Statistical Inference

Lecture 02b

ANU - RSFAS

Last Updated: Fri Feb 23 17:58:48 2018

General Set-Up

- Suppose we have a data $x_1, \ldots, x_n$ which we model as coming from a particular probability distirbution with density $f(x; \theta)$.
- I determine (through some means) a statistic to learn about θ.

$$\hat{\theta} = T(x_1, \ldots, x_n)$$

- Since $x_1, \ldots, x_n$ are considered a draw (draw 1) from $f(\boldsymbol{x}; \theta)$, we could get another draw (draw 2) and calculate $T(\boldsymbol{x})$. Continue in this manner we have:

$$\hat{\theta}_{\text{draw 1}}, \hat{\theta}_{\text{draw 2}}, \ldots, \hat{\theta}_{\text{draw } k}$$

- So $\hat{\theta}$ is random because we model the data as being random $X_1, \ldots, X_n$.

Evaluating Estimators

- Generally, we may have many types of estimators $T_1(\cdot)$, $T_2(\cdot)$, etc.
- We need approaches to compare and contrast estimators!

Definition 2.1: $\hat{\theta} = T(X_1, \ldots, X_n)$ is an unbiased estimator for θ if $E[T(\boldsymbol{X})] = \theta$. The bias of an estimator is defined as:

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

- **Eg.:** For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$ where $E[X] = \mu,\ V(X) = \sigma^2$.

$$E[\bar{X}] = \mu, \quad E[S^2] = \sigma^2$$

- $\text{Bias}(\bar{X}) = E[\bar{X}] - \mu = 0$. We say that the estimator is unbiased.

- Unbiasedness seems like a good idea . . . but don't we many times want to minimimize the squared difference?

Definition 2.1: $\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$

$$
\begin{aligned}
E\left[(\hat{\theta} - \theta)^2\right] &= E\left[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2\right] \\
&= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2] \\
&= E[(\hat{\theta} - E(\hat{\theta}))^2] + 2(E(\hat{\theta}) - \theta)E[(\hat{\theta} - E(\hat{\theta}))] + E[(E(\hat{\theta}) - \theta)^2] \\
&= E[(\hat{\theta} - E(\hat{\theta}))^2] + 0 + E[(E(\hat{\theta}) - \theta)^2] \\
&= E[(\hat{\theta} - E(\hat{\theta}))^2] + (E(\hat{\theta}) - \theta)^2 \\
&= V(\hat{\theta}) + \text{Bias}(\hat{\theta})^2
\end{aligned}
$$

Evaluating Estimators

Eg.: Let's consider $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}\,(\mu, \sigma^2)$.

- We know the $E[S^2] = \sigma^2$. So it is unbiased.
- I actually think I like the estimator:

$$\frac{1}{n+1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

What do you think?

Evaluating Estimators

Eg.: Let's reconsider our question for $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}\,(\mu, \sigma^2)$.

- Also consider the following estimator (MLE) for σ^2:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=i}^{n} (X_i - \bar{X})^2$$

- The unbiased estimator is $S^2 = \frac{1}{n-1} \sum_{i=i}^{n} (X_i - \bar{X})^2$

$$\text{MSE}(S^2) = \text{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

$$\text{MSE}(\hat{\sigma}^2) = \frac{(2n-1)\sigma^4}{n^2}$$

$$\text{MSE}(S^2) > \text{MSE}(\hat{\sigma}^2)$$

Evaluating Estimators

- Let's now consider $\tilde{\sigma}^2 = \frac{1}{n+1}\sum_{i=1}^{n}(x_i - \bar{x})^2$:
- First let's derive the variance of S^2:

$$
\begin{aligned}
(n-1)S^2/\sigma^2 \sim \chi^2_{n-1} \quad \Rightarrow \quad & V\left((n-1)S^2/\sigma^2\right) = 2(n-1) \\
& \left[(n-1)/\sigma^2\right]^2 V\left(S^2\right) = 2(n-1) \\
= \quad & V\left(S^2\right) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{(n-1)}
\end{aligned}
$$

$$
\begin{aligned}
\tilde{\sigma}^2 &= \frac{1}{n+1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{n-1}{n+1}S^2
\end{aligned}
$$

Evaluating Estimators

$$E[\tilde{\sigma}^2] = E\left[\frac{n-1}{n+1}S^2\right] = \frac{n-1}{n+1}E\left[S^2\right] = \frac{n-1}{n+1}\sigma^2$$

$$V[\tilde{\sigma}^2] = V\left[\frac{n-1}{n+1}S^2\right] = \left[\frac{n-1}{n+1}\right]^2 V\left[S^2\right] = \left[\frac{n-1}{n+1}\right]^2 \frac{2\sigma^4}{(n-1)}$$

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

Evaluating Estimators

- Now let's compare to the MLE:

$$\begin{aligned}
\frac{2\sigma^4}{n+1} &\overset{?}{<} \frac{(2n-1)\sigma^4}{n^2} \\
\frac{2}{n+1} - \frac{(2n-1)\sigma^4}{n^2} &\overset{?}{<} 0 \\
\frac{(1-n)\sigma^4}{n^2(n+1)} &< 0 \quad \text{for n } \geq 1
\end{aligned}$$

$$MSE(\tilde{\sigma}^2) < MSE(\hat{\sigma}^2) < MSE(S^2)$$

Evaluating Estimators

Eg.: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$.

- The MLE is $\hat{p} = \bar{X}$. The bias: $Bias(\hat{p}) = E[\hat{p}] - p = p - p = 0$.
- The variance of $\hat{p} = V(\bar{X}) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$.

$$MSE(\hat{p}) = \frac{p(1-p)}{n} + 0^2 = \frac{p(1-p)}{n}$$

Evaluating Estimators

- Let's consider a Bayesian estimator (we will show how to derive this later):

$$\hat{p}_B = \frac{y + a}{a + b + n}$$

- Now compare the MSE of the Bayesian and ML estimators.

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

Evaluating Estimators

$$MSE[\hat{p}_B] = MSE\left[\frac{Y+a}{a+b+n}\right] = \left[\frac{np(1-p)}{(a+b+n)^2}\right] + \left[\frac{np+a}{a+b+n} - p\right]^2$$

- To compare with the MLE we need specific values for a and b. Let's pick these to make the $MSE[\hat{p}_B]$ constant for p. You of course can try other values!
- It turns out that if we set $a = b = \sqrt{n/4}$ then we get (which is constant for p):

$$\hat{p}_B = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

$$MSE[\hat{p}_B] = \frac{1}{(4(1+\sqrt{n})^2)}$$

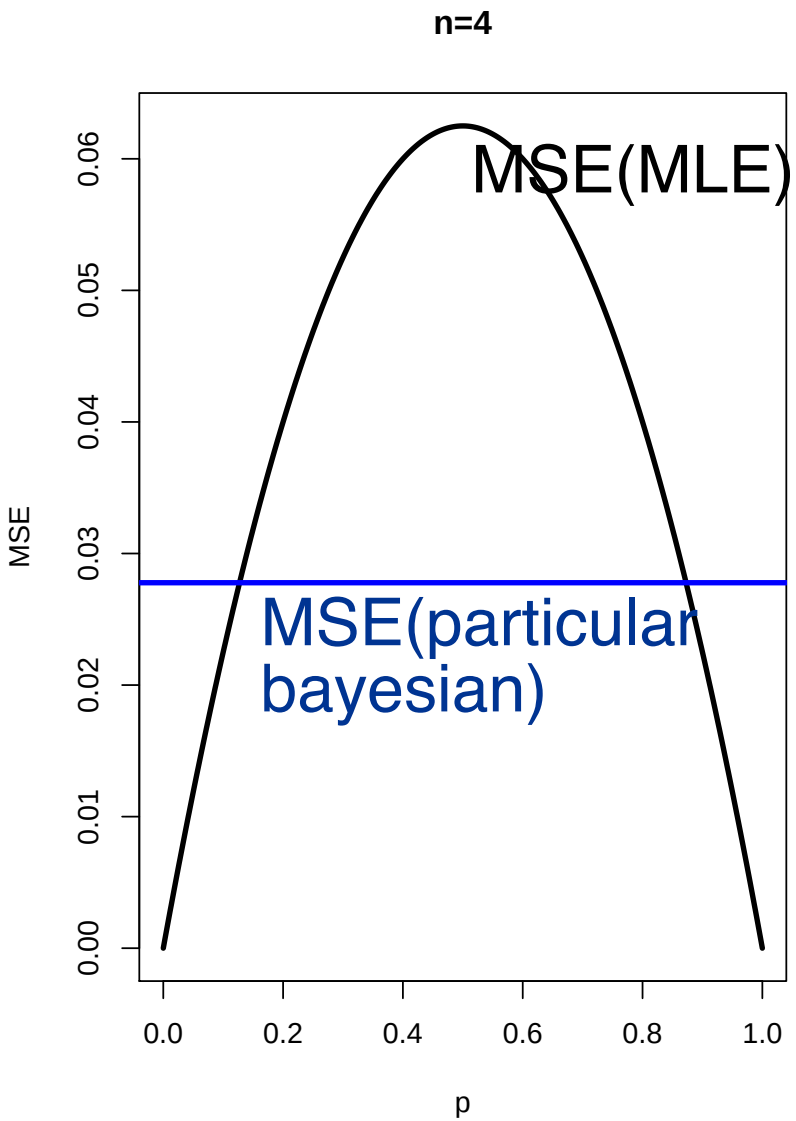
Evaluating Estimators

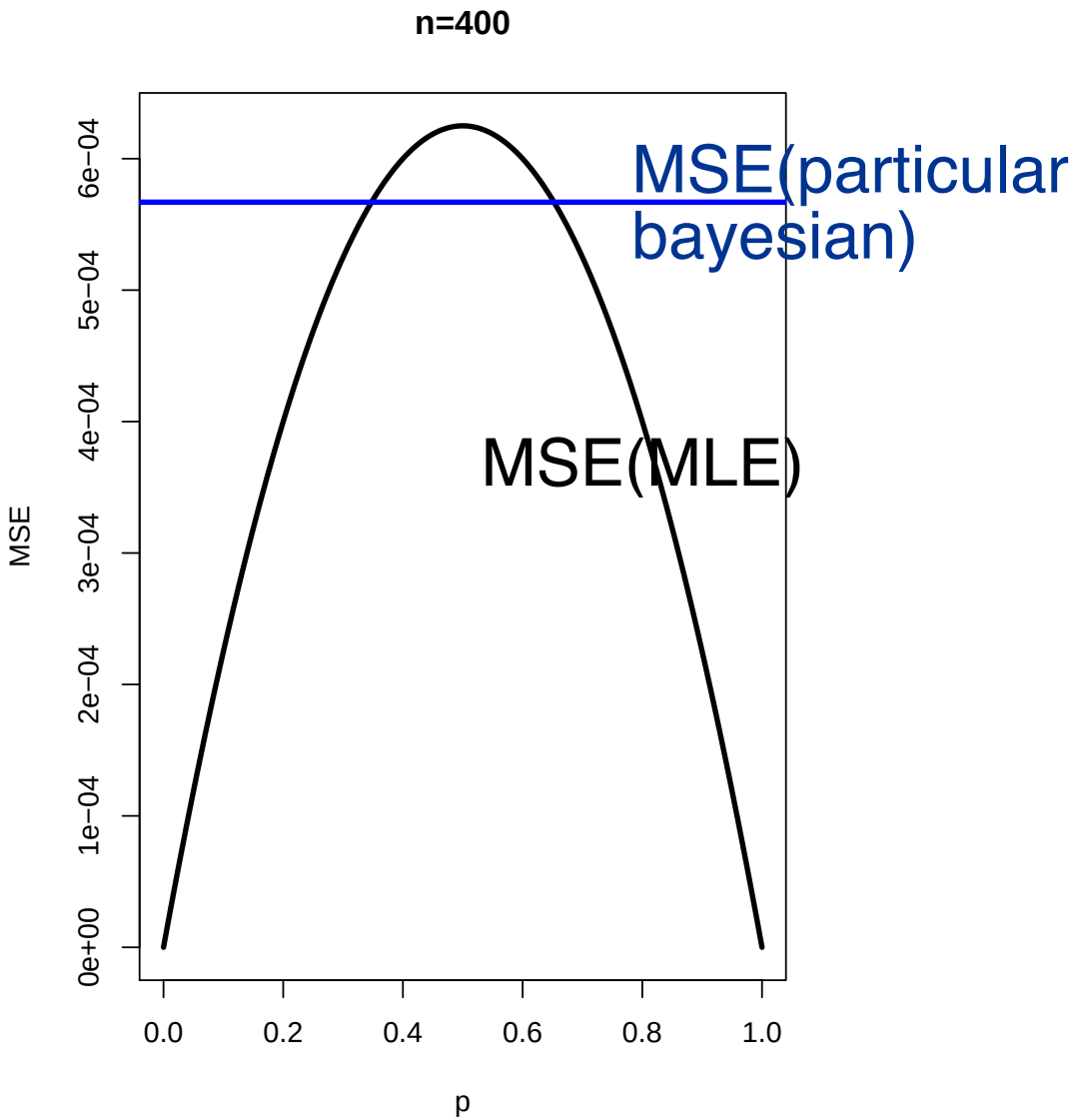
```
p <- seq(0,1, by=0.001)

par(mfrow=c(1,2))
n <- 4
MSE.mle <- p*(1-p)/n
plot(p, MSE.mle, type="l", lwd=3, main="n=4", ylab="MSE")
abline(h = 1/(4*(sqrt(n) + 1)^2), lwd=3, col="blue")

n <- 400
MSE.mle <- p*(1-p)/n
plot(p, MSE.mle, type="l", lwd=3, main="n=400", ylab="MSE")
abline(h = 1/(4*(sqrt(n) + 1)^2), lwd=3, col="blue")
```

Evaluating Estimators





Evaluating Estimators

- For small n the Bayesian estimator is the better choice, unless we believe p is near 0 or 1.
- For large n, then the MLE is the better choice, unless we believe p is near $1/2$.

$\tilde{P}_{con} = 0.5$ for all x

constant estimator.

$$MSE(\tilde{P}_{con}) = Var(\tilde{P}_{con}) + Bias(\tilde{P}_{con})^2$$

$$= 0 + (E(\tilde{P}_{con}) - P)^2$$

$$= (0.5 - P)^2$$

Evaluating Estimators

Definition 2.1: An estimator $\hat{\theta}$ is **weakly consistent** if

$$P(|\hat{\theta} - \theta| > \epsilon) \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty$$

for any $\epsilon > 0$.

Proof: Use Chebyshev's inequality:

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

sufficient but not necessary.

- Thus $V(\hat{\theta}) \rightarrow 0$ and $bias(\hat{\theta}) \rightarrow 0$ implies that $\hat{\theta}$ is consistent.

Evaluating Estimators

- Is the following estimator consistent?

$$\hat{p}_B = \frac{y + a}{a + b + n}$$

$$\begin{aligned} V\left[\frac{Y + a}{a + b + n}\right] &= \left[\frac{1}{a + b + n}\right]^2 np(1 - p) \\ &= \left[\frac{n^{1/2}}{a + b + n}\right]^2 p(1 - p) \to 0 \text{ as } n \to \infty \end{aligned}$$

$$\begin{aligned} V\left[\frac{Y + a}{a + b + n}\right] &= \left[\frac{1}{a + b + n}\right]^2 np(1 - p) \\ &= \left[\frac{n^{1/2}}{a + b + n}\right]^2 p(1 - p) \to 0 \text{ as } n \to \infty \end{aligned}$$

$$
\begin{aligned}
bias(\hat{p}) &= \frac{np + a}{a + b + n} - p \\
&= \frac{np}{a + b + n} - p + \frac{a}{a + b + n} \\
&= p - p = 0 \text{ as } n \to \infty
\end{aligned}
$$

- There we have a consistent estimator.

Evaluating Estimators

Eg.: An estimator T^* is a best unbiased estimator of $\tau(\theta)$ [notice I am being a bit more general here] if it satisfies $E[T^*] = \tau(\theta)$ for all θ and, for any other estimator T with $E[T] = \tau(\theta)$ we have

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

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

- Finding a best unbiased estimator can be difficult.

Evaluating Estimators

Eg.: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. $E(X) = \lambda = V(X)$

- $E[\bar{X}] = \lambda$ sample mean
- $E[S^2] = \lambda$ sample variance
- Through some work we can show:

$$V(\bar{X}) < V(S^2)$$

- But we could also consider (infinite number of choices for $a \in [0, 1]$):

linear combination also

$$T(\bar{X}, S^2) = a\bar{X} + (1-a)S^2$$

⇒ unbiased estimator!

- Are there other unbiased estimators? yes, maybe.
- We would like to know if we have unbaised estimator whether its variance is the smallest possible! Thus it is the most efficient **Definition 2.4**.

Cramer-Rao Inequality

Section 2.4 (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 θ is a scalar parameter. Also, let $T = t(X_1, \ldots, X_n)$ be an unbiased estimator for $\tau(\theta)$. Then under certain regularity (smoothness) conditions:

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

- $\tau'(\theta) = \frac{d}{d\theta}\tau(\theta)$
- Where $I(\theta) = ni(\theta)$ is called the Expected Fisher Information.
- $I(\theta) = E\left[\left(\frac{\partial l(\theta)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 l(\theta)}{\partial \theta^2}\right]$

Cramer-Rao Inequality

C-R Inequality Extended: Let $X_1, \ldots, X_n$ be a sample [note we don't have to have iid] with pdf $f(\boldsymbol{x}|\theta)$ and let $T(\boldsymbol{X})$ be an estimator [doesn't have to be unbiased] then based on regularity conditions we have:

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

Cramer-Rao Inequality [lower bound]

- If $E[T(\boldsymbol{X})] = \tau(\theta)$, so $T(\boldsymbol{X})$ is an unbiased estimator for $\tau(\theta)$,

$$V[T(\boldsymbol{X})] \geq \frac{[\tau'(\theta)]^2}{I(\theta)}$$

- If we have iid samples:

$$V[T(\boldsymbol{X})] \geq \frac{[\tau'(\theta)]^2}{ni(\theta)}$$

Cramer-Rao Inequality [lower bound]

Eg: Poisson distribution: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$.

$$f(x|\lambda) = \frac{\lambda^x exp(-\lambda)}{x!}$$

- $\tau(\lambda) = \lambda \Rightarrow \frac{d}{d\theta}\tau(\lambda) = 1$.

Cramer-Rao Inequality [lower bound]

Eg: Poisson distribution.

$$
\begin{aligned}
i(\lambda) &= E\left[\left(\frac{d}{d\lambda} ln\{f(x|\lambda)\}\right)^2\right] \\
&= E\left[\left(\frac{d}{d\lambda}\{x ln(\lambda) - \lambda - x!\ \}\right)^2\right] \\
&= E\left[\left(\frac{X}{\lambda} - 1\right)^2\right] \\
&= \frac{1}{\lambda^2} E\left[(X-\lambda)^2\right] \\
&= \frac{1}{\lambda^2} Var(X) \\
&= \frac{1}{\lambda^2}\lambda = \frac{1}{\lambda}
\end{aligned}
$$

Cramer-Rao Inequality [lower bound]

$$V_\lambda(T) \geq \frac{1}{n\lambda^{-1}} = \frac{\lambda}{n}$$

We will show that a sufficient statistic for λ to be $\sum_{i=1}^{n} X_i$ so $\bar{X}$ is also a sufficient statistic.

$$V(\bar{X}) = \frac{\lambda}{n}$$

$V\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n^2} V(x_1 + \cdots + x_n)$

$= \frac{1}{n^2}(Vx_1 + Vx_2 + \ldots)$

$= \frac{n\lambda}{n^2} = \frac{\lambda}{n}$

We see that the lower bound is achieved by $\bar{X}$ thus it is the Minimum Variance Unbiased Estimator (MVUE) of λ.

Cramer-Rao Inequaility - Regularity Conditions

- $\frac{\partial}{\partial\theta} ln\{f(x|\theta)\}$ exists for all x and θ;
- interchange of integration and differentiation is permissible;
- The expectation $i(\theta) = E\left[\left[\frac{\partial}{\partial\theta} ln\{f(x|\theta)\}\right]^2\right]$, where X is a generic random variable having distribution with density $f(x|\theta)$, is finite for all $\theta \in \Theta$.

Cramer-Rao Inequaility - Proof

- The proof is based on the Cauchy-Schwarz Inequality: For any two random variables Y and Z:

$$[Cov(Y,Z)]^2 \leq V(Y)V(Z) \;\Rightarrow\; V(Y) \geq \frac{[Cov(Y,Z)]^2}{V(Z)}$$

- We choose Y in this equation to be $T(\boldsymbol{X})$ and Z to be $\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)$.
- Recall:

$$V(Z) = E[Z^2] - (E[Z])^2$$

$$Cov(Y,Z) = E[YZ] - E[Y]\ E[Z]$$

Cramer-Rao Inequaility - Proof

- First note:

$$
\begin{aligned}
\frac{\partial}{\partial\theta} E[T(\boldsymbol{X})] &= \frac{\partial}{\partial\theta} \int_{\mathcal{X}} T(\boldsymbol{x}) f(\boldsymbol{x}|\theta) dx \\
&= \int_{\mathcal{X}} T(\boldsymbol{x}) \left[\frac{\partial}{\partial\theta} f(\boldsymbol{x}|\theta) \right] dx \\
&= \int_{\mathcal{X}} T(\boldsymbol{x}) \left[\frac{\partial}{\partial\theta} f(\boldsymbol{x}|\theta) \frac{f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta)} \right] dx \\
&= E\left[T(\boldsymbol{X}) \frac{\frac{\partial}{\partial\theta} f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta)} \right] \\
&= E\left[T(\boldsymbol{X}) \frac{\partial}{\partial\theta} \log f(\boldsymbol{x}|\theta) \right]
\end{aligned}
$$

Likelihood?

$x_1, \ldots, x_n$ $\overset{iid}{\sim}$ $f_x(x; \theta)$

- Joint den.

$$P(X = x_1, \ldots, X = x_n)$$

$$= \prod_{i=1}^{n} P(X_i = x_i; \theta)$$

- likelihood.

$$L(\theta; \vec{x}) = \prod_{i=1}^{n} P(X_i = x_i; \theta)$$

Cramer-Rao Inequaility - Proof

- Second note:

$$
\begin{aligned}
E\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right] &= \int_{\mathcal{X}}\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right] f(\boldsymbol{x}|\theta)dx \\
&= \int_{\mathcal{X}}\left[\frac{\frac{\partial}{\partial\theta}f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta)}\right] f(\boldsymbol{x}|\theta)dx \\
&= \int_{\mathcal{X}}\frac{\partial}{\partial\theta}f(\boldsymbol{x}|\theta)dx \\
&= \frac{\partial}{\partial\theta}\int_{\mathcal{X}} f(\boldsymbol{x}|\theta)dx \\
&= \frac{\partial}{\partial\theta}(1) = 0
\end{aligned}
$$

Cramer-Rao Inequaility - Proof

$$
\begin{aligned}
Cov\left[T(\boldsymbol{X}),\ \frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right] &= E\left[T(\boldsymbol{X})\ \frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right] \\
&\quad -E\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right] E\left[T(\boldsymbol{X})\right] \\
&= E\left[T(\boldsymbol{X})\ \frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right] - [0]E\left[T(\boldsymbol{X})\right] \\
&= E\left[T(\boldsymbol{X})\ \frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right] \\
&= \frac{\partial}{\partial\theta}E[T(\boldsymbol{X})]
\end{aligned}
$$

Cramer-Rao Inequaility - Proof

$$
\begin{aligned}
V\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right] &= E\left[\left(\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right)^2\right] - \left(E\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right]\right)^2 \\
&= E\left[\left(\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right)^2\right] - (0)^2 \\
&= E\left[\left(\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right)^2\right]
\end{aligned}
$$

Cramer-Rao Inequaility - Proof

$$
\begin{aligned}
V(Y) &\geq \frac{[Cov(Y,Z)]^2}{V(Z)} \\
V[T(\boldsymbol{X})] &\geq \frac{\left[Cov\left[T(\boldsymbol{X}),\ \frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right]\right]^2}{V\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right]} \\
&\geq \frac{\left[\frac{\partial}{\partial\theta}E[T(\boldsymbol{X})]\right]^2}{E\left[\left(\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right)^2\right]}
\end{aligned}
$$

Cramer-Rao Inequaility

- So we have: UMVUE

$$V\left[T(\boldsymbol{X})\right] \geq \frac{\left[\frac{\partial}{\partial\theta}E[T(\boldsymbol{X})]\right]^2}{E\left[\left(\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right)^2\right]}$$

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

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

↓ a single data pt.

Cramer-Rao Inequaility - Proof

- We need to show

joint distribution

single data point

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

Cramer-Rao Inequaility - Proof

expand this

$$
\begin{aligned}
E\left[\left(\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right)^2\right] &= E\left[\left(\frac{\partial}{\partial\theta}\ log \prod_{i=1}^{n} f(x_i|\theta)\right)^2\right] \\
&= E\left[\left(\frac{\partial}{\partial\theta}\ \sum_{i=1}^{n} log\ f(x_i|\theta)\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} \frac{\partial}{\partial\theta}\ log\ f(x_i|\theta)\right)^2\right] \\
&= \sum_{i=1}^{n} E\left[\left(\frac{\partial}{\partial\theta}\ log\ f(x_i|\theta)\right)^2\right] \quad \text{individual terms} \\
&\quad + \sum_{i\neq j} E\left(\frac{\partial}{\partial\theta}\ log\ f(x_i|\theta)\ \frac{\partial}{\partial\theta}\ log\ f(x_j|\theta)\right)
\end{aligned}
$$

& cross terms

Cramer-Rao Inequaility - Proof

$$
\begin{aligned}
&= \sum_{i=1}^{n} E\left[\left(\frac{\partial}{\partial\theta}\, log\ f(x_i|\theta)\right)^2\right] \\
&\quad + \sum_{i\neq j} E\left(\frac{\partial}{\partial\theta}\, log\ f(x_i|\theta)\ \frac{\partial}{\partial\theta}\, log\ f(x_j|\theta)\right) \\
&= \sum_{i=1}^{n} E\left[\left(\frac{\partial}{\partial\theta}\, log\ f(x_i|\theta)\right)^2\right] \qquad \text{independent.} \\
&\quad + \sum_{i\neq j} E\left(\frac{\partial}{\partial\theta}\, log\ f(x_i|\theta)\right)\ E\left(\frac{\partial}{\partial\theta}\, log\ f(x_j|\theta)\right) \\
&= \sum_{i=1}^{n} E\left[\left(\frac{\partial}{\partial\theta}\, log\ f(x_i|\theta)\right)^2\right] + [0][0] \\
&= nE\left[\left(\frac{\partial}{\partial\theta}\, log\ f(x|\theta)\right)^2\right] = ni(\theta)
\end{aligned}
$$

Fisher Information

- The Fisher information, or expected Fisher information, or the information number is:

most of time, ↓ is easier to work with

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

- For one data point we have:

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

- For iid data:

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

Fisher Information - Proof

Proof of expression of Fisher Information.

- First note:

$$
\begin{aligned}
\frac{\partial^2}{\partial\theta^2} \log f(\boldsymbol{x}|\theta) &= \frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta} \log f(\boldsymbol{x}|\theta) \\
&= \frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta} f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta)} \\
&= \frac{\left[\frac{\partial^2}{\partial\theta^2} f(\boldsymbol{x}|\theta)\right] f(\boldsymbol{x}|\theta) - \left[\frac{\partial}{\partial\theta} f(\boldsymbol{x}|\theta)\right]\left[\frac{\partial}{\partial\theta} f(\boldsymbol{x}|\theta)\right]}{[f(\boldsymbol{x}|\theta)]^2} \\
&= \frac{\frac{\partial^2}{\partial\theta^2} f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta)} - \left[\frac{\frac{\partial}{\partial\theta} f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta)}\right]^2 \\
&= \frac{\frac{\partial^2}{\partial\theta^2} f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta)} - \left[\frac{\partial}{\partial\theta} \log f(\boldsymbol{x}|\theta)\right]^2
\end{aligned}
$$

Fisher Information - Proof

- Now let's take expectations:

$$
\begin{aligned}
E\left[\frac{\partial^2}{\partial\theta^2}\ log\ f(\boldsymbol{x}|\theta)\right] &= E\left[\frac{\frac{\partial^2}{\partial\theta^2}f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta)}\right] - E\left[\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right]^2\right] \\
&= -E\left[\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right]^2\right] + \int_{\mathcal{X}} \frac{\frac{\partial^2}{\partial\theta^2}f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta)} f(\boldsymbol{x}|\theta)dx \\
&= -E\left[\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right]^2\right] + \int_{\mathcal{X}} \frac{\partial^2}{\partial\theta^2} f(\boldsymbol{x}|\theta)dx \\
&= -E\left[\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right]^2\right] + \frac{\partial^2}{\partial\theta^2}\int_{\mathcal{X}} f(\boldsymbol{x}|\theta)dx \\
&= -E\left[\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right]^2\right] + \frac{\partial^2}{\partial\theta^2}[1] \\
&= -E\left[\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right]^2\right] + 0
\end{aligned}
$$

Fisher Information - Proof

- So we have:

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

- Note: We already showed $E\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right] = 0$. We also showed the following (but just to make it clear)

$$V\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right] = E\left[\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right]^2\right] - \left(E\left[\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x}|\theta)\right]\right)^2$$

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