# Math 37600 PR - (19364) - Lectures 03

Ethan Akin
Email: eakin@ccny.cuny.edu

Fall, 2024

# Sec. 7.1: Measurements of Estimator Quality

We continue the study of point estimation: estimating the true value $\theta_0$ for a parametrized family of distributions $X$ with pdf or pmf $f(x;\theta)$. We consider the choice among different estimators $Y = \delta(\mathbf{X})$ for $\theta_0$.

HMC Definition 7.1.1: $Y$ is called a *minimum variance unbiased estimator* (MVUE)when among unbiased estimators it has the minimum variance.

Given the regularity conditions of the previous chapter, an unbiased estimator which is efficient in the sense of the Rao-Cramer Theorem is an MVUE. For example, $\bar{X}$ is an efficient unbiased estimator for $\theta$ when $X \sim \mathcal{N}(\theta, \sigma^2)$ with $\sigma^2$ known, see Exercise 6.2.1. Thus, $\bar{X}$ is an MVUE. See Example 7.1.1.

We think of $\delta(\mathbf{X})$ as a choice or *decision* about $\theta$ upon which we will act. We imagine that if the true value is $\theta$ and we choose $\delta$ different from $\theta$ the result will be a cost measured by a *loss function* $\mathcal{L}(\theta, \delta)$. The most common loss function is just the square of the difference $\mathcal{L}(\theta, \delta) = (\delta - \theta)^2$.

The expected loss is called the *risk function*:

$$R(\theta, \delta) = E_\theta(\mathcal{L}(\theta, \delta(\mathbf{X}))).$$

HMC Example 7.1.2: For the family $\{\mathcal{N}(\theta, 1)\}$ and a sample of size $n$, we consider two estimators for $\theta$, $\theta_1 = \hat{\theta} = \bar{X}$ and $\theta_2 = 0$.

$$E_\theta((\theta_1 - \theta)^2) = Var_\theta(\theta_1) = \frac{1}{n}$$
$$E_\theta((\theta_2 - \theta)^2) = \theta^2.$$

So if $\theta$ is in fact very close to 0, the guess $\theta_2$ has a smaller risk. This illustrates that we usually cannot find a choice which minimizes $R(\theta, \delta)$ uniformly for all $\theta$.

What is often then used as $\delta$ varies over some class $\mathcal{D}$ of decision functions is a *minimax decision function* which minimizes the worst possible risk. That is,

$$\max_{\theta} R(\theta, \delta_0) \;\leq\; \max_{\theta} R(\theta, \delta) \quad \text{for all} \;\; \delta \in \mathcal{D}.$$

This is a very conservative decision rule as it is minimizing the worst case, which may be quite rare.

# Sec. 7.2: Sufficient Statistics

In using a sample $\mathbf{X} = X_1, \ldots, X_n$ to estimate $\theta$, it is useful to extract some statistic $Y = u(\mathbf{X})$ which condenses the information we need to estimate $\theta$ allowing us to discard the rest. $Y$ is a sufficient statistic when, given its value, the remaining information in $\mathbf{X}$ provides no additional help in estimating $\theta$. This is because the conditional distribution of $\mathbf{X}$ given $Y$ is the same for all $\theta$.

HMC Example 7.2.1: Let $\mathbf{X} = X_1, \ldots, X_n$ be an iid sequence of $Bern(\theta)$ rvs with $f(x, \theta) = \theta^x (1-\theta)^{1-x}$ for $x = 0, 1$.

$$f(\theta, \mathbf{x}) = \theta^{\sum x_i} (1-\theta)^{n - \sum x_i}.$$

We see that the number of successes $\sum x_i$ provides all the information we can use to estimate $\theta$. The additional information of which successes occurred when provides no additional useful information. $S = \sum X_i$ is a $Bin(n, \theta)$ rv and so $P(S = k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$. Hence

$$P(\mathbf{X} = \mathbf{x} | S = k) = \begin{cases} 0 & \text{if } \sum x_i \neq k, \\ 1/\binom{n}{k} & \text{if } \sum x_i = k. \end{cases}.$$

The conditional distribution is a discrete uniform distribution with all rearrangements equally likely.

HMC Definition 7.2.1: $Y = u(\mathbf{X}) = u(X_1, \ldots, X_n)$ is a sufficient statistic when the conditional distribution

$$\frac{f(x_1;\theta)\ldots f(x_1;\theta)}{f_Y(u(x_1,\ldots,x_n);\theta)} = H(x_1,\ldots,x_n)$$

does not depend on $\theta$.

The definition required that we find the distribution of $Y$ , i.e. $f_Y(y;\theta)$. Luckily we can avoid that.

HMC Theorem 7.2.1 (**Neyman Factorization Theorem**):
$Y = u(X_1, \dots, X_n)$ is a sufficient statistic if and only if there exist nonnegative functions $k_1$ and $k_2$ such that

$$L(\theta; \mathbf{x}) = f(x_1; \theta) \dots f(x_n; \theta) = k_1(u(x_1, \dots, x_n), \theta) \cdot k_2(x_1, \dots, x_n)$$

and $k_2(x_1, \dots, x_n)$ does not depend on $\theta$.

Proof: If $Y$ is a sufficient statistic we can use $H$ for $k_2$ and $f_Y$ for $k_1$.

For the reverse direction we will just do the proof in the discrete case. For $t$ a value of $u(x_1, \ldots, x_n)$

$$f_Y(t;\theta) = \sum_{\mathbf{x}} f(x_1;\theta)\ldots f(x_n;\theta) = k_1(t,\theta)\cdot\sum_{\mathbf{x}} k_2(\mathbf{x}),$$

where we sum over the set of $n-$tuples $\mathbf{x}$ such that $u(\mathbf{x}) = t$.

So for such an $n-$tuple

$$f(x_1;\theta)\ldots f(x_n;\theta)/f_Y(t;\theta) = k_2(x_1,\ldots,x_n)/\sum_{\mathbf{x}} k_2(\mathbf{x}),$$

which does not depend on $\theta$. So $Y$ is a sufficient statistic.

Example 7.2.2: The family is $\Gamma(2, \theta)$

$$L(\theta) = \frac{\prod_i x_i}{\Gamma(2)^n} \cdot \frac{1}{\theta^{2n}} exp[-\frac{1}{\theta} \sum_i x_i].$$

So $\sum_i X_i$ or, equivalently, $\bar{X}$ is a sufficient statistic.

Example 7.2.3, 7.2.6: The family is $f(\theta, x) = e^{-(x-\theta)}$ for $\theta < x$ or,equivalently, $= e^{-(x-\theta)} I_{(\theta,\infty)}(x)$, where $I_{(\theta,\infty)}$ is the indicator function.

$$L(\theta) = exp(-\sum_i x_i) \cdot exp(n\theta) I_{(\theta,\infty)}(\min_i x_i).$$

The first factor does not depend on $\theta$ and so it is $\min_{i=1}^n \ X_i$ which is the sufficient statistic.

Example 7.2.4: Back in Example 6.3.2, we considered the family $\{\mathcal{N}(\theta, \sigma^2)\}$ with $\sigma^2$ known, From the formulae there it follows that $\bar{X}$ is a sufficient statistic.

Example 7.2.5: The family is $\theta x^{\theta-1}$ for $0 < x < 1, 0 < \theta$.

$$L(\theta) = \theta^n (\prod_i x_i)^{\theta-1} \cdot 1.$$

So $\prod_{i=1}^n X_i$ is a sufficient statistic, as is its log.

# Sec. 7.3: Sufficient Statistic Properties

We need to recall some facts about conditional expectation when $X$ and $Y$ are rv's.

The conditional expectation $E(Y|X = x)$ is a function of the values of $X$, i.e. for $x$ in the support of $X$. Composing with $X$ itself we obtain the rv $E(Y|X)$ which is a function of $X$, namely $E(Y|X = x)$ applied to $X$.

We defined the conditional variance

$$Var(Y|X) \; = \; E((Y - E(Y|X))^2|X) \; = \; E(Y^2|X) - E(Y|X)^2.$$

We showed that

$$\begin{aligned} E(Y) \; &= \; E[E(Y|X)], \\ Var(Y) \; &= \; E[Var(Y|X)] \; + \; Var[E(Y|X)] \; \geq \; Var[E(Y|X)]. \end{aligned}$$

$Y = u(\mathbf{X})$ is a sufficient statistic when the conditional distribution of $\mathbf{X}$ given $Y$ does not depend of $\theta$.

This mean that for any function $g(\mathbf{X})$ the conditional expectation

$$E_\theta(g(X_1, \ldots, X_n)|Y) \quad \text{does not depend on} \quad \theta.$$

That is, we can compute this conditional expectation without knowing the true value of $\theta$. So if $Z = g(X_1, ..., X_n)$ is any statistic and $Y = u(X_1, \ldots, X_n)$ is a sufficient statistic, then $E_\theta(Z|Y)$ is also a statistic, computable from the sample data $X_1, \ldots, X_n$. So we will then write $E(Z|Y)$ for it, omitting the redundant $\theta$ subscript.

HMC Theorem 7.3.1 (**Rao-Blackwell Theorem**): If $Y = u(X_1, \ldots, X_n)$ is a sufficient statistic and $Z = g(X_1, ..., X_n)$ is an unbiased estimator for $\theta$, then $Z_1 = E(Z|Y)$ is an unbiased estimator with variance less than or equal to that of $Z$.

Proof: $E_\theta(Z_1) = E_\theta(E(Z|Y)) = E_\theta(Z) = \theta$.

$$Var_\theta(Z_1) \;=\; Var_\theta(E(Z|Y)) \;\leq\; Var_\theta(Z).$$

HMC Theorem 7.3.2: If $Y = u(X_1, \ldots, X_n)$ is a sufficient statistic and if $L(\theta; \mathbf{X})$ has a unique maximizer $\hat{\theta}(\mathbf{X})$, then $\hat{\theta}$ is a function of $Y$.

Proof: $L(\theta; \mathbf{x}) = f_Y(y; \theta) \cdot H(\mathbf{x})$. Since $H$ does not depend on $\theta$, the location of the unique maximizer for $L$ is the location of the unique maximizer for $f_Y(y; \theta)$. That is, $\hat{\theta}$ can be obtained from $f_Y(y; \theta)$ and so is a function of $y = u(\mathbf{x})$.

Example 7.3.1: The family has density $\{f(x; \theta) = \theta e^{-\theta x} \quad x > 0$ and so $X \sim Exp(\theta) \sim Gamma(1, \frac{1}{\theta})$.

$$\begin{aligned} L(\theta; \mathbf{X}) &= \theta^n exp(-\theta \sum_i X_i) \\ \ell(\theta) &= n \ln \theta - \theta Y_1, \end{aligned}$$

with $Y_1 = \sum_{i=1}^n X_i$ a sufficient statistic and with MLE $Y_2 = \frac{n}{Y_1} = \frac{1}{\bar{X}}$.

Since $Y_1 \sim \Gamma(n, \frac{1}{\theta})$, it follows that for any positive integer $k$ smaller than $n$ (letting $z = \theta t$)

$$E(Y_2^k) = E((\frac{1}{\bar{X}})^k) \; = \; n^k \int_0^\infty \frac{\theta^n}{\Gamma(n)} t^{-k} t^{n-1} e^{-\theta t} dt \; =$$
$$\frac{\theta^k n^k}{\Gamma(n)} \int_0^\infty z^{n-k-1} e^{-z} dt = \frac{\theta^k n^k (n-k-1)!}{(n-1)!}.$$

With $k = 1$ this is $\frac{\theta n}{n-1}$ and so $Z = \frac{(n-1)Y_2}{n}$ is unbiased.

With $k = 2$ this is $\frac{\theta^2 n^2}{(n-1)(n-2)}$.

$$Var(Z) \; = \; (\frac{(n-1)}{n})^2 \cdot \frac{\theta^2 n^2}{(n-1)(n-2)} \; - \; \theta^2 \; = \; \frac{\theta^2}{(n-2)}.$$

With $n = 1$

$$\ell'(\theta) = \theta^{-1} - X, \quad -\ell''(\theta) = \frac{1}{\theta^2}$$

and so $I(\theta) = \frac{1}{\theta^2}$.

The efficiency is $1/Var(Z)nI(\theta) = \frac{n-2}{n} = 1 - \frac{2}{n}$.

The unbiased estimator $Z$ is only asymptotically efficient. That is, its efficiency is less than 1 but tends to 1 as $n$ tends to $\infty$. That $Z$ is nonetheless an MVUE requires additional argument, using results from the next two sections. It is an unbiased estimate which is a function of a sufficient statistic which is complete in the sense of the next section.

# Sec. 7.4: Completeness and Uniqueness

With respect to a probability distribution we say that a phenomenon occurs *almost everywhere* (or ae) when it holds except on a set of probability zero. We saw this when we described convergence ae in Chapter 5. For distributions described by a parametrized family $f(x;\theta)$ of pmf's or pdf's, we say that something is true ae when the exceptions form a subset which has probability zero with respect to every $f(x;\theta)$.

HMC Definition 7.4.1: If $Y$ is an rv with distribution from a family $\{f_Y(y;\theta) : \theta \in \Omega\}$, then the family is called a *complete family* when for a real function $g$, the equation $E_\theta(g(Y)) = 0$ for all $\theta \in \Omega$ only when $g(Y) = 0$ ae.

In particular, if $Y = u(X_1, \ldots, X_n)$ is a sufficient statistic for a family $f(x;\theta)$, then we call $Y$ a *complete sufficient statistic*, when the induced family of distributions $f_Y(y;\theta)$ is complete.

As with sufficiency itself, we can test for completeness of a sufficient statistic without actually computing the induced family $f_Y(y;\theta)$ because we can instead use $f(\mathbf{X};\theta)$ and observe a sufficient statistic $Y$ is a complete when for a real function $g$

$$E_\theta[g(u(X_1,\ldots,X_n))] \quad \text{for all } \theta\in\Omega \quad \Longrightarrow \quad g\circ u \;=\; 0 \;\; ae.$$

The importance of completeness comes from the following result. Up to now the only way we could recognize an unbiased estimator had minimum variance (an MVUE) was when it was an efficient estimator.

For example, in Example 7.3.1 we showed that the unbiased estimator obtained from the sufficient statistic has an efficiency of $1-\frac{1}{n}$. This is only asymptotic efficiency. However, HMC asserted that it was an MVUE. This does in fact follow from the next theorem.

HMC Theorem 7.4.1 (**Lehman and Scheffé Theorem**) Assume that $Y = u(X_1, \ldots, X_n)$ is a complete sufficient statistic for a family $f(x;\theta)$. Up to equality ae, there is at most one unbiased estimator for $\theta$ which is a function of $Y$. If it exists it is the unique MVUE for $\theta$.

Proof: If $\phi(Y)$ and $\psi(Y)$ are unbiased estimators then $E_\theta[\phi(Y) - \psi(Y)] = \theta - \theta = 0$ and so $\phi(Y) - \psi(Y) = 0$ ae by completeness. That is, $\phi(Y) = \psi(Y)$ ae.

If $Z$ is any unbiased estimator, then the Rao-Blackwell Theorem says that $E(Z|Y)$ is an unbiased estimator which is a function of $Y$ and with a variance no greater than that of $Z$. So $E(Z|Y)$ is the unique unbiased estimator which is a function of $Y$ and so it is an MVUE.

Notice that if $Y$ is a sufficient statistic which is not complete then there exists a nonzero function $g$ such that $E_\theta(g(Y)) = 0$ for all $\theta$. If $\phi(Y)$ is an unbiased estimator, then for each real $t$, $\psi_t(Y) = \phi(Y) + tg(Y)$ is a different unbiased estimator which is a function of $Y$.

If $Cov(g(Y), \phi(Y)) \neq 0$ then

$$t = -2Cov(g(Y), \phi(Y))/Var(\phi(Y)) \implies Var(\phi) = Var(\psi_t).$$

Example 7.4.1 uses the uniqueness of the Laplace transform to show that $\sum_{i=1}^n X_i$ is a complete sufficient statistic for the family $h(\theta; x) = \frac{1}{\theta}e^{-x/\theta}$ and so $\bar{X}$ is the MVUE for $\theta$. This estimator was the mle and was efficient.

In Example 7.3.1 the family considered was $f(\theta; z) = \theta e^{-\theta x}$. Exchanging the parameter for its reciprocal does not affect completeness or the choice of a sufficient statistic. On the other hand, the estimators are changed. We saw that the mle $1/\bar{X} = n/\sum_i X_i$, the reciprocal of the mle for the reciprocal parameter. This estimate was, however, biased. The unbiased estimator was shown to be $(n-1)/\sum_i X_i$. By completeness this is indeed the MVUE.

Example 7.4.2: For the family of uniform distributions with $f(x;\theta) = \frac{1}{\theta}, \quad 0 < x < \theta,\ 0 < \theta$, we have seen that $Y = \max(X_1, \ldots, X_n)$ is the MLE and is a sufficient statistic. We also saw that it is biased and that $Z = \frac{n+1}{n}Y$ is an unbiased estimate for $\theta$.

To prove completeness we must show that for any function $u, \quad E_\theta(u(Y)) = 0$ for all $\theta > 0$ implies $u(y) = 0$.

We have seen that the pdf of $Y$ is $g(y;\theta) = \frac{ny^{n-1}}{\theta^n} \quad 0 < y < \theta$ and so

$$E_\theta(u(Y)) \;=\; \int_0^\theta \; u(t)\frac{nt^{n-1}}{\theta^n}dt.$$

This is $= 0$ for all $\theta$ if and only if $\int_0^\theta \; u(t) \cdot t^{n-1}dt = 0$ for all $\theta$. Differentiating with respect to $\theta$ this is equivalent to $u(\theta)\theta^{n-1} = 0$ for all $\theta > 0$. That is, the function $u(Y)$ is identically zero.

Since the unbiased estimate $Z$ is a function of the sufficient statistic, it is the MVUE. Notice that because this family is not regular we cannot use efficiency to show this result.

# Sec. 7.5: Exponential Families

For sufficient statistics a convenient class of examples come from exponential families. Let $f(x;\theta)$ be a family of pdf's or pmf's for an rv $X$ with the parameter $\theta$ varying in a real interval. We call the family an *exponential family* when

$$f(x;\theta) = exp[p(\theta)K(x) + H(x) + q(\theta)],$$

for $x$ in the support.

We make the following regularity assumptions.

- The support $\mathcal{S}$ does not depend on $\theta$.
- The functions $p(\theta)$ and $q(\theta)$ are twice continuously differentiable functions with $p(\theta)$ not a constant.
- If $X$ is a continuous rv, then $K(x)$ is continuously differentiable with $K'(x)$ never zero and $H(x)$ is continuous.

  If $X$ is discrete, then $K(x)$ is a non-constant function on the support $\mathcal{S}$.

For a sample $\mathbf{X} = X_1, \ldots, X_n$ we see that

$$L(\theta; \mathbf{x}) = f(\mathbf{x}; \theta) = exp[p(\theta) \sum_{i=1}^{n} K(x_i) \, + \, \sum_{i=1}^{n} H(x_i) \, + \, nq(\theta)],$$

From the Factorization Theorem, it is clear that
$Y = \sum_{i=1}^{n} K(X_i)$ is a sufficient statistic.

HMC Theorem 7.5.1: The pdf or pmf of $Y = \sum_{i=1}^{n} K(X_i)$ is given by

$$f_Y(y; \theta) = R(y) exp[p(\theta) y + nq(\theta)]$$

where neither the support $S_Y$ nor the function $R$ depend on $\theta$. In addition,

$$E_\theta(Y) = -nq'(\theta)/p'(\theta), \qquad Var_\theta(Y) = \frac{n}{p'(\theta)^3} \begin{vmatrix} q'(\theta) & p'(\theta) \\ q''(\theta) & p''(\theta) \end{vmatrix}$$

We derive the formula for $f_Y(y;\theta)$ in the discrete case where

$$f_Y(y;\theta) = \sum_{\mathbf{x}\in S_n(y)} f(\mathbf{x};\theta).$$

Here $S_n(y) = \{\mathbf{x} \in \mathcal{S}^n : \sum_{i=1}^n K(x_i) = y\}$. So $exp[p(\theta)y + nq(\theta)]$ pulls out as a common factor. The remaining factor is

$$R(y) = \sum_{\mathbf{x}\in S_n(y)} exp[\sum_{i=1}^n H(x_i)].$$

For the expectation and variance we differentiate the identity $\int_{\mathcal{S}^n} f(\mathbf{x};\theta)d\mathbf{x} = 1$ with respect to $\theta$ and then differentiate again to get:

$$0 = \int (p'(\theta)y + nq'(\theta))R(y)exp[p(\theta)y + nq(\theta)]dy$$

$$= \int (p'(\theta)y + nq'(\theta))f_Y(y;\theta)dy,$$

$$0 = \int (p''(\theta)y + nq''(\theta)) + (p'(\theta)y + nq'(\theta))^2 f_Y(y;\theta)dy.$$

That is,

$$p'(\theta)E(Y) + nq'(\theta) = 0,$$

$$p''(\theta)E(Y) + nq'(\theta) + E[(p'(\theta)Y + nq'(\theta))^2] = 0.$$

Therefore, $E(Y) = -nq'/p'$ and

$$(p')^2 E(Y^2) - (p')^2 E(Y)^2 \; = \\ -p'' E(Y) - nq'' - 2np'q'E(Y) - (nq')^2 - (nq')^2.$$

The last three terms the right add to zero and the remaining two add to $(np''q' - nq''p')/p'$.

Thus, $Var(Y) = (np''q' - nq''p')/(p')^3$.

Using the uniqueness of the Laplace transform it is possible to show

HMC Theorem 7.5.2: For the above exponential family the sufficient statistic $Y = \sum_{i=1}^{n} K(X_i)$ is complete.

Thus, if we can find an unbiased estimate which is a function of $Y$, i.e. $\phi(Y)$ such that $E_\theta(\phi(Y)) = \theta$ then $\phi(Y)$ is ae unique and is an MVUE.

HMC Example 7.5.1: The pmf for the Poisson distribution $Poiss(\theta)$ can be written

$$f(x, \theta) = exp[\ln(\theta)x + \ln(1/x!) + (-th)]$$

and so it is an exponential family.

HMC Example 7.5.2: With variance $\sigma^2$ known the pdf for the normal distribution $\mathcal{N}(\theta, \sigma^2)$ can be written

$$f(x, \theta) = exp[\frac{\theta}{\sigma^2}x - \frac{x^2}{\sigma^2} - \ln(\sqrt{2\pi\sigma^2}) - \frac{\theta^2}{\sigma^2}].$$

and so it is an exponential family.

Finally, we observe, following HMC Sec. 8.2, that if $p(\theta)$ is an increasing function then the likelihood ratio is:

$$\Lambda(\theta_1, \theta_2, \mathbf{x}) = \frac{exp[p(\theta_1)\sum_i K(x_i) + \sum_i H(x_i) + nq(\theta_1)]}{exp[p(\theta_2)\sum_i K(x_i) + \sum_i H(x_i) + nq(\theta_2)]}$$

$$= exp[(p(\theta_1) - p(\theta_2))\sum_i K(x_i) + n(q(\theta_1) - q(\theta_2))].$$

and for $\theta_1 < \theta_2$ this is a decreasing function of $\sum_{i=1}^n K(x_i)$.

If $p(\theta)$ is a decreasing function then for $\theta_1 < \theta_2$ the likelihood ratio is an increasing function of $\sum_{i=1}^n K(x_i)$.

Thus, when $p(\theta)$ is a monotone function, the exponential family has monotone likelihood ratio in the sufficient statistic $\sum_{i=1}^n K(x_i)$.