1 Evaluating the performance of estimators

Example: Suppose we observe $X_1, \ldots, X_n$ iid $\mathrm{N}(\theta, \sigma_0^2)$, with σ_0^2 known, and wish to estimate θ.
Two possible estimators are:
$\hat{\theta} = \bar{X}$ and sample mean and $\hat{\theta} = M \equiv$ sample median.
Which is better? How to measure performance?
Some possibilities:

1. Compare $E|\bar{X} - \theta|$ with $E|M - \theta|$.
2. Compare $E(\bar{X} - \theta)^2$ with $E(M - \theta)^2$
3. Compare $EL(\theta, \bar{X})$ with $EL(\theta, M)$

where $L(\cdot, \cdot)$ is an appropriate "loss function": the value $L(\theta, a)$ is some measure of the loss incurred when the true value is θ and our estimate is a.

$$\begin{aligned} \text{absolute error loss:} \quad L(\theta, a) &= |a - \theta| \\ \text{squared error loss:} \quad L(\theta, a) &= (a - \theta)^2 \\ \text{game show loss:} \quad L(\theta, a) &= I(|a - \theta| > c) \\ \text{Stein's loss (for } \theta, a > 0) \quad L(\theta, a) &= \frac{a}{\theta} - 1 - \log\left(\frac{a}{\theta}\right) \end{aligned}$$

Historically, estimators have been most frequently compared using Mean Squared Error: $\mathrm{MSE}(\theta) = E_\theta(\hat{\theta} - \theta)^2$. This is because the MSE can often be calculated or approximated (for large samples), and has nice mathematical properties.

1.1 Admissible and Inadmissible Estimators

Let $W = W(\mathbf{X})$ be an estimator of $\tau = \tau(\theta)$. Define $\mathrm{MSE}_W(\theta) = E_\theta(W - \tau(\theta))^2$.

Definition 1. *An estimator W is inadmissible (w.r.t. squared error loss) if there exists another estimator $V = V(\mathbf{X})$ such that*

$$MSE_V(\theta) \leq MSE_W(\theta), \quad \forall\, \theta \in \Theta$$

with strict inequality for at least one value of θ. (An estimator is inadmissible if there is another estimator that "beats" it.) An estimator which is not inadmissible is called admissible.

Note: An admissible estimator may actually be very bad. An inadmissible estimator can sometimes be pretty good.
Note: If we are using a loss function $L(\tau, a)$, we also define inadmissible and admissible estimators in the same way, replacing the MSE by the more general notion of a risk function $R(\theta, W) = E_\theta L(\tau(\theta), W(\mathbf{X}))$.
Examples: Again, suppose we observe $X_1, \ldots, X_n$ iid $\mathrm{N}(\theta, \sigma_0^2)$, with σ_0^2 known, and wish to estimate θ. Consider the estimator $W \equiv 0$ that always estimates θ by 0 regardless of the data $\mathbf{X}$. This is a very bad estimator, but it is admissible because it is great when $\theta = 0$. No non-degenerate estimator V can possibly beat W since it would have to satisfy

$$\begin{aligned} &\mathrm{MSE}_V(0) \le \mathrm{MSE}_W(0) \\ &\implies E_0(V-0)^2 \le E_0(W-0)^2 \\ &\implies E_0 V^2 \le 0 \implies P_0(V=0) = 1 \implies V \equiv 0. \end{aligned}$$

Now consider the estimator $M \equiv$ sample median. We show later that the sample mean $\bar{X}$ has a uniformly smaller MSE than M so that M is inadmissible. (The two MSE functions are constant, i.e., flat.) However, M is not a bad estimate of θ, and might be used if there were doubts about the normality assumption (perhaps the true distribution has thicker tails) or concern about outliers.

1.2 Bias, Variance, and MSE (for an estimator W of $\tau(\theta)$)

$$\begin{aligned} \mathrm{Bias}_W(\theta) &= E_\theta(W - \tau(\theta)) \\ \mathrm{Var}_W(\theta) &= E_\theta(W - E_\theta W)^2 \equiv \mathrm{Var}_\theta(W) \end{aligned}$$

Fact: $\mathrm{MSE}_W(\theta) = \mathrm{Bias}_W^2(\theta) + \mathrm{Var}_W(\theta)$

Proof. For any rv Y with finite second moment, we know

$$EY^2 = (EY)^2 + \mathrm{Var}(Y)$$

Taking $Y = W - \tau$ leads to

$$\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Var}$$

Since $\mathrm{Var}(W - \tau) = \mathrm{Var}(W)$ □

Definition 2. *An estimator W with $\mathrm{Bias}_W(\theta) \equiv 0$, that is,*

$$E_\theta(W) = \tau(\theta), \quad \theta \in \Theta$$

is said to be unbiased. If not, it is biased.

For an unbiased estimator, MSE = Var.
Example: Coin-tossing. $X_1, X_2, \ldots, X_n$ iid Bernoulli(θ). $\hat{\theta}_{MLE} = T/n, T = \sum X_i$.
$\hat{\theta}_{Bayes} = (1-p)a + p(T/n)$ where $p = n/(\alpha+\beta+n)$ and a is the prior mean $\alpha/(\alpha+\beta)$.
Now observe,

$$\begin{aligned} \mathrm{MSE}_{\mathrm{MLE}}(\theta) &= Bias^2 + \mathrm{Var} \\ &= 0 + \frac{\theta(1-\theta)}{n} \end{aligned}$$

Also

$$\begin{aligned} \mathrm{MSE}_{\mathrm{Bayes}}(\theta) = (\theta) &= Bias^2 + \mathrm{Var} \\ &= [(1-p)a - (1-p)\theta]^2 + p^2\frac{\theta(1-\theta)}{n} \\ &= (1-p)^2(a-\theta)^2 + p^2\frac{\theta(1-\theta)}{n} \end{aligned}$$

Which is better? (according to MSE). Answer: Neither dominates the other.

$$\mathrm{MSE}_{\mathrm{Bayes}}(a) = p^2\frac{a(1-a)}{n} < \frac{a(1-a)}{n} \mathrm{MSE}_{\mathrm{Bayes}}(0) = (1-p)^2a^2 > 0 = \mathrm{MSE}_{\mathrm{MLE}}(1)$$

Thus, the Bayes estimate is superior in the neighborhood of $\theta = a$, and the MLE is superior near $\theta = 0$ and $\theta = 1$.
Note: Both $\mathrm{MSE}_{\mathrm{MLE}}(\theta)$ and $\mathrm{MSE}_{\mathrm{Bayes}}(\theta)$ are parabolas (quadratic functions of θ).
Note: Regarding (in)admissibility, the above remarks prove nothing. But it can be shown that both the Bayes estimate and MLE are admissible here. Typically, Bayes estimates (with proper priors) are admissible.
Example: Estimating the variance in $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$, $\tau(\theta) = \sigma^2$.
$W = c\sum_{i=1}^n (X_i - \bar{X})^2 = cSS$. $c = 1/(n-1)$ usual unbiased estimator, $c = 1/n$ is MLE. Which is better?

$$\frac{SS}{\sigma^2} \sim \chi^2_{n-1}$$

Thus

$$\begin{aligned} E\{cSS\} &= c\sigma^2(n-1) \\ \mathrm{Var}\{cSS\} &= 2c^2\sigma^4(n-1) \end{aligned}$$

$$\begin{aligned} MSE_{cSS}(\mu,\sigma^2) = E\{(cSS-\sigma^2)^2\} &= \text{Bias}^2 + \text{Var} \\ &= (c\sigma^2(n-1)-\sigma^2)^2 + 2c^2\sigma^4(n-1) \\ &= \sigma^4\{(c(n-1)-1)^2 + 2c^2(n-1)\} \\ &= \sigma^4\psi(c) \end{aligned}$$

with $\psi(c) = (c(n-1)-1)^2 + 2(n-1)c^2$.

$$\begin{aligned} \psi'(c) &= 2\{c(n-1)-1\}(n-1) + 4(n-1)c \\ &= 2(n-1)\{c(n+1)-1\}. \end{aligned}$$

$\psi''(c) = 2(n-1)(n+1) > 0, n \geq 2.$

$\psi'(c) = 0$ when $c = 1/(n+1)$. Both $c = 1/(n-1)$ and $c = 1/n$ give inadmissible estimators. $c = 1/(n-1)$ is the best c with Stein's loss function: (page 351)

$$L(\sigma^2, a) = \frac{a}{\sigma^2} - \log\frac{a}{\sigma^2} - 1$$

Other plausible loss functions:

$$\begin{aligned} L(\sigma^2, a) &= \frac{\sigma^2}{a} - \log\frac{\sigma^2}{a} - 1, \text{ or} \\ &= \frac{\sigma^2}{a} - \log\frac{a}{\sigma^2} - 2, \text{ or} \\ &= (\log a - \log\sigma^2)^2 \end{aligned}$$

Comments:

1. Estimator with best c might not be admissible !
2. MSE inappropriate (or dubious, anyway) for estimation of σ^2.

Theorem 1. *(Rao - Blackwell Theorem) If $T = T(X)$ is a sufficient statistic for θ, $E_\theta S(X) = \tau(\theta)$ for all θ, and $E_\theta(S(X) - \tau(\theta))^2 < \infty$ for all θ, then*

$$S^*(X) = E(S(X) \mid T(X))$$

satisfies

$$\begin{aligned} E_\theta S^*(X) &= \tau(\theta), \quad \forall\theta, \quad \textit{and} \\ E_\theta(S^*(X) - \tau(\theta))^2 &\leq E_\theta(S(X) - \tau(\theta)), \quad \forall\theta, \end{aligned}$$

Notes:

1. $E(S(X) \mid T(X))$ does not depend on θ necause $T(X)$ is sufficient so that $\mathcal{L}(X \mid T)$ (and $\mathcal{L}(S \mid T)$) does not depend on θ.

2. Equality of MSE's for a particular θ can occur iff $P_\theta(S(X) = S^*(X)) = 1$.

3. $S^*(X)$ is a function of $T(X)$ (i.e., $\exists\, \psi$ such that $S^*(X) = \psi(T(X))$).

Proof. Recall for any rv's X, Y with $EX^2 < \infty$, we have

$$\begin{aligned} EX &= E(E(X \mid Y)) \\ \mathrm{Var}(X) &= E(\mathrm{Var}(X \mid Y)) + \mathrm{Var}(E(X \mid Y)). \end{aligned}$$

Now apply these facts:

$$\begin{aligned} E_\theta[S^*(\mathbf{X})] &= E_\theta[E_\theta(S(\mathbf{X}) \mid T(\mathbf{X})] \\ &= E_\theta S(\mathbf{X}) = \tau(\theta). \\ E_\theta(S-\tau)^2 = \mathrm{Var}_\theta(S) &= \underbrace{E[\mathrm{Var}(S \mid T)]}_{\geq 0} + \mathrm{Var}[\underbrace{E(S \mid T)}_{S^*}] \\ &\geq \mathrm{Var}(S^*) = E_\theta(S^* - \tau(\theta))^2. \end{aligned}$$

Equality can occur only when $E_\theta \mathrm{Var}(S \mid T) = 0$. But

$$\begin{aligned} E_\theta \mathrm{Var}(S \mid T) &= E_\theta\{E[(S - E(S \mid T))^2 \mid T]\} \\ &= E_\theta\{(S - E(S \mid T))^2\} = E_\theta\{(S - S^*)^2\} \\ &= 0 \quad \text{iff} \quad P_\theta(S = S^*) = 1 \end{aligned}$$

Arguing more loosely, $E_\theta \mathrm{Var}(S \mid T) = 0 \implies \mathrm{Var}(S \mid T) = 0 \implies S$ is a function of $T \implies S^* \equiv E(S \mid T) = S$. □

Example: $X_1, X_2, \ldots, X_n$ iid Bernoulli(p). $T = \sum X_i$ is a sufficient statistic for p. $\mathcal{L}(X \mid T)$ puts equal probability of $1/\binom{n}{T}$ on all strings wit T 1's and $n - T$ 0's. Generate from $\mathcal{L}(X \mid T = t)$ by placing t1's and $n - t$ 0's in an urn, and randomly drawing (without replacement) until the urn is empty.

1. Estimation of p: $EX_1 = p$ for all p so that $S = X_1$ is an unbiased estimator of p. X_1 is *not* a function of T so it can be improved by conditioning (as in the Rao-Blackwell Theorem).

$$S^* = E(S \mid T) = E(X_1 \mid T) = P(X_1 = 1 \mid T) = \frac{T}{n}.$$

$S^* = T/n$ is the usual estimator (the sample proportion). Clearly

$$ES^* = p \quad \forall\, p$$

$$\mathrm{Var}(S^*) = \frac{p(1-p)}{n} < \mathrm{Var}(S) = p(1-p) \quad \forall\, p$$

verifying the conclusion of the R-B Theorem.

2. Estimation of p^2: $EX_1X_2 = p^2$ for all p so that $S = X_1X_2$ is an unbiased estimator of p. X_1X_2 is *not* a function of T so it can be improved by conditioning (as in the Rao-Blackwell Theorem).

$$\begin{aligned} S^* &= E(X_1X_2 \mid T) = P(X_1X_2 = 1 \mid T) \\ &= P(X_1 = 1, X_2 = 1 \mid T) = \frac{T}{n} \cdot \frac{T-1}{n-1}. \end{aligned}$$

By R-B Thm, S^* is an unbiased estimate of p^2 with smaller variance than S. This can be verified by straightforward calculations. For comparison, what is the MLE of p^2?
The MLE of p is T/n, so the invariance principle for MLEs says the MLE of p^2 is $(T/n)^2$. Clearly $\frac{T}{n} \cdot \frac{T-1}{n-1}$ and $(T/n)^2$ are very close when n is large. Which is better? Neither dominates. $(T/n)^2$ is biased, but the bias is negligible for large n.

3. Estimation of p^3: $S = X_1X_2X_3$ is an unbiased estimator of p^3.

$$\begin{aligned} S^* &= E(X_1X_2X_3 \mid T) = P(X_1 = X_2 = X_3 = 1 \mid T) \\ &= \frac{T}{n} \cdot \frac{T-1}{n-1} \cdot \frac{T-2}{n-2}. \end{aligned}$$

is the Rao-Blackwell improvement on S. The pattern is now clear for p^4, etc.

Suppose $T = T(X)$ is a complete and sufficient statistic for θ. Then

1. For any parameter $\tau(\theta)$, there is at most one unbiased estimator which is a function of T.

2. If $S = S(X)$ is unbiased for $\tau(\theta)$ (and $\mathrm{Var}(S) < \infty$) for all θ, then

$$S^*(X) = S^* = E(S \mid T)$$

is the UMVUE for $\tau(\theta)$.

Definition 3. *$S = S(X)$ is the UMVUE (uniformly minimum variance unbiased estimator) for $\tau(\theta)$ if*

$$\begin{aligned} E_\theta S &= \tau(\theta), \forall\, \theta, \text{ and} \\ Var_\theta(S) &\leq Var_\theta(S'), \forall\, \theta. \end{aligned}$$

for any other unbiased estimator S'.

Terminalogy: UMVUE = "best unbiased estimator".

3. An unbiased estimator (with finite variance) which is a function of T is the UMVUE.

1. Proof of 1. Suppose $S_1(X) = \psi_1(T(X))[S_1 = \psi_1(T)]$, $S_2(X) = \psi_2(T(X))[S_2 = \psi_2(T)]$ and $ES_1 = ES_2 = \tau(\theta)$ for all θ. Then $E(S_1 - S_2) = 0$ for all θ or $E_\theta g(T) = 0$ for all θ where $g(t) = \psi_1(t) - \psi_2(t)$. By completeness $P_\theta(g(T) = 0) = P_\theta(S_1 = S_2) = 1$ for all θ. Thus $S_1 = S_2$ will probability 1 for all θ.

2. Proof of 2. S^* is unbiased and a function of T. Suppose W is any unbiased estimator for $\tau(\theta)$. Then R-B Theorem says $W^* = E(W \mid T)$ is an unbiased estimator of $\tau(\theta)$ and $\mathrm{Var}_\theta(W^*) \leq \mathrm{Var}_\theta(W)$ for all θ. But W^* is unbiased and a function of T, so 1. implies $W^* = S^*$. Hence $\mathrm{Var}_\theta S^* \leq \mathrm{Var}_\theta W$ for all θ, and S^* is the UMVUE.

3. Proof of 3. Suppose $E_\theta S(X) = \tau(\theta)$ for all θ and $S(X) = \psi(T(X))$. Then $S^* = E(S \mid T)$ is the UMVUE by 2. But S is a function of T so that $E(S \mid T) = S$. Thus S is the UMVUE.

Example: Observe $X_1, \ldots, X_n$ iid Bernoulli(p).

- Find the UMVUE of p: $T = \sum_i X_i$ is a CSS. $E(T/n) = p$. Since T/n is an unbiased estimator of p which is a function of the CSS T, it is the UMVUE.

- Find the best unbiased estimator of p^2. Since, from R-B Theorem,

$$E\left(\frac{T(T-1)}{n(n-1)}\right) = p^2,$$

it is an unbiased estimator of p^2 which is a function of the CSS T. Hence it is the UMVUE. Once can check unbiasedness directly:

$$\begin{aligned} ET(T-1) = E(T^2) - ET &= \mathrm{Var}(T) + (ET)^2 - ET \\ &= np(1-p) + (np)^2 - np = n(n-1)p^2 \end{aligned}$$

Comment: "Estimate a parameter by its UMVUE" is another approach to estimation, but not a very good one. Often, no unbiased estimator exists, or the only one that exists is bad.

Example: Observe $X_1, \ldots, X_n$ iid N(μ, σ^2) $\theta = (\mu, \sigma^2)$ unknown. Here $T = (\bar{x}, s^2)$ is a CSS. (Recall the derivation: T is a 1-1 function of the natural SS for a 2pef.)

- **Estimation of $\tau(\mu, \sigma^2) = \mu$:** $\bar{x}$ is unbiased ($E\bar{x} = \mu$) and a function of $T \implies \bar{x}$ is UMVUE. MLE of θ is $\hat{\theta} = (\bar{x}, n^{-1}\sum_i (X_i - \bar{x})^2)$. So invariance principle says MLE

of μ is $\tau(\hat{\theta}) = \bar{x}$. MOM estimate is also $\bar{x}$ since $E\bar{x} = \mu$.
Note: For estimating μ, the MLE, MOM, UMVUE all agree on $\bar{X}$. But Bayes estimate is different. What about the sample median M? M is an unbiased estimator of μ. (Proof?) But it is not a function of the CSS T. Thus Rao-Blackwelizing M leads to the UMVUE (which we know is $\bar{x}$) which has a strictly smaller variance than M. Thus: $E(M \mid T) = \bar{x}$ and $\text{Var}(M) > \text{Var}(\bar{x}) = \sigma^2/n$.

- **Estimation of $\tau(\mu, \sigma^2) = \sigma^2$:** Let $SS = \sum_i (X_i - \bar{x})^2$. $s^2 = SS/(n-1)$ is an unbiased estimator σ^2 and a function of the CSS T. Therefore s^2 is the UMVUE. By the invariance principle, the MLE of σ^2 is SS/n. This is slightly biased.

- **Estimation of $\tau(\mu, \sigma^2) = \mu^2$:** The MLE of μ^2 is $(\bar{x})^2$ by invariance of MLEs. $\bar{x}^2$ is biased for μ^2:

$$E(\bar{x}^2) = \text{Var}(\bar{x}) + (E\bar{x})^2 = \frac{\sigma^2}{n} + \mu^2 > \mu^2.$$

An unbiased estimate of μ^2 is $W \equiv \bar{x}^2 - s^2/n$: since

$$E\left(\bar{x}^2 - \frac{s^2}{n}\right) = \left(\frac{\sigma^2}{n} + \mu^2\right) - \frac{\sigma^2}{n} = \mu^2.$$

Substracting s^2/n removes (or corrects for) the bias in the MLE. W is the UMVUE since it is unbiased and a function of T. Which is better: $\bar{x}^2$ or W? For $n > 3$, W has slightly smaller MSE than $\bar{x}^2$. (Verify?). Thus $\bar{x}^2$ is inadmissible for $n > 3$ (but it is a perfectly reasonable estimator). But W is also inadmissible because it sometimes takes on "impossible" values. $\mu^2 \geq 0$, but W can be negative ! $P(W < 0)$ is positive and will be sizeable when μ is small ($\approx 1/2$ when $\mu = 0$). A better estimate is clearly $W_+ = \max\{W, 0\}$. Whenever $W_+ \neq W$, we know W_+ is closer to the true value of μ^2. More formally

$$\begin{aligned} E(W-\mu^2)^2 - E(W_+ - \mu^2)^2 &= E[(W-\mu^2)^2 - (W_+ - \mu^2)^2] = \\ &= E[\underbrace{\{(W-\mu^2)^2 - (0-\mu^2)^2\} I(W<0)}_{\text{always} \geq 0 \text{ and sometimes} > 0}] > 0 \end{aligned}$$

But W_+ is biased ! No unbiased estimator of μ^2 exists which does not take on negative values.

Fact: There are situations where there are no unbiased estimators (and hence, no UMVUE exists).
Example: Observe $X_1, X_2, \ldots, X_n$ iid Poisson(λ). There exists no unbiased estimator of $1/\lambda$.
We know that $T = \sum_{i=1}^n X_i$ is a CSS. Suppose $\exists S = S(\mathbf{X})$ with $ES = 1/\lambda$ for all $\lambda >$

0. Then $\psi(T) = E(S \mid T)$ also satisfies $E\psi(T) = 1/\lambda$ for all $\lambda > 0$. This implies $E[\lambda\psi(T) - 1] = 0$ so that (multiplying by n) $E[n\lambda\psi(T) - n] = 0$. Now apply the following Lemma to $T \sim \text{Poisson}(n\lambda)$.

Lemma 1. *(See Theorem 3.6.8(a) on page 126). If $Y \sim$ Poisson(λ) then $E[\lambda g(Y)] = E[Yg(Y-1)]$.*

Thus $E[n\lambda\psi(T) - n] \equiv 0$ so that $P\{T\psi(T-1) - n = 0\} \equiv 1$ by completeness of T. Thus $T\psi(T-1) = n \implies \psi(T) = n/(T+1)$. But $E\left(\frac{n}{T+1}\right) = \frac{1}{\lambda}(1 - e^{-n\lambda}) \neq 1/\lambda$. Contradiction ! The R-B Theorem plus completeness sometimes gives easy proofs of otherwise difficult facts.

Example: Suppose $X_1, \ldots, X_n$ iid Poisson(λ). Let $\bar{x}$ and s^2 be the sample mean and variance.

Note: $\bar{x}$ is a 1-1 function of the CSS $T = \sum_i X_i$ and is therefore also a CSS.

Note: $E\bar{x} = EX_i = \lambda$ and $Es^2 = Var(X_i) = \lambda$.

Since $\bar{x}$ is unbiased for λ and a function of the CSS $\bar{x}$, we know it is best unbiased. But s^2 is also unbiased for λ. Since $\bar{x}$ is a CSS, $E(s^2 \mid \bar{x})$ must be the best unbiased estimator which is $\bar{x}$. We conclude that $E(s^2 \mid \bar{x}) = \bar{x}$.