STAT260 - Lecture 4 July 14, 2025

Bias - variance trade-off

Criteria for goodness of estimators

1) Principle of unbiasedness $E[\hat{\theta}] = \theta$

2) Consistency (Convergence in prob.) $\hat{\theta} \xrightarrow{P} \theta$, $\lim_{n \to \infty} Var(\hat{\theta}) = 0$

3) Principle of minimum variance: an estimator with low variance is preferable

4) Known sampling dist.: for most distributions, we have approximate distributions

Def. $\hat{\theta}$ is unbiased estimator of θ if $E[\hat{\theta}] = \theta$.

Otherwise we call it biased.

The bias of $\hat{\theta}$ is $B(\hat{\theta}) = E[\hat{\theta}] - \theta$

In reality sample space can't go to $n \to \infty$,

Def. Mean Square Error (→ used in machine learning)

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Bias – Variance Trade-off (decomposition)

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + [B(\hat{\theta})]^2$$

≥ 0 ≥ 0

Prove: $MSE(\hat{\theta}) = E[(\hat{\theta}-\theta)^2]$

$$= E\left[\left((\hat{\theta} - E[\hat{\theta}]) + (E[\hat{\theta}]-\theta)\right)^2\right]$$

$$= E\left[(\theta - E[\hat{\theta}])^2 + (E[\hat{\theta}]-\theta) + 2E\left[E(\hat{\theta})-\theta\right)(\hat{\theta}-E[\hat{\theta}])\right]$$

$$= Var(\hat{\theta}) + B^2(\hat{\theta}) + 2B(\theta)\underbrace{E\left[(\theta - E[\hat{\theta}])\right]}_{0}$$

$$= Var(\hat{\theta}) + B^2(\hat{\theta}) + 2B(\theta)\left(E[\hat{\theta}] - \cancel{E[E[\hat{\theta}]]}\right)$$

$E[\hat{\theta}]$

Can always convert biased est. into an unbiased one

$E[\hat{\theta}] = a\theta + b, \quad a, b > 0$

what is $B(\hat{\theta})$? $B(\hat{\theta}) = E[\hat{\theta}] - \theta = a\theta + b - \theta = (a-1)(\theta) + b$

$\Rightarrow B(\hat{\theta}) \neq 0 \Rightarrow \hat{\theta}$ is biased

- Find a function of $\hat{\theta}$ which is unbiased estimator of θ

$E[\hat{\theta}] = a\theta + b \Rightarrow E[\hat{\theta}] - b = a\theta \Rightarrow \frac{1}{a} E[\hat{\theta} - b] = \theta$

$\hat{\theta}^* = \frac{\hat{\theta} - b}{a}$ unbiased for θ $\Rightarrow E\left[\frac{\hat{\theta} - b}{a}\right] = \theta$

Suppose we have two independent estimators

$E[\hat{\theta}_1] = \theta \qquad E[\hat{\theta}_2] = \theta$

$\text{var}(\hat{\theta}_1) = \sigma_1^2 \qquad \text{var}(\hat{\theta}_2) = \sigma_2^2$

linear combination $\hat{\theta} = w\hat{\theta}_1 + (1-w)\hat{\theta}_2, \quad 0 < w < 1$

$w_1 x_1 + w_2 x_2$

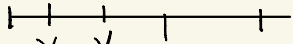
Convex combination

$w_1 x_1 + w_2 x_2$

$w_1 + w_2 = 1 \Rightarrow w_2 = 1 - w_1$

$0 < w_1, w_2 < 1$

$Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} Unif(0, \theta)$



Convex Combination of estimators gives lower var. unbiased results.

↑ use $Y_{(n)} = \max(Y_1, \ldots, Y_n)$ to estimate θ

$Y_{(1)}$ $\bar{Y}$

$E[\hat{\theta}] = wE[\hat{\theta}_1] + (1-w)E[\hat{\theta}_2]$

$= w\theta + (1-w)\theta = w\theta$

$+\theta - w\theta = \theta \Rightarrow$

$E[\hat{\theta}] = \theta$

$\Rightarrow \hat{\theta}$ is unbiased estimator

Find the optimal value of w s.t. variance of $\hat{\theta}$ is minimized

$var(\hat{\theta}) = var(w\hat{\theta}_1) + var((1-w)\hat{\theta}_2)$

$= w^2 var(\hat{\theta}_1) + (1-w)^2 var(\hat{\theta}_2)$

$= w^2\sigma_1^2 + (1-w)^2\sigma_2^2$

$g(w) = w^2\sigma_1^2 + (1-w)\sigma_2^2$

$\frac{d}{dw}(g(w)) = 2w\sigma_1^2 - (1-w)\sigma_2^2 = 0$

$\Rightarrow 2w\sigma_1^2 - 2(1-w)\sigma_2^2 = 0$

$\Rightarrow w\sigma_1^2 + \sigma_2^2 + w\sigma_2^2 = 0 \Rightarrow$

$w^* = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$

$\frac{d^2[g(w)]}{dw^2} = \sigma_1^2 + \sigma_2^2 > 0 \Rightarrow w^*$ is global minimum

ex: $\hat{\theta} = 0.8\hat{\theta}_1 + 0.2\hat{\theta}_1$

Note: All machine learning algorithms is an optimization problem

One more example:

$Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} Exp(\theta)$ $\quad f(x) = \frac{1}{\theta} e^{-y/\theta}, y > 0,$

$F(x) = 1 - e^{-y/\theta}$

Show that $\hat{\theta} = n\hat{Y}_{(1)}$ is unbiased for θ and find $MSE(\hat{\theta})$

Use dist. function technique to find sampling dist of $\hat{\theta}$

$F_{Y_{(1)}}(y) = P(Y_{(1)} < y) = P(\min(Y_1, \ldots, Y_n) \le y)$

$= 1 - P[\min(Y_1, \ldots, Y_n) \ge y]$

$= 1 - P[Y_1 > y, Y_2 > y, \ldots, Y_n > y]$

$$= 1 - P[Y_1 > y]\, P[Y_2 > y] \cdots P[Y_n > y]$$

$$= 1 - \left(P[Y_1 > y]\right)^n = 1 - \left(1 - F_y(y)\right)^n$$

$$\frac{d}{dy} F_{y_{(1)}}(y) = n\left(1 - F(y)\right)^{n-1} f(y)$$

using dist: $f_{y_{(1)}}(y) = n\left(1 - \left(1 - e^{-y/\theta}\right)\right)^{n-1} \cdot \frac{1}{\theta} e^{-y/\theta}$

$$= n e^{\frac{-(n-1)y}{\theta}} \cdot \frac{1}{\theta} e^{-y/\theta}$$

$$= \frac{n}{\theta} e^{-\frac{ny}{\theta}}$$

$$f_{y_{(1)}}(y) = \frac{n}{\theta} e^{-\left(\frac{n}{\theta}\right)y} \Rightarrow Y_1 \sim Exp\left(\frac{\theta}{n}\right)$$

$$E[Y_{(1)}] = \frac{\theta}{n} \Rightarrow E[nY_{(1)}] = (n)\left(\frac{\theta}{n}\right) = \theta$$

$\Rightarrow \hat{\theta} = nY_{(1)}$ is unbiased for θ

$$MSE(\hat{\theta}) = Var\left(n y_{(1)}\right) + \left[B\left(n y_{(1)}\right)\right]^2$$

$$= n^2 Var\left(y_{(1)}\right) = n^2 \left(\frac{\theta}{n}\right)^2 = \theta^2$$

$\therefore MSE(\hat{\theta}) = \theta^2$ → variance is constant

$E[\bar{y}] = \theta$ unbiased $\quad MSE(\hat{\theta}) = Var(\bar{y})$

$$= \frac{Var(y_i)}{n} = \frac{\theta^2}{n}$$

$$\therefore MSE(\hat{\theta}) = \frac{\theta^2}{n}$$

After break:

$$y_1, y_2, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$S_2^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

a) Find bias of each estimator

b) Find MSE of each estimator

c) Compare them

$$\frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2_{(n-1)} \Rightarrow E\left[\frac{(n-1)S_1^2}{\sigma^2}\right] = n-1$$

$$\Rightarrow E[S_1^2] = \sigma^2$$

$$B(S_1^2) = E[S_1^2] - \sigma^2 = \sigma^2 - \sigma^2 = 0 \Rightarrow B(S_1^2) = 0$$

$$S_2^2 = \left(\frac{n-1}{n}\right) S_1^2 \Rightarrow E[S_2^2] = \left(\frac{n-1}{n}\right) E[S_1^2]$$

$$= \left(\frac{n-1}{n}\right)\sigma^2$$

$$B(S_2^2) = E[S_2^2] - \sigma^2 = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$$

$$\Rightarrow B(S_2^2) = -\frac{1}{n}\sigma^2$$

1st one is better, according to principle of unbiasedness

S_1^2 is unbiased, whereas S_2^2 is biased.

However, the bias of S_2^2 vanishes as $n \to \infty$

$$\frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2_{(n-1)} \Rightarrow \operatorname{var}\left(\frac{(n-1)S_2^2}{\sigma^2}\right) = 2(n-1)$$

$$\Rightarrow \left(\frac{n-1}{\sigma^2}\right)^2 \operatorname{var}(S_1^2) = 2(n-1)$$

$$\Rightarrow \operatorname{var}(S_1^2) = \left(\frac{2}{n-1}\right)\sigma^4$$

$$MSE(S_1^2) = \operatorname{var}(S^2) + \underbrace{\left[B(S_1^2)\right]^2}_{0}$$

$$= \left(\frac{2}{n-1}\right)\sigma^4$$

$$S_2^2 = \left(\frac{n-1}{n}\right) S_1^2 \Rightarrow \operatorname{var}(S_2^2) = \left(\frac{n-1}{n}\right)^2 \operatorname{var}(S_1^2) =$$

$$\left(\frac{n-1}{n}\right)^2 \cdot \left(\frac{2}{n-1}\right)\sigma^4 = \left(\frac{2(n-1)}{n^2}\right)\sigma^4$$

$$MSE(S_2^2) = \left(\frac{2(n-1)}{n^2}\right)\sigma^4 + \left(\frac{1}{n}\right)^2\sigma^4 = \left(\frac{2n-1}{n^2}\right)\sigma^4$$

$$\therefore\ MSE(S_1^2) = \left(\frac{2}{n-1}\right)\sigma^4$$

$$MSE(S_2^2) = \left(\frac{2n-1}{n^2}\right)\sigma^4$$

we have $B(S_1^2) = 0$ and $B(S_2^2) \neq 0$, so the first estimator is preferable according to principle of unbiasedness

$n>1 \Rightarrow \left(\frac{2}{n-1}\right)\sigma^4 > \left(\frac{2n-1}{n^2}\right)\sigma^4 \Rightarrow MSE(S_1^2) > MSE(S_2^2)$

So, S_2^2 is preferable according to principle of minimum Variance

whichever one you use depends on sample size n.

$n \to \infty \Rightarrow \left(\frac{n}{n-1}\right) \approx \left(\frac{2n-1}{n^2}\right) \Rightarrow MSE(S_1^2) \approx MSE(S_2^2)$

So S_1^2 is preferable

$y \sim Poisson(\theta)$. Show that $(-1)^y$ is unbiased estimator for $e^{-2\theta}$

Compute $E[(-1)^y] = \sum_{i=1}^{\infty} (-1)^y \frac{\theta^y e^{-\theta}}{y!} = e^{-\theta} \sum_{i=0} \frac{(-\theta)^y}{y!}$

$= e^{-\theta} \cdot e^{-\theta} = e^{-2\theta}$

$(-1)^y$ is unbiased

Q1. why isn't this a good estimator, $(-1)^y \to -1, 1, -1, 1, \ldots$ -2θ bc $\theta < 1$

Q2. Are we violating principle of unbiasedness? Yes.

Principles are not theorems to hold true in all cases, and seldom do they yield satisfactory result in all cases.

Standard Error

The standard deviation of the sampling dist. of $\hat{\theta}$ is called standard error and is denoted by $\sigma_{\hat{\theta}} = \sqrt{\text{var}(\hat{\theta})}$

$$Var(\bar{y}) = \frac{\sigma^2}{n} \Rightarrow \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

$$y \sim Binomial(n, p) \Rightarrow Var(\hat{p}) = Var\left(\frac{Y}{n}\right)$$

$$= \frac{1}{n^2} Var(y) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

$$\therefore \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Error of Estimation:

$$\epsilon = |\hat{\theta} - \theta|$$

$y_1, \ldots, y_n \overset{iid}{\sim} Exp(\theta)$

Suggest an unbiased estimator for θ and find its standard error.

$E[\bar{y}] = \theta$ unbiased estimator

$$Var(\bar{y}) = \frac{Var(y_i)}{n} = \frac{\theta^2}{n} \Rightarrow \sigma_{\bar{y}} = \frac{\theta}{\sqrt{n}} \approx \frac{\hat{\theta}}{\sqrt{n}}$$

$$\Rightarrow \sigma_{\bar{y}} = \frac{\bar{y}}{\sqrt{n}}$$

plug-in principle

An engineer observes $n \geq 10$ independent length-of-life measurements for an electronic component. The average of these observations is 1020 hours. If these observations come from exponential dist. with mean θ, then estimate θ and place a 2-standard-error bound on the error of estimation.

$n \geq 10$

$\hat{\theta} = \bar{Y} = 1020 \Rightarrow$ 2-Standard-error $= (2)\hat{\sigma}_{\hat{\theta}} \approx \frac{2 \cdot \bar{Y}}{\sqrt{n}} = 2\left(\frac{1020}{\sqrt{10}}\right) = 645.1$

Last example:

The # of grain nucleation sites per unit volume is modeled as poisson dist with mean λ.

50 samples are collected with average 20 sites per unit volume. Estimate mean λ and place a 2-standard-error bound on the error of estimation.

Let y_i = r.v. denoting # of grain nucleation sites

$y_i \sim \text{Poisson}(\lambda)$, $i = 1, \ldots, n$ $\quad E[y_i] = \lambda$ $\quad \text{Var}(y_i) = \lambda$

$$E[\bar{y}] = \frac{1}{50}\sum_{i=1}^{50} E[y_i] = \frac{1}{50}\sum_{i=1}^{50} \lambda = \lambda$$ unbiased estimator

$$\text{Var}(\bar{y}) = \frac{\text{Var}(y_1)}{n} = \frac{\lambda}{50} \Rightarrow \sigma_{\bar{y}} = \sqrt{\frac{\lambda}{50}} \approx \sqrt{\frac{\lambda}{10}}$$

$$= \sqrt{\frac{\bar{y}}{50}} = 0.6324$$

bound on error:

$$20 \pm 2 \times 0.6324 = 20 \pm 1.265 = (18.7, 21.2)$$