Bootstrap Confidence Intervals

Probability and Statistics for Data Science

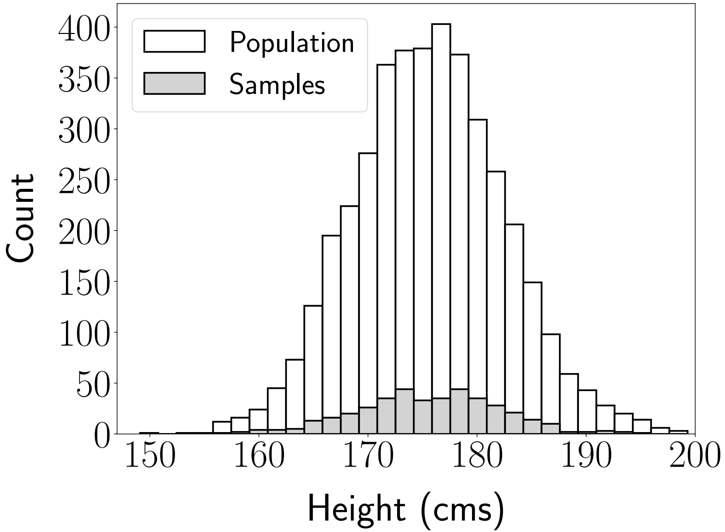
Carlos Fernandez-Granda

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

Random sampling

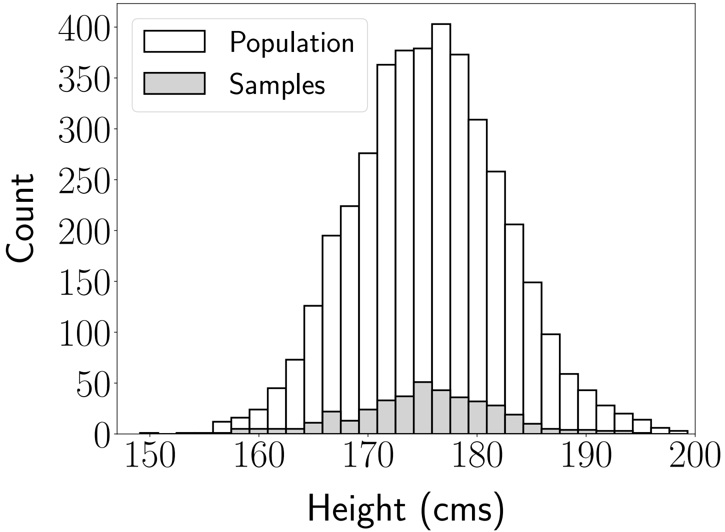
$n := 400$ random samples

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



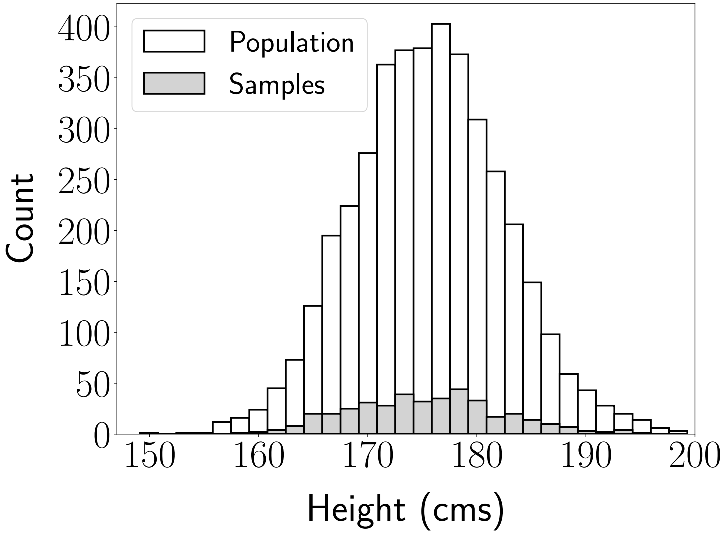
n := 400 random samples

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



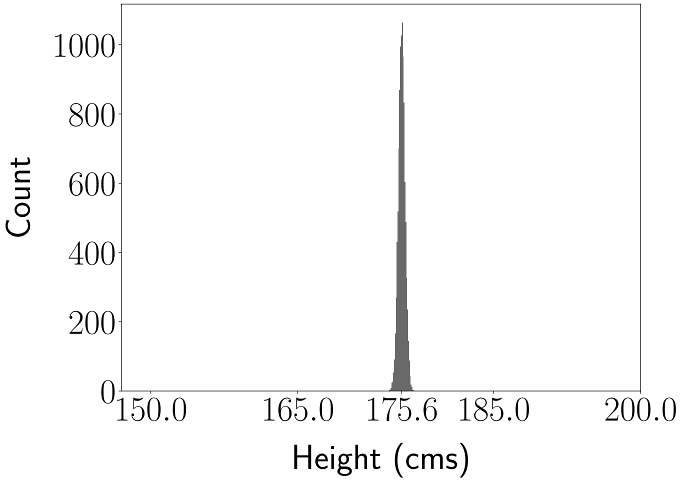
$n := 400$ random samples

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



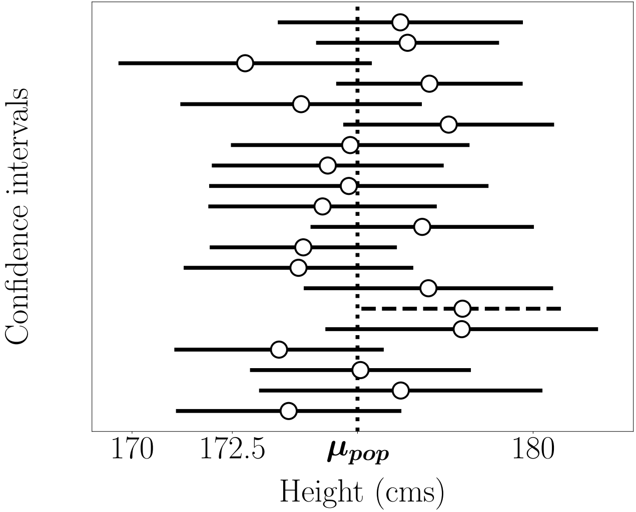
Sample means of 10,000 datasets of size $n := 400$

Goal: Quantify uncertainty from available data



Confidence interval

Range of values that contain parameter with high probability (e.g. 95%)



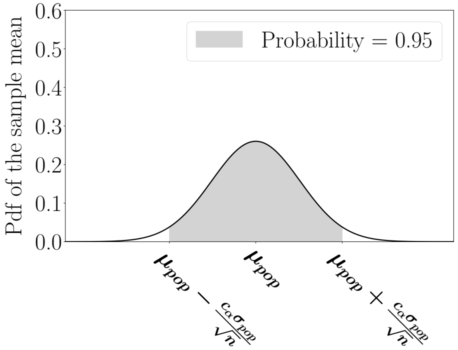
Confidence intervals for unbiased Gaussian estimator

Estimator $\tilde{g}$ with standard error se

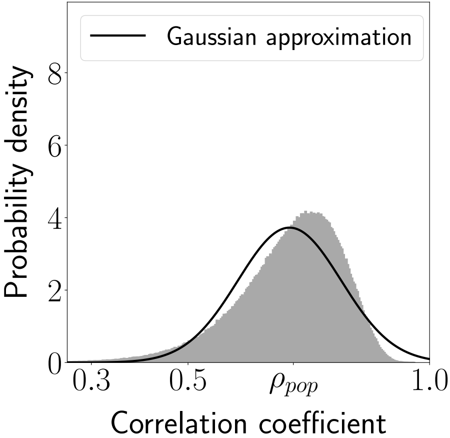
$$\widetilde{\mathcal{I}}_{0.95} := [\tilde{g} - 1.96\,\mathrm{se}, \tilde{g} + 1.96\,\mathrm{se}]$$

Sample mean $\tilde{m}$ (population variance $= \sigma_{\mathrm{pop}}^2$)

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



What if estimator is not Gaussian?



Bootstrap percentile confidence intervals are valid under certain assumptions

Plan

Bootstrap percentile confidence intervals

Gaussian estimator: Sample mean

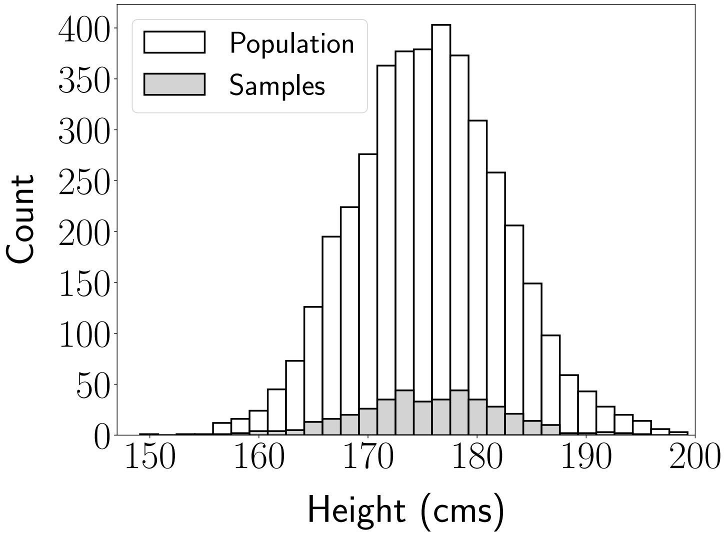
Non-Gaussian estimator: Sample correlation coefficient

Bootstrap percentile confidence intervals

Gaussian estimator: Sample mean

Non-Gaussian estimator: Sample correlation coefficient

Random sampling

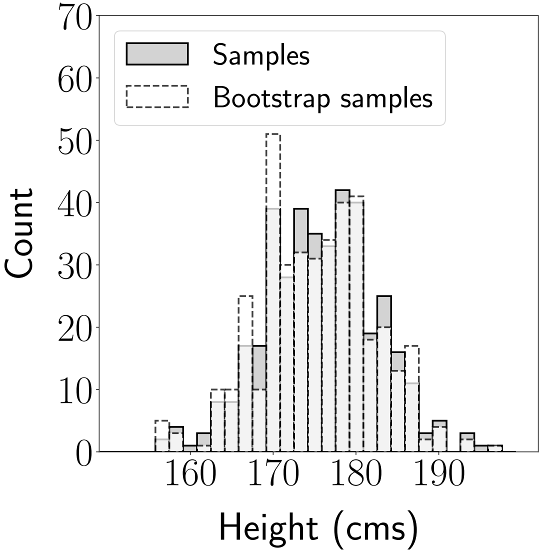


The bootstrap

Resample available data to form new *datasets* with *n samples*

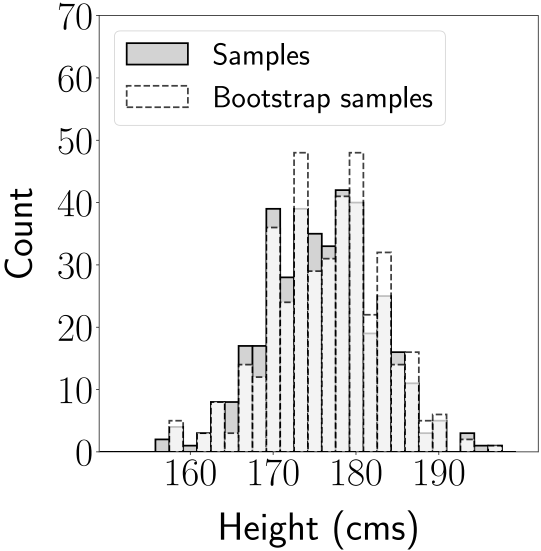
Bootstrap samples

Bootstrap sample mean: 175.3



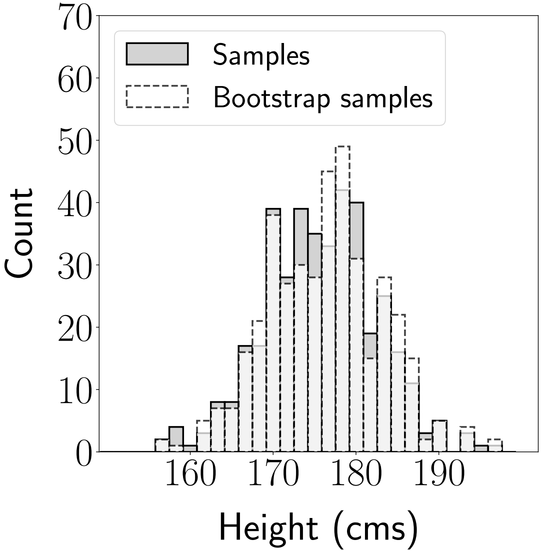
Bootstrap samples

Bootstrap sample mean: 176.6

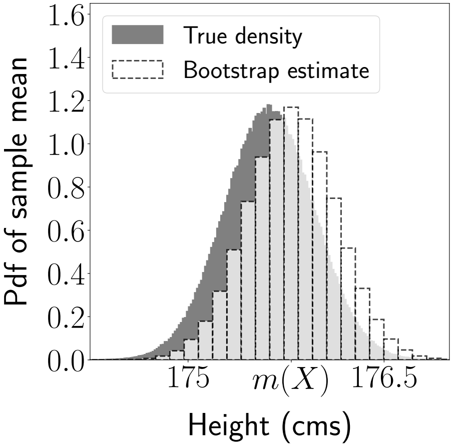


Bootstrap samples

Bootstrap sample mean: 176.2

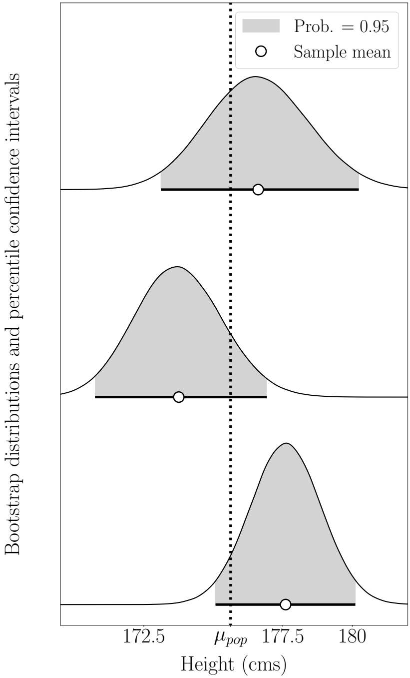


Distribution of bootstrap estimator

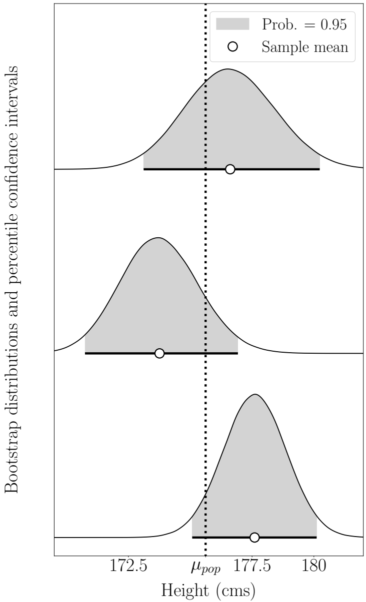


0.95 percentile confidence interval:

[2.5 percentile, 97.5 percentile]


Prob. = 0.95
Sample mean
Bootstrap distributions and percentile confidence intervals
172.5
μ_{pop}
177.5
180
Height (cms)

Are these valid confidence intervals?



Bootstrap percentile confidence intervals

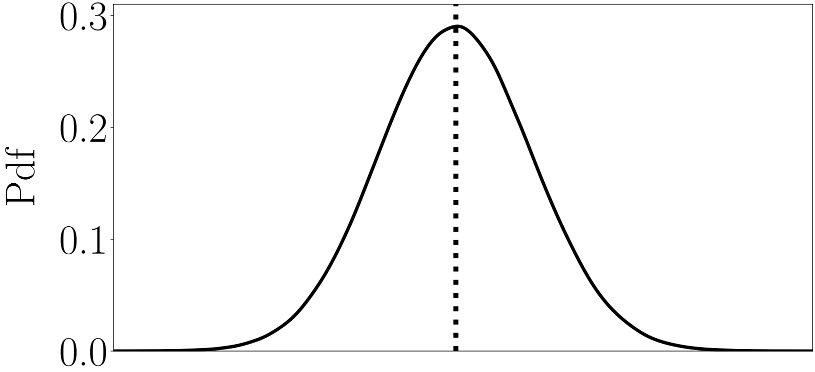
Gaussian estimator: Sample mean

Non-Gaussian estimator: Sample correlation coefficient

Sample mean (n random samples)

Mean = Population mean μ_{pop}

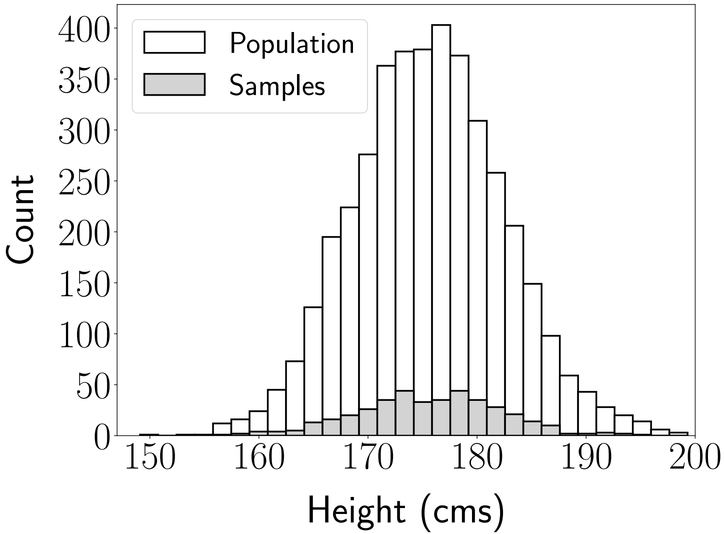
Std = $\sqrt{\text{Population variance / sample size}} = \sigma_{\text{pop}} \; / \; \sqrt{n}$



Approximately Gaussian by central limit theorem

Bootstrap mean (n bootstrap samples)

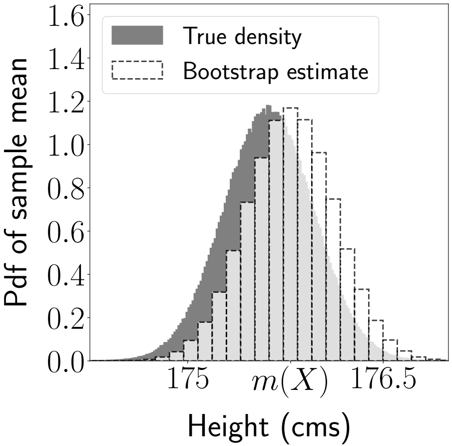
Analogous to sample mean, but the sample X is now the population

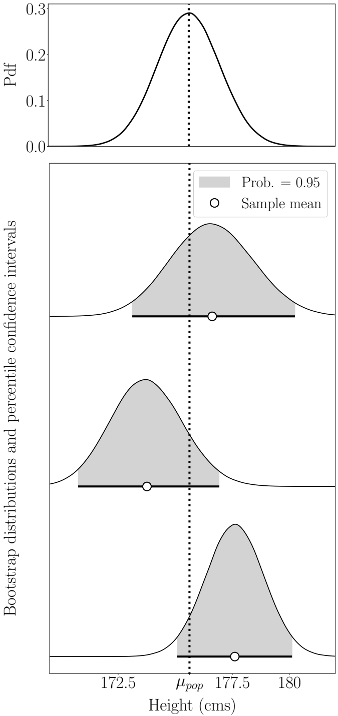


Bootstrap mean (n bootstrap samples)

Mean = Sample mean $m(X)$

$$\text{Std} = \sqrt{\text{Sample variance / sample size}} = \sqrt{v(X)/n}$$
$$\approx \sigma_{\text{pop}} \, / \, \sqrt{n}$$





If bootstrap mean is Gaussian,

2.5 percentile:
Mean - 1.96 std

97.5 percentile:
Mean + 1.96 std

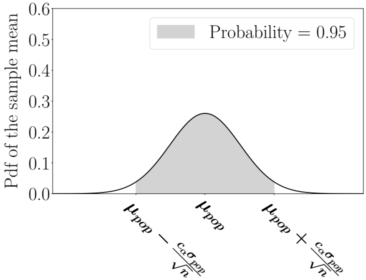
Bootstrap percentile confidence intervals:

$$\left[m(X) - 1.96 \frac{\sigma_{\text{pop}}}{\sqrt{n}}, m(X) + 1.96 \frac{\sigma_{\text{pop}}}{\sqrt{n}} \right]$$

Confidence interval for the population mean

Sample mean $\tilde{m}$ (population variance $= \sigma^2_{\text{pop}}$)

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



Realization for a sample $X := \{x_1, x_2, ..., x_n\}$

$$\left[m(X) - 1.96\frac{\sigma_{\text{pop}}}{\sqrt{n}}, m(X) + 1.96\frac{\sigma_{\text{pop}}}{\sqrt{n}}\right]$$

Conclusion

Bootstrap percentile confidence intervals are valid as long as:

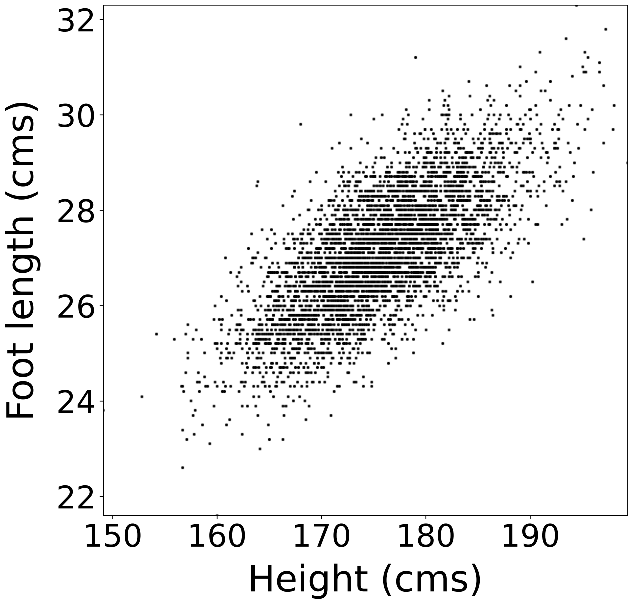
- The estimator is Gaussian and unbiased
- The bootstrap distribution is Gaussian and centered at the estimator
- The bootstrap standard error approximates the true standard error

Bootstrap percentile confidence intervals

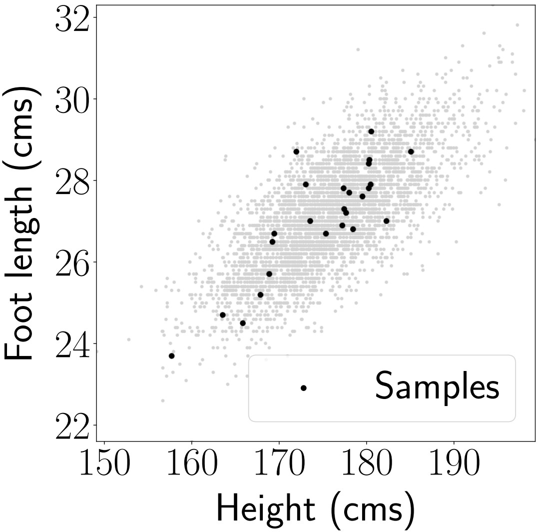
Gaussian estimator: Sample mean

Non-Gaussian estimator: Sample correlation coefficient

Population correlation coefficient: 0.718

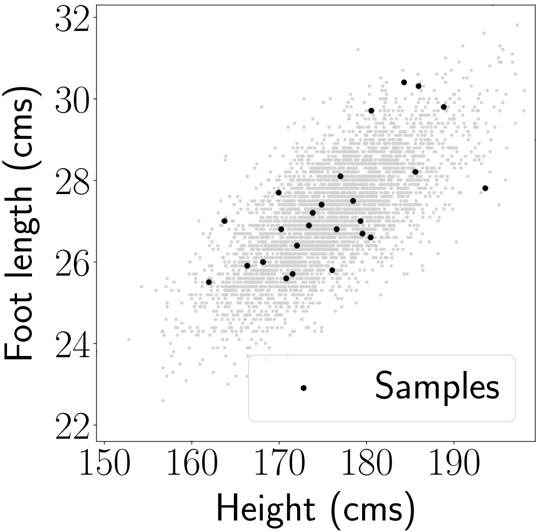


$n := 25$ random samples



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

$n := 25$ random samples



Sample correlation coefficient: $\rho_{\text{sample}} = 0.687$

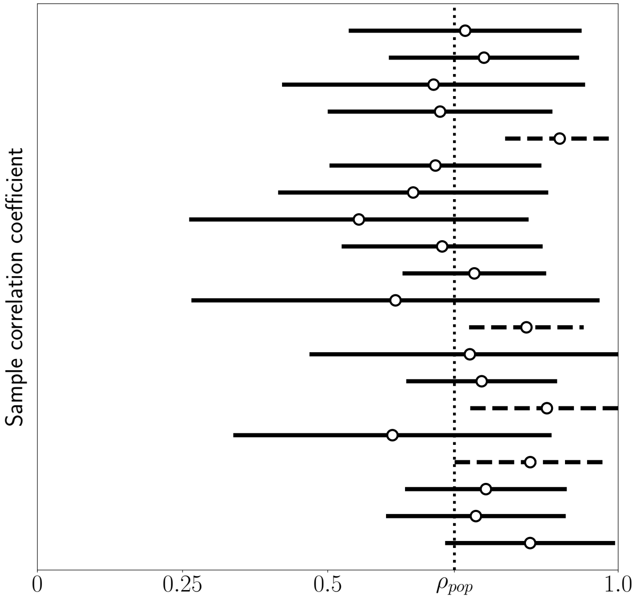
Gaussian confidence intervals

Assuming estimator is unbiased and approximately Gaussian

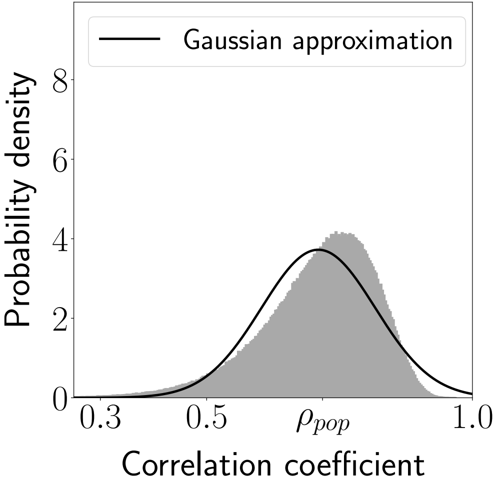
$$[\rho_{\text{sample}} - 1.96\,\text{se}, \rho_{\text{sample}} + 1.96\,\text{se}]$$

Standard error estimated via bootstrapping

Coverage: 90.7% (out of 10^4)

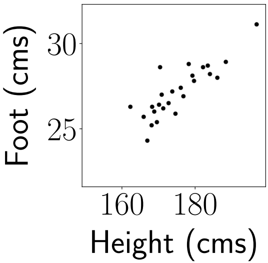


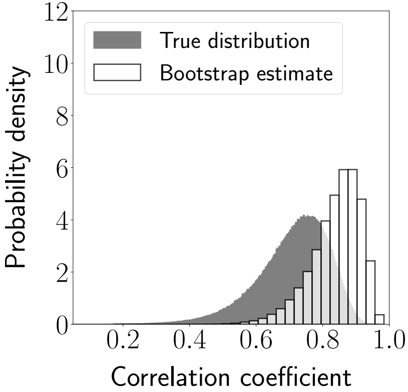
Distribution of sample correlation coefficient

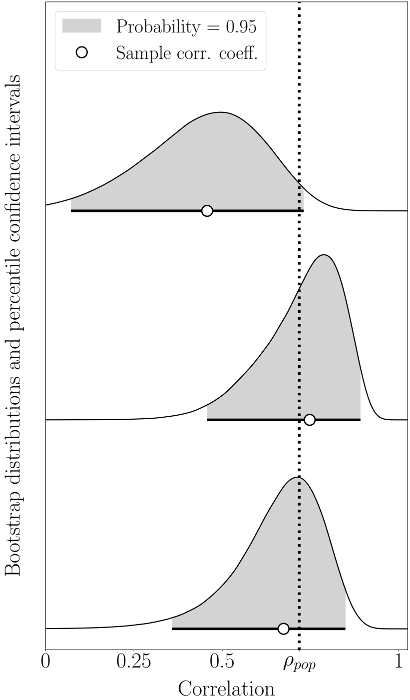


What about percentile confidence intervals?

Bootstrap distribution

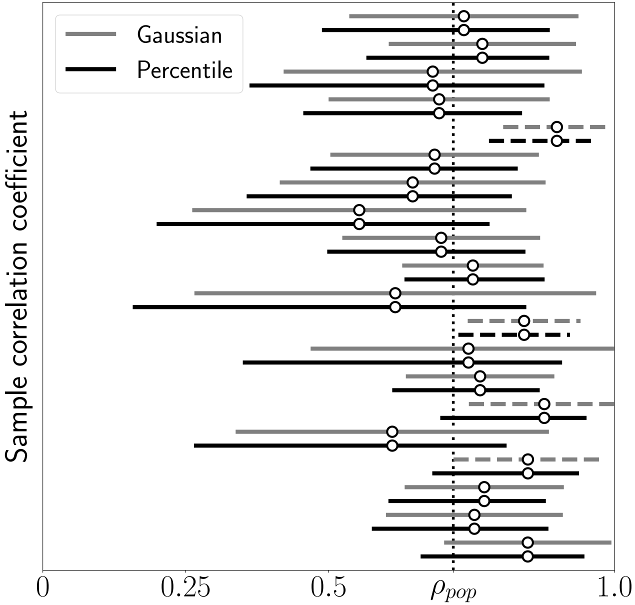






Probability = 0.95
Sample corr. coeff.
Bootstrap distributions and percentile confidence intervals
0
0.25
0.5
ρ_{pop}
1
Correlation

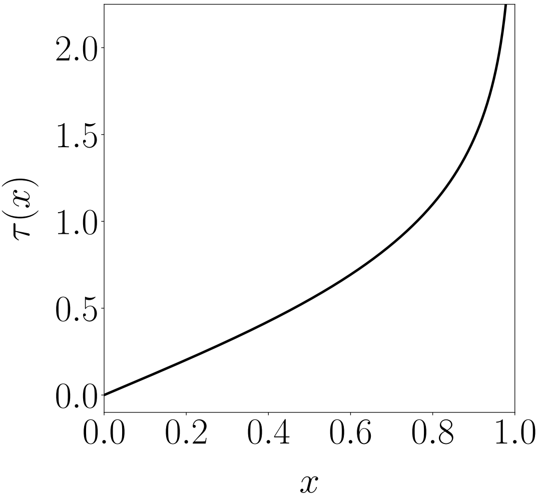
Coverage: 92.8%



Assumption

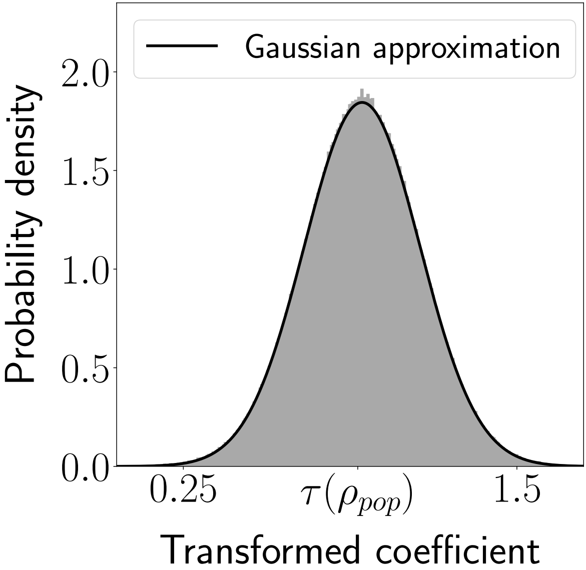
Estimator is approximately Gaussian after a monotonic transformation

Fisher's transformation



$$\tau(\rho) := \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right)$$

Transformed distribution



Assumptions

After applying the transformation τ:

- The estimator is Gaussian and unbiased with respect to transformed parameter $\tau(\gamma)$
- The bootstrap distribution is Gaussian and centered at the estimator
- The bootstrap standard error approximates the true standard error

$$\mathrm{P}\left(\tilde{t}_{0.025} \leq \tau(\gamma) \leq \tilde{t}_{0.975}\right) = 0.95$$

$\tilde{t}_{0.025}$ / $\tilde{t}_{0.975}$ are percentiles of transformed bootstrap samples

Key insight

$$\mathrm{P}\left(\tilde{t}_{0.025} \leq \tau(\gamma) \leq \tilde{t}_{0.975}\right) = 0.95$$

implies

$$\mathrm{P}\left(\tilde{q}_{0.025} \leq \gamma \leq \tilde{q}_{0.975}\right) = 0.95$$

for bootstrap percentiles $\tilde{q}_{0.025}$ / $\tilde{q}_{0.975}$

Monotone transformation

$$\mathrm{P}\left(\tilde{t}_{0.025} \leq \tau(\gamma) \leq \tilde{t}_{0.975}\right) = 0.95$$

implies

$$\mathrm{P}\left(\tau^{-1}(\tilde{t}_{0.025}) \leq \gamma \leq \tau^{-1}(\tilde{t}_{0.975})\right) = 0.95$$

Since $\tau(\tilde{q}_{0.975}) = \tilde{t}_{0.975}$ and $\tau(\tilde{q}_{0.025}) = \tilde{t}_{0.025}$

$$\mathrm{P}\left(\tilde{q}_{0.025} \leq \gamma \leq \tilde{q}_{0.975}\right) = 0.95$$

Monotone transformations preserve percentiles

Estimator computed from bootstrap samples: $\tilde{w}_{\text{bs}}$

$$\mathrm{P}\left(\tau(\tilde{w}_{\text{bs}}) \leq \tau(\tilde{q}_{0.975})\right) = \mathrm{P}\left(\tilde{w}_{\text{bs}} \leq \tilde{q}_{0.975}\right) = 0.975$$

$$\tau(\tilde{q}_{0.975}) = \tilde{t}_{0.975} \qquad \tau(\tilde{q}_{0.025}) = \tilde{t}_{0.025}$$

Conclusion

Bootstrap percentile confidence intervals are valid as long as there exists a monotonic transformation after which

- The estimator is Gaussian and unbiased
- The bootstrap distribution is Gaussian and centered at the estimator
- The bootstrap standard error approximates the true standard error

We do not need to know the transformation!

What have we learned

How to build bootstrap percentile intervals

Why they work for unbiased Gaussian estimators

Why they work for (some) non-Gaussian estimators