Bootstrapping

⋆ Schematic view of bootstrapping

Real world

$$\begin{array}{ccc} P & \longrightarrow & x = (x_1, \ldots, x_n) \\ & & \downarrow \\ & & \widehat{\theta} = s(x) \end{array}$$

$$\Longrightarrow$$

Bootstrap world

$$\begin{array}{ccc} \widehat{P} & \longrightarrow & x^\star = (x_1^\star, \ldots, x_n^\star) \\ & & \downarrow \\ & & \widehat{\theta}^\star = s(x^\star) \end{array}$$

Bootstrapping

- ⋆ Schematic view of bootstrapping

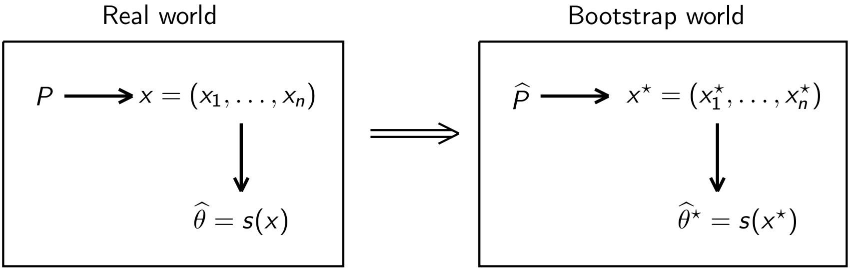
Real world $\qquad\qquad\qquad\qquad$ Bootstrap world

$$
\boxed{\begin{tikzcd} P \arrow[r] & x = (x_1, \ldots, x_n) \arrow[d] \\ & \widehat{\theta} = s(x) \end{tikzcd}}
\quad \Longrightarrow \quad
\boxed{\begin{tikzcd} \widehat{P} \arrow[r] & x^\star = (x_1^\star, \ldots, x_n^\star) \arrow[d] \\ & \widehat{\theta}^\star = s(x^\star) \end{tikzcd}}
$$

- ⋆ We have discussed bootstrapping for estimating
  - standard deviation
  - bias

Bootstrapping

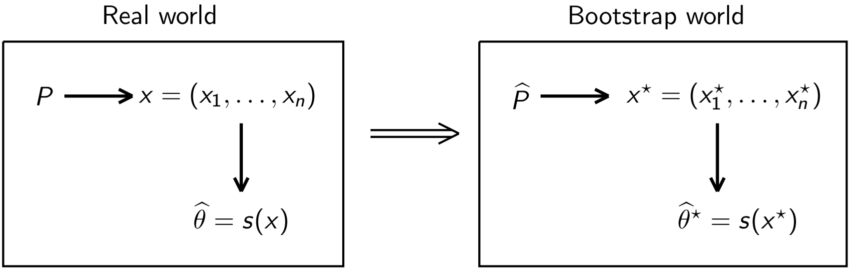
- ⋆ Schematic view of bootstrapping



- ⋆ We have discussed bootstrapping for estimating
  - standard deviation
  - bias
- ⋆ We start by defining the ideal bootstrap estimator
  - estimate it via Monte Carlo sampling

Bootstrapping

- ⋆ Schematic view of bootstrapping



- ⋆ We have discussed bootstrapping for estimating
  - standard deviation
  - bias
- ⋆ We start by defining the ideal bootstrap estimator
  - estimate it via Monte Carlo sampling
- ⋆ Today: bias-correction, confidence intervals, permutation tests

Bootstrapping of bias

- Observed data: $x_1, \ldots, x_n \sim F$ iid
- Parameter of interest: $\theta = t(F)$
- Estimator: $\widehat{\theta} = s(x)$
- Bias of $\widehat{\theta}$:
$$\text{bias}_F = \text{bias}_F(\widehat{\theta}, \theta) = E_F[s(x)] - t(F)$$
- Ideal bootstrap estimator for bias:
$$\text{bias}_{\widehat{F}} = E_{\widehat{F}}[s(x^\star)] - t(\widehat{F})$$
- boostrap estimate for the bias:
$$\widehat{\text{bias}}_B = \widehat{\theta}^\star(\cdot) - t(\widehat{F})$$

Bootstrapping of bias

- Observed data: $x_1, \ldots, x_n \sim F$ iid
- Parameter of interest: $\theta = t(F)$
- Estimator: $\widehat{\theta} = s(x)$
- Bias of $\widehat{\theta}$:
$$\text{bias}_F = \text{bias}_F(\widehat{\theta}, \theta) = E_F[s(x)] - t(F)$$
- Ideal bootstrap estimator for bias:
$$\text{bias}_{\widehat{F}} = E_{\widehat{F}}[s(x^\star)] - t(\widehat{F})$$
- boostrap estimate for the bias:
$$\widehat{\text{bias}}_B = \widehat{\theta}^\star(\cdot) - t(\widehat{F})$$
- Bias corrected estimator:
$$\widehat{\theta}_c = \widehat{\theta} - \widehat{\text{bias}}_B$$

Bootstrapping of bias

- Observed data: $x_1, \ldots, x_n \sim F$ iid
- Parameter of interest: $\theta = t(F)$
- Estimator: $\widehat{\theta} = s(x)$
- Bias of $\widehat{\theta}$:

$$\text{bias}_F = \text{bias}_F(\widehat{\theta}, \theta) = E_F[s(x)] - t(F)$$

- Ideal bootstrap estimator for bias:

$$\text{bias}_{\widehat{F}} = E_{\widehat{F}}[s(x^\star)] - t(\widehat{F})$$

- boostrap estimate for the bias:

$$\widehat{\text{bias}}_B = \widehat{\theta}^\star(\cdot) - t(\widehat{F})$$

- Bias corrected estimator:

$$\widehat{\theta}_c = \widehat{\theta} - \widehat{\text{bias}}_B = \widehat{\theta} - \left[\widehat{\theta}^\star(\cdot) - t(\widehat{F})\right]$$

Bootstrapping of bias

- Observed data: $x_1, \ldots, x_n \sim F$ iid
- Parameter of interest: $\theta = t(F)$
- Estimator: $\widehat{\theta} = s(x)$
- Bias of $\widehat{\theta}$:
$$\text{bias}_F = \text{bias}_F(\widehat{\theta}, \theta) = E_F[s(x)] - t(F)$$
- Ideal bootstrap estimator for bias:
$$\text{bias}_{\widehat{F}} = E_{\widehat{F}}[s(x^\star)] - t(\widehat{F})$$
- boostrap estimate for the bias:
$$\widehat{\text{bias}}_B = \widehat{\theta}^\star(\cdot) - t(\widehat{F})$$
- Bias corrected estimator:
$$\widehat{\theta}_c = \widehat{\theta} - \widehat{\text{bias}}_B = \widehat{\theta} - \left[\widehat{\theta}^\star(\cdot) - t(\widehat{F})\right]$$
  - if $\widehat{\theta} = t(\widehat{F})$ we get: $\widehat{\theta}_c = 2\widehat{\theta} - \widehat{\theta}^\star(\cdot)$

Bootstrapping of bias

- Observed data: $x_1, \ldots, x_n \sim F$ iid
- Parameter of interest: $\theta = t(F)$
- Estimator: $\widehat{\theta} = s(x)$
- Bias of $\widehat{\theta}$:
$$\text{bias}_F = \text{bias}_F(\widehat{\theta}, \theta) = E_F[s(x)] - t(F)$$
- Ideal bootstrap estimator for bias:
$$\text{bias}_{\widehat{F}} = E_{\widehat{F}}[s(x^\star)] - t(\widehat{F})$$
- boostrap estimate for the bias:
$$\widehat{\text{bias}}_B = \widehat{\theta}^\star(\cdot) - t(\widehat{F})$$
- Bias corrected estimator:
$$\widehat{\theta}_c = \widehat{\theta} - \widehat{\text{bias}}_B = \widehat{\theta} - \left[\widehat{\theta}^\star(\cdot) - t(\widehat{F})\right]$$
  - if $\widehat{\theta} = t(\widehat{F})$ we get: $\widehat{\theta}_c = 2\widehat{\theta} - \widehat{\theta}^\star(\cdot)$
- Note: $\text{Var}[\widehat{\theta}_c] \geq \text{Var}[\widehat{\theta}]$ so bias correction will not always give a better estimator. If the bias is small we should not do bias correction.

Recall: Derivation of the t-interval

- ⋆ Assume: $x_1, \ldots, x_n \sim F$ (iid)
  - notation: $\mathrm{E}[x_i] = \theta$, $\mathrm{Var}[x_i] = \sigma^2$
- ⋆ Estimator for θ:

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx N\left(\theta, \mathrm{se}^2\right) \quad \text{where} \quad \mathrm{se}^2 = \mathrm{Var}\left[\widehat{\theta}\right] = \frac{\sigma^2}{n}$$

- ⋆ Thereby

$$\frac{\widehat{\theta} - \theta}{\mathrm{se}} \approx N(0,1)$$

Recall: Derivation of the t-interval

- Assume: $x_1, \ldots, x_n \sim F$ (iid)
  - notation: $\mathrm{E}[x_i] = \theta$, $\mathrm{Var}[x_i] = \sigma^2$
- Estimator for θ:

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i \approx N\left(\theta, \mathrm{se}^2\right) \quad \text{where} \quad \mathrm{se}^2 = \mathrm{Var}\left[\widehat{\theta}\right] = \frac{\sigma^2}{n}$$

- Thereby

$$\frac{\widehat{\theta} - \theta}{\mathrm{se}} \approx N(0,1) \qquad \Rightarrow \qquad \frac{\widehat{\theta} - \theta}{\widehat{\mathrm{se}}} \approx t_{n-1}$$

Recall: Derivation of the t-interval

- Assume: $x_1, \ldots, x_n \sim F$ (iid)
  - notation: $\mathrm{E}[x_i] = \theta$, $\mathrm{Var}[x_i] = \sigma^2$
- Estimator for θ:

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx N\left(\theta, \mathrm{se}^2\right) \quad \text{where} \quad \mathrm{se}^2 = \mathrm{Var}\left[\widehat{\theta}\right] = \frac{\sigma^2}{n}$$

- Thereby

$$\frac{\widehat{\theta} - \theta}{\mathrm{se}} \approx N(0,1) \qquad \Rightarrow \qquad \frac{\widehat{\theta} - \theta}{\widehat{\mathrm{se}}} \approx t_{n-1}$$

- Confidence interval based on

$$P\left(-t_{\frac{\alpha}{2},n-1} \leq \frac{\widehat{\theta} - \theta}{\widehat{\mathrm{se}}} \leq t_{\frac{\alpha}{2},n-1}\right) \approx 1 - \alpha$$

Recall: Derivation of the t-interval

- Assume: $x_1, \ldots, x_n \sim F$ (iid)
  - notation: $\mathrm{E}[x_i] = \theta$, $\mathrm{Var}[x_i] = \sigma^2$
- Estimator for θ:

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i \approx N\left(\theta, \mathrm{se}^2\right) \quad \text{where} \quad \mathrm{se}^2 = \mathrm{Var}\left[\widehat{\theta}\right] = \frac{\sigma^2}{n}$$

- Thereby

$$\frac{\widehat{\theta} - \theta}{\mathrm{se}} \approx N(0,1) \qquad \Rightarrow \qquad \frac{\widehat{\theta} - \theta}{\widehat{\mathrm{se}}} \approx t_{n-1}$$

- Confidence interval based on

$$P\left(-t_{\frac{\alpha}{2},n-1} \leq \frac{\widehat{\theta} - \theta}{\widehat{\mathrm{se}}} \leq t_{\frac{\alpha}{2},n-1}\right) \approx 1 - \alpha$$

- Confidence interval

$$\left[\widehat{\theta} - t_{\frac{\alpha}{2},n-1}\widehat{\mathrm{se}}, \widehat{\theta} + t_{\frac{\alpha}{2},n-1}\widehat{\mathrm{se}}\right]$$

Recall: The z-confidence interval

- The z-confidence interval (assuming n to be large):

$$\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\widehat{\mathsf{se}}, \widehat{\theta} + z_{\frac{\alpha}{2}}\widehat{\mathsf{se}}\right]$$

Recall: The z-confidence interval

- The z-confidence interval (assuming n to be large):

$$\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\widehat{\mathrm{se}}, \widehat{\theta} + z_{\frac{\alpha}{2}}\widehat{\mathrm{se}}\right]$$

- An interpretation of this interval is:
  - if we draw a sample $\theta \sim N\left(\widehat{\theta}, \widehat{\mathrm{se}}\right)$
  - the lower and upper limits in the confidence interval are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ percentiles in the normal distribution

Recall: The z-confidence interval

- The z-confidence interval (assuming n to be large):

$$\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\widehat{\text{se}}, \widehat{\theta} + z_{\frac{\alpha}{2}}\widehat{\text{se}}\right]$$

- An interpretation of this interval is:
  - if we draw a sample $\theta \sim N\left(\widehat{\theta}, \widehat{\text{se}}\right)$
  - the lower and upper limits in the confidence interval are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ percentiles in the normal distribution
- Note: This interpretation is ideal for bootstrapping