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

6 Constructing MVUEs using Completeness and Sufficiency **12**
6.1 Sufficiency . . . 13

7 Feb 10 **13**

8 Feb 12 **14**

9 Completeness **15**

10 Returning to MVUEs **17**

11 Feb 24 **18**

12 Asymptotic Evaluation of Estimators **19**

13 Feb 26 **19**

1 Point Estimation

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

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 $\mu_k(\theta) = \mathbb{E}_\theta(X^k) = \int\limits_{\mathbb{X}} x^k f(x|\theta)dx$

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

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

<u>Definition</u>: The method of moments estimator of θ is the solution to the system of equations:

$$\mu_1(\theta) = \hat{\mu}_1 \tag{1}$$
$$. \tag{2}$$
$$. \tag{3}$$
$$\mu_p(\theta) = \hat{\mu}_p \tag{4}$$

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

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} Beta(\alpha, \beta), \theta = (\alpha, \beta)$

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

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

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

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

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

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

<u>Example</u>: $x_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$

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

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

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

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 (intervtible). Then, $\hat{\eta}_{MM} = \eta(\hat{\theta}_{MM})$

Let's say $x_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ 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.

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

<u>Properties</u>:

- 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 $\mu_k(\theta) = \hat{\mu}_k \approx \mu_k(\theta^*)$

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} Bin(m, \theta), i = 1, ..., n$ (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. $\hat{\tau}_{MM}(\theta) = ln\frac{\overline{X}/m}{1-\overline{X}/m}$

2 Maximum Likelihood

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

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

<u>Likelihood Function</u>: 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 $x_i \overset{\text{iid}}{\sim} f(x_i|\theta)$ (f is marginal dist.), then $l(\theta; x) = \sum_{i=1}^{n} ln[f(x_i|\theta)]$ (because $x_i \overset{\text{iid}}{\sim}$, the sum is just a transformation and we can apply LLN, CLT, etc.)

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

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

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

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

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

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

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

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

It would make sense that $\hat{\theta}(x) = \underset{\theta \in \Theta}{argmax}[l(X|\theta)] \approx \underset{\theta \in \Theta}{argmax}\, \mathbb{E}_{\theta_0}(ln[f(X|\theta)])$ 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 $f_0 \equiv f_1$.

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

<u>Proof (Discrete Case)</u>: 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

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

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

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

<u>Definition</u>: If we observed $x \in X$, the posterior distribution is: $\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int\limits_{\Theta} f(\theta|x)\pi(\theta)d\theta}$

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

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

1. Estimate θ using a measure of 'center':
2. $\hat{\theta} = \mathbb{E}(\theta|x) = \int\limits_{\Theta} \theta\pi(\theta|x)d\theta$
3. $0.5 = \int\limits_{0}^{\hat{\theta}} pi(\theta|x)d\theta$
4. $\hat{\theta}(x) = \underset{\theta \in \Theta}{argmax}[\pi(\theta|x)]$

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

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

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

$\pi(\theta|x) = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)\int\limits_{0}^{1}\frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}d\theta}$

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

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

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

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 smapling (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 infitinty, get more weight on the $\bar{X}$

3.2 Conjugate Priors

<u>Definition</u>: Let $f(x|theta)$ be a gamily 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 eponential family, $f(x|\theta) = (\pi h(x_i))e^{\sum_{j=1}^{k} T_j(x)w_j(\theta)+nln[c(\theta)]}, T_j(x) = \sum_{i=1}^{n} t_j(x_i)$

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

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

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} N(\theta, \sigma^2), \sigma^2$ 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 distirbution, 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). $t_2 = \frac{-1}{2\tau^2}, t_1 = \frac{\nu}{\tau^2}$

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

$\implies \theta|x \sim N(\nu^*, \tau^{*2})$ 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

<u>Definition</u>: 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

<u>Example</u>: $l(t,\theta) = (t-\theta)^2$

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

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) $\implies w^*(x) = \mathbb{E}_{\tilde{\pi}}(\theta|X=x)$

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} Exp(\theta)$

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

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

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

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} N(\mu,\sigma^2)$

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

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

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

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

In general:

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

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} N(\mu,\sigma^2)$

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

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

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

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

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

<u>Definition</u>: 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 $\tau(\theta), Var(w) \geq Var_\theta(w^*)$

<u>Example</u>: $x_1, ..., x_n \overset{\text{iid}}{\sim}, \mathbb{E}[x_i] = \theta$

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 ws

<u>Example</u>: $x_i \overset{\text{iid}}{\sim} Pois(\lambda)$

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 characterstic 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 -

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

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

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

<u>Definition</u>: The Information Number (Fisher Information) is:

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

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

<u>Fact</u>: If CR condition holds, then the Fisher Information is the variance of the score:

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

<u>Theorem</u>: 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) $Var_\theta(W) < \infty \forall \theta$

then $Var_\theta(W) \geq \frac{(\tau'(\theta))^2}{I_n(\theta)}$ (for scalar θ)

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 $x = ay$

<u>Proof</u>: CRLB

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

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

<u>Comments</u>:

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 $I_1(\theta) = Var_\theta(\frac{d}{d\theta}ln(f(x_1|\theta)))$. Then,

$$\begin{aligned}
I_n(\theta) &= Var_\theta(\frac{d}{d\theta}ln(f(\underset{\sim}{\boldsymbol{x}}|\theta))) \\
&= Var_\theta(\sum_{i=1}^n \frac{\partial}{\partial\theta}ln(f(\underset{\sim}{\boldsymbol{x}}|theta))) \\
&= \sum_{i=1}^n Var_\theta(\frac{\partial}{\partial\theta}ln(f(x_i|\theta))) \\
&= nI_1(\theta)
\end{aligned}$$

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, $Var_\theta(W) \succeq \tau'(\theta)^\top I_n^{-1}(\theta)\tau'(\theta)$ ($A \succeq B$ means $A - B$ is Positive Definite)

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} Pois(\theta), i = 1, ..., n$

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

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

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

Done in class on paper: $\bar{x}^2$ is a biased estimator: $\mathbb{E}[\bar{x}^2] = Var(\bar{x}) + \mathbb{E}[\bar{x}]^2 = Var(\bar{x}) + \theta^2$. The bias is $Var(\bar{x}) = \theta/n$

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2), \theta = (\mu, \sigma^2)$

$$
I_n(\theta) = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & n/(2\sigma^4) \end{bmatrix}
$$

$\tau(\theta) = \mu, \tau'(\theta) = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$

If W is unbiased for μ, $Var_\theta(W) \geq [\tau'(\theta)]^\top I_n^{-1}(\theta)\tau'(\theta) = \left(\begin{smallmatrix} 1 & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right) = \sigma^2/n \implies \bar{x}$ is MVUE.

$\tau(\theta) = \sigma^2 \implies \tau'(\theta) = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$

If W is unbiased for σ^2: $Var_\theta(W) \geq 2\sigma^2/n$

$Var_\theta(s^2) = 2\sigma^4/(n-1) > 2\sigma^4/n$

<u>Example</u>: (CR Condition Fails)

$X_i \overset{\text{iid}}{\sim} Unif(0, \theta)($ implicitly assuming $\tau(theta) = \theta)$

$\psi(x;\theta) = \frac{\partial}{\partial\theta} ln[f(x|\theta)] = \frac{\partial}{\partial\theta} ln[\theta^{-n} I(0 < X_{(n)} \leq \theta)]$

We can still compute $\mathbb{E}_\theta[\psi^2(x;\theta)] = \mathbb{E}_\theta[(\frac{\partial}{\partial\theta}(-nln(\theta)))^2] = \frac{n^2}{\theta^2} \implies$ The CRLB would be: $\frac{1}{n^2/\theta^2} = \frac{\theta^2}{n^2}$

$W(x) = \frac{n+1}{n} X_{(n)}$ is unbiased for θ. However, $Var_\theta(W) = \frac{\theta^2}{n(n+2)} < \frac{\theta^2}{n^2}$. This is possible because the CR Condition didn't hold in the first place, so the CRLB wasn't actually a valid representation of the lower bound.

6 Constructing MVUEs using Completeness and Sufficiency

Let W be an unbiased estimator of $\tau(\theta)$ and T be any statistic. Consider $\mathbb{E}_\theta[W|T]$ as a new random variable.

$\mathbb{E}_\theta[\mathbb{E}_\theta(W|T)] = \mathbb{E}_\theta[W]$

If I condition in a way that I still get a statistic, it remains an unbaised estimator.

$Var_\theta(W) = Var_\theta(\mathbb{E}_\theta[W|T]) + \mathbb{E}_\theta[Var_\theta(W|T)] \geq Var_\theta(\mathbb{E}_\theta[W|T])$

Questions moving forward:

1. How can we choose T so that $\mathbb{E}_\theta[W|T]$ doesn't depend on θ? (sufficiency)
2. How can we choose T to maximally reduce the variance? (completeness)

6.1 Sufficiency

Because W is a statistic, if $x|T$ does not depend on θ, then neither will W.

<u>Definition</u>: A statistic T is sufficient if the distribution of $X|T$ does not depend on θ.

Interpretation: If T is sufficient, then two data points x, y should lead to the same inference if $T(x) = T(y)$.

Obvious Sufficient Statistics include:

- $x = (x_1, ..., x_n)$
- $(x_{(1)}, ..., x_{(n)})$

<u>Fact</u>: If I know $T(x)$ I can generate a new dataset Y with the exact same distributionas X and $T(x) = T(y)$.

<u>Example</u>: $x_1, ..., x_n \sim Bern(\theta)$. Then $\sum x_i$ is sufficient.

Generate Y using $\sum x_i$ such that $Y_i \overset{\text{iid}}{\sim} Bern(\theta)$

$$P(Y = y|\sum x_i = t) = \begin{cases} 0 & \sum y_i \neq t \\ \frac{1}{\binom{n}{t}} & \sum y_i = t \end{cases}$$

$$P_\theta(Y = y) = \sum_{t=0}^{n} \frac{1}{\binom{n}{t}} I(\sum y_i = t) P_\theta(\sum x_i = t) = \sum_{t=0}^{n} I(\sum y_i = t)\binom{n}{t}\theta^t(1-\theta)^{n-t} = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}$$

7 Feb 10

Sufficiency: If we take an unbiased estimator, and condition it with a sufficient statistic, it will stay an unbiased estimator that does not depend on θ.

If T is sufficient (discrete case), $P_\theta(X = x|T(X) = t) = \frac{P_\theta(X=x \text{ and } T(X)=t)}{P_\theta(T(x)=t)} = \frac{P_\theta(X=x)I(t=T(x))}{P_\theta(T(X)=T(x))} = \begin{cases} 0 & t \neq T(x) \\ \frac{p(x|\theta)}{q(T(x)|\theta)} & t = T(x) \end{cases}$

<u>Theorem</u>: Factorization

A statistic $T(X)$ is sufficient iff there are functions $h(X)$ and $g(t, \theta)$ such that $f(X|\theta) = h(x)g(T(x), \theta)$.

<u>Proof</u> (Discrete):

If T is sufficient, take $g(t, \theta) = q(t|\theta), h(x) = \frac{p(x|\theta)}{q(T(x)|\theta)}$. That yields the same as above.

Now, assume the factorization holds: $q(t|\theta) = P_\theta(T(X) = t) = \sum_{y:T(y)=t} p(y|\theta) = \sum_{y:T(y)=t} h(y)g(T(y), \theta) = [\sum_{y:T(y)=t} h(y)]g(t, \theta)$.

Now, $\frac{p(x|\theta)}{q(T(x)|\theta)} = \frac{h(x)g(T(x),\theta)}{[\sum\limits_{y:T(y)=T(x)} h(y)]g(T(x),\theta)} = \frac{h(x)}{\sum\limits_{y:T(y)=T(x)} h(y)}$ which does not involve θ.

<u>Example</u>:

$X_i \overset{\text{iid}}{\sim} N(\theta, 1)$

$f(x|\theta) = \prod\limits_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\theta)^2} = (2\pi)^{-n/2} exp(-1/2 \sum\limits_{i=1}^{n} (x_i - \theta)^2)$

Example factorizations

1. $(2\pi)^(- n/2) : h(x)$ or could be absorbed into g (doesn't really matter)

$g(T(x), \theta)$ is the exponential: $T(x) = (x_1, ..., x_n)$

2. $(2\pi)^{-n/2} exp(-1/2 \sum\limits_{i=1}^{n} (x_{(i)} - \theta)^2)$

$T(x) = (x_{(1)}, ..., x_{(n)})$

3. $(2\pi)^{-n/2} exp(-1/2 \sum x_i^2) exp(n\bar{x}\theta - \frac{n\theta^2}{2})$

$\implies T(X) = \sum x_i$ or $T(X) = \bar{x}$ are sufficient

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2), \theta = (\mu, \sigma^2)$

$f(x|\theta) = (2\pi\sigma^2)^{-n/2} exp(-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2) = (2\pi)^{-n/2} \sigma^{-n} exp(-\frac{1}{2\sigma^2} \sum (x_i^2) + \frac{\mu}{\sigma^2} \sum x_i - \frac{n\mu^2}{2\sigma^2})$

$T(x) = (\sum x_i, \sum x_i^2)$ is sufficient.

<u>Note</u>: Any 1:1 transformation of a sufficient statistic is sufficient!

Set $t_1 = \sum x_i, t_2 = \sum x_i^2$: $\bar{x} = \frac{t_1}{n}, s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{1}{n-1}(\sum x_i^2 - n\bar{x}^2) = \frac{1}{n-1}(t_2 - \frac{t_1^2}{n}) \implies (\bar{x}, s^2)$ is also sufficient.

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} Unif(\theta, \theta + 1)$

$f(x|\theta) = \prod\limits_{i=1}^{n} I(x_i > \theta) I(x_i < \theta + 1) = I(X_{(1)} > \theta) I(X_{(n)} < \theta) \implies T(X) = (X_{(1)}, X_(n))$ is sufficient.

<u>Example</u> (Exponential Families)

$f(x|\theta) = (\prod h(x_i)) exp[\sum\limits_{i=1}^{n} \sum\limits_{j=1}^{k} w_i(\theta) t_j(x_i) + nln(c(\theta))] = \tilde{h}(x) exp[\sum\limits_{j=1}^{k} w_j(\theta) T_j(x) + nln(c(\theta))]$

$\implies (T_1(x), ..., T_k(x))$ is a sufficient statistic.

8 Feb 12

<u>Definition</u>: A sufficient statistic is is said to be minimal sufficient if it is a function of any other sufficient statistic.

Note: a function $g(x)$ is a function of $f(x)$ iff $f(x) = f(y) \implies g(x) = g(y)$

Let $T(X)$ be any sufficient statistic. Then by the factorization theorem, $f(x|\theta) = h(x)g(T(x), \theta)$. Therefore, when $T(x) = T(y)$, $\frac{f(x|\theta)}{f(y|\theta)} = \frac{h(x)g(T(x),\theta)}{h(y)g(T(y),\theta)} = \frac{h(x)}{h(y)}$: doesn't depend on θ. When minimally sufficient, this implies the ratio will not have a

dependence on θ. We also want this to go the other way.

<u>Theorem</u>: A sufficient statistic is minimal sufficient if $\frac{f(x|\theta)}{f(y|\theta)}$ is independent of θ implies $T(x) = T(y)$.

<u>Proof</u>: Define T^* to be any sufficient statistic, and T satisfies the above theorem condition. If $T^*(x) = T^*(y)$ then by factorization theorem, $\frac{f(x|\theta)}{f(y|\theta)} = \frac{h^*(x)g^*(T^*(x),\theta)}{h^*(y)g^*(T^*(y),\theta)} = \frac{h^*(x)}{h^*(y)}$ which doesn't depend on θ.

By the above theorem, $T(x) = T(y)$, so $T(x)$ is minimal sufficient.

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$

1) σ^2 known: $f(x|\mu) = (2\pi\sigma^2)^{-n/2} exp(-\frac{1}{2\sigma^2}\sum(x_i - \mu)^2) = (2\pi\sigma^2)^{n/2} exp(1\frac{1}{2\sigma^2} + \frac{n\mu\bar{x}}{\sigma^2} - \frac{n\mu^2}{2\sigma^2})$

$\frac{f(x|\mu)}{f(y|\mu)} = exp(-\frac{1}{\sigma^2}[\sum x_i^2 - \sum y_i^2])exp(\frac{n\mu}{\sigma^2}(\bar{x} - \bar{y}))$. The only way to have independence from μ is if $\bar{x} = \bar{y}$, so $T(x) = \bar{(}x)$ is minimal sufficient.

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} Unif(0, 1)$

$f(x|\theta) = I(\theta < x_{(1)})I(\theta + 1 > x_{(n)})$

$\frac{f(x|\theta)}{f(y|\theta)} = \frac{I(\theta<x_{(1)})I(\theta+1>x_{(n)})}{I(\theta<y_{(1)})I(\theta+1>y_{(n)})}$. For this to be independent of θ, the corresponding indicators must 'change' at the same values, meaning $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$. Therefore, $T = (x_{(1)}, x_{(n)})$ is minimal sufficient.

In this family, $T'(x) = (\frac{x_{(n)}-1+x_{(1)}}{2}, x_{(n)} - x_{(1)})$ is minimal sufficient. Notice $X_{(n)} - X_{(1)} = (X_{(n)} - \theta) - (X_{(1)} - \theta) \overset{D}{=} z_{(n)} - z_{(1)}, Z_i \overset{\text{iid}}{\sim} Unif(0, 1)$ meaning that part of the minimal statistic does not depend on θ: but the minimal statistic is supposed to be the minimal information to describe θ: uh oh

<u>Definition</u>: A statistic S is ancillary to θ if its distribution does not depend on θ.

<u>Finding Ancillary Statistics</u>:

1. Location Families: $f(x|\theta) = g(x - \theta)$. $X_i - \theta \overset{D}{=} Z \sim g \implies X_i - X_j$ is ancillary. Or $X_{(i)} - X_{(j)}$, etc. Spacing between the data is anciallry in a location family

2. Scale Families: $f(x|\theta) = \frac{1}{\theta}g(x/\theta), \theta > 0$. $\frac{X_i}{\theta} \overset{D}{=} Z \sim g \implies , \frac{X_i}{X_j}, \frac{X_{(i)}}{X_{(j)}}$.

9 Completeness

Want statistics that are "completely about θ"

<u>Definition</u>: A statistic T is complete for θ if, for a transformation $g(T), \mathbb{E}_\theta[g(T)] = 0 \implies g(T) = 0$ with probability 1.

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} f(x_i|\theta)$

$T(X) = X$ is not complete. Let $g(T) = x_1 - x_2$. Then $P(g(T) \neq 0) > 0$, but $\mathbb{E}_\theta[g(T)] = \mathbb{E}_\theta[x_1] - \mathbb{E}_\theta[x_2] = 0$.

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} Unif(\theta, \theta + 1)$

$T = (x_{(1)}, x_{(n)})$ is not complete. Let $g(T) = x_{(n)} - x_{(1)} - \mathbb{E}[X_{(n)} - x_{(1)}]$. Since $P(g(T) \neq 0) > 0, \mathbb{E}_\theta(g(T)) = 0$, T is not complete.

<u>Feb 19</u>

Example: $X_i \overset{\text{iid}}{\sim} Unif(0,\theta), \theta > 0$

$T = X_{(n)}$ is minimal sufficient. Is it complete?

To show completeness, assume there exists a function $g(t)$ such that $\mathbb{E}_\theta[g(T)] = 0 \forall \theta > 0$. We then need to show that this only happens when $g(t) = 0$.

Assume $g(t)$ is continuous. Assume $0 = \mathbb{E}_\theta[g(T)] = \int_0^\theta g(t) X_{(n)} dt = \int_0^\theta g(t) n \frac{t^{n-1}}{\theta^n} dt$

Derivative of both sides:

$$0 = n \frac{d}{d\theta} \theta^{-n} \int_0^\theta g(t) t^{n-1} dt = n \left[-n\theta^{-n+1} \int_0^\theta g(t) t^{n-1} dt + \theta^{-n} g(\theta) \theta^{n-1} \right] = -n^2 \theta^{-1} \int_0^\theta g(t) \frac{t^{n-1}}{\theta^n} dt + \frac{ng(\theta)}{\theta} = \frac{ng(\theta)}{\theta}$$

$\implies 0 = \frac{ng(\theta)}{\theta} \implies 0 = g(\theta) \forall \theta > 0$, so $T = X_{(n)}$ is complete.

Example: $X_i \overset{\text{iid}}{\sim} Bern(\theta), 0 < \theta < 1$

$T = \sum X_i$ is sufficient (+ minimal).

Assume $g(t)$ satisfies $\mathbb{E}_\theta[g(T)] = 0$

$$0 = \mathbb{E}_\theta[g(T)] = \sum_{i=1}^n g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} = (1-\theta)^n \sum_{i=1}^n g(t) \binom{n}{t} \left(\frac{\theta}{1-\theta} \right)^t \ \forall 0 < \theta < 1$$

This is a polynomial of degree n in $r = \frac{\theta}{1-\theta} \in (0, \infty)$

$\implies$ the polynomial must be 0, meaning all of its coefficients must be 0.

$\implies g(t) \binom{n}{t} = 0 \implies g(t) = 0, t = 0, ..., n$

$\implies T$ is complete.

Theorem:

Assume $X_i \overset{\text{iid}}{\sim} f(x|\theta)$ where $f(x|\theta)$ is an exponential family:

$$F(x|\theta) = h(x) exp\{ \sum_{j=1}^k w_j(\theta) t_j(x) + ln(c(\theta)) \}$$

The sufficient statistic $T = (\sum_{i=1}^n t_1(x_i), ..., \sum_{i=1}^n t_k(x_i))$ is also complete iff the parameter space $\Theta \subset \mathbb{R}^k$ contains an open set.

Interval $(0,1)$ is open. All intergers is not open. Open: can you 'draw' an infitessmally small 'circle' around some point in your set and still only contain things in your set?

Example: $X_i \overset{\text{iid}}{\sim} N(\theta_1, \theta_2), T = (\bar{X}, S^2)$

i) $\Theta = \mathbb{R} x (0, \infty)$ is open. Then T is complete

ii) $\Theta = \mathbb{R} x \{1, 2\}$ is not open. The second set is not open, so T is not complete.

iii) $\Theta = \{(\theta_1, \theta_2) : \theta_2 = \theta_1^2 + 1\}$ which is not an open set, so T is not complete.

For (ii) and (iii), try drawing out the support and see if you can draw a circle around any point and still be in the set.

Facts: Facts about complete statistics:

1) If a minimal sufficient statistic exists, then any complete sufficient statistic is also minimal sufficient

2) Any complete sufficient statistic is independent of any ancillary statistic

Example: $X_i \overset{\text{iid}}{\sim} N(\theta, 1), \theta \in \mathbb{R}$

Then $T = \bar{X}$ is complete and sufficient.

Since $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2 = \frac{1}{n-1}\sum[(x_i - \theta) - (\bar{x} - \theta)]^2 \overset{D}{=} \frac{1}{n-1}\sum(z_i - \bar{z})^2, z_i \overset{\text{iid}}{\sim} N(0, 1)$

$\implies S^2$ is anciallary, $\implies \bar{X} \perp\!\!\!\perp S^2$

If instead $Y_i \overset{\text{iid}}{\sim} N(\theta, \sigma^2), \sigma^2 > 0$ now not known.

What about $(\bar{Y}, S_Y^2)$

$X_i oversetD{=}Z_i + \theta \implies \bar{X} \overset{D}{=} \bar{Z} + \theta$

$Y_i \overset{D}{=} \sigma Z_i + \theta \implies \bar{Y} \overset{D}{=} \sigma\bar{Z} + \theta \implies \bar{Y} \overset{D}{=} \sigma(\bar{X} - \theta) + \theta$

$S^2 \overset{D}{=} S_Z^2$

$S_Y^2 \overset{D}{=} \sigma^2 S_Z^2 \implies (\bar{Y}, S_Y^2) \overset{D}{=} (\sigma(\bar{X} - \theta) + \theta, \sigma^2 S^2) \implies$ these are independent.

10 Returning to MVUEs

Assume W is an unbiased estimator for $\tau(\theta)$ and that T is an sufficient statistic.

Theorem: Rao-Blackwell

The statistic $\phi(T) = \mathbb{E}[W|T]$ is unbiased for $\tau(\theta)$ and $Var_\theta(\phi(T)) \leq Var_\theta(W)$

Proof:

i) $\phi(T)$ is a statistic because T is sufficient

ii) $\phi(T)$ is unbiased because $\mathbb{E}_\theta[\phi(T)] = \mathbb{E}_\theta[\mathbb{E}(W|T)] = \mathbb{E}_\theta[W] = \tau(\theta)$

iii) $Var_\theta(W) = Var_\theta(\phi(T)) + \mathbb{E}_\theta[Var_\theta(W|T)] \geq Var_\theta(\phi(T))$

Theorem:

If an MVUE exists, then it is unique

Proof:

Let W_1, W_2 be MVUEs. Define $W_3 = \frac{1}{2}(W_1 + W_2)$ (unbiased), $Var_\theta(W_3) = \frac{1}{4}[Var_\theta(W_1) + Var_\theta(W_2) + 2Cov_\theta(W_1, W_2)] = \frac{1}{2}[Var_\theta(W_1) + 2Cov_\theta(W_1, W_2)] \leq Var_\theta(W_1)$

Must have equality on the last step (otherwise W_1, W_2 would not be MVUEs), implying $W_1 = W_2$.

Theorem:

An unbiased estimator W is the MVUE iff it is uncorrelated with any U satistfying $\mathbb{E}_\theta[U] = 0, \forall\theta$ (unbiased estimators of 0).

11 Feb 24

<u>Theorem</u>:

If W is unbiased for $\tau(\theta)$, then W is the MVUE iff it is uncorrelated with any unbiased estimator of 0.

<u>Proof</u>:

Assume W is uncorrelated with any U satisfying $\mathbb{E}_\theta[U] = 0$. Then if W' is any unbiased estimator of $\tau(\theta)$, then $U = W - W'$ is unbiased for 0. $Var_\theta(W') = Var_\theta(W - U) = \text{Var}_\theta(W) + \text{Var}_\theta(U) - 2Cov_\theta(W, U) = \text{Var}_\theta(W) + \text{Var}_\theta(U) \geq \text{Var}_\theta(W)$

Hint for HW: Assume the $Cov_\theta(W, U) \neq 0$ for some unbiased U for 0. Then consider $W' = W + aU$. Then use contrapositive to show we can find something with lower variance.

<u>Theorem</u>: (Lehman-Scheffe)

Suppose W is a function of a complete sufficient statistic. Then it is the MVUE of $\tau(\theta) = \mathbb{E}_\theta(W)$.

<u>Proof</u>:

Let U be unbiased for 0. $Cov_\theta(W, U) = \mathbb{E}_\theta(WU) = \mathbb{E}_\theta[\mathbb{E}_\theta[WU|T]] = \mathbb{E}_\theta[W\,\mathbb{E}_\theta[U|T]]$. Because T is sufficient, $g(T) = \mathbb{E}_\theta(U|T)$ doe snot depend on θ. However, $\mathbb{E}_\theta[g(T)] = \mathbb{E}_\theta[U] = 0$. By completeness, $P(g(T) = 0) = 1$.

$\implies Cov_\theta(W, U) = \mathbb{E}_\theta[W(0)] = 0$

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} Pois(\theta), \tau(\theta) = \theta^2, \theta > 0$

We showed previously that $W = \bar{X}^2 - \frac{\bar{X}}{n}$ is unbiased for θ^2.

Why is it the MVUE? Because $\bar{X}$ is complete (exponential family, open set) and sufficient (factorization theorem).

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} N(\theta, 1), \theta \in \mathbb{R}$. Find MVUE for $\tau(\theta) = \theta^2$.

$\bar{X}$ was the MVUE for $\tau(\theta) = \theta$, so we'll start with a guess of $W = \bar{x}^2$. $\mathbb{E}_\theta[W] = \text{Var}_\theta(\bar{x}) + \mathbb{E}_\theta[\bar{x}]^2 = \frac{1}{n} + \theta^2$. So the bias is $\frac{1}{n}$.

So, $W = \bar{x}^2 - \frac{1}{n}$ is the MVUE.

Remember, for any distribution, $\mathbb{E}_\theta[\bar{x}] = \theta$ (population mean) and $\text{Var}_\theta(\bar{x}) = \frac{\sigma^2}{n}$.

<u>Example</u>: $X_i \overset{\text{iid}}{\sim} Exp(\theta), \theta > 0$. $P(X_1 \leq 1) = \tau(\theta) = 1 - e^{-\frac{1}{\theta}}$.

1. What is the complete sufficient statistic?

$\theta^{-n} e^{-\frac{\sum X_i}{\theta}} \implies \sum X_i$ is CSS (Complete Sufficient Statistic).

2. Start with any unbiased estimator

$W = I(X_1 \leq 1)$ is unbiased, so $W^* = \mathbb{E}[W|T]$ is the MVUE.

$\frac{X_1}{\sum X_i} \sim Beta(1, n-1)$ is independent of T

$$\mathbb{E}[I(X_1 \leq 1)|T] = P(\tfrac{X_1}{T} \leq \tfrac{1}{T}|T) = \begin{cases} 1 & T < 1 \\ 1 - (1 - \frac{1}{T})^{n-1} & T \geq 1 \end{cases}$$

Summary:

Can use Cramer-Rao lower bound to find MVUE if you're able to factor the Score. Otherwise, dont get a whole lot.

Exponential Family: Sufficiency is automatic, completeness depends on support

Suffifiency: Factorization Theorem. Simplify correctly, end with Minimal Sufficient

Lehman-Scheffe: If something is Complete and Suffificient, it is the MVUE

12 Asymptotic Evaluation of Estimators

Looking at 'as n goes to infinity', or large sample theory.

<u>Definition</u>: Let W_n be an estimator based on sample of size n of $\tau(\theta)$. W_n is said to be consistent for $\tau(\theta)$ if $W_n \overset{P_\theta}{\to} \tau(\theta) \forall \theta$

$P_\theta(|W_n - \tau(\theta)| > \epsilon) \underset{n\to\infty}{\to} 0$ for all θ and any $\epsilon > 0$.

13 Feb 26

Methods for showing consistency:

1. Tail/Concentration Bounds to control the probabilities

Markov's Inequality, Chebychev, Chernoff

2. If $W_n = g(\bar{x})$, we can apply LLN ($\bar{x} \overset{P}{\to} \mu$) then CMT (Continous Mapping Theorem) ($G(\bar{x} \overset{P}{\to} g(\mu))$) if g is continous.

3. M-estimation (Maiximizer/Minimizer of some function, ie. MLE)

$\frac{1}{n} l(\theta|x) \overset{P_{\theta_0}}{\to} E_{\theta_0}[ln(f(x_1|\theta))]$

Basically, we know the right side is min/max at θ_0, so we need to show that the left side approcahes that same min/max as the overeall distirbutions converge

<u>Example</u>:

Suppose W_n satisfies:

1) $\mathbb{E}_\theta[W_n] \underset{n\to\infty}{\to} \tau(\theta)$ (W_n is asymptocially unbiased for $\tau(\theta)$)

2) $Var_\theta \underset{n\to\infty}{\to} 0$

In this case, W_n is consistent:

$P_\theta(|W_n - \tau(\theta)| > \epsilon) \leq \frac{\mathbb{E}_\theta[(W_n - \tau(\theta))^2]}{\epsilon^2}$ Chebychev's! $= \frac{MSE_{\tau(\theta)}(W_n)}{\epsilon^2} \to 0$ by (1) and (2).

<u>Example</u>: $X \overset{\text{iid}}{\sim} Beta(\alpha, 1)$

1) Method of Moments: $\bar{x} = \frac{\alpha}{\alpha+1} \leftrightarrow \hat{\alpha} = \frac{\bar{x}}{1-\bar{x}}$

$\bar{X} \overset{P}{\to} \frac{\alpha}{\alpha+1}$ as $n \to \infty$

Define $g(x) = \frac{x}{1-x}$, continious on (0, 1). By CMT: $g(\bar{x}) \overset{P}{\to} g(\frac{\alpha}{\alpha+1}) = \frac{\frac{\alpha}{\alpha+1}}{1-\frac{\alpha}{\alpha+1}} = \alpha$

Thus, MoM estimator converges to α so it is consistent.

2) The MLE is $\hat{\alpha} = \frac{n}{-\sum ln(x_i)} = \frac{-1}{\bar{Y}}, \bar{Y} = \frac{-1}{n}\sum ln(x_i)$

$\bar{Y} \xrightarrow{P} \mathbb{E}_\alpha[-ln(x_i)], -ln(x_i) \sim Exp(\frac{a}{\alpha}) \implies \bar{Y} \xrightarrow{P} \frac{1}{\alpha}$

$g(x) = \frac{1}{x}$ continous at $\frac{1}{\alpha}$, so $g(\bar{Y}) \xrightarrow{P} \alpha$

The CLT says as $\sqrt{n}(\bar{X} - \mu) \xrightarrow{D} N(0, \sigma^2)$

$P(|\bar{X} - \mu| > \frac{C}{\sqrt{n}}) = P(\sqrt{n}|\bar{X} - \mu| > C) \to 1 - \Phi(\frac{C}{\sigma})$

<u>Example</u>: $X \overset{\text{iid}}{\sim} Unif(0, \theta)$

$\frac{X_{(n)}}{\theta} \sim Beta(n, 1)$, so $P_\theta(|X_1 - \theta| > Cn^{-1}) = P_\theta(\theta - X_{(n)} < Cn^{-1}) = P_\theta(\frac{X_{(n)}}{\theta} < 1 - \frac{C}{n\theta}) = (1 - \frac{C}{n})^n \to e^{-C}$

<u>Definition</u>: If $\lim_{n\to\infty} b_n Var(W_n) \to \gamma^2$, then γ^2 is called the limiting variance of W_n

$Var(W_n) \approx \frac{\gamma^2}{n}$

eg. $Var(\bar{X}) = \frac{\sigma^2}{n} \implies b_n = n, b_n Var(\bar{X}) = \sigma^2 \to \sigma^2$

<u>Example</u>: $X \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$

$Var(\frac{1}{\bar{X}}) = \infty$

<u>Definition</u>: Suppose $\sqrt{n}(W_n - \tau(\theta)) \xrightarrow{D} Y, E(Y) = 0$, and $Var(Y) < \infty$. Then, $\sigma^2 = Var(Y)$ is the asymptotic variance of W_n