

Contents

1	Point Estimation	2
1.1	Method of Moments	2
2	Maximum Likelihood	4
3	Decision Theory	6
3.1	Bayesian Estimation	6
3.2	Conjugate Priors	7
3.3	Bayes Rules	7
4	Jan 29	8
5	Best Unbiased Estimation	9
5.1	Cramer-Rao Lower Bound	10

1 Point Estimation

Definition: A point estimator is any scalar (or vector) -valued function of the sample. $(x_1, \dots, x_n) \sim f(x|\theta)$

A point estimator for $\tau(\theta)$ is a statistic $T(x)$ with the purpose of approximating $\tau(\theta)$

1.1 Method of Moments

The k-th moment of a r.v. X is $\mu_k(\theta) = \mathbb{E}_\theta(X^k) = \int_{\mathbb{X}} x^k f(x|\theta) dx$

Given an iid sample $x_1, \dots, x_n \stackrel{\text{iid}}{\sim} f(x|\theta)$ we have sample moments: $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n x_i^k$

Suppose $\theta \in \Theta \in \mathbb{R}^P$, and that $\mu_k(\theta)$ exists and is finite for $k = 1, \dots, p$.

Definition: The method of moments estimator of θ is the solution to the system of equations:

$$\mu_1(\theta) = \hat{\mu}_1 \quad (1)$$

$$\cdot \quad (2)$$

$$\cdot \quad (3)$$

$$\mu_p(\theta) = \hat{\mu}_p \quad (4)$$

We call it $\hat{\theta}_{MM}$

Example: $X_i \stackrel{\text{iid}}{\sim} \text{Beta}(\alpha, \beta), \theta = (\alpha, \beta)$

$$\mu_1(\theta) = \frac{\alpha}{\alpha + \beta}$$

$$\mu_2(\theta) = \text{Var}_\theta(x_1) + \left(\frac{\alpha}{\alpha + \beta}\right)^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} + \frac{\alpha^2}{(\alpha + \beta)^2}$$

$$\frac{\alpha}{\alpha + \beta} = \hat{\mu}_1, \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} + \frac{\alpha^2}{(\alpha + \beta)^2} = \hat{\mu}_2$$

$\beta = \left(\frac{1 - \hat{\mu}_1}{\hat{\mu}_1}\right) \alpha$ from the first equation. Plug this into the second expression and solve:

$$\hat{\alpha}_{MM} = \hat{\mu}_1 \left[\frac{\hat{\mu}_1(1 - \hat{\mu}_1)}{\hat{\mu}_2 - \hat{\mu}_1^2} - 1 \right]$$

$$\implies \hat{\beta}_{MM} = \frac{(1 - \hat{\mu}_1)}{\hat{\mu}_1} \hat{\alpha}_{MM}$$

Example: $x_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$

$$\mu_1(\theta) = \hat{\mu}_1, \mu_1(\theta) = \mu, \implies \hat{\mu}_{MM} = \mu$$

$$\mu_2(\theta) = \text{Var}_\theta(x_1) + \mu_1^2 = \sigma^2 + \mu^2 = \hat{\mu}_2$$

$$\sigma_{MM}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{n-1}{n} s^2$$

Note: If μ, σ^2 are the mean/variance of any family, their MM estimators are $\overline{X}, \frac{n-1}{n} s^2$.

For families where parameters are not just the mean and variance, you can find it via two ways: use mean/variance in MM, then calculate parameters, or use parameters in MM then calculate mean/variance. Both yield same results.

Fact: MM estimators are invariant of re-parameterizations.

Let $\eta = \eta(\theta)$ be a 1:1 mapping (invertible). Then, $\hat{\eta}_{MM} = \eta(\hat{\theta}_{MM})$

Let's say $x_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ and we want to estimate $\tau(\theta) = \frac{\mu}{\sigma}$. This isn't 1:1. What can we do? We can do the 'Transformations' method and create a second value τ_2 , etc. We can also just plug in the estimators.

Definition: The MM estimator for a parametric function $\tau(\theta)$ is just $\hat{\tau}_{MM}(\theta) = \tau(\hat{\theta}_{MM})$

Properties:

- MM equations may have a unique solution, no solution, or many solutions
- Often, MM estimators are used as initial values for another estimation technique (ie. a root finding method)
- Why should it work? Let's say θ^* is the true value of θ . Then Law of Large Numbers says: $\hat{\mu}_k \xrightarrow{P} \mu_k(\theta^*)$. Then we are solving $\mu_k(\theta) = \hat{\mu}_k \approx \mu_k(\theta^*)$

Example: $X_i \stackrel{\text{iid}}{\sim} \text{Bin}(m, \theta), i = 1, \dots, n$ (m is known)

Find the MM estimator of $\tau(\theta) = \ln \frac{\theta}{1-\theta}$

1. Find MM for θ ($\hat{\theta}_{MM}$)
2. Plug in ($\hat{\tau}_{MM}(\theta) = \tau(\hat{\theta}_{MM})$)
3. $\hat{\tau}_{MM}(\theta) = \ln \frac{\bar{X}/m}{1-\bar{X}/m}$

2 Maximum Likelihood

$X = (X_1, \dots, X_n) \sim f(x|\theta), \theta \in \Theta \subset \mathbb{R}^k$

Notation: $f(x; \theta)$: function of x indexed at θ . Basically, given some set value of the θ .

Likelihood Function: The likelihood function is: $L(\theta; x) = f(x|\theta)$

Notes:

- L is a function of θ for each $x \in \mathbb{X}$ (in sample space)
- Plugging in X for x gives $L(\theta; X)$, a stochastic process (ie. plug in a random X makes this a random function for θ)
- The log-likelihood function is $l(\theta; x) = \ln[L(\theta; x)]$
- If $x_i \stackrel{\text{iid}}{\sim} f(x_i|\theta)$ (f is marginal dist.), then $l(\theta; x) = \sum_{i=1}^n \ln[f(x_i|\theta)]$ (because $x_i \stackrel{\text{iid}}{\sim}$, the sum is just a transformation and we can apply LLN, CLT, etc.)

Maximum Likelihood Estimate: If $x \in \mathbb{X}$ is observed, a maximum likelihood estimate of $\theta, \hat{\theta}(x)$, is any value $\theta \in \Theta$ that maximizes $L(\theta|x)$.

$$\hat{\theta}(x) = \underset{\theta \in \Theta}{\operatorname{argmax}} [L(\theta|x)]$$

This is a function of observed data (an estimate, not an estimator).

Maximum Likelihood Estimator: A maximum likelihood estimator (MLE) is $\hat{\theta} = \hat{\theta}(X)$

If an ML estimate exists, then $\hat{\theta}(x) = \underset{\theta \in \Theta}{\operatorname{argmax}} [l(\theta; x)]$. This is because $\ln(x)$ is a strictly increasing function.

Why does maximum likelihood work? Can we show that $\hat{\theta} \approx \theta_0$ (true parameter)?

Assume $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x_i|\theta)$. $l(\theta|x) = \sum_{i=1}^n \ln[f(x_i|\theta)]$

$$\frac{1}{n} l(\theta|x) = \frac{1}{n} \sum_{i=1}^n \ln[f(x_i|\theta)] \xrightarrow{P} \mathbb{E}_{\theta_0}(\ln[f(x|\theta)]) = \int_{\mathbb{X}} \ln[f(x|\theta)] f(x|\theta_0) dx$$

It would make sense that $\hat{\theta}(x) = \underset{\theta \in \Theta}{\operatorname{argmax}} [l(X|\theta)] \approx \underset{\theta \in \Theta}{\operatorname{argmax}} \mathbb{E}_{\theta_0}(\ln[f(X|\theta)])$ which we hope $= \theta_0$.

Define $D(\theta; \theta_0) = \mathbb{E}_{\theta_0}(\ln[f(X|\theta)])$. We will show that $D(\theta_0; \theta_0) - D(\theta; \theta_0) \geq 0 \forall \theta$.

Kullback-Liebler Divergence: Let f_0 and f_1 be any two PDFs/PMFs. The Kullback-Liebler divergence from f_0 to f_1 is $K(f_0, f_1) = -\mathbb{E}_{f_0}[\ln \frac{f_1(x)}{f_0(x)}]$

$$D(\theta_0; \theta_0) - D(\theta; \theta_0) = \mathbb{E}_{\theta_0}[\ln(f(X|\theta_0)) - \ln(f(X|\theta))] = -\mathbb{E}_{\theta_0}[\ln \frac{f(X|\theta)}{f(X|\theta_0)}].$$

Lemma: For any two PDFs/PMFs f_0, f_1 , $K(f_0, f_1) \geq 0$, with equality iff $f_0 \equiv f_1$.

Remember Jensen's Inequality: When $g(x)$ is convex (happy), $\mathbb{E}[g(x)] \geq g(\mathbb{E}[x])$.

Proof (Discrete Case): Suppose $X \sim f_0$ and set $Z = \frac{f_1(x)}{f_0(x)}$. Let $S_j = \{x : f_j(x) > 0\}$

Since $g(z) = -\ln(z)$ is convex and $\mathbb{E}_{f_0}(z) = \sum_{x \in S_0} \frac{f_1(x)}{f_0(x)} f_0(x) = \sum_{x \in S_0} f_1(x) \leq 1$.

By Jensen's Inequality: $K(f_0, f_1) = -\mathbb{E}_{f_0}[\ln(Z)] = \mathbb{E}_{f_0}[g(z)] \underset{1}{\geq} g(\mathbb{E}_{f_0}(z)) \underset{2}{\geq} 0$.

This is only 'equal' when g is linear. Since $g(z)$ is not linear, equality in 1 only happens iff $Z = \frac{f_1(x)}{f_0} = c \neq 0, \forall x \in S_0, [S_0 \subset S_1]$. Equality in 2 only happens iff $\sum_{x \in S_0} f_1(x) = 1, [S_1 \subset S_0]$.

Suppose 1 and 2 are equalities. $1 = \sum_{x \in S_1} f_1(x) = \sum_{x \in S_1} c f_0(x) = c \sum_{x \in S_1} f_0(x) = c \sum_{x \in S_0} f_0(x) = c$

3 Decision Theory

3.1 Bayesian Estimation

$$x_1, \dots, x_n \sim f(x|\theta), \theta \in \Theta$$

Definition: A priori distribution π for θ is a PDF/PMf over Θ : $\int_{\Theta} \pi(\theta) d\theta = 1, \pi(\theta) \geq 0, \theta \in \Theta$

Main Idea: π tells us what θ s are "important". $\pi(\theta|x)$ tells us which are important after knowing x .

Definition: If we observed $x \in X$, the posterior distribution is: $\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta) d\theta}$

Here, $m(x) = \int_{\Theta} f(\theta|x)\pi(\theta) d\theta$ is the "marginal" distribution of x .

Some logical things to do with $\pi(\theta|x)$:

1. Estimate θ using a measure of 'center':

$$2. \hat{\theta} = \mathbb{E}(\theta|x) = \int_{\Theta} \theta \pi(\theta|x) d\theta$$

$$3. 0.5 = \int_0^{\hat{\theta}} \pi(\theta|x) d\theta$$

$$4. \hat{\theta}(x) = \underset{\theta \in \Theta}{\operatorname{argmax}} [\pi(\theta|x)]$$

Example: $x_i \stackrel{\text{iid}}{\sim} \text{Bern}(\theta), \theta \sim \text{Beta}(\alpha, \beta)$ (the common prior for the Bernoulli is the Beta: also defined on 0:1)

$$\begin{aligned} \pi(\theta) &= \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}, 0 < \theta < 1 \\ &\propto \theta^{\alpha-1} (1-\theta)^{\beta-1} \end{aligned}$$

$$f(x|\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}$$

$$\pi(\theta|x) = \frac{\theta^{\sum x_i} (1-\theta)^{n-\sum x_i} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha, \beta) \int_0^1 \frac{\theta^{\sum x_i} (1-\theta)^{n-\sum x_i} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha, \beta)} d\theta}$$

Denominator is just a function of x (integrating out θ) and some constants.

$$\implies \pi(\theta|x) \propto \theta^{\alpha+\sum x_i-1} (1-\theta)^{\beta+n-\sum x_i-1}$$

$$\text{i.e., } \theta|x \sim \text{Beta}(\alpha + \sum x_i, \beta + n - \sum x_i)$$

$$\text{The posterior mean is } \hat{\theta}(x) = \frac{\alpha + \sum x_i}{\alpha + \beta + n}$$

Take-aways from this example:

1. Both $\pi(\theta)$ and $\pi(\theta|x)$ were Beta distributions. We say then that the Beta is conjugate for the Bernoulli likelihood
2. We did not need to compute the $m(x)$. In practice, usually can't compute it and instead rely on sampling (MCMC, etc.)

3. The usual frequentist estimate is \bar{x} (this would be the limiting case of $\alpha = \beta = 0$). In fact, $\hat{\theta}_B(x) = \frac{\alpha}{\alpha+\beta+n} + \frac{\sum x_i}{\alpha+\beta+n} = \left(\frac{\alpha}{\alpha+\beta}\right) \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right) + \bar{X} \left(\frac{n}{\alpha+\beta+n}\right)$. That is, the posterior mean is a weighted average of the prior mean and the MLE. As α or β go to infinity, result is dominated by prior. As n goes to infinity, get more weight on the \bar{X}

3.2 Conjugate Priors

Definition: Let $f(x|\theta)$ be a family of PMFs/PDFs indexed by $\theta \in \Theta$. A family of distributions is $\Pi = \{\pi(\theta)\}$ is said to be conjugate for $f(x|\theta)$ if $\pi \in \Pi \implies \pi(\theta|x) \in \Pi$.

Note: Number of parameters in conjugate family is always going to be 1 more than the base distribution.

If $f(x|\theta)$ is an exponential family, $f(x|\theta) = (\pi h(x_i)) e^{\sum_{j=1}^k T_j(x) w_j(\theta) + n \ln[c(\theta)]}$, $T_j(x) = \sum_{i=1}^n t_j(x_i)$

A conjugate family with hyperparameter $t \in \mathbb{R}^{k+1}$ is $\pi_t(\theta) \propto e^{\sum_{j=1}^k t_j w_j(\theta) + t_{k+1} \ln[c(\theta)]}$

where t must satisfy $\int_{\theta} e^{\sum_{j=1}^k t_j w_j(\theta) + t_{k+1} \ln[c(\theta)]} < \infty$

Example: $X_i \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$, σ^2 is known

$$\begin{aligned} f(x|\theta) &= (2\pi\sigma^2)^{-n/2} e^{-1/(2\sigma^2) \sum (x_i - \theta)^2} \\ &= h(x) e^{\frac{n\bar{x}}{\sigma^2} \theta - \frac{n}{2\sigma^2} \theta^2} \end{aligned}$$

Thus the conjugate prior has the form $\pi_t(\theta) \propto e^{t_1 \theta + t_2 \theta^2}$. This is e to a quadratic which must be a Normal. We know t_2 must be negative so the tails die out.

Since this must be a normal distribution, find the mean and variance in terms of t_1 and t_2 :

Let ν, τ^2 represent mean and variance that correspond to (t_1, t_2) . $t_2 = \frac{-1}{2\tau^2}, t_1 = \frac{\nu}{\tau^2}$

$$\pi_t(\theta) \propto e^{\frac{n\bar{x}}{\sigma^2} \theta - \frac{n}{2\sigma^2} \theta^2} e^{t_1 \theta + t_2 \theta^2} = e^{(t_1 + \frac{n\bar{x}}{\sigma^2}) \theta + (t_2 - \frac{n}{2\sigma^2}) \theta^2} = e^{t_1^* \theta + t_2^* \theta^2}$$

$$\implies \theta|x \sim N(\nu^*, \tau^{*2}) \text{ where } \tau^{*2} = -\frac{1}{2t_2^*} = \frac{\sigma^2 \tau^2}{n\tau^2 + \sigma^2}, \nu^* = \frac{\tau^2}{\tau^2 + \sigma^2/n} \bar{X} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n} \nu$$

3.3 Bayes Rules

Definition: A Bayes rule for a loss function $l(t|\theta)$ and a prior π is an estimator W^* for which the Bayes risk is minimized: $r_{\pi}(W^*) \leq r_{\pi}(W) \forall W$

$$\begin{aligned}
r_\pi(W) &= \int_{\Theta} R(\theta; W) \pi(\theta) d\theta \\
&= \int_{\Theta} \mathbb{E}[l(W(x), \theta) | \theta] \pi(\theta) d\theta \\
&= \int_{\mathbb{X}} \mathbb{E}[l(W(x), \theta) | X = x] m(x) dx \\
&= \int_{\mathbb{X}} r_\pi(W(x) | x) m(x) dx
\end{aligned}$$

If $r_\pi(W^*(x) | x) \leq r_\pi(W(x) | x)$, then W^* is the Bayes Rule

Example: $l(t, \theta) = (t - \theta)^2$

$$\begin{aligned}
r_\pi(w(x) | x) &= \mathbb{E}[(w(x) - \theta)^2 | X = x] \\
&= \int_{\Theta} (w(x) - \theta)^2 \pi(\theta | x) d\theta \\
&= \int_{\Theta} (w^2(x) - 2w(x)\theta + \theta^2) \pi(\theta | x) d\theta \\
&= w^2(x) - 2w(x) \mathbb{E}[\theta | X = x] + \mathbb{E}[\theta^2 | X = x] \implies w^*(x) \\
&= \frac{2 \mathbb{E}[\theta | X = x]}{2(1)} \\
&= \mathbb{E}[\theta | X = x]
\end{aligned}$$

When using squared-error loss, the Bayes Rule is the posterior mean, when using absolute value squared-error loss it's the posterior median.

4 Jan 29

REVIEW

$$r_\pi(w) = \int_{\mathbb{X}} r_\pi(W(x) | x) m(x) dx$$

$$r_\pi(w(x) | x) = \mathbb{E}[l(W(x), \theta) | X = x]$$

If $l(t, \theta) = (t - \theta)^2 \implies$ The Bayes Rule is $\mathbb{E}[\theta | X = x]$

If $l(t, \theta) = |t - \theta| \implies w^*(x)$ is the posterior median.

A reasonable loss smight be realted to SEL or AEL (Squared/Absolute Error Loss): $l(t, \theta) = g(\theta)(t - \theta)^2$

$$\begin{aligned}
r_\pi(w(x)|x) &= \int_{\Theta} l(t, \theta) \pi(\theta|x) d\theta \\
&= \int_{\Theta} g(\theta) (t - \theta)^2 \pi(\theta|x) d\theta \\
&= \int_{\Theta} (t - \theta)^2 [g(\theta) \pi(\theta|x)] d\theta
\end{aligned}$$

Provided $\int_{\Theta} [g(\theta) \pi(\theta|x)] d\theta$ is finite (integrable), it becomes the Kernel of another function: define: $\tilde{\pi}(\theta|x) \propto g(\theta) \pi(\theta|x)$

Then $r_\pi(w(x)|x) \propto \int_{\Theta} (t - \theta)^2 \tilde{\pi}(\theta|x) d\theta = r_{\tilde{\pi}}(w(x)|x)$ (under SEL) $\implies w^*(x) = \mathbb{E}_{\tilde{\pi}}(\theta|X = x)$

Example: $X_i \stackrel{\text{iid}}{\sim} \text{Exp}(\theta)$

$\pi(\theta)$ is the $\text{Gamma}(\alpha, \beta)$

$$l(t, \theta) = \left(\frac{t}{\theta} - 1\right)^2 = \frac{(t - \theta)^2}{\theta^2} \implies g(\theta) = \frac{1}{\theta^2}$$

$g(\theta) \pi(\theta|x) \propto \frac{1}{\theta^2} \theta^n e^{\theta \sum x_i} \theta^{\alpha-1} e^{-\theta/\beta} = \theta^{\alpha+n-3} e^{-\theta(\sum x_i + \beta^{-1})}$ which is a $\text{Gamma}(\alpha + n - 2, (\sum x_i + \beta^{-1})^{-1})$ (note that we need $n \geq 2$ to have valid parameters).

The Bayes Rule is $w^*(x) = \frac{\alpha+n-2}{\sum x_i + \beta^{-1}}$

5 Best Unbiased Estimation

Recall: $MSE_\theta(w) = \mathbb{E}[(w - \theta)^2]$ (this is a risk function, so common it has its own name!)

Example: $X_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$

$$1. w(x) = \bar{x}, MSE_{\mu, \sigma^2}(\bar{x}) = \mathbb{E}_{\mu, \sigma^2}[(\bar{x} - \mu)^2]$$

Since $\bar{x} \sim N(\mu, \sigma^2/n)$, $MSE_{\mu, \sigma^2}(\bar{x}) = \text{Var}(\bar{x}) = \frac{\sigma^2}{n}$

$$2. \text{ Let } s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2$$

$$MSE_{\mu, \sigma^2}(s^2) = \mathbb{E}_{\mu, \sigma^2}[(s^2 - \sigma^2)^2] = \text{Var}(s^2) = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{n-1}$$

In general:

$$\begin{aligned}
MSE_\theta(w) &= \mathbb{E}_\theta[(w - \theta)^2] \\
&= \mathbb{E}_\theta[(w - \mathbb{E}_\theta[w] + \mathbb{E}_\theta[w] - \theta)^2] \\
&= \mathbb{E}_\theta[(w - \mathbb{E}_\theta[w])^2] + (\mathbb{E}_\theta[w] - \theta)^2 + 2 \mathbb{E}_\theta[(w - \mathbb{E}_\theta[w])(\mathbb{E}_\theta[w] - \theta)] \\
&= \text{Var}_\theta(w) + \text{Bias}_\theta^2(w), \text{Bias}_\theta(w) = \mathbb{E}_\theta[w] - \theta
\end{aligned}$$

Example: $X_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

$$\hat{\theta} = \frac{1}{n} \sum (x_i - \bar{x})^2 \text{ (MLE/MM)}$$

$$\begin{aligned}
MSE_{\hat{\sigma}^2}(\hat{\sigma}^2) &= Var_{\hat{\sigma}^2}(\hat{\sigma}^2) + Bias_{\hat{\sigma}^2}^2(\hat{\sigma}^2) \\
&= Var_{\sigma^2}\left(\frac{n-1}{n}s^2\right) + Bias_{\sigma^2}^2\left(\frac{n-1}{n}s^2\right) \\
&= \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} + [\mathbb{E}_{\sigma^2}\left(\frac{n-1}{n}s^2\right) - \sigma^2]^2 \\
&= \frac{2\sigma^4(n-1)}{n^2} + \left[\frac{n-1}{n}\sigma^2 - \sigma^2\right]^2 \\
&= \frac{2\sigma^4}{n-1}\left(\frac{n-1}{n}\right)^2 + \frac{\sigma^2}{n^2} \\
&= \frac{2\sigma^4}{n-1}\left[\frac{(n-1/2)(n-1)}{n^2}\right] < MSE_{\sigma^2}(s^2)
\end{aligned}$$

Let $\tau(\theta)$ be some estimand:

Define: w is said to be unbiased for $\tau(\theta)$ if $\mathbb{E}_\theta(w) = \tau(\theta) \forall \theta \in \Theta$

Definition: An estimator w^* is said to be a (uniformly) minimum variance unbiased estimator (MVUE) of $\tau(\theta)$ if:

1. w^* is unbiased for $\tau(\theta)$
2. For any w unbiased for $\tau(\theta)$, $Var(w) \geq Var_\theta(w^*)$

Example: $x_1, \dots, x_n \stackrel{\text{iid}}{\sim} \mathbb{E}[x_i] = \theta$

Let $w_1 = x_1, w_2 = \bar{x}$. Both are unbiased, but \bar{x} has smaller variance. To prove the MVUE, we'd have to compare w_2 to all other w s

Example: $x_i \stackrel{\text{iid}}{\sim} Pois(\lambda)$

Both \bar{x}, s^2 are unbiased. \bar{x} is simpler, and also, $Var(\bar{x}) = \frac{\lambda}{n} < Var_\lambda(s^2)$

Two ways to establish that an estimator is an MVUE:

1. Find a lower bound $L(\theta)$ such that $Var_\theta(w) \geq L(\theta)$ for all unbiased w . Then, $Var_\theta(w^*) = L(\theta)$, w^* is MVUE
2. Show that the MVUE has characteristic properties and construct w^* with those properties

5.1 Cramer-Rao Lower Bound

The score function is $\psi(x; \theta) = \frac{\partial}{\partial \theta} \ln(f(x|\theta))$

CR Condition: For any statistic W with finite expectation -

$$(|\mathbb{E}_\theta[W]| < \infty \forall \theta), \quad \frac{d}{d\theta} \mathbb{E}_\theta[W] = \frac{d}{d\theta} \int_{\mathbb{X}} W(x) f(x|\theta) dx = \int_{\mathbb{X}} \frac{\partial}{\partial \theta} W(x) f(x|\theta) dx$$

$$\text{Notice: } \frac{\partial f(x|\theta)}{\partial \theta} = \frac{\partial f(x|\theta)}{\partial \theta} \frac{f(x|\theta)}{f(x|\theta)} = \frac{\partial \ln(f(x|\theta))}{\partial \theta} f(x|\theta) = \psi(x; \theta) f(x|\theta)$$

$$\text{Thus: } \frac{d}{d\theta} \mathbb{E}_\theta[W] = \int_{\mathbb{X}} W(x) \psi(x; \theta) f(x|\theta) dx$$

Definition: The Information Number (Fisher Information) is:

$$I_n(\theta) = \mathbb{E}_\theta[\psi^2(x; \theta)] \text{ (Scalar)}$$

$$I_n(\theta) = \mathbb{E}_\theta[\psi(x; \theta)\psi(x; \theta)^\top] \text{ (Vector)}$$

Fact: If CR condition holds, then the Fisher Information is the variance of the score:

$$I_n(\theta) = \text{Var}_\theta(\psi(x; \theta)) \text{ (note: this means that } \mathbb{E}_\theta[\psi(x; \theta)] = 0)$$

Theorem: CRLB

Suppose the CR Condition holds and $0 < I_n(\theta) < \infty$. If W is a statistic satisfying:

- i) $\mathbb{E}_\theta[W] = \tau(\theta), |\tau(\theta)| < \infty$ (unbiased estimator for $\tau(\theta)$) and
- ii) $\text{Var}_\theta(W) < \infty \forall \theta$

$$\text{then } \text{Var}_\theta(W) \geq \frac{(\tau'(\theta))^2}{I_n(\theta)} \text{ (for scalar } \theta)$$

Remember: Cauchy-Schwartz: $|x^\top y| \leq \|x\| \|y\| \implies |\mathbb{E}[xy]| \leq \mathbb{E}[x^2]^{1/2} \mathbb{E}[y^2]^{1/2}$. Equality only happens when $x = ay$

Proof: CRLB

Apply CS inequality to $Y_\theta = W - \tau(\theta), z_\theta = \psi(x; \theta)$

$$\begin{aligned} |\tau'(\theta)| &= \left| \frac{d}{d\theta} \mathbb{E}_\theta[W] \right| = \left| \int_{\mathbb{X}} W(x) \psi(x; \theta) f(x|\theta) dx \right| = |\mathbb{E}_\theta[W \psi(x; \theta)]| \text{ (here, because } \psi \text{ has mean zero we can add constant on } W) \\ &= |\mathbb{E}_\theta[Y_\theta z_\theta]| \leq [\text{Var}_\theta(Y_\theta) \text{Var}_\theta(z_\theta)]^{1/2} = [\text{Var}_\theta(W) I_n(\theta)]^{1/2} \implies \text{Var}_\theta(W) \geq \frac{(\tau'(\theta))^2}{I_n(\theta)} \end{aligned}$$

Comments:

1) To apply this result, recall that we have equality in the CRLB iff $W(X) - \tau(\theta) = k(\theta)\psi(x; \theta)$ (some constant times the score, refer to Cauchy Schwartz again). **THIS IS VERY IMPORTANT!!!!**

2) In the iid case, let $I_1(\theta) = \text{Var}_\theta(\frac{d}{d\theta} \ln(f(x_1|\theta)))$. Then,

$$\begin{aligned} I_n(\theta) &= \text{Var}_\theta\left(\frac{d}{d\theta} \ln(f(\mathbf{x}|\theta))\right) \\ &= \text{Var}_\theta\left(\sum_{i=1}^n \frac{\partial}{\partial \theta} \ln(f(\mathbf{x}|\theta))\right) \\ &= \sum_{i=1}^n \text{Var}_\theta\left(\frac{\partial}{\partial \theta} \ln(f(x_i|\theta))\right) \\ &= nI_1(\theta) \end{aligned}$$

3) If $\frac{d}{d\theta} \mathbb{E}_\theta[\psi(x; \theta)] = \int_{\mathbb{X}} \frac{\partial}{\partial \theta} [\psi(x; \theta) f(x|\theta)] dx$ then: $I_n(\theta) = -\mathbb{E}_\theta[\frac{\partial}{\partial \theta} \psi(x; \theta)]$ (when iid:) $= -n \mathbb{E}_\theta[\frac{\partial^2}{\partial \theta^2} \ln(f(x_1|\theta))]$

4) If $\tau(\theta)$ is a vector, $\text{Var}_\theta(W) \succeq \tau'(\theta)^\top I_n^{-1}(\theta) \tau'(\theta)$ ($A \succeq B$ means $A - B$ is Positive Definite)

Example: $X_i \stackrel{\text{iid}}{\sim} \text{Pois}(\theta), i = 1, \dots, n$

a) Consider $\tau(\theta) = \theta$.

$$\begin{aligned}
\psi(x; \theta) &= \frac{\partial}{\partial \theta} \ln \left[e^{-n\theta} \frac{\theta^{\sum x_i}}{\prod x_i!} \right] \\
&= \frac{\partial}{\partial \theta} [-n\theta + \ln(\theta) \sum x_i - \sum \ln(x_i!)] \\
&= -n + \frac{\sum x_i}{\theta} \\
&= \frac{n}{\theta} (\bar{x} - \theta) \\
&\implies \frac{\theta}{n} \psi(x; \theta) \\
&= \bar{x} - \theta \implies \bar{x} \text{ is the MVUE of } \theta.
\end{aligned}$$

To find $Var_{\theta}(\bar{x})$, we know that $Var_{\theta}(\bar{x}) = \frac{\theta}{n}$. We can also use the CRLB since it's an equality: $\tau'(\theta) = 1, I_n(\theta) = nI_1(\theta) = n(-\mathbb{E}_{\theta}[\frac{\partial}{\partial \theta} \ln(f(x_1|\theta))]) = -n\mathbb{E}_{\theta}[\frac{\partial}{\partial \theta} (-1 + \frac{x_1}{\theta})] = n\mathbb{E}_{\theta}[\frac{x_1}{\theta^2}] = \frac{n\theta}{\theta^2} = \frac{n}{\theta} \implies CRLB = \frac{1}{n/\theta} = \frac{\theta}{n}$

b) Consider $\tau(\theta) = \theta^2$