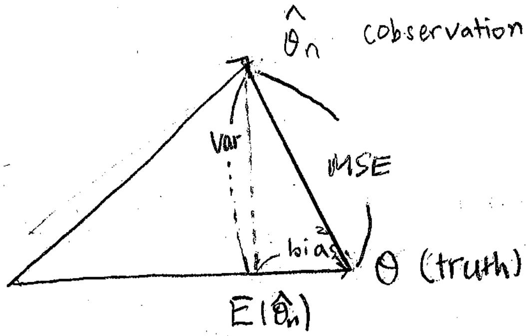
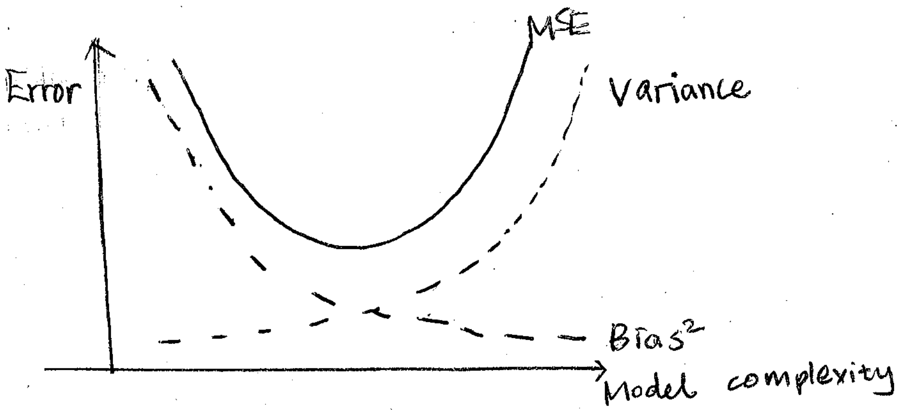
Week 3



$$MSE = E_\theta (\hat{\theta}_n - \theta)^2$$
$$= bias^2 + var$$

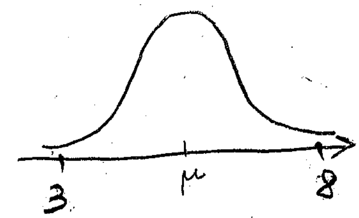


The bootstrap

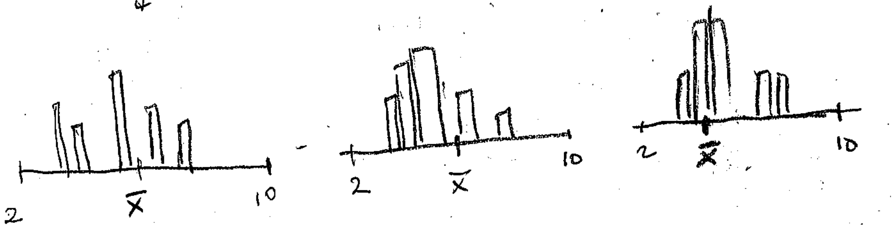
[Ideal World]

Population F

Sampling distribution of $\hat{\mu} = \bar{X}$

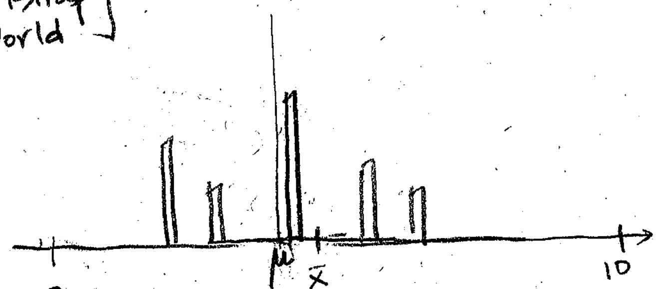


samples

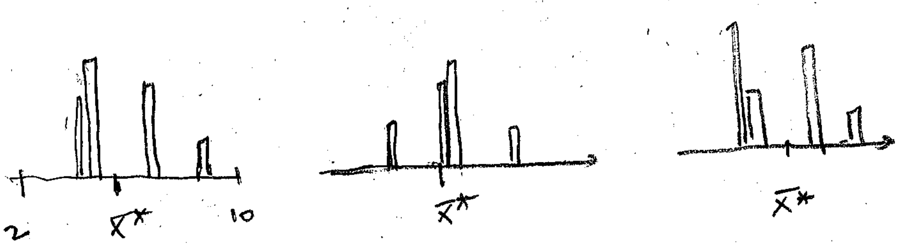


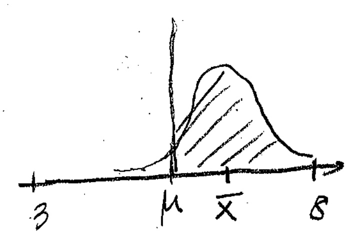
[Bootstrap World]

$\hat{F}$



Bootstrapped samples





The idea of the nonparametric bootstrap: empirical CDF + MC integration

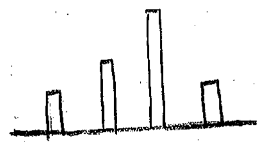
sample $X_1, \ldots, X_n \sim F$

Statistic $T_n = g(X_1, \ldots X_n)$

want: $\mathbb{V}_F(T_n)$

$$\approx \mathbb{V}_{\hat{F}_n}(T_n) \approx \widehat{\mathbb{V}}_{\hat{F}_n}(T_n)$$

CDF MC integration



$\rightarrow T^*_{n,1} = g(X_1^*, \ldots X_n^*)$

$\rightarrow$

$\rightarrow$

$\rightarrow T^*_{n,B} = g(X_1^*, \ldots, X_n^*)$

$F \sim X_1, \ldots X_n$

$T_n = g(X_1, \ldots X_n)$

$\mathbb{V}_{\hat{F}_n}(T_n)$

= variance of T_n

if the distribution of the data is $\hat{F}_n$

Approximate $\mathbb{V}_{\hat{F}_n}(T_n)$ by MC integration

$$v_{boot} = \widehat{\mathbb{V}}_{\hat{F}_n}(T_n) = \frac{1}{B}\sum_{j=1}^{B}\left(T^*_{n,j} - \frac{1}{B}\sum_{k=1}^{B} T^*_{n,k}\right)^2$$

<u>MC integration</u>: if h is any function with finite mean,

$$E[h(Y)] = \int h(y)\, dF_Y(y) \approx \frac{1}{B}\sum_{j=1}^{B} h(Y_j), \quad (Y_1, \ldots Y_B) \overset{iid}{\sim} F_Y$$

by the law of large numbers.

want: $\mathbb{V}(Y) = \int y^2 dF(y) - \left(\int y\, dF(y)\right)^2$

$$\Uparrow$$

$$\frac{1}{B}\sum_{j=1}^{B} Y_j^2 - \left(\frac{1}{B}\sum_{j=1}^{B} Y_j\right)^2$$

* MC integration

Expectation

$$E[h(X)] = \int h(x)\, dF_x = \int h(x) f(x)\, dx$$

average value of $h(x)$ over distribution f

variance

$$V[h(X)] = E[h(X) - E[h(X)]]^2$$

expected deviation from the expected value

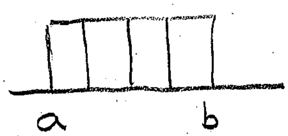
$$= E[h(X)^2] - [E[h(X)]]^2$$

$\int_a^b h(x)dx$ = The integral of $h(x)$ over $[a, b]$

$X_i \sim \text{Unif}[a,b]$

$$H_N = \frac{b-a}{N} \sum_{i=1}^{N} h(X_i) \text{ : Monte Carlo integration}$$

$$E[H_N] = \int_{[a,b]} h(X)\, dF_x$$



$$f(x) = \frac{1}{b-a} I_{[a,b]}(x)$$

$$E[H_N] = \frac{b-a}{N} \sum_{i=1}^{N} E[h(X_i)]$$

$$= \frac{b-a}{N} \sum_{i=1}^{N} \int_a^b h(x) \frac{1}{b-a}\, dx$$

$$= \frac{1}{N} \sum_{i=1}^{N} \int_a^b h(x)\, dx$$

$$= \int_a^b h(x)\, dx$$

* General MC

$$H_N = \frac{1}{N} \sum_{i=1}^{N} \frac{h(X_i)}{f(X_i)}, \text{ where } X_i \sim F, \text{ pdf: } f$$

$$E[H_N] = E\left[\frac{1}{N} \sum_{i=1}^{N} \frac{h(X_i)}{f(X_i)}\right] = \frac{1}{N} \sum_{i=1}^{N} \int \frac{h(x)}{f(x)} f(x)\, dx = \int h(x)\, dx$$

Method 1 Normal interval

T_n is close to Normal

$$T_n \pm z_{\alpha/2} \sqrt{V_{boot}}$$

Method 2 Pivotal interval

$\theta = T(F)$

$\hat{\theta}_n = T(\hat{F}_n)$

pivot $R_n = \hat{\theta}_n - \theta$: whose distribution doesn't depend on θ.

boot strap replication of $\hat{\theta}_n$: $(\hat{\theta}^*_{n,1}, \ldots, \hat{\theta}^*_{n,B})$

CDF of the pivot

$$H(r) = \mathbb{P}_F(R_n \le r)$$

$$a = \hat{\theta}_n - H^{-1}(1-\tfrac{\alpha}{2})$$

$$b = \hat{\theta}_n - H^{-1}(\tfrac{\alpha}{2})$$

$$\begin{aligned}
\mathbb{P}(a \le \theta \le b) &= \mathbb{P}(a - \hat{\theta}_n \le \theta - \hat{\theta}_n \le b - \hat{\theta}_n) \\
&= \mathbb{P}(\hat{\theta} - b \le \hat{\theta}_n - \theta \le \hat{\theta}_n - a) \\
&= \mathbb{P}(\hat{\theta}_n - b \le R \le \hat{\theta}_n - a) \\
&= H(\hat{\theta}_n - a) - H(\hat{\theta}_n - b) \\
&= (1 - \tfrac{\alpha}{2}) - (\tfrac{\alpha}{2}) = 1 - \alpha
\end{aligned}$$

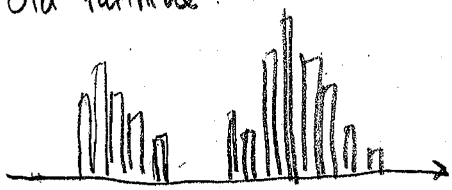
$$\widehat{CDF} = \hat{H}(r) = \frac{1}{B} \sum_{b=1}^{B} \Sigma\, I(R^*_{n,b} \le r)$$

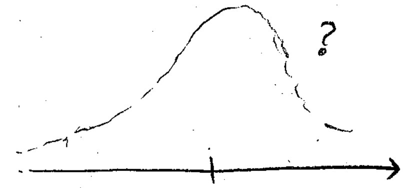
$$\hat{a} = \hat{\theta}_n - r^*_{1-\alpha/2} = 2 \cdot \hat{\theta}_n - \theta^*_{1-\alpha/2}$$

$$\hat{b} = \hat{\theta}_n - r^*_{\alpha/2} = 2 \cdot \hat{\theta}_n - \theta^*_{\alpha/2}$$

Method 3 Percentile interval $(\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2})$

Old faithful:

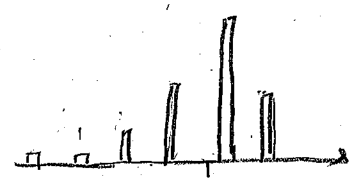




X_{med}

Data: $x_1, \ldots x_n$, x_{med} : sample median

1. draw bootstrap sample $x_1^*, \ldots, x_n^*$
2. compute median x^*_{med}
3. Repeat 1 & 2 B times $x^*_{med,1}, \ldots, x^*_{med,B}$



X^*_{med}

(1) Normal interval

$$v_{Boot} = \frac{1}{B}\sum_{b=1}^{B}\left(x^*_{med,b} - \frac{1}{B}\sum_{r=1}^{B} x^*_{med,r}\right)$$

$$\left(x_{med} - z_{\alpha/2}\sqrt{v_{Boot}},\ x_{med} + z_{\alpha/2}\sqrt{v_{Boot}}\right)$$

where $z_{\alpha/2} = \mathbb{P}(Z > z_{\alpha/2})$

$$(76 - 1.96 \times \sqrt{1.018},\ 76 + 1.96 \times \sqrt{1.018})$$

(2) Pivotal interval

$\hat{\theta}_n^* = x^*_{med}$

θ^*_β : the β sample quantile of

$(\hat{\theta}^*_{n,1}, \ldots, \hat{\theta}^*_{n,B}) = (x^*_{med,1}, \ldots, x^*_{med,B})$

$$(2x_{med} - \hat{\theta}^*_{1-\alpha/2},\ 2x_{med} - \hat{\theta}^*_{\alpha/2})$$

$$(2 \times 76 - 77,\ 2 \times 76 - 73)$$

(3) percentile interval

$$(\hat{\theta}^*_{\alpha/2},\ \hat{\theta}^*_{1-\alpha/2})$$

$$(73,\ 77)$$