde{a}}(a)\, p_{\tilde{b}}(s-a) = p_{\tilde{a}} * p_{\tilde{b}}(s)$$

Independent discrete random variables $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$ with integer values

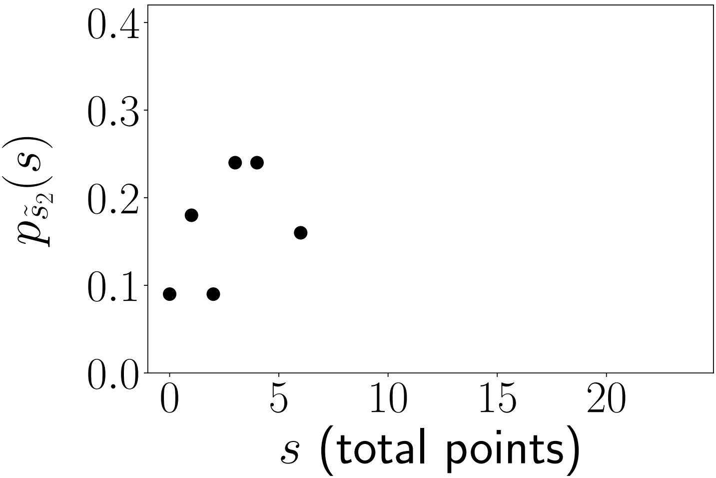
The pmf of $\tilde{s}_n = \sum_{i=1}^{n} \tilde{a}_i$ is

$$p_{\tilde{s}_n}(s) = p_{\tilde{a}_1} * p_{\tilde{a}_2} * \cdots * p_{\tilde{a}_n}(s)$$

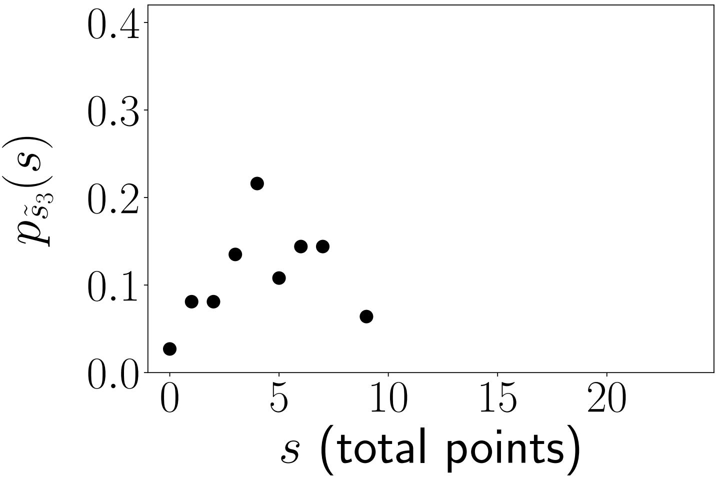
Soccer league: 1 game



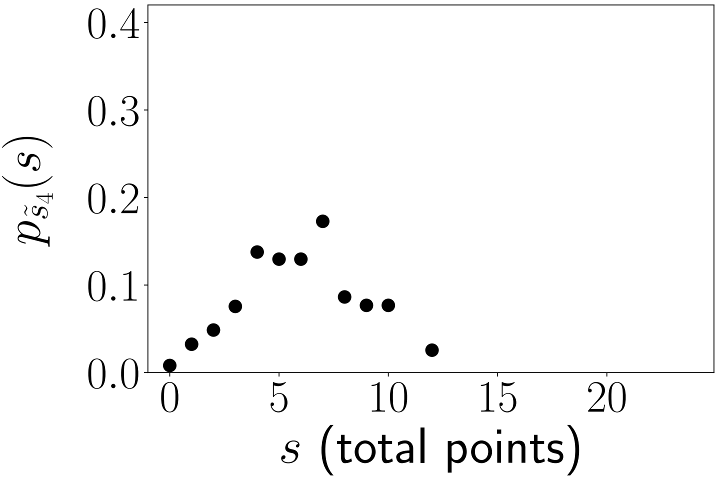
Soccer league: 2 games



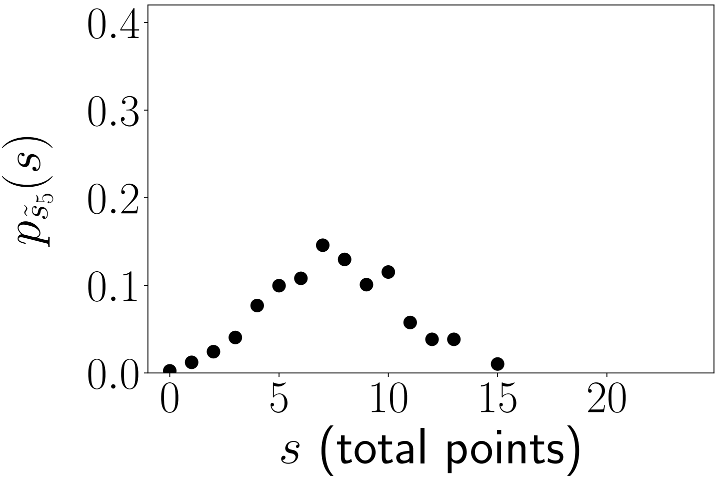
Soccer league: 3 games



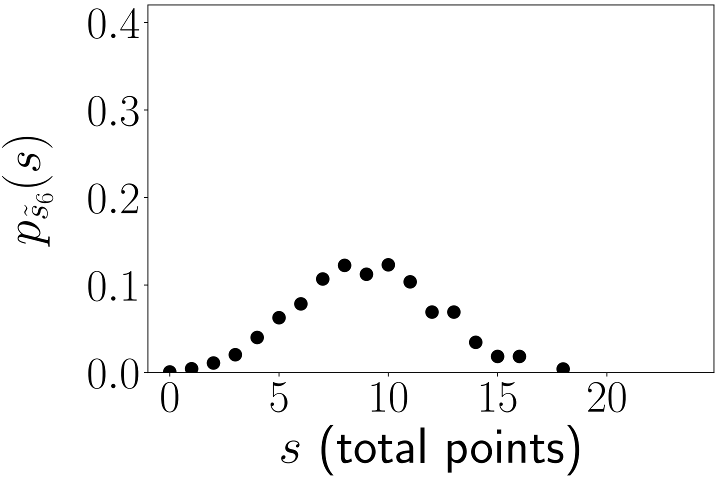
Soccer league: 4 games



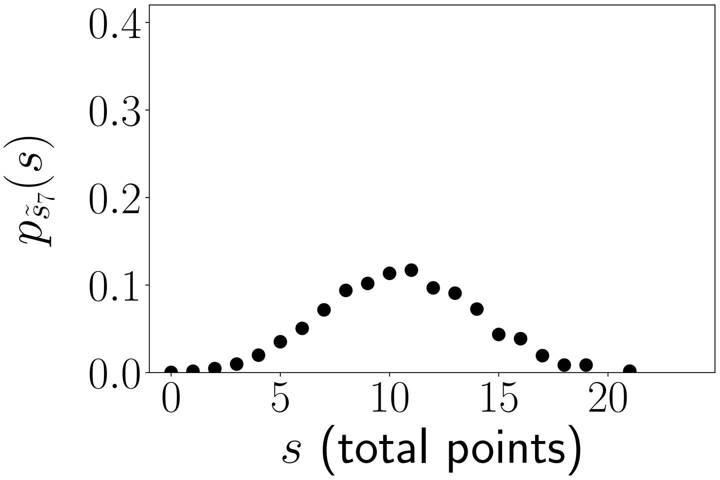
Soccer league: 5 games



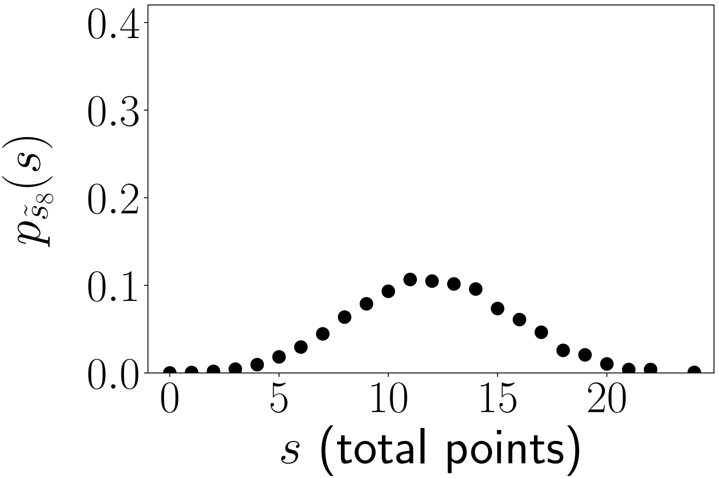
Soccer league: 6 games



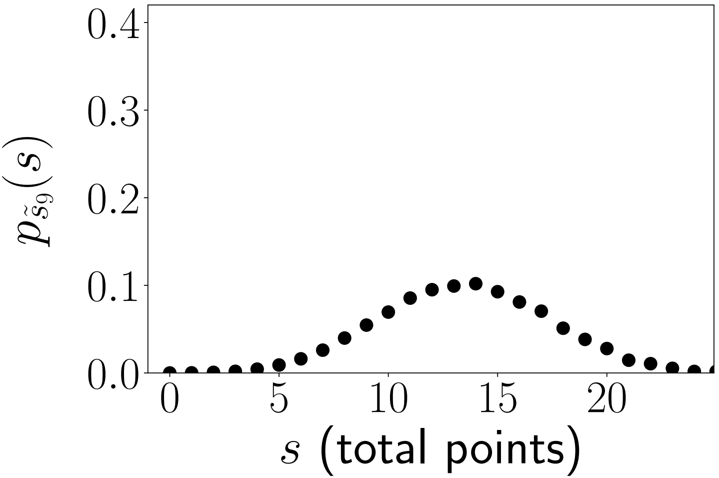
Soccer league: 7 games



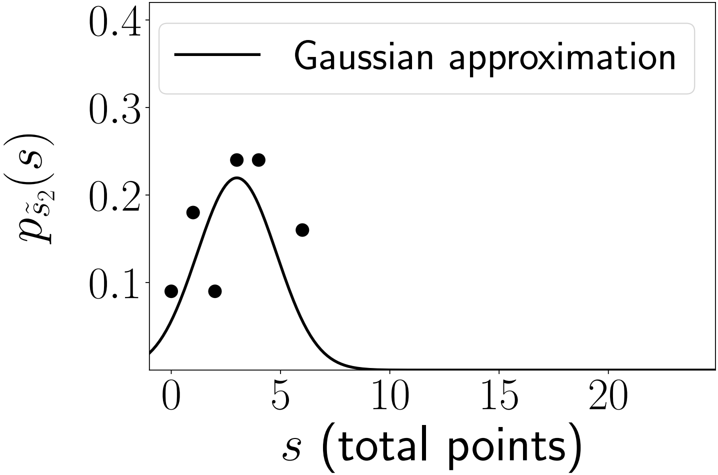
Soccer league: 8 games



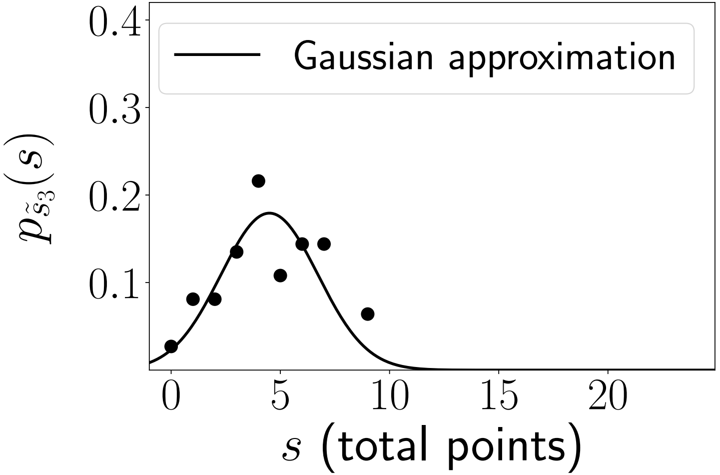
Soccer league: 9 games



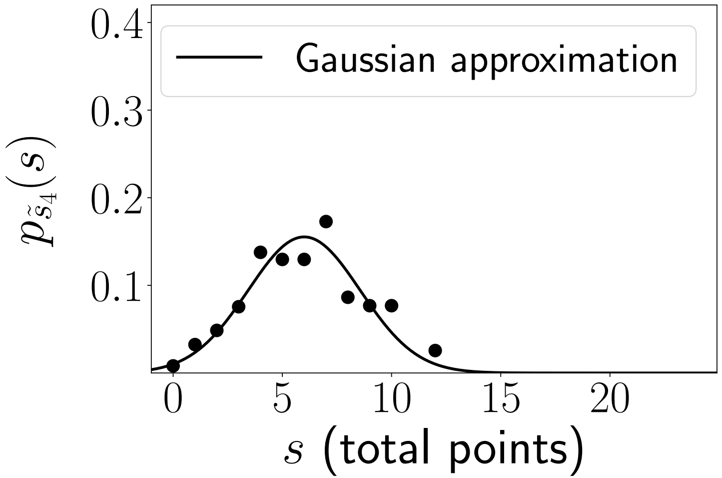
Soccer league: 2 games



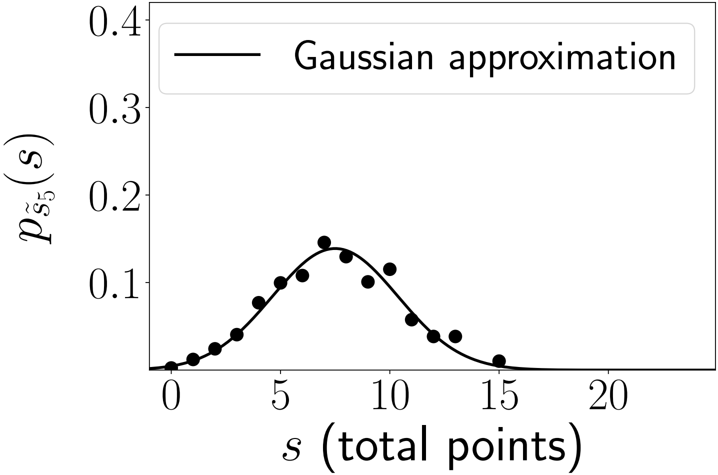
Soccer league: 3 games



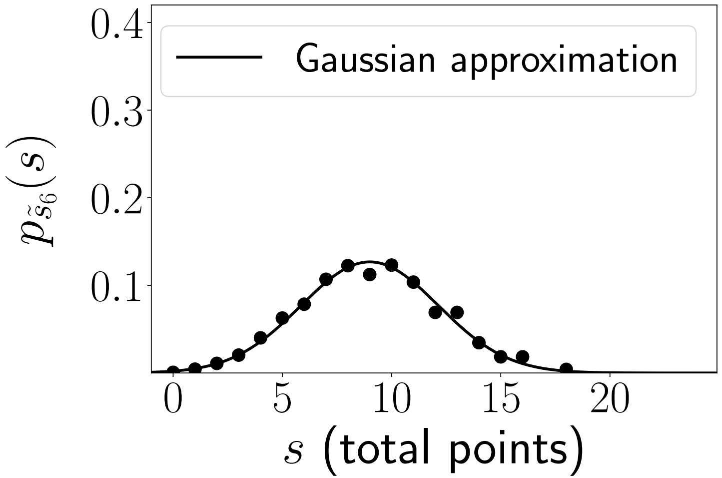
Soccer league: 4 games



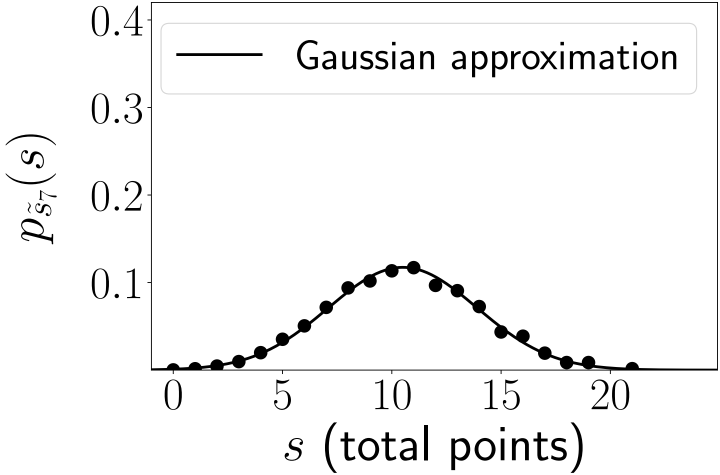
Soccer league: 5 games



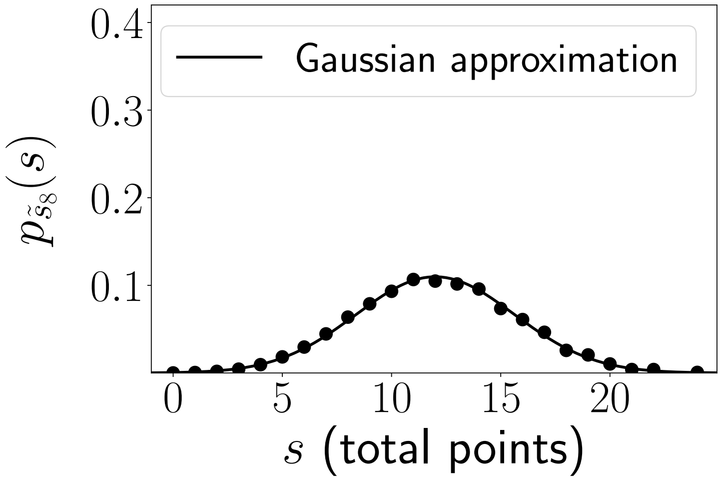
Soccer league: 6 games



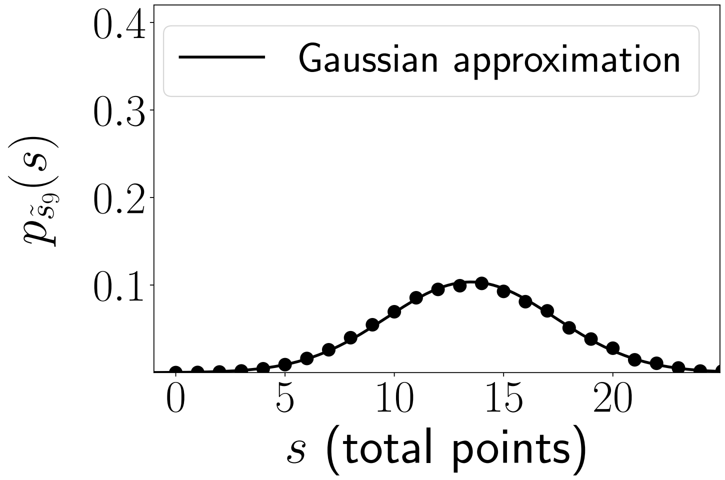
Soccer league: 7 games



Soccer league: 8 games



Soccer league: 9 games



Sum of independent continuous random variables

Independent continuous random variables $\tilde{a}$ and $\tilde{b}$

The pdf of $\tilde{s} = \tilde{a} + \tilde{b}$ is

$$\begin{aligned} f_{\tilde{s}}(s) &= \int_{a=-\infty}^{\infty} f_{\tilde{a}}(a)\, f_{\tilde{b}}(s-a)\, \mathrm{d}a \\ &= f_{\tilde{a}} * f_{\tilde{b}}(s) \end{aligned}$$

Independent continuous random variables $\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n$

The pdf of $\tilde{s}_n = \sum_{i=1}^{n} \tilde{a}_i$ is

$$f_{\tilde{s}_n}(s) = f_{\tilde{a}_1} * f_{\tilde{a}_2} * \cdots * f_{\tilde{a}_n}(s)$$

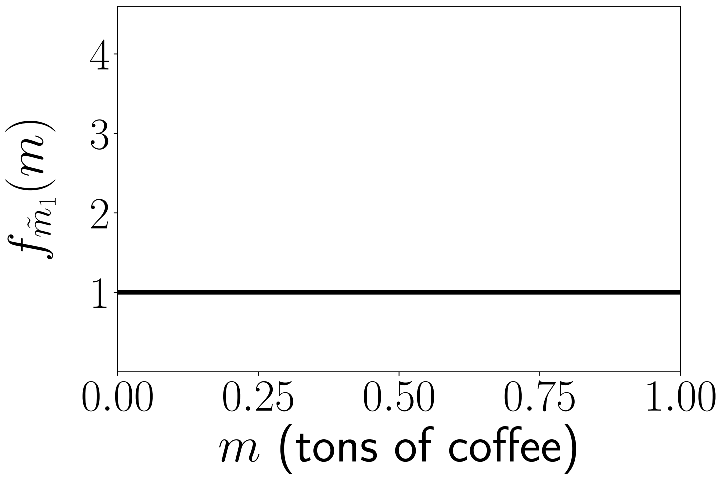
Sample mean

Independent continuous random variables $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$

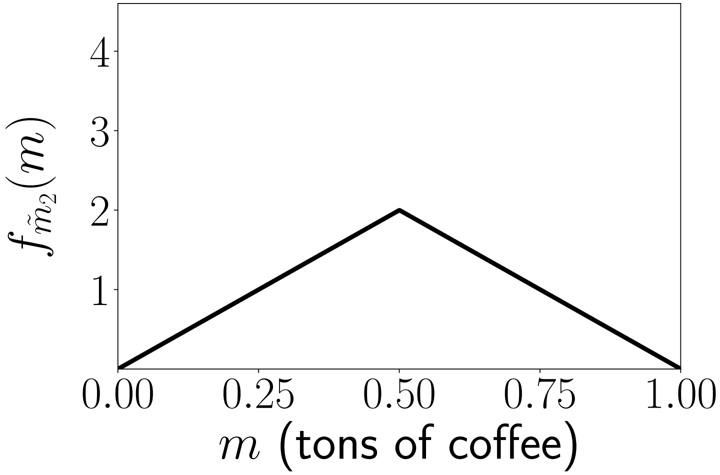
$\tilde{m}_n := \frac{1}{n}\tilde{s}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{a}_i$

$$f_{\widetilde{m}_n}(m) = n \left(f_{\tilde{a}_1} * f_{\tilde{a}_2} * \cdots * f_{\tilde{a}_n} \right)(nm)$$

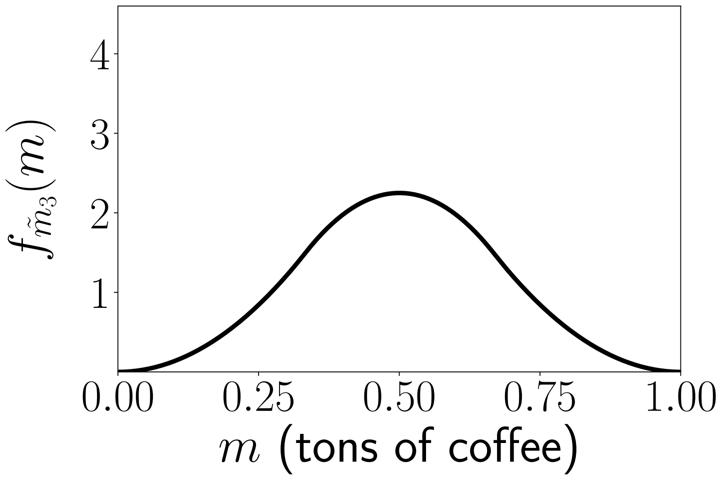
Purchased coffee: 1 supplier



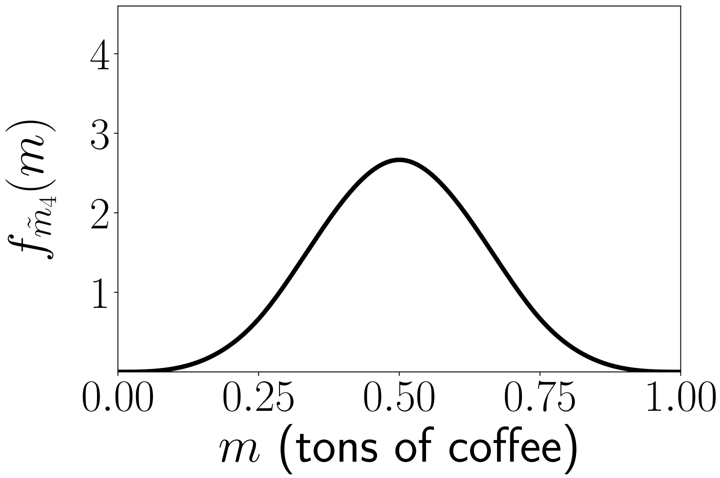
Purchased coffee: 2 suppliers



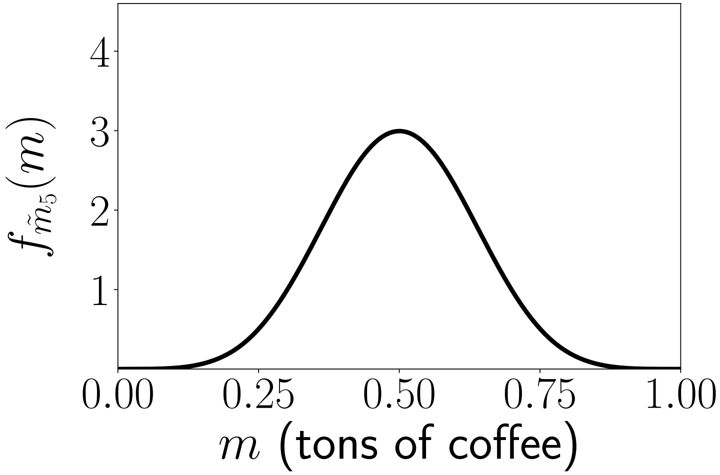
Purchased coffee: 3 suppliers



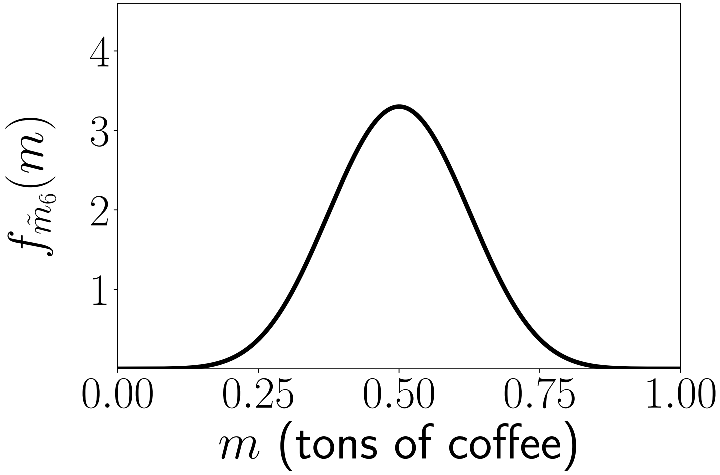
Purchased coffee: 4 suppliers



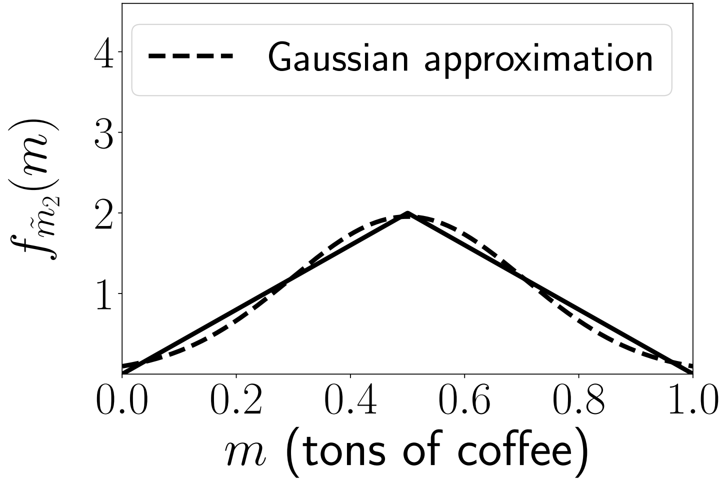
Purchased coffee: 5 suppliers



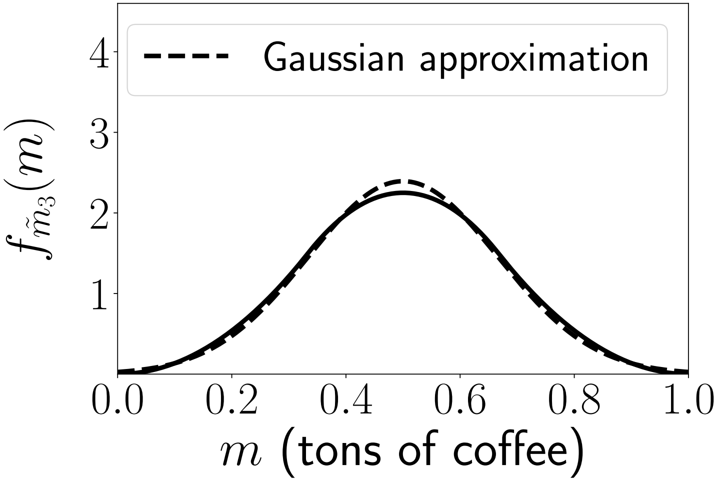
Purchased coffee: 6 suppliers



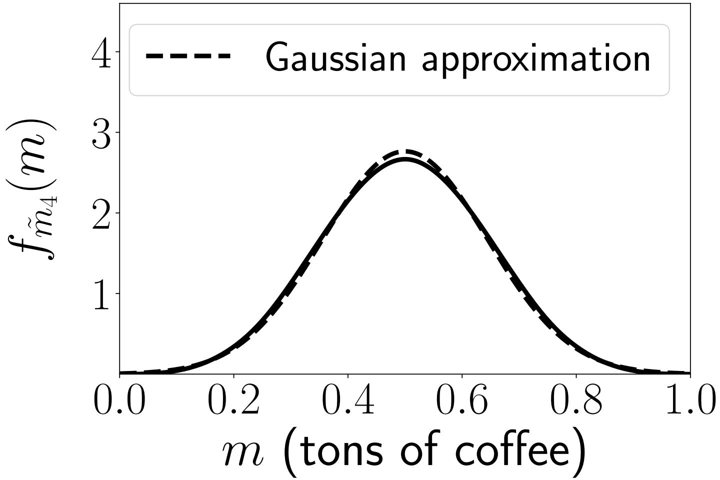
Purchased coffee: 2 suppliers



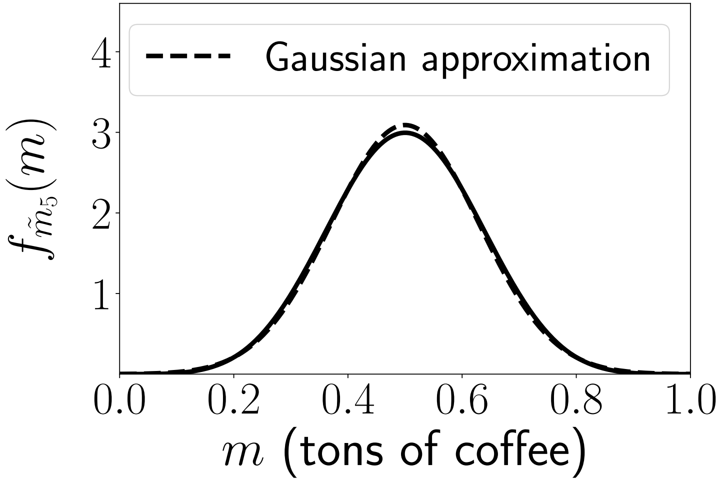
Purchased coffee: 3 suppliers



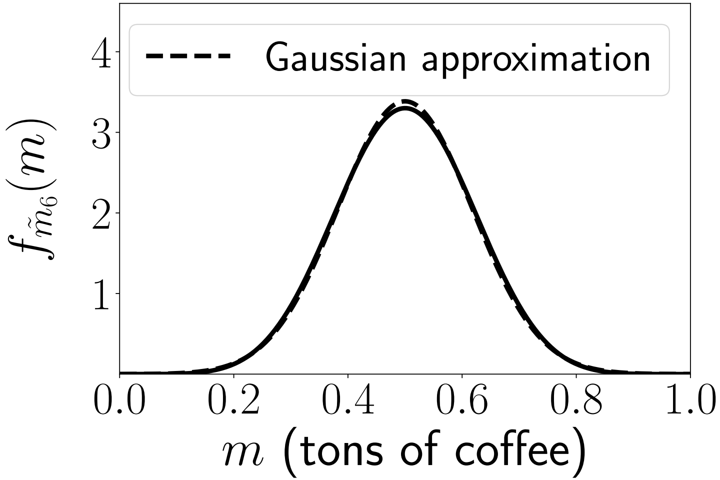
Purchased coffee: 4 suppliers



Purchased coffee: 5 suppliers



Purchased coffee: 6 suppliers



Central limit theorem

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

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

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

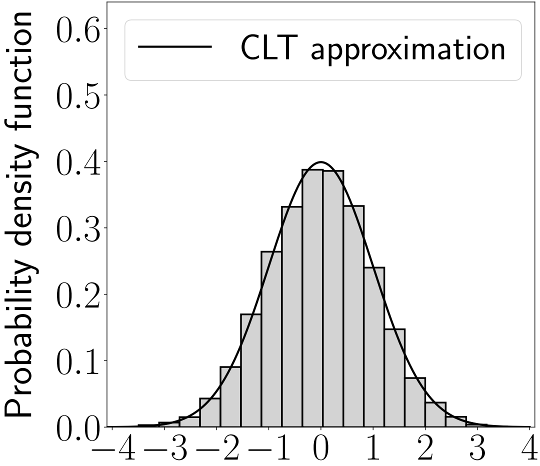
$$\mathrm{E}\left[\tilde{m}_n\right] = \mu_{\text{pop}}$$

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

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

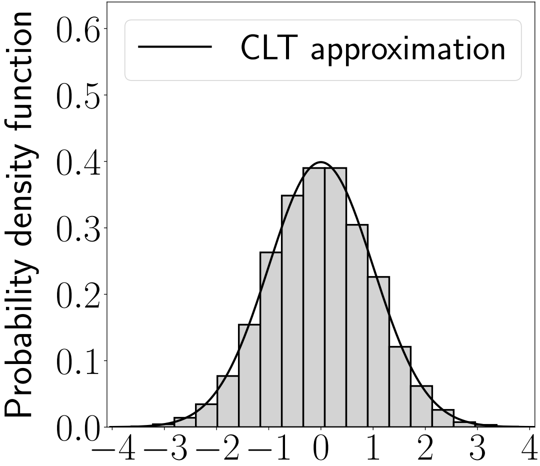
Height data: $n = 20$

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



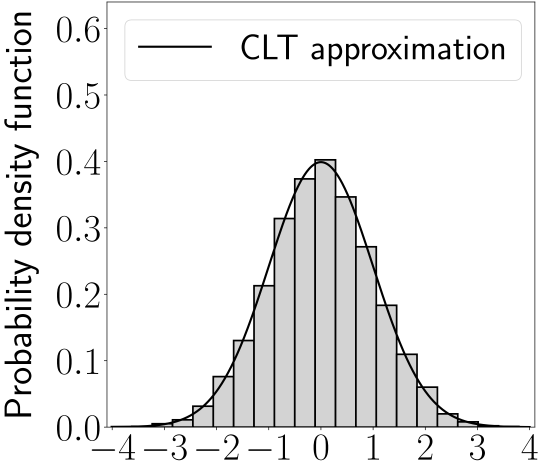
Height data: $n = 100$

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



Height data: n =1,000

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



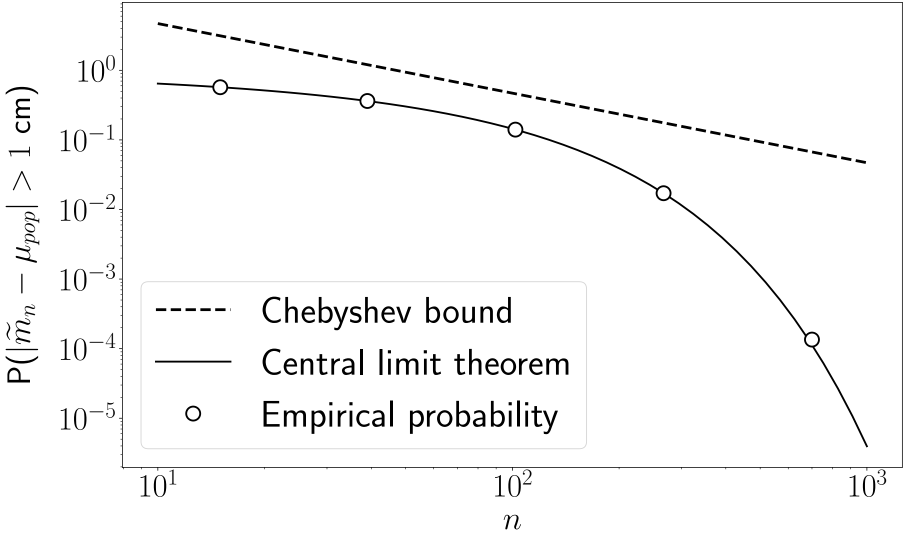
Chebyshev bound

$$\mathrm{P}\left(|\tilde{m}_n - \mu_{\text{pop}}| > \epsilon\right) \leq \frac{\sigma_{\text{pop}}^2}{n\epsilon^2}$$

Terrible approximation...

Do we get a better approximation from the central limit theorem?

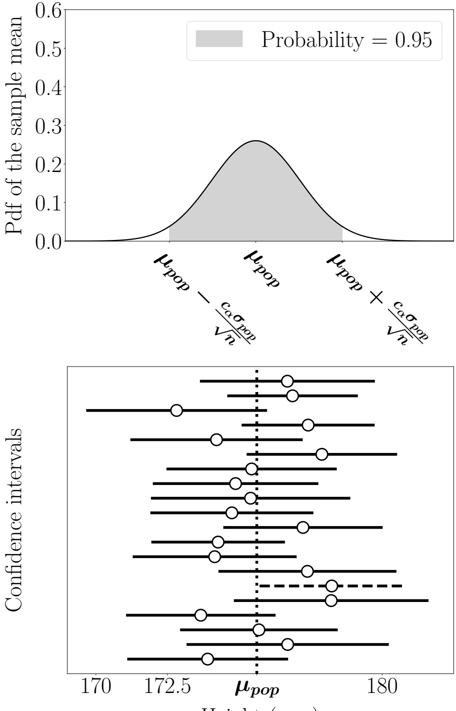
Much better



How can we exploit this to quantify uncertainty?

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}
$\mu_{pop} \times \frac{c_\alpha \sigma_{pop}}{\sqrt{n}}$
Confidence intervals
170
172.5
μ_{pop}
180

Confidence interval for the population mean

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

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

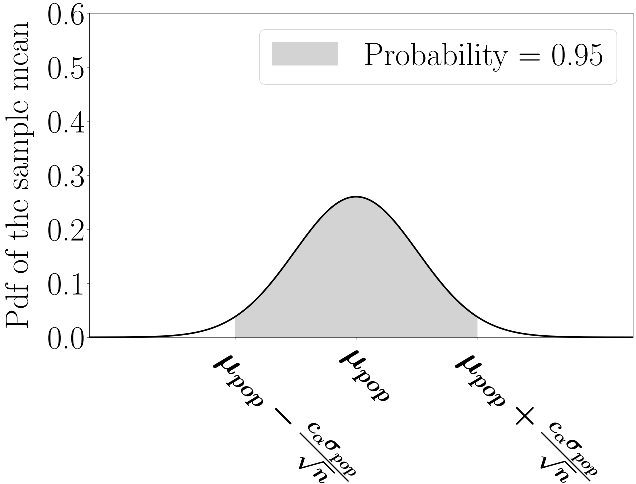
We don't know σ_{pop}!

Solution: Use sample standard deviation or an upper bound

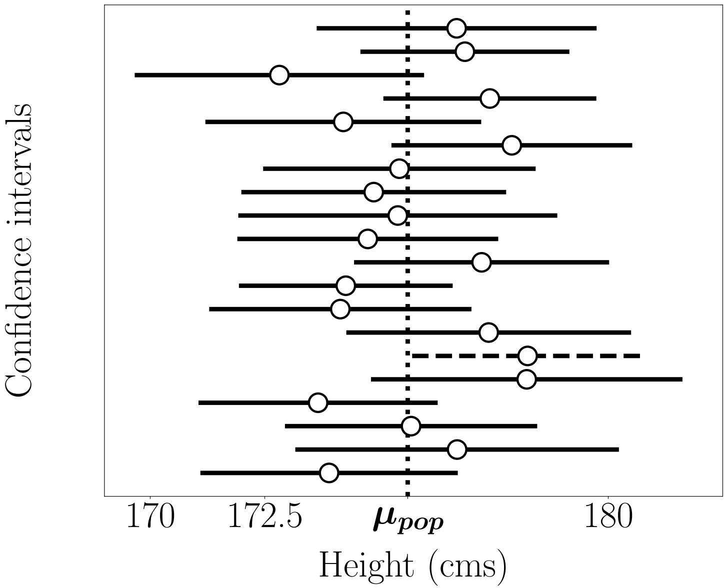
Height data: $n = 20$

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

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



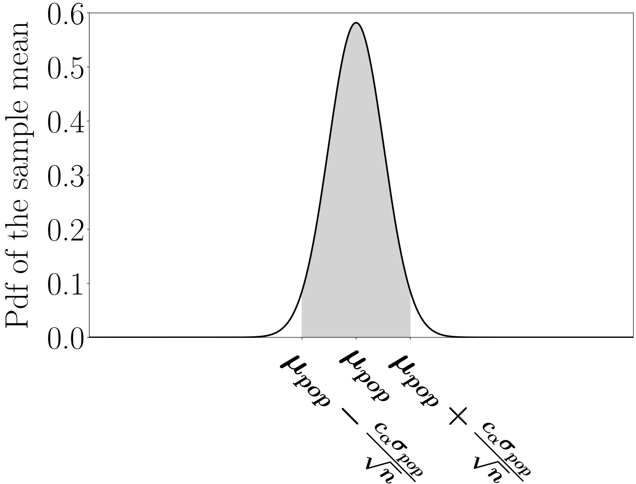
0.95 confidence intervals ($n = 20$)



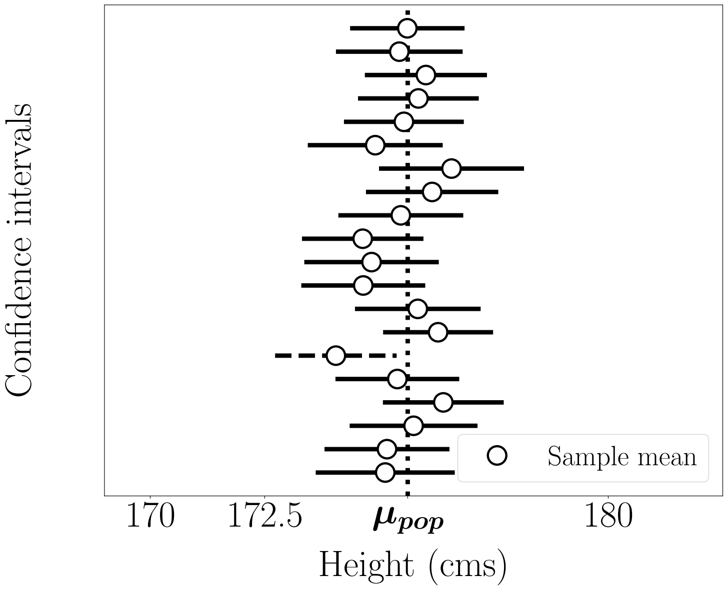
Height data: $n = 100$

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

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



0.95 confidence intervals ($n = 100$)



Confidence interval for the population mean

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

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

What if we don't know formula for standard error?

Challenge

How to estimate standard error computationally?

Sample from the data, as *if it were the population*

The bootstrap

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

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

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

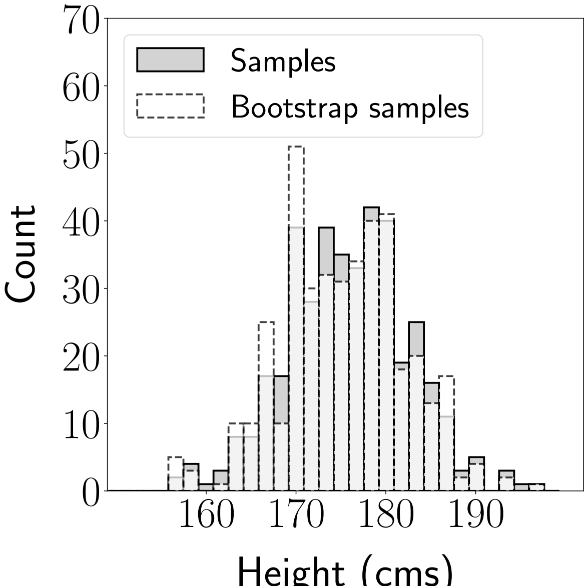
Bootstrap standard error

The bootstrap standard error of h is

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

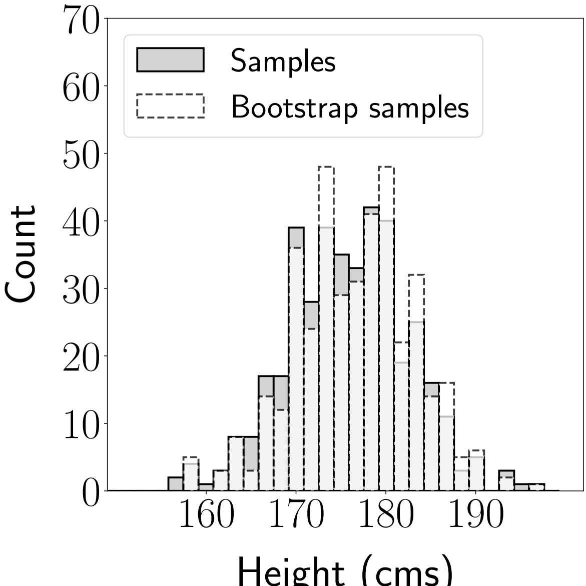
Bootstrap samples

Bootstrap sample mean: 175.3



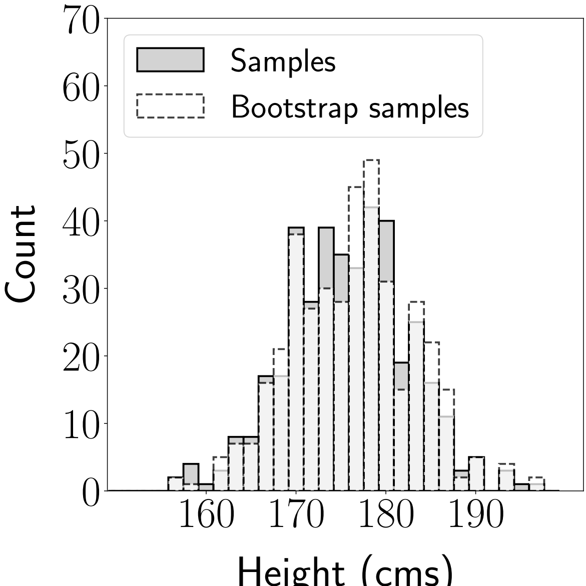
Bootstrap samples

Bootstrap sample mean: 176.6



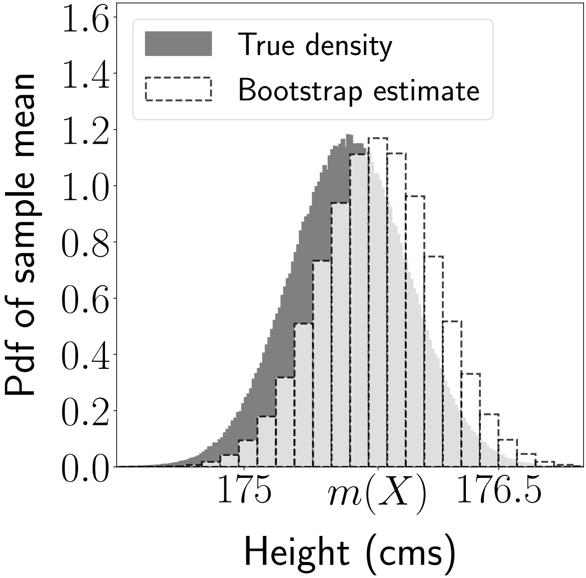
Bootstrap samples

Bootstrap sample mean: 176.2



Distribution of bootstrap samples

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



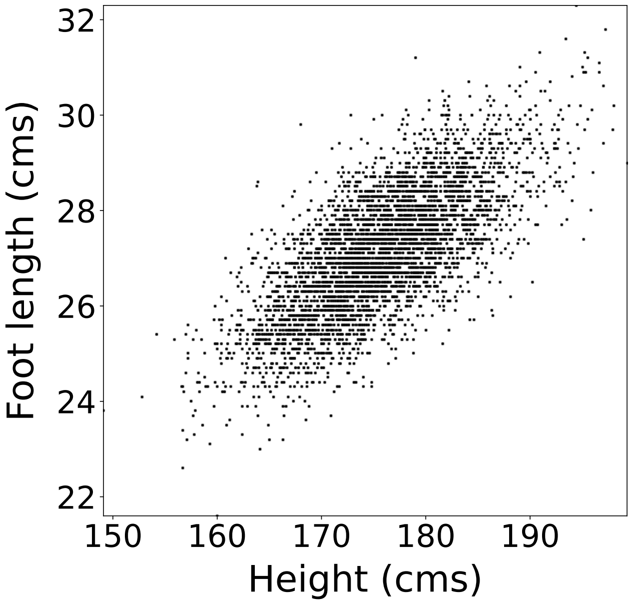
Bootstrap Gaussian confidence interval

1-α bootstrap Gaussian confidence interval

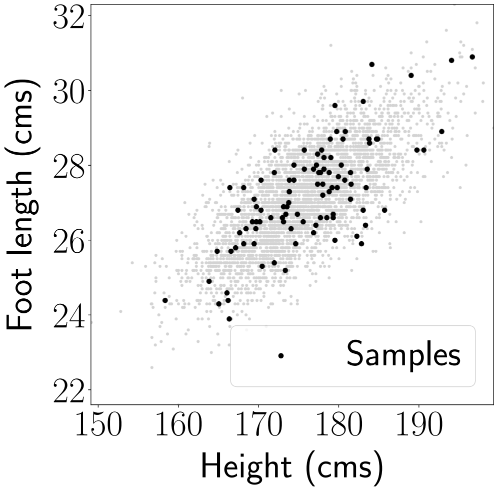
$$\mathcal{I}^{\mathsf{BSG}}_{1-\alpha} := [h(X) - c_\alpha \mathsf{se}_{\mathsf{bs}}, h(X) + c_\alpha \mathsf{se}_{\mathsf{bs}}] \qquad c_\alpha := F_{\tilde{z}}^{-1}\left(1 - \frac{\alpha}{2}\right)$$

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

Population correlation coefficient: 0.718



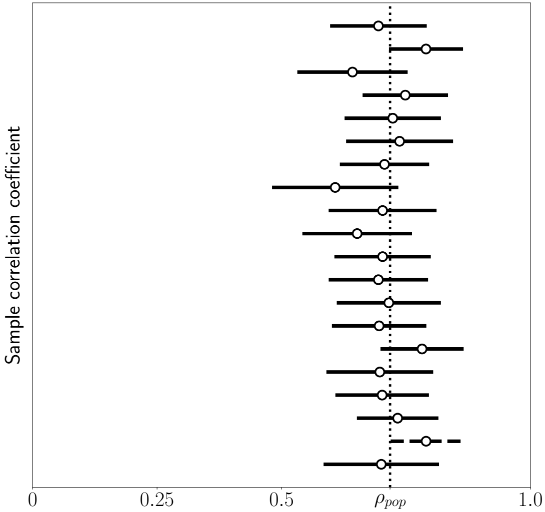
100 samples



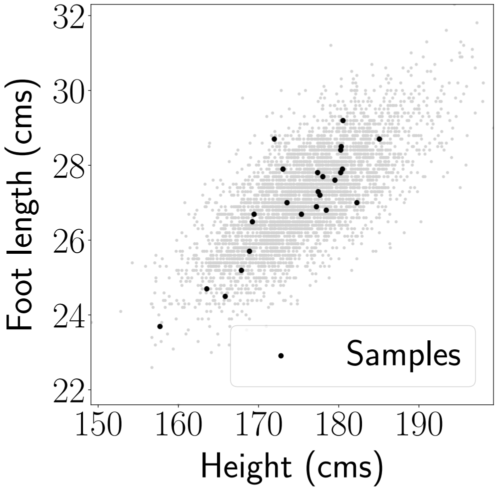
Sample correlation coefficient: $\rho_{\text{sample}} = 0.727$

Bootstrap Gaussian confidence intervals

Coverage: 93.7% (out of 10^4)

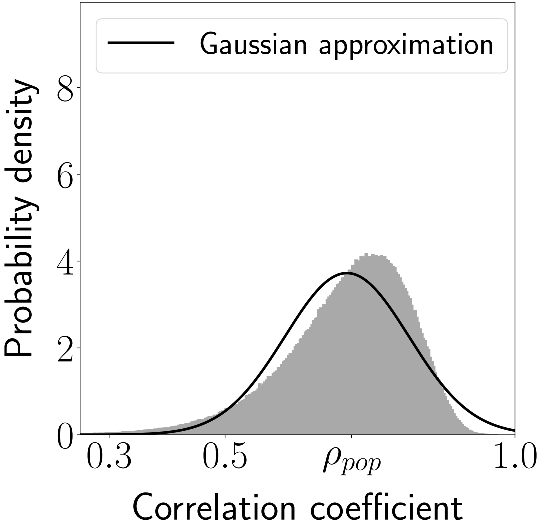


25 samples

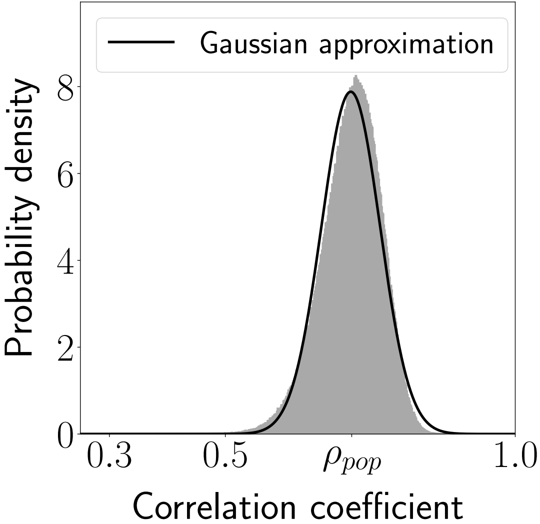


Sample correlation coefficient: $\rho_{\text{sample}} = 0.842$

Distribution of sample correlation coefficient ($n := 25$)

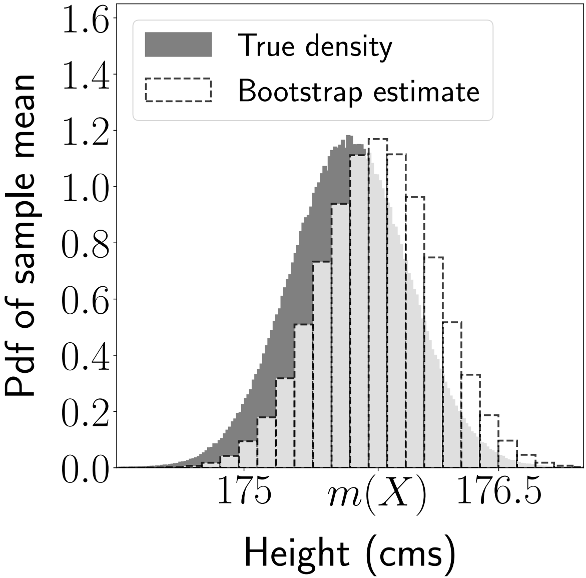


Distribution of sample correlation coefficient ($n := 100$)



True vs bootstrap distribution

True standard error: 0.343 Bootstrap standard error: 0.339



Bootstrap percentile confidence interval

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

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

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

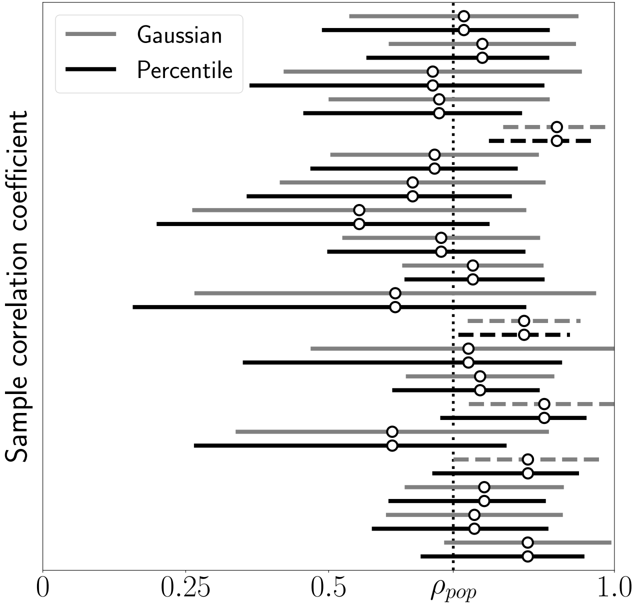
Bootstrap percentiles

$$\mathrm{P}\left(h(\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_n) \leq q_{\alpha/2}\right) = \frac{\alpha}{2}$$
$$\mathrm{P}\left(h(\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_n) \leq q_{1-\alpha/2}\right) = 1 - \frac{\alpha}{2}$$

1-α bootstrap percentile confidence interval

$$\mathcal{I}_{1-\alpha}^{\mathrm{BSP}} := [q_{\alpha/2}, q_{1-\alpha/2}]$$

Bootstrap confidence intervals



What have we learned

1. Random sampling
2. The bias
3. The standard error
4. The law of large numbers
5. The central limit theorem
6. Confidence intervals
7. The bootstrap