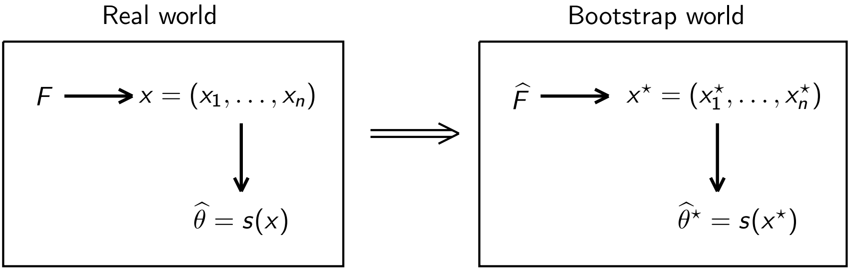
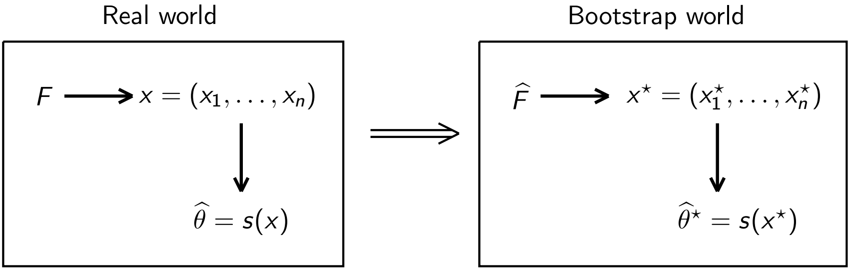
Bootstrapping

- ⋆ Schematic view of bootstrapping



Bootstrapping

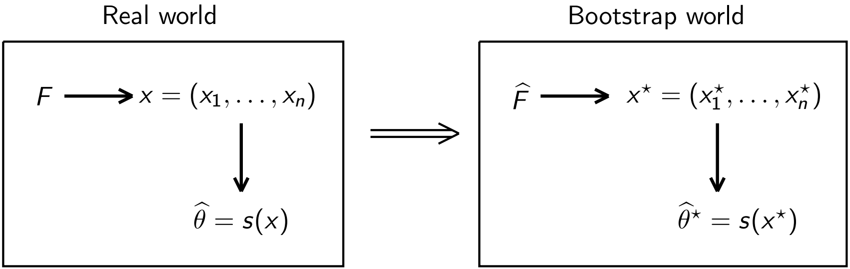
- Schematic view of bootstrapping



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

Bootstrapping

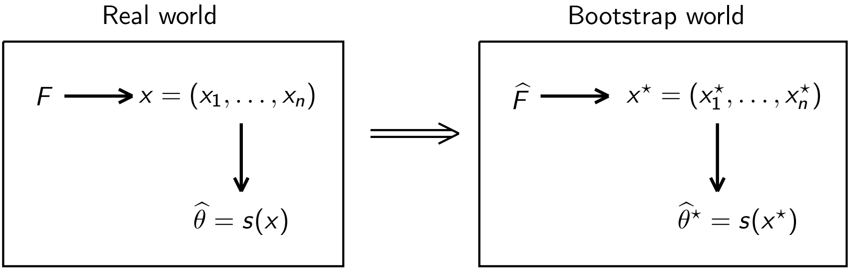
- Schematic view of bootstrapping



- 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: confidence intervals, prediction error, permutation tests

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):

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

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

Recall: Classification problem

- ⋆ Training data: $x_1, \ldots, x_n \sim F$ (iid), where $x_i = (z_i, y_i)$
- ⋆ Estimated classification rule

$$\widehat{y}(z; x) \quad \text{where} \quad x = (x_1, \ldots, x_n)$$

- ⋆ Define

$$\text{err}(x, F) = P_F\left(y \neq \widehat{y}(z; x)\right)$$

Recall: Classification problem

- ⋆ Training data: $x_1, \ldots, x_n \sim F$ (iid), where $x_i = (z_i, y_i)$
- ⋆ Estimated classification rule

$$\widehat{y}(z; x) \text{ where } x = (x_1, \ldots, x_n)$$

- ⋆ Define

$$\text{err}(x, F) = P_F\left(y \neq \widehat{y}(z; x)\right)$$

- ⋆ Misclassification rate

$$\theta = \mathsf{E}_F\left[\text{err}(x, F)\right]$$

Recall: Classification problem

- ⋆ Training data: $x_1, \ldots, x_n \sim F$ (iid), where $x_i = (z_i, y_i)$
- ⋆ Estimated classification rule

$$\widehat{y}(z; x) \;\; \text{where} \;\; x = (x_1, \ldots, x_n)$$

- ⋆ Define

$$\text{err}(x, F) = P_F\left(y \neq \widehat{y}(z; x)\right)$$

- ⋆ Misclassification rate

$$\theta = \mathsf{E}_F\left[\text{err}(x, F)\right]$$

- ⋆ Ideal bootstrap estimate of misclassication rate

$$\widehat{\theta} = \mathsf{E}_{\widehat{F}}\left[\text{err}\left(x^\star, \widehat{F}\right)\right]$$

Recall: Classification problem

- Training data: $x_1, \ldots, x_n \sim F$ (iid), where $x_i = (z_i, y_i)$
- Estimated classification rule

$$\widehat{y}(z; x) \quad \text{where} \quad x = (x_1, \ldots, x_n)$$

- Define

$$\text{err}(x, F) = P_F\left(y \neq \widehat{y}(z; x)\right)$$

- Misclassification rate

$$\theta = \mathsf{E}_F\left[\text{err}(x, F)\right]$$

- Ideal bootstrap estimate of misclassication rate

$$\widehat{\theta} = \mathsf{E}_{\widehat{F}}\left[\text{err}\left(x^\star, \widehat{F}\right)\right]$$

- Apparent misclassification rate

$$\text{err}(x, \widehat{F}) = \frac{1}{n}\sum_{i=1}^{n} I\left(y_i \neq \widehat{y}(z_i; x)\right)$$