Lecture 3: Confidence intervals

EMAT30007 Applied Statistics

Nikolai Bode & Filippo Simini

Department of Engineering Mathematics

Outline of the lecture

In this lecture you will learn:

- How to calculate z **and** t **confidence intervals** for a statistical estimate.
- How to interpret a confidence interval.
- How to use the **bootstrap method** to estimate a quantity of a population and its uncertainty.

Uncertainty of an estimate

- We have learned how to estimate the parameter of a PDF using the Maximum Likelihood method and the Method of Moments.
  We have learned how to estimate the mean and the variance of a RV.
- We have learned that these estimators are (asymptotically) unbiased: their average over an infinite number of observations is equal to the true value.

But how much can we trust an estimate made using only one sample of finite size?

In fact, if we compute another estimate based on a different sample of the same finite size, we will find a different value. The variability of the estimates originates from the variability (randomness) of the samples and depends on the size of the samples (large sample size, less variability and vice versa).

How can we quantify the uncertainty of an estimate?

Sampling distribution

The sampling distribution is the (theoretical) distribution of the estimates made using an infinite number of independent samples of the same size. The mean of this distribution is the true value of the estimated quantity (if the estimator is asymptotically unbiased) and the spread around the mean denotes the typical range of variability (or uncertainty) of our estimates.

A possible way to quantify the uncertainty of an estimate is to compute a confidence interval, which specifies the expected distance of an estimate from the true value (the mean of the sampling distribution).

We would like to make statements like this: *"If we computed this estimate on an infinite number of independent samples of size n, we expect that a fraction $(1-\alpha)$ of our estimates will be contained inside the interval $CI_\alpha(n)$."*

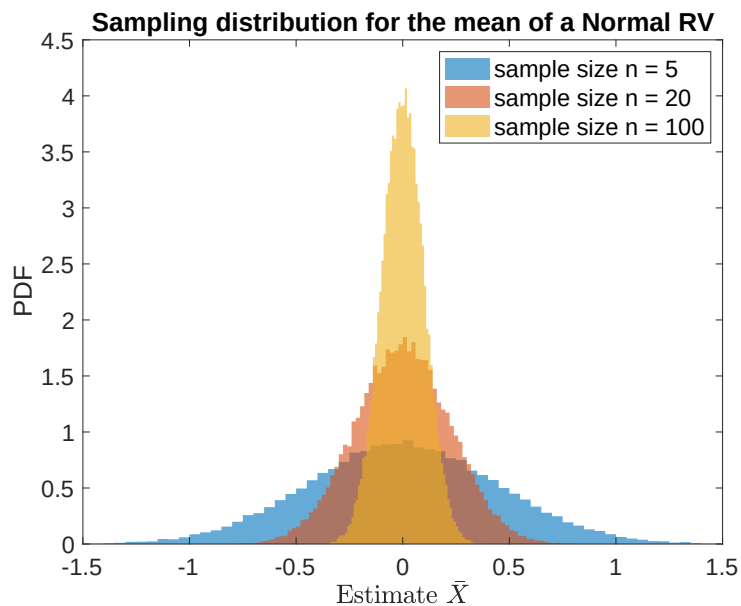
In this lecture we'll see how to compute $CI_\alpha(n)$ for a specified α.

Sampling distribution of the mean of Normal RVs

An interesting quantity of a population that is typically estimated is the mean. The estimator of the mean μ from a sample of n independent observations $X_1, ..., X_n$ is $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

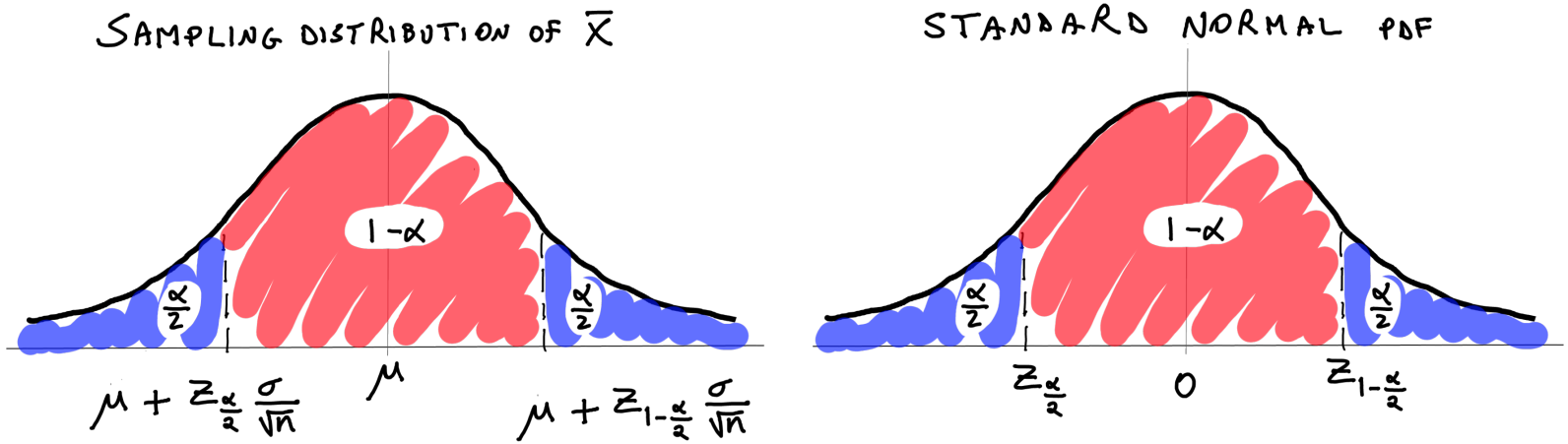
When the X_i are Normal RVs, the sampling distribution of the mean can be computed exactly, because the mean of independent Normal RVs is a Normal RV. It can be shown that the sampling distribution of the mean of n identical and independent Normal RVs with PDF $N(\mu, \sigma^2)$ is Normal with mean μ and variance σ^2/n: $\bar{X} \sim N(\mu, \sigma^2/n)$.

As we know, the estimator $\bar{X}$ is unbiased: its expected value (mean) is the true mean ($E(\bar{X}) = \mu$). Note that the larger is n (the sample size) the smaller is the variance of the distribution of the estimates $\bar{X}$ around the true value, which means that we expect a random estimate to be closer to the true value when n is large.



Estimating confidence intervals

Since we know the distribution of $\bar{X}$, we can compute the size of the interval $[q_{\alpha/2}, q_{1-\alpha/2}]$ centered on the true value μ such that with probability $(1-\alpha)$ a random estimate $\bar{X}$ falls inside $[q_{\alpha/2}, q_{1-\alpha/2}]$. We could estimate $[\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}]$ using the sample's empirical quantiles, but this method is not very accurate.



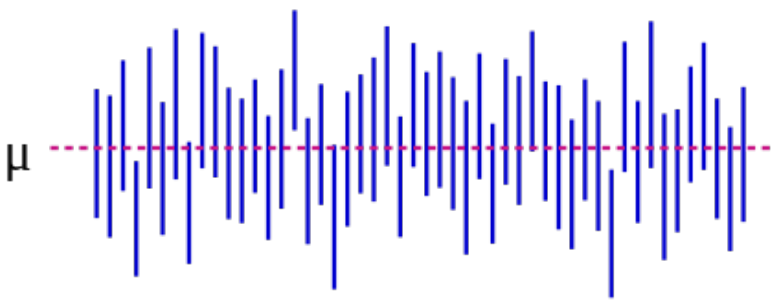
It can be shown that $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ is a Standard Normal RV and $Prob[\mu + z_{\alpha/2}(\frac{\sigma}{\sqrt{n}}) < \bar{X} < \mu + z_{1-\alpha/2}(\frac{\sigma}{\sqrt{n}})] = (1-\alpha)$ where $\boldsymbol{z_{\alpha/2}}$ is the $\alpha/2$-quantile of the Standard Normal distribution, i.e. the value such that the probability to draw an equal or smaller number from a Standard Normal distribution is $\alpha/2$: $Prob(Z \leq z_{\alpha/2}) = F_Z(z_{\alpha/2}) = \alpha/2$, hence $z_{\alpha/2} = F_Z^{-1}(\alpha/2)$ with $Z \sim N(0,1)$.

z-Confidence Interval for the mean

- If with probability $(1-\alpha)$ a random estimate $\bar{X}$ falls inside the interval $[\mu + z_{\alpha/2}(\frac{\sigma}{\sqrt{n}}), \mu + z_{1-\alpha/2}(\frac{\sigma}{\sqrt{n}})]$ then with probability $(1-\alpha)$ the interval $[\bar{X} - z_{1-\alpha/2}(\frac{\sigma}{\sqrt{n}}), \bar{X} - z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})]$ centered on a random estimate $\bar{X}$ will contain the true mean μ. That is, we expect that $100 \cdot (1-\alpha)\%$ of the random intervals of size $2|z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})|$ centered on $\bar{X}$ contain the true mean.
- We say that

$$\left[\bar{x} - z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \quad \bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] \tag{1}$$

  is the z-confidence interval for the mean at confidence level $100 \cdot (1-\alpha)\%$.
- Note that to compute the z-interval we must know the variance σ^2.



https://en.wikipedia.org/wiki/Confidence_interval

The blue vertical line segments represent 50 realizations of confidence intervals of Eq. (1) for the population mean, μ, represented as a red horizontal dashed line. Note that some intervals do not contain the population mean; their expected number is 50α.

Example: z-CI for mean student heights

We want to now the mean height of the students in the University. We measure the height of $n = 10$ students chosen at random and we observe the sample $x = [176, 165, 189, 180, 172, 169, 162, 161, 183, 170]$ cm.

Assuming that a student's height is a Normal RV with standard deviation $\sigma = 9$ cm, what is the z-confidence interval for the unknown mean height μ at a confidence level of 95%?

To use the definition of Eq. (1), $\left[\bar{x} - z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \quad \bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$ we need to compute:

- the sample mean $\bar{x} = \frac{1}{n}\sum_i x_i = 172.7$.
- $\alpha = 0.05$ from $100 \cdot (1 - \alpha) = 95$.
- $z_{\alpha/2} = -1.96 = -z_{1-\alpha/2}$ using MATLAB's `norminv(0.025)`.

We obtain $|z_{\alpha/2}|\frac{\sigma}{\sqrt{n}} = 1.96\frac{9}{\sqrt{10}} = 5.58$ and confidence interval $[167.12, 178.28]$.

t-Confidence Interval for the mean

In order to compute the z-confidence interval we must know the variance σ^2.

Usually if we don't know μ we also don't know σ.

When we don't know σ^2 we could use the observed data to compute the unbiased estimator of the variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

It can be shown that

- $T = \frac{\bar{X}-\mu}{\sqrt{S^2/n}}$ is a RV following a t-distribution with $n-1$ degrees of freedom
- $Prob\left[t_{\alpha/2,n-1} < \frac{\bar{X}-\mu}{\sqrt{S^2/n}} < t_{1-\alpha/2,n-1}\right] = (1-\alpha)$

  where $\boldsymbol{t_{\alpha/2,n-1}}$ is the $\alpha/2$-quantile of the t-distribution with $n-1$ degrees of freedom, i.e. the value such that the probability to draw an equal or smaller number from a t-distribution with $n-1$ degrees of freedom is $\alpha/2$:
  $Prob(T \leq t_{\alpha/2,n-1}) = F_T(t_{\alpha/2,n-1}) = \alpha/2 \quad \Rightarrow \quad t_{\alpha/2,n-1} = F_T^{-1}(\alpha/2)$.
- The t-confidence interval for the mean at confidence level $100 \cdot (1-\alpha)\%$ is

$$\left[\bar{x} - t_{1-\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right), \quad \bar{x} - t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)\right] \tag{2}$$

Example: t-CI for mean student heights

We want to know the mean height of the students in the University. We measure the height of $n = 10$ students chosen at random and we observe the sample $x = [176, 165, 189, 180, 172, 169, 162, 161, 183, 170]$ cm.

Assuming that a student's height is a Normal RV *with unknown standard deviation*, what is the t-confidence interval for the unknown mean height μ at a confidence level of 95%?

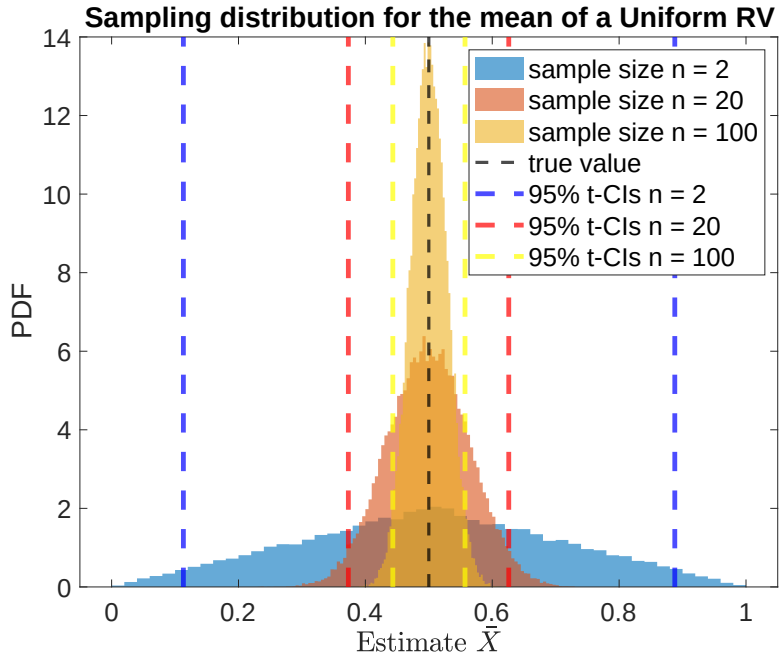
To use the definition of Eq. (2), $\left[\bar{x} - t_{1-\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right), \quad \bar{x} - t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)\right]$ we need to compute:

- the sample mean $\bar{x} = \frac{1}{n}\sum_i x_i = 172.7.$
- the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2} = 9.24.$
- $\alpha = 0.05$ from $100 \cdot (1 - \alpha) = 95.$
- $t_{\alpha/2,n-1} = -2.23 = -t_{1-\alpha/2,n-1}$ using MATLAB's `tinv(0.025, 10)`.

We obtain $|t_{\alpha/2,n-1}|\frac{s}{\sqrt{n}} = 2.23\frac{9.24}{\sqrt{10}} = 6.52$ and confidence interval $[166.18, 179.22]$.

Pros of z and t confidence intervals

- z and t confidence intervals (CIs) are exact under the assumption of normal RVs. They are approximately correct for non-Normal RVs with finite variance and large sample size (e.g. $n > 30$) because of the Central Limit Theorem.



Example: CIs for the mean of a Uniform distribution between 0 and 1.

Central Limit Theorem

Consider n independent RVs $X_1, ..., X_n$ that have identical distribution P_X with finite variance. The Central Limit Theorem tells us that the distribution of the RV

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tends to a Standard Normal distribution $N(0, 1)$ for large samples ($n \to \infty$). Here $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$; μ and σ are the mean and standard deviation of P_X.

Example: *z-interval for the mean of a Bernoulli distribution*

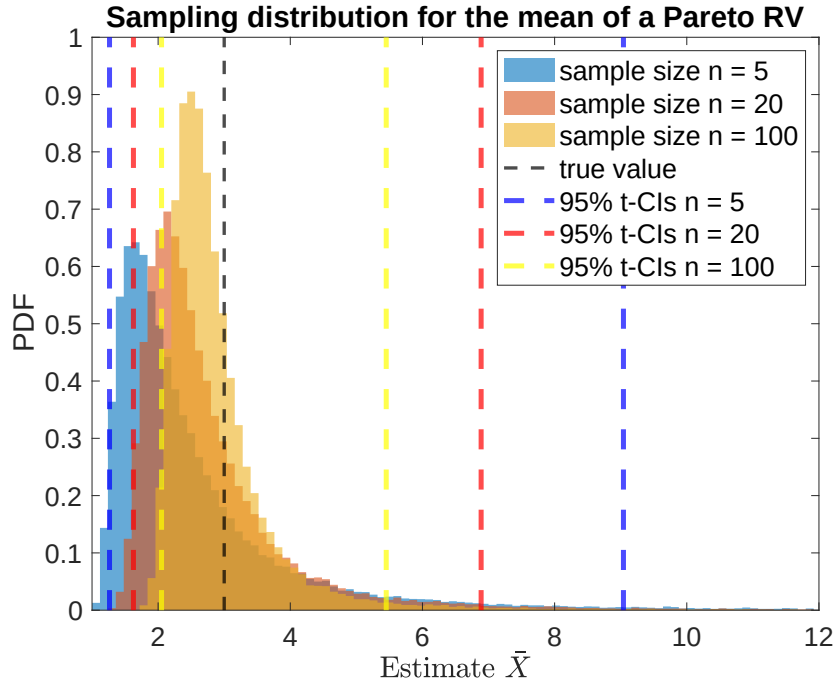
For large random samples, a $(1-\alpha)100\%$ confidence interval for the parameter p of a Bernoulli distribution (i.e. a population proportion) is

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{3}$$

where $\hat{p} = \sum_{i=1}^{n} x_i/n$ and $\sigma = \sqrt{p(1-p)}$ has been estimated as $\sqrt{\hat{p}(1-\hat{p})}$.

Cons of z and t confidence intervals

- For non-Normal RV, if sample size n is small or the RV's distribution is highly skewed (i.e. infinite variance), the z and t CIs are not accurate.
- theoretical CIs are difficult to obtain for quantities other than μ, the population mean (e.g. PDF parameters).



Example: CIs for the mean of a Pareto distribution $P_X(x) = \frac{\theta}{x^{1+\theta}}$, $(x > 1)$ with parameter $\theta = 1.5$.

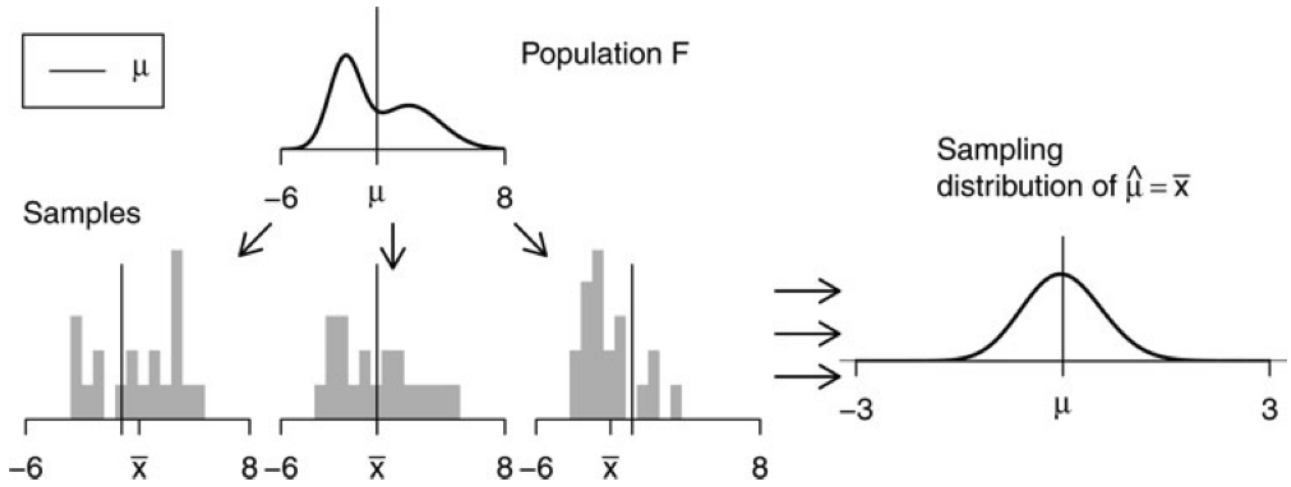
Bootstrapping

The approach of resampling methods is to use the available observations to generate new data and estimate parameters and confidence intervals.

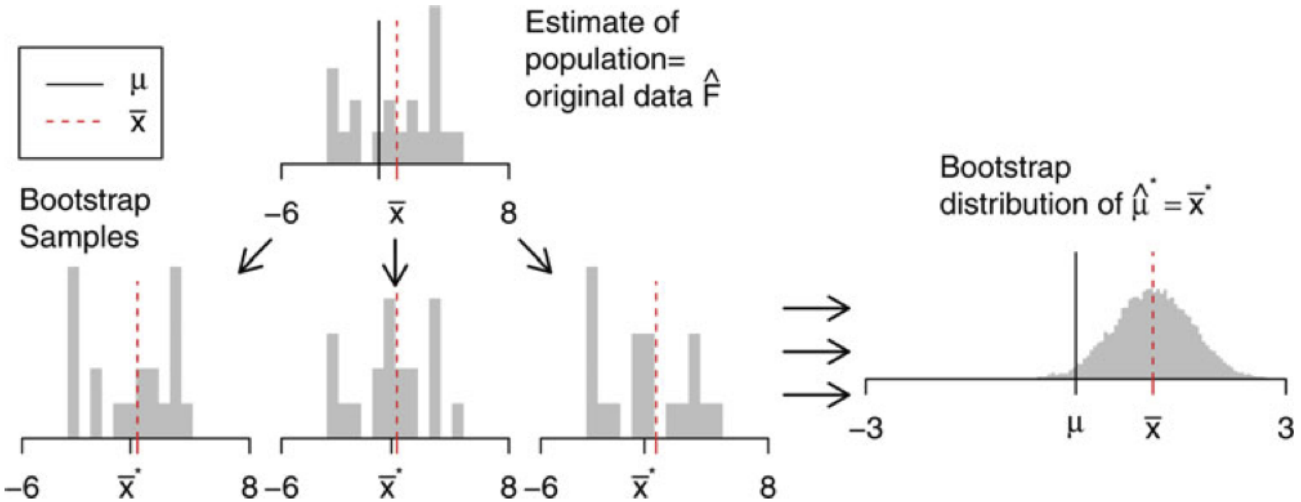
The idea of bootstrapping is to treat the original sample as the 'population' and generate new samples drawing with replacement from the original sample.

Bootstrap algorithm to estimate the bootstrap distribution of a quantity θ:

1. Generate a new sample by sampling with replacement from the original dataset of available observations. Such sample is called bootstrap sample. The size of the bootstrap sample is the same as the original dataset.
2. Compute the quantity of interest using the bootstrap sample, $\hat{\theta}^*$.
3. Repeat steps 1 and 2 r times (e.g. $r = 10000$). The empirical PDF of the r bootstrap samples is called the bootstrap distribution.



Ideal sampling



Bootstrap sampling

Hesterberg, T. C. (2015). The American Statistician, 69(4), 371-386.

Bootstrap estimate and confidence intervals

From the bootstrap distribution we can extract various information about the quantity of interest, θ:

- The bootstrap estimate of θ is the mean of the bootstrap distribution, $\overline{\hat{\theta}^*}$.
- The bootstrap standard error is the sample standard deviation of the bootstrap distribution, $s_b = \sqrt{\frac{1}{r-1}\sum_{i=1}^{r}\left(\hat{\theta}_i^* - \overline{\hat{\theta}^*}\right)^2}$.
- The bootstrap percentile interval at confidence level $100 \cdot (1-\alpha)\%$ (with $\alpha \in [0,1]$) is the range of the middle $100 \cdot (1-\alpha)\%$ of the bootstrap distribution, $[q_{\alpha/2}, q_{1-\alpha/2}]$, where $q_{\alpha/2}$ is the empirical $\alpha/2-$quantile of the bootstrap distribution: $Prob(\theta \leq q_{\alpha/2}) = \alpha/2$.

Bootstrap percentile confidence interval

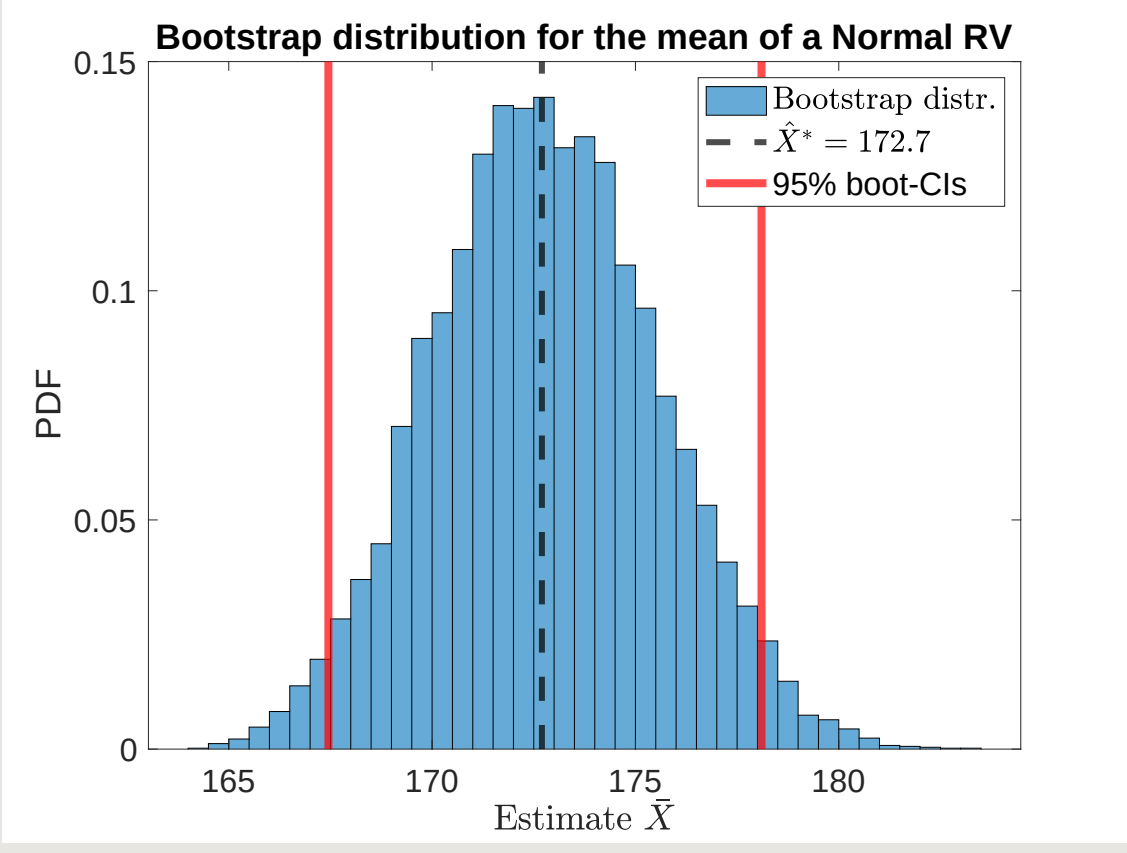
To compute the bootstrap percentile interval at confidence level $100 \cdot (1-\alpha)\%$ (with $\alpha \in [0,1]$) we need to compute $q_{\alpha/2}$ and $q_{1-\alpha/2}$, the empirical quantiles of the bootstrap distribution.

How to compute the empirical quantiles

For a 90% confidence interval ($\alpha = 0.9$) keep the middle 90%, leaving 5% in each tail and 5% in the head. The 90% confidence interval boundaries would be 5th percentile and 95th percentile.
In case we have $r = 10000$ bootstrap replications: $\theta^*_1, \theta^*_2, \theta^*_3, ..., \theta^*_{10000}$ the 90% CI is: $[\theta^*_{500}, \theta^*_{9500}]$.

- Disadvantages: Can be too narrow for small samples and skewed sampling distributions.
- Advantages: A very intuitive and easy to implement method. Can estimate the CI of any quantity (not just means).

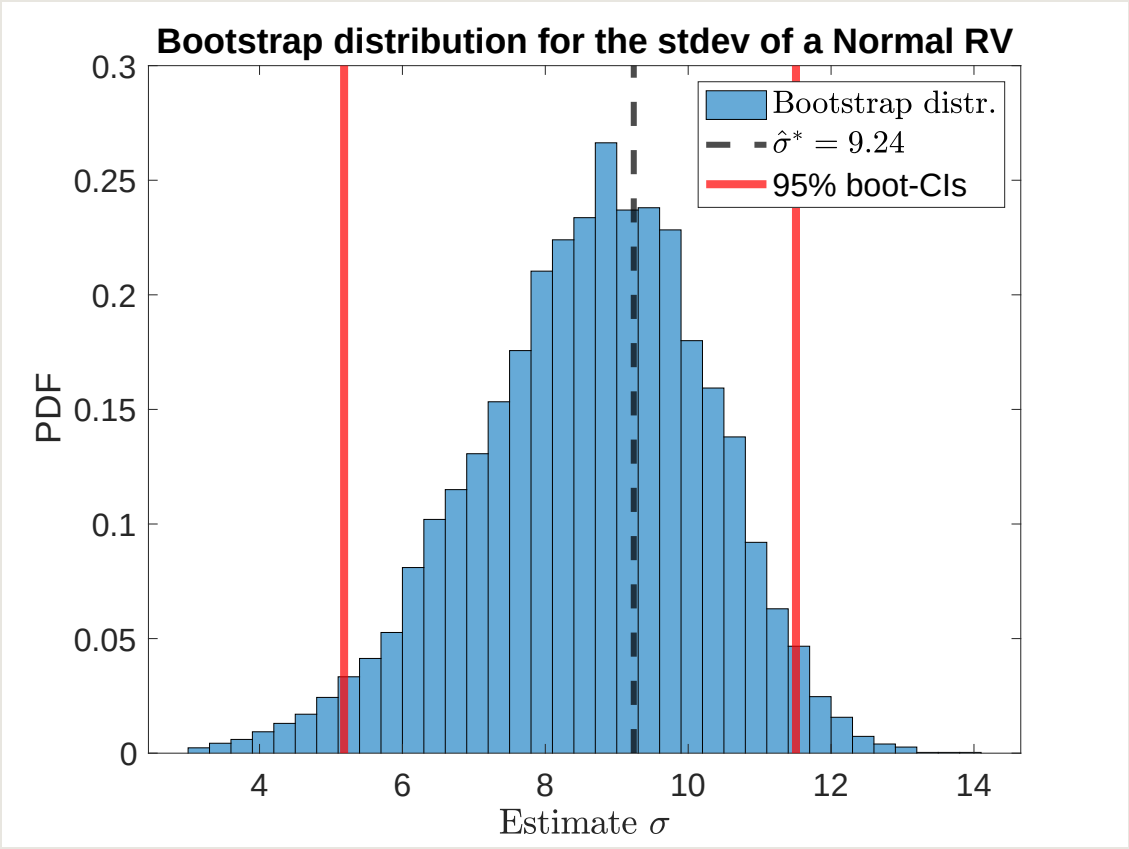
Example: *Bootstrap CI of the mean student heights*



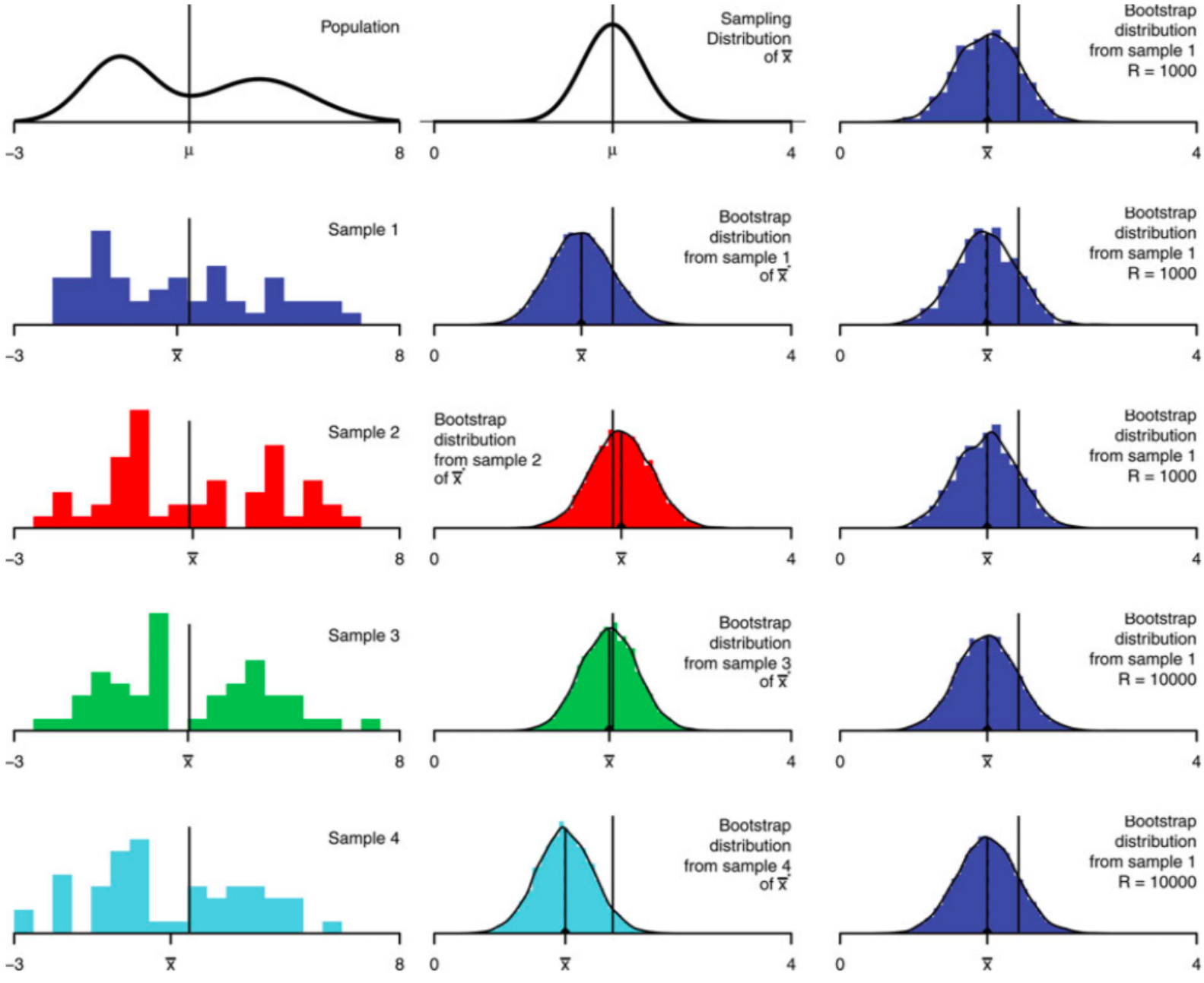
boot-CI at 95%: $[167.5, 178.2]$ estimated using $r = 10000$ bootstrap samples from original data: $x = [176, 165, 189, 180, 172, 169, 162, 161, 183, 170]$ cm.

Bootstrap confidence intervals for other quantities

Example: *Bootstrap CI of the standard deviation of student heights*



boot-CI at 95%: $[5.23, 11.49]$ estimated using $r = 10000$ bootstrap samples from original data: $x = [176, 165, 189, 180, 172, 169, 162, 161, 183, 170]$ cm.


Population
Sampling Distribution of $\bar{x}$
Bootstrap distribution from sample 1 R = 1000
Sample 1
Bootstrap distribution from sample 1 of $\bar{x}^*$
Bootstrap distribution from sample 1 R = 1000
Sample 2
Bootstrap distribution from sample 2 of $\bar{x}^*$
Bootstrap distribution from sample 1 R = 1000
Sample 3
Bootstrap distribution from sample 3 of $\bar{x}^*$
Bootstrap distribution from sample 1 R = 10000
Sample 4
Bootstrap distribution from sample 4 of $\bar{x}^*$
Bootstrap distribution from sample 1 R = 10000
−3
μ
8
0
4
$\bar{x}$

Properties, pros and cons of bootstrapping

- The bootstrap distribution is centered at the observed statistic, not the population parameter, for example, at $\bar{x}$, not μ. The bootstrap does not provide a better estimate of the population parameter: no matter how many bootstrap samples we take, they are centered at $\bar{x}$, not μ.
- Instead, the bootstrap distributions are useful for estimating the spread and shape of the sampling distribution.
- The spread and shape of bootsrap distributions from different samples are very similar if sample size is large, but they are different if sample size is small (e.g. $n = 10$).
- We sample with the same size as the original data because by doing so the standard errors reflect the actual data, rather than a hypothetical larger or smaller dataset.
- The bootstrap distribution can be different from the sampling distribution if it is very skewed. But bootstrap CIs can be significantly more accurate than t-CIs in these cases.