# Inference Notes

Rui Qiu

2018-03-05

## 1 Lecture 02b (2018-03-05) / properties of estimators

Suppose $x_1, \ldots, x_n$ with density $f(x;\theta)$, a statistics $\hat{\theta} = T(x_1, \ldots, x_n)$, e.g. $x_i \sim N(\mu, \sigma^2)$ then $\hat{\mu} = \bar{X}$.

**Estimates** are exact values, while **estimators** are random variables.

**Definition:** $\hat{\theta} = T(X_1, \ldots, X_n)$ is an unbiased estimator of $\theta$ if $E(T(\mathbf{X})] = \theta$, the bias of an estimator is

$$\text{bias}(\hat{\theta}) = E[T(\mathbf{X})] - \theta$$

**Definition:**

$$\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = V(\hat{\theta}) + \text{Bias}(\hat{\theta})^2.$$

Accepting some slight bias in reduction of variance. Bias is related to accuracy, variance is related to spread.

**Example:** Consider $X_1, \ldots, X_n \sim i.i.d. N(\mu, \sigma^2)$, we have three estimators:

$$\begin{aligned}
\tilde{\sigma}^2 &= \frac{1}{n+1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \text{ MLE} \\
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \text{ also an unbiased estimator}
\end{aligned}$$

The corresponding MSEs are:

$$
\begin{aligned}
\mathrm{MSE}[\tilde{\sigma}^2] &= \frac{2\sigma^4}{n+1} \\
\mathrm{MSE}[\hat{\sigma}^2] &= \frac{(2n-1)\sigma^4}{n^2} \\
\mathrm{MSE}[S^2] &= V(S^2) = \frac{2\sigma^4}{n-1} \\
\mathrm{MSE}[\tilde{\sigma}^2] &< \mathrm{MSE}[\hat{\sigma}^2] < \mathrm{MSE}[S^2]
\end{aligned}
$$

**Example:** $X_1, \ldots, X_n \sim i.i.d.\text{Bernoulli}(p)$ with the MLE $\hat{p} = \bar{X}$.

$$
\begin{aligned}
\mathrm{Bias}(\hat{p}) &= E[\hat{p}] - p = 0 \\
V(\hat{p}) &= V(\bar{X}) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n} \\
\mathrm{MSE}(\hat{p}) &= V(\hat{p}) + \mathrm{Bias}(\hat{p})^2 = \frac{p(1-p)}{n}.
\end{aligned}
$$

**Bayesian Estimator Example:** $\hat{p}_B = \frac{y+a}{a+b+n}$.

$$
\begin{aligned}
E\left[\frac{Y+a}{a+b+n}\right] &= \frac{E[Y]+a}{a+b+n} = \frac{np+a}{a+b+n} \\
V\left[\frac{Y+a}{a+b+n}\right] &= \left[\frac{1}{a+b+n}\right]^2 V(Y) = \left[\frac{1}{a+b+n}\right]^2 np(1-p) \\
\mathrm{MSE}[\hat{p}_B] = \mathrm{MSE}\left[\frac{Y+a}{a+b+n}\right] &= \left[\frac{np(1-p)}{(a+b+n)^2}\right] + \left[\frac{np+a}{a+b+n} - p\right]^2
\end{aligned}
$$

It turns out if we set $a = b = \sqrt{n/4}$, we get $\hat{p}_B$ and $\mathrm{MSE}[\hat{p}_B]$ respectively. Then we can plot the MSEs plot and compare them. (For small $n$, ... for large $n$ ...)

## 2 Lecture 02b (2018-03-07) / properties of estimators

**Example:** For constant estimator, $\tilde{p}_{con} = 0.5$ for all $x$, then

$$\begin{aligned} \text{MSE}(\tilde{p}_{con}) &= V(\tilde{p}_{con}) + \text{Bias}(\tilde{p}_{con})^2 \\ &= 0 + (E(\tilde{p}_{con}) - p)^2 \\ &= (0.5 - p)^2 \end{aligned}$$

**Definition:** An estimator $\theta$ is **weakly consistent** if

$$P(|\hat{\theta} - \theta| > \epsilon) \to 0 \text{ as } n \to \infty, \forall \epsilon > 0$$

**Proof with Chebyshev's inequality:**

$$\begin{aligned} P(|\hat{\theta} - \theta| > \epsilon) &\leq \frac{E[(\hat{\theta} - \theta)^2]}{\epsilon^2} \\ &= \frac{\text{MSE}(\hat{\theta})}{\epsilon} \\ &= \frac{1}{\epsilon^2}\left[V(\hat{\theta}) + \text{bias}(\hat{\theta})^2\right] \end{aligned}$$

$V(\hat{\theta}) \to 0$ and $\text{bias}(\hat{\theta}) \to 0 \implies \hat{\theta}$ is consistent. (sufficient but not necessary)

For far, there is no uniform way to find a best estimator. But instead, we say an estimator $T^*$ is a **best unbiased estimator** of $\tau(\theta)$ if it satisfies $E[T^*] = \tau(\theta)$ for all $\theta$ and for any other estimator $T$ with $E[T] = \tau(\theta)$ we have:

$$V(T^*) \leq V(T) \ \forall \ \theta.$$

$T^*$ also called **minimum variance unbiased estimator (MVUE)** for $\tau(\theta)$.

Suppose we have a bunch of unbiased estimators, maybe even more. How can we be so sure that we find the one with smallest variance?

Introducing...

**Definition (Cramer-Rao Inequality [lower bound]:** Let $X_1, \ldots, X_n$ be a random sample from a distribution family with density function $f_X(x;\theta)$, where $\theta$ is a scalar parameter. Also let $T = t(X_1, \ldots, X_n)$ be an unbiased estimator of $\tau(\theta)$, then under certain regularity (smoothness) conditions:

$$V(T) \geq \frac{\{\tau'(\theta)\}^2}{ni(\theta)} = \{\tau'(\theta)\}^2 \cdot I(\theta)^{-1}$$

Note:
$\tau'(\theta) = \frac{d}{d\theta}\tau(\theta)$
$I(\theta) = ni(\theta)$, expected Fisher Information
$I(\theta) = E\left[\left(\frac{d \ln \theta}{d\theta}\right)^2\right] = -E\left[\frac{d^2 \ln \theta}{d\theta^2}\right]$

**C-R inequality extended:** $X_1, \ldots, X_n$ be a sample (not necessarily to be iid), with pdf $f(\mathbf{x}|\theta)$ and let $T(\mathbf{X})$ be an estimator (not necessarily to be unbiased) then based on regularity conditions:

$$V[T(\mathbf{X})] \geq \frac{\left[\frac{d}{d\theta}E[T(\mathbf{X})]\right]^2}{E\left[\left(\frac{d}{d\theta}\log f(\mathbf{x} \mid \theta)\right)^2\right]} = \frac{\left[\frac{d}{d\theta}E[T(\mathbf{X})]\right]^2}{I(\theta)}$$

Note:
$\frac{d}{d\theta}\ln\{f(x\theta)\}$ exists for all $x$ and $\theta$.
...
...

**Proof by Cauchy-Schwarz Inequality:**

**Corollary (iid case):** If the regularity conditions hold and $T(\mathbf{X})$ is an unbiased estimator for $\tau(\theta)$ and we have $X_1, \ldots, X_n \sim iid\ f(x|\theta)$, then

$$V[T(\mathbf{X})] \geq \frac{\left[\frac{d}{d\theta}E[T(\mathbf{X})]\right]^2}{nE\left[\left(\frac{d}{d\theta}\log f(x|\theta)\right)^2\right]} = \frac{[\tau'(\theta)]^2}{ni(\theta)} = \{\tau'(\theta)\}^2 I(\theta)^{-1}$$

**Definition:** The Fisher information, or expected Fisher information, or the information number is

$$I(\theta) = E\left[\left(\frac{d}{d\theta}\log f(\mathbf{x}|\theta)\right)^2\right] = -E\left[\left(\frac{d^2}{d\theta^2}\log f(\mathbf{x}|\theta)\right)\right]$$

For one data point we have

$$i(\theta) = E\left[\left(\frac{d}{d\theta}\log f(x|\theta)\right)^2\right]$$

For iid data

$$ni(\theta) = I(\theta)$$