# Rao-Cramér lower bound and asymptotic normality of the maximum likelihood estimator

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
Department of Diagnostic Radiology
McGill University

sahirbhatnagar.com

April 14, 2021

- Recap
    - Point Estimation
    - Maximum Likelihood Estimator
    - MLE example

- Methods of Evaluating Estimators
    - Uniformly Minimum Variance Unbiased Estimator
    - Challenges in finding the UMVUE
    - Cramér–Rao Lower Bound (CRLB)
    - Fisher Information
    - Correlation review
    - CRLB proof
    - Attainment of CRLB

- Efficiency and Asymptotic Normality of MLE
    - Efficiency

- Recap
  - Point Estimation
  - Maximum Likelihood Estimator
  - MLE example
- Methods of Evaluating Estimators
  - Uniformly Minimum Variance Unbiased Estimator
  - Challenges in finding the UMVUE
  - Cramér–Rao Lower Bound (CRLB)
  - Fisher Information
  - Correlation review
  - CRLB proof
  - Attainment of CRLB
- Efficiency and Asymptotic Normality of MLE
  - Efficiency

# Point Estimation

- When sampling is from a population described by a pdf or a pmf $f(x|\theta)$, knowledge of $\theta$ yields knowledge of the entire population.

# Point Estimation

- When sampling is from a population described by a pdf or a pmf $f(x|\theta)$, knowledge of $\theta$ yields knowledge of the entire population.
- Hence it is natural to seek a method of finding a good estimator of the point $\theta$, i.e., a good *point estimator*

# Point Estimation

- When sampling is from a population described by a pdf or a pmf $f(x|\theta)$, knowledge of $\theta$ yields knowledge of the entire population.
- Hence it is natural to seek a method of finding a good estimator of the point $\theta$, i.e., a good *point estimator*

**Definition 1.1 (Point estimator).**

A *point estimator* is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic is a point estimator.

# Point Estimation

- When sampling is from a population described by a pdf or a pmf $f(x|\theta)$, knowledge of $\theta$ yields knowledge of the entire population.
- Hence it is natural to seek a method of finding a good estimator of the point $\theta$, i.e., a good *point estimator*

**Definition 1.1 (Point estimator).**

A *point estimator* is any function $W(X_1, \dots, X_n)$ of a sample; that is, any statistic is a point estimator.

- An *estimator* is a function of the sample, while an *estimate* is the realized value of an estimator (a number) that is obtained when a sample is actually taken

# Point Estimation

- When sampling is from a population described by a pdf or a pmf $f(x|\theta)$, knowledge of $\theta$ yields knowledge of the entire population.
- Hence it is natural to seek a method of finding a good estimator of the point $\theta$, i.e., a good *point estimator*

**Definition 1.1 (Point estimator).**

A *point estimator* is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic is a point estimator.

- An *estimator* is a function of the sample, while an *estimate* is the realized value of an estimator (a number) that is obtained when a sample is actually taken
- Notationally, when a sample is taken, an estimator is a function of the random variables $X_1, \ldots, X_n$, while an estimate is a function of the realized values $x_1, \ldots, x_n$.

# Methods of Finding Estimators

- In some cases, there will be an obvious or natural candidate for a point estimator of a particular parameter, e.g., the sample mean ($\bar{X}$) as a point estimator of the population mean ($\mu$)

# Methods of Finding Estimators

- In some cases, there will be an obvious or natural candidate for a point estimator of a particular parameter, e.g., the sample mean ($\bar{X}$) as a point estimator of the population mean ($\mu$)
- However, when we leave a simple case like this, intuition may not only desert us, it may also lead us astray. Therefore, it is useful to have some techniques that will at least give us some reasonable candidates for consideration.

# Methods of Finding Estimators

- In some cases, there will be an obvious or natural candidate for a point estimator of a particular parameter, e.g., the sample mean ($\bar{X}$) as a point estimator of the population mean ($\mu$)
- However, when we leave a simple case like this, intuition may not only desert us, it may also lead us astray. Therefore, it is useful to have some techniques that will at least give us some reasonable candidates for consideration.
- Maximum Likelihood Estimators is, by far, the most popular technique for deriving estimators

# Maximum Likelihood Estimator

- Let $X_1, \ldots, X_n$ be an iid sample from a population with pdf or pmf $f(x|\Theta)$ where $\Theta \equiv (\theta_1, \ldots, \theta_k)$ have unknown values and $\mathbf{x} = x_1, \ldots, x_n$ are the observed sample values. The likelihood function is defined by

$$L(\Theta|\mathbf{x}) = L(\theta_1, \ldots, \theta_k | x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k) \qquad (1)$$

### Definition 1.2 (Maximum Likelihood Estimator).

For each sample point $\mathbf{x}$, let $\widehat{\Theta}$ be a parameter value at which $L(\Theta|\mathbf{x})$ attains its maximum as a function of $\Theta$, with $\mathbf{x}$ held fixed. A *maximum likelihood estimator (MLE)* of the parameter $\Theta$ based on a sample $\mathbf{X}$ is $\widehat{\Theta}(\mathbf{X})$.

- $\widehat{\Theta}(\mathbf{x})$ is called the maximum likelihood estimate of $\Theta$ based on data $\mathbf{x}$
- $\widehat{\Theta}(\mathbf{X})$ is the maximum likelihood estimator (MLE) of $\Theta$

## Example 1 (Poisson MLE).

*Let $X_1, \dots, X_n$ be iid Poisson($\lambda$). Then the likelihood function is*

$$L(\lambda|\mathbf{x}) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \tag{2}$$

*The log-likehood is given by*

$$\ell(\lambda|\mathbf{x}) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log x_i! \tag{3}$$

# Example 1 (Poisson MLE).

*Let $X_1, \ldots, X_n$ be iid Poisson($\lambda$). Then the likelihood function is*

$$L(\lambda|\mathbf{x}) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \tag{2}$$

*The log-likehood is given by*

$$\ell(\lambda|\mathbf{x}) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log x_i! \tag{3}$$

*Taking the derivative of (3) with respect to $\lambda$ we get:*

$$\frac{\partial \ell(\lambda|\mathbf{x})}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n \tag{4}$$

## Example 1 (Poisson MLE).

*Let $X_1, \ldots, X_n$ be iid Poisson($\lambda$). Then the likelihood function is*

$$L(\lambda|\mathbf{x}) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \tag{2}$$

*The log-likehood is given by*

$$\ell(\lambda|\mathbf{x}) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log x_i! \tag{3}$$

*Taking the derivative of (3) with respect to $\lambda$ we get:*

$$\frac{\partial \ell(\lambda|\mathbf{x})}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n \tag{4}$$

*Setting (4) equal to zero for the first order condition, and solving for $\lambda$, yields $\hat{\lambda} = \bar{x}$.*

## Example 1 (Poisson MLE).

*Let $X_1, \ldots, X_n$ be iid Poisson($\lambda$). Then the likelihood function is*

$$L(\lambda|\mathbf{x}) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \tag{2}$$

*The log-likehood is given by*

$$\ell(\lambda|\mathbf{x}) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log x_i! \tag{3}$$

*Taking the derivative of (3) with respect to $\lambda$ we get:*

$$\frac{\partial \ell(\lambda|\mathbf{x})}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n \tag{4}$$

*Setting (4) equal to zero for the first order condition, and solving for $\lambda$, yields $\hat{\lambda} = \bar{x}$. In order to verify that this is the MLE for $\lambda$ we take the second derivative of (3) with respect to $\lambda$:*

$$\frac{\partial^2 \ell(\lambda|\mathbf{x})}{\partial \lambda^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2} \tag{5}$$

*Since (5) is negative and the log-likelihood is concave, $\hat{\lambda} = \bar{x}$ solves for the global maximum.*

# Poisson Unbiased Estimation

## Theorem 1.3 (Relationships between a statistic and population parameter).

*Let $X_1, \ldots, X_n$ be a random sample from a population with mean $\mu$ and variance $\sigma^2 < \infty$.*

$$\begin{aligned} &a.\ \mathrm{E}\bar{X} = \mu \\ &b.\ \operatorname{Var}\bar{X} = \frac{\sigma^2}{n} \\ &c.\ \mathrm{E}S^2 = \sigma^2 \end{aligned}$$

*where $\bar{X}$ and $S^2$ are the sample mean and sample variance, respectively.*

# Poisson Unbiased Estimation

### Theorem 1.3 (Relationships between a statistic and population parameter).

*Let $X_1, \ldots, X_n$ be a random sample from a population with mean $\mu$ and variance $\sigma^2 < \infty$.*

$$
\begin{aligned}
&a.\ \mathrm{E}\bar{X} = \mu \\
&b.\ \operatorname{Var}\bar{X} = \tfrac{\sigma^2}{n} \\
&c.\ \mathrm{E}S^2 = \sigma^2
\end{aligned}
$$

*where $\bar{X}$ and $S^2$ are the sample mean and sample variance, respectively.*

Applying Theorem 1.3 to $X_1, \ldots, X_n$ iid Poisson($\lambda$) we have

$$
\begin{aligned}
\mathrm{E}_\lambda \bar{X} &= \lambda, &\qquad \text{for all } \lambda, \\
\mathrm{E}_\lambda S^2 &= \lambda, &\qquad \text{for all } \lambda,
\end{aligned}
$$

so both $\bar{X}$ and $S^2$ are unbiased estimators of $\lambda$.

# Poisson MLE using `optim`

We can use the `stats::optim` function in `R` to find the MLE, provided we have a likelihood function. The `optim` can maximize (or minimize) an objective function using many different algorithms. This is referred to as **solving the objective function numerically**.

```r
# define the objective function
ll.poisson <- function(lambda, x) {
  sum(x) * log(lambda) - length(x) * lambda
}

# generate some data
data <- rpois(1e6, 5)

# by default optim finds the min, but the
# negative min is the max therefore we need
# to use list(fnscale = -1)
opt <- optim(par = 2,
             fn = ll.poisson,
             method = "BFGS",
             control = list(fnscale = -1),
             x = data)

# compare numeric vs. analytical solutions
c(numerical = opt$par,
  xbar = mean(data),
  samplevar = var(data))

## numerical      xbar samplevar
##  5.000532  5.000535  5.015680
```

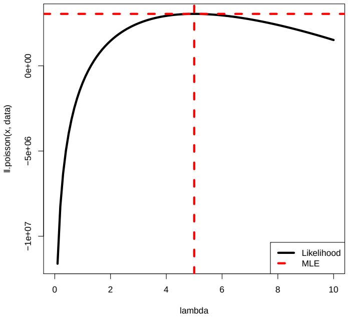
# Poisson MLE using optim

We can use the stats::optim function in R to find the MLE, provided we have a likelihood function. The optim can maximize (or minimize) an objective function using many different algorithms. This is referred to as **solving the objective function numerically**.

```r
# define the objective function
ll.poisson <- function(lambda, x) {
  sum(x) * log(lambda) - length(x) * lambda
}

# generate some data
data <- rpois(1e6, 5)

# by default optim finds the min, but the
# negative min is the max therefore we need
# to use list(fnscale = -1)
opt <- optim(par = 2,
             fn = ll.poisson,
             method = "BFGS",
             control = list(fnscale = -1),
             x = data)

# compare numeric vs. analytical solutions
c(numerical = opt$par,
  xbar = mean(data),
  samplevar = var(data))
```

```
## numerical      xbar samplevar
##  5.000532  5.000535  5.015680
```

```r
curve(ll.poisson(x, data), 0,10, xlab = "lambda",lwd = 4)
abline(h = opt$value, v = opt$par, lty = 2, lwd = 4, col = "red")
```



- Recap
    - Point Estimation
    - Maximum Likelihood Estimator
    - MLE example
- Methods of Evaluating Estimators
    - Uniformly Minimum Variance Unbiased Estimator
    - Challenges in finding the UMVUE
    - Cramér–Rao Lower Bound (CRLB)
    - Fisher Information
    - Correlation review
    - CRLB proof
    - Attainment of CRLB
- Efficiency and Asymptotic Normality of MLE
    - Efficiency

# Methods of Evaluating Estimators

- As we saw in the Poisson example, $\bar{X}$ and $S^2$ are both unbiased estimators of $\lambda$. *How do we choose between these estimators?*

# Methods of Evaluating Estimators

- As we saw in the Poisson example, $\bar{X}$ and $S^2$ are both unbiased estimators of $\lambda$. *How do we choose between these estimators?*
- The mean squared error (MSE) of an estimator $W$ of a parameter $\theta$ is the function of $\theta$ defined by

$$\mathrm{E}_\theta(W-\theta)^2 = \mathrm{Var}_\theta W + (Bias_\theta W)^2$$

# Methods of Evaluating Estimators

- As we saw in the Poisson example, $\bar{X}$ and $S^2$ are both unbiased estimators of $\lambda$. *How do we choose between these estimators?*
- The mean squared error (MSE) of an estimator $W$ of a parameter $\theta$ is the function of $\theta$ defined by

$$\mathrm{E}_\theta(W-\theta)^2 = \mathrm{Var}_\theta W + (Bias_\theta W)^2$$

- If $W_1$ and $W_2$ are both unbiased estimators of a parameter $\theta$, that is, $\mathrm{E}_\theta W_1 = \mathrm{E}_\theta W_2 = \theta$, then their MSE is equal to their variances $\rightarrow$ we should choose the estimator with the smaller variance

# Methods of Evaluating Estimators

- As we saw in the Poisson example, $\bar{X}$ and $S^2$ are both unbiased estimators of $\lambda$. *How do we choose between these estimators?*
- The mean squared error (MSE) of an estimator $W$ of a parameter $\theta$ is the function of $\theta$ defined by

$$\mathrm{E}_\theta (W - \theta)^2 = \mathrm{Var}_\theta W + (Bias_\theta W)^2$$

- If $W_1$ and $W_2$ are both unbiased estimators of a parameter $\theta$, that is, $\mathrm{E}_\theta W_1 = \mathrm{E}_\theta W_2 = \theta$, then their MSE is equal to their variances $\rightarrow$ we should choose the estimator with the smaller variance
- If we can find an unbiased estimator with uniformly smallest variance – a best unbiased estimator – then our task is done.

# Uniformly Minimum Variance Unbiased Estimator

- The goal of this section is to investigate a method for finding a “best” unbiased estimator defined in the following way:

**Definition 1.4 (UMVUE).**

$W^*(\mathbf{X})$ is the best unbiased estimator, or uniformly minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$ if

1. $\mathrm{E}_\theta\left[W^*(\mathbf{X}) \mid \theta\right] = \tau(\theta)$ for all $\theta$ (unbiased)
2. $\mathrm{Var}\left[W^*(\mathbf{X}) \mid \theta\right] \leq \mathrm{Var}[W(\mathbf{X}) \mid \theta]$ for all $\theta$, where $W$ is any other unbiased estimator of $\tau(\theta)$ (minimum variance).

# Uniformly Minimum Variance Unbiased Estimator

- The goal of this section is to investigate a method for finding a "best" unbiased estimator defined in the following way:

**Definition 1.4 (UMVUE).**

$W^*(\mathbf{X})$ is the best unbiased estimator, or uniformly minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$ if

1. $\mathrm{E}_\theta\left[W^*(\mathbf{X}) \mid \theta\right] = \tau(\theta)$ for all $\theta$ (unbiased)
2. $\mathrm{Var}\left[W^*(\mathbf{X}) \mid \theta\right] \leq \mathrm{Var}[W(\mathbf{X}) \mid \theta]$ for all $\theta$, where $W$ is any other unbiased estimator of $\tau(\theta)$ (minimum variance).

- Finding a best unbiased estimator (if one exists) is not an easy task as we'll see in the next example

# Poisson Unbiased Estimation Revisited

- Recall that by applying Theorem 1.3 to $X_1, \ldots, X_n$ iid Poisson($\lambda$)

$$\begin{aligned} \mathrm{E}_\lambda \bar{X} &= \lambda, \qquad \text{for all } \lambda, \\ \mathrm{E}_\lambda S^2 &= \lambda, \qquad \text{for all } \lambda, \end{aligned}$$

  so both $\bar{X}$ and $S^2$ are unbiased estimators of $\lambda$.

# Poisson Unbiased Estimation Revisited

- Recall that by applying Theorem 1.3 to $X_1, \ldots, X_n$ iid Poisson$(\lambda)$

$$\begin{aligned} \mathrm{E}_\lambda \bar{X} &= \lambda, \qquad && \text{for all } \lambda, \\ \mathrm{E}_\lambda S^2 &= \lambda, \qquad && \text{for all } \lambda, \end{aligned}$$

  so both $\bar{X}$ and $S^2$ are unbiased estimators of $\lambda$.

- Again from Theorem 1.3, we have $\mathrm{Var}_\lambda \bar{X} = \lambda / n$

# Poisson Unbiased Estimation Revisited

- Recall that by applying Theorem 1.3 to $X_1, \dots, X_n$ iid Poisson($\lambda$)

$$\begin{aligned} \mathrm{E}_\lambda \bar{X} &= \lambda, \qquad \text{for all } \lambda, \\ \mathrm{E}_\lambda S^2 &= \lambda, \qquad \text{for all } \lambda, \end{aligned}$$

  so both $\bar{X}$ and $S^2$ are unbiased estimators of $\lambda$.
- Again from Theorem 1.3, we have $\mathrm{Var}_\lambda \bar{X} = \lambda/n$
- $\mathrm{Var}_\lambda \left[S^2\right] = \frac{1}{n}\mu_4 + \frac{\mu_2^2(n-3)}{n(n-1)}$ where $\mu_j$ is the $j$ th centered moment → lengthy calculation

# Poisson Unbiased Estimation Revisited

- Even if we can establish that $\bar{X}$ is better than $S^2$, consider the class of estimators
$$W_a(\bar{X}, S^2) = a\bar{X} + (1-a)S^2.$$
  For every constant $a$, $\mathrm{E}_\lambda W_a(\bar{X}, S^2) = \lambda$, so we now have infinitely many unbiased estimators of $\lambda$.

# Poisson Unbiased Estimation Revisited

- Even if we can establish that $\bar{X}$ is better than $S^2$, consider the class of estimators
$$W_a(\bar{X}, S^2) = a\bar{X} + (1-a)S^2.$$
For every constant $a$, $\mathrm{E}_\lambda W_a(\bar{X}, S^2) = \lambda$, so we now have infinitely many unbiased estimators of $\lambda$.

- Even if $\bar{X}$ is better than $S^2$, is it better than every $W_a(\bar{X}, S^2)$?

# Poisson Unbiased Estimation Revisited

- Even if we can establish that $\bar{X}$ is better than $S^2$, consider the class of estimators
$$W_a(\bar{X}, S^2) = a\bar{X} + (1-a)S^2.$$
For every constant $a$, $\mathrm{E}_\lambda W_a(\bar{X}, S^2) = \lambda$, so we now have infinitely many unbiased estimators of $\lambda$.

- Even if $\bar{X}$ is better than $S^2$, is it better than every $W_a(\bar{X}, S^2)$?
- Furthermore, how can we be sure that there are not other, better, unbiased estimators lurking about?

# Poisson Unbiased Estimation Revisited

- Even if we can establish that $\bar{X}$ is better than $S^2$, consider the class of estimators

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

  For every constant $a$, $\mathrm{E}_\lambda W_a(\bar{X}, S^2) = \lambda$, so we now have infinitely many unbiased estimators of $\lambda$.

- Even if $\bar{X}$ is better than $S^2$, is it better than every $W_a(\bar{X}, S^2)$?
- Furthermore, how can we be sure that there are not other, better, unbiased estimators lurking about?
- This example shows some of the problems that might be encountered in trying to find a best unbiased estimator, and perhaps that a more comprehensive approach is desirable.

# How to find the Best Unbiased Estimator?

- Find the lower bound of variances of any unbiased estimator of $\tau(\theta)$, say $B(\theta)$.
- If $W^*$ is an unbiased estimator of $\tau(\theta)$ and satisfies $\text{Var}\left[W^*(\mathbf{X}) \mid \theta\right] = B(\theta)$, then $W^*$ is the best unbiased estimator.
- This is the appraoch taken with the use of the Cramér–Rao Lower Bound. The names Cramér and Rao are often interchanged depending on the textbook and professor's training.

# Cramér–Rao Inequality

**Theorem 1.5 (Cramér–Rao Lower Bound (CRLB)).**

*Let* $X_1, \cdots, X_n$ *be iid with common pdf* $f(\mathbf{x} \mid \theta)$. *Let* $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ *be a statistic with mean* $\mathrm{E}_\theta W(\mathbf{X}) = k(\theta)$ *satisfying*

$$\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X}) = \frac{d}{d\theta}\int_{x\in\mathcal{X}} W(\mathbf{x})f(\mathbf{x} \mid \theta)\,d\mathbf{x} = \int_{x\in\mathcal{X}} \frac{\partial}{\partial\theta} W(\mathbf{x})f(\mathbf{x} \mid \theta)\,d\mathbf{x}$$

*and*

$$\mathrm{Var}_\theta\, W(\mathbf{X}) < \infty$$

# Cramér–Rao Inequality

**Theorem 1.5 (Cramér–Rao Lower Bound (CRLB)).**

*Let* $X_1, \cdots, X_n$ *be iid with common pdf* $f(\mathbf{x} \mid \theta)$*. Let* $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ *be a statistic with mean* $\mathrm{E}_\theta W(\mathbf{X}) = k(\theta)$ *satisfying*

$$\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X}) = \frac{d}{d\theta}\int_{x\in\mathcal{X}} W(\mathbf{x})f(\mathbf{x} \mid \theta)\,d\mathbf{x} = \int_{x\in\mathcal{X}} \frac{\partial}{\partial\theta} W(\mathbf{x})f(\mathbf{x} \mid \theta)\,d\mathbf{x}$$

*and*

$$\mathrm{Var}_\theta\ W(\mathbf{X}) < \infty$$

*Then, a lower bound of* $\mathrm{Var}_\theta\ W(\mathbf{X})$ *is*

$$\mathrm{Var}_\theta\ W(\mathbf{X}) \geq \frac{\left[k'(\theta)\right]^2}{nI(\theta)}$$

*where* $I(\theta)$ *is the* ***Fisher information***

# Fisher Information

- Before we prove the CRLB, we must first derive the Fisher information, and recall the formula for the covariance of two random variables as well as the Cauchy-Schwarz Inequality

# Fisher Information

- Before we prove the CRLB, we must first derive the Fisher information, and recall the formula for the covariance of two random variables as well as the Cauchy-Schwarz Inequality

**Theorem 1.6 (Fisher information).**

*If $X$ is a random variable with pdf $f(x \mid \theta)$ which satisfies certain regularity assumptions then*

$$\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right] = 0$$

*and*

$$\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right] = -\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right]$$

*The quantity $I(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right]$ is called the information number or Fisher information.*

# Proof of Theorem 1.6 (1/4)

## Proof.

First, we see that

$$\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right] = \int \left\{\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right\} f(x \mid \theta)\, dx$$

# Proof of Theorem 1.6 (1/4)

## Proof.

First, we see that

$$\begin{aligned}
\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right] &= \int\left\{\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right\} f(x \mid \theta)\,dx \\
&= \int \frac{\frac{\partial}{\partial\theta}f(x \mid \theta)}{f(x \mid \theta)} f(x \mid \theta)\,dx
\end{aligned}$$

# Proof of Theorem 1.6 (1/4)

### Proof.

First, we see that

$$\begin{aligned}
\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right] &= \int \left\{\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right\} f(x \mid \theta)\,dx \\
&= \int \frac{\frac{\partial}{\partial\theta} f(x \mid \theta)}{f(x \mid \theta)} f(x \mid \theta)\,dx \\
&= \int \frac{\partial}{\partial\theta} f(x \mid \theta)\,dx
\end{aligned}$$

# Proof of Theorem 1.6 (1/4)

## Proof.

First, we see that

$$\begin{aligned}
\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right] &= \int\left\{\frac{\partial}{\partial\theta}\log f(x\mid\theta)\right\}f(x\mid\theta)dx \\
&= \int\frac{\frac{\partial}{\partial\theta}f(x\mid\theta)}{f(x\mid\theta)}f(x\mid\theta)dx \\
&= \int\frac{\partial}{\partial\theta}f(x\mid\theta)dx \\
&= \frac{d}{d\theta}\int f(x\mid\theta)dx = \frac{d}{d\theta}1 = 0.
\end{aligned}$$

# Proof of Theorem 1.6 cont'd (2/4)

Note that

$$\frac{\partial^2}{\partial\theta^2}\left(\log f(x \mid \theta)\right) = \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right\}$$

# Proof of Theorem 1.6 cont'd (2/4)

Note that

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\left(\log f(x \mid \theta)\right) &= \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right\} \\ &= \frac{\partial}{\partial\theta}\left\{f(x \mid \theta)^{-1}\frac{\partial}{\partial\theta}f(x \mid \theta)\right\}\end{aligned}$$

# Proof of Theorem 1.6 cont'd (2/4)

Note that

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\left(\log f(x \mid \theta)\right) &= \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right\} \\
&= \frac{\partial}{\partial\theta}\left\{f(x \mid \theta)^{-1}\frac{\partial}{\partial\theta}f(x \mid \theta)\right\} \\
&= f(x \mid \theta)^{-1}\frac{\partial^2}{\partial\theta^2}f(x \mid \theta) + \frac{\partial}{\partial\theta}\left[f(x \mid \theta)^{-1}\right]\frac{\partial}{\partial\theta}f(x \mid \theta)
\end{aligned}$$

# Proof of Theorem 1.6 cont'd (2/4)

Note that

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\left(\log f(x \mid \theta)\right) &= \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right\} \\
&= \frac{\partial}{\partial\theta}\left\{f(x \mid \theta)^{-1}\frac{\partial}{\partial\theta}f(x \mid \theta)\right\} \\
&= f(x \mid \theta)^{-1}\frac{\partial^2}{\partial\theta^2}f(x \mid \theta) + \frac{\partial}{\partial\theta}\left[f(x \mid \theta)^{-1}\right]\frac{\partial}{\partial\theta}f(x \mid \theta) \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(x \mid \theta)}{f(x \mid \theta)} - \frac{\frac{\partial}{\partial\theta}f(x \mid \theta)}{(f(x \mid \theta))^2}\frac{\partial}{\partial\theta}f(x \mid \theta)
\end{aligned}$$

# Proof of Theorem 1.6 cont'd (2/4)

Note that

$$
\begin{aligned}
\frac{\partial^2}{\partial\theta^2}(\log f(x \mid \theta)) &= \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right\} \\
&= \frac{\partial}{\partial\theta}\left\{f(x \mid \theta)^{-1}\frac{\partial}{\partial\theta}f(x \mid \theta)\right\} \\
&= f(x \mid \theta)^{-1}\frac{\partial^2}{\partial\theta^2}f(x \mid \theta) + \frac{\partial}{\partial\theta}\left[f(x \mid \theta)^{-1}\right]\frac{\partial}{\partial\theta}f(x \mid \theta) \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(x \mid \theta)}{f(x \mid \theta)} - \frac{\frac{\partial}{\partial\theta}f(x \mid \theta)}{(f(x \mid \theta))^2}\frac{\partial}{\partial\theta}f(x \mid \theta) \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(x \mid \theta)}{f(x \mid \theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(x \mid \theta)}{(f(x \mid \theta))}\right)^2
\end{aligned}
$$

# Proof of Theorem 1.6 cont'd (2/4)

Note that

$$
\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\left(\log f(x \mid \theta)\right) &= \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right\} \\
&= \frac{\partial}{\partial\theta}\left\{f(x \mid \theta)^{-1}\frac{\partial}{\partial\theta}f(x \mid \theta)\right\} \\
&= f(x \mid \theta)^{-1}\frac{\partial^2}{\partial\theta^2}f(x \mid \theta) + \frac{\partial}{\partial\theta}\left[f(x \mid \theta)^{-1}\right]\frac{\partial}{\partial\theta}f(x \mid \theta) \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(x \mid \theta)}{f(x \mid \theta)} - \frac{\frac{\partial}{\partial\theta}f(x \mid \theta)}{(f(x \mid \theta))^2}\frac{\partial}{\partial\theta}f(x \mid \theta) \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(x \mid \theta)}{f(x \mid \theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(x \mid \theta)}{(f(x \mid \theta))}\right)^2 \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(x \mid \theta)}{f(x \mid \theta)} - \left(\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right)^2
\end{aligned}
$$

# Proof of Theorem 1.6 cont'd (3/4)

Then, we have

$$\mathrm{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X\mid\theta)}{f(X\mid\theta)}\right]=\int\frac{\frac{\partial^2}{\partial\theta^2}f(X\mid\theta)}{f(X\mid\theta)}f(x\mid\theta)dx$$

# Proof of Theorem 1.6 cont'd (3/4)

Then, we have

$$\begin{aligned}
\mathrm{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X \mid \theta)}{f(X \mid \theta)}\right] &= \int \frac{\frac{\partial^2}{\partial\theta^2}f(X \mid \theta)}{f(X \mid \theta)} f(x \mid \theta)\,dx \\
&= \int \frac{\partial^2}{\partial\theta^2} f(x \mid \theta)\,dx
\end{aligned}$$

# Proof of Theorem 1.6 cont'd (3/4)

Then, we have

$$
\begin{aligned}
\mathrm{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X \mid \theta)}{f(X \mid \theta)}\right] &= \int \frac{\frac{\partial^2}{\partial\theta^2}f(X \mid \theta)}{f(X \mid \theta)} f(x \mid \theta)\,dx \\
&= \int \frac{\partial^2}{\partial\theta^2} f(x \mid \theta)\,dx \\
&= \int \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta} f(x \mid \theta)\right\} dx
\end{aligned}
$$

# Proof of Theorem 1.6 cont'd (3/4)

Then, we have

$$
\begin{aligned}
\mathrm{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2} f(X \mid \theta)}{f(X \mid \theta)}\right] &= \int \frac{\frac{\partial^2}{\partial\theta^2} f(X \mid \theta)}{f(X \mid \theta)} f(x \mid \theta)\, dx \\
&= \int \frac{\partial^2}{\partial\theta^2} f(x \mid \theta)\, dx \\
&= \int \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta} f(x \mid \theta)\right\} dx \\
&= \frac{\partial}{\partial\theta} \int \frac{\partial}{\partial\theta} f(x \mid \theta)\, dx
\end{aligned}
$$

# Proof of Theorem 1.6 cont'd (3/4)

Then, we have

$$
\begin{aligned}
\mathrm{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X \mid \theta)}{f(X \mid \theta)}\right] &= \int \frac{\frac{\partial^2}{\partial\theta^2}f(X \mid \theta)}{f(X \mid \theta)} f(x \mid \theta)\,dx \\
&= \int \frac{\partial^2}{\partial\theta^2} f(x \mid \theta)\,dx \\
&= \int \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta} f(x \mid \theta)\right\} dx \\
&= \frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta} f(x \mid \theta)\,dx \\
&= \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\int f(x \mid \theta)\,dx\right\} = \frac{\partial}{\partial\theta}[0] = 0
\end{aligned}
$$

# Proof of Theorem 1.6 cont'd (4/4)

So, it follows that

$$\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X\mid\theta)\right]=\mathrm{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X\mid\theta)}{f(X\mid\theta)}\right]-\mathrm{E}_\theta\left[\frac{\left(\frac{\partial}{\partial\theta}f(X\mid\theta)\right)^2}{(f(X\mid\theta))^2}\right]$$

# Proof of Theorem 1.6 cont'd (4/4)

So, it follows that

$$\begin{aligned}
\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right] &= \mathrm{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X \mid \theta)}{f(X \mid \theta)}\right] - \mathrm{E}_\theta\left[\frac{\left(\frac{\partial}{\partial\theta}f(X \mid \theta)\right)^2}{(f(X \mid \theta))^2}\right] \\
&= 0 - \mathrm{E}_\theta\left[\left(\frac{\frac{\partial}{\partial\theta}f(X \mid \theta)}{f(X \mid \theta)}\right)^2\right]
\end{aligned}$$

# Proof of Theorem 1.6 cont'd (4/4)

So, it follows that

$$
\begin{aligned}
\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X\mid\theta)\right] &= \mathrm{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X\mid\theta)}{f(X\mid\theta)}\right] - \mathrm{E}_\theta\left[\frac{\left(\frac{\partial}{\partial\theta}f(X\mid\theta)\right)^2}{(f(X\mid\theta))^2}\right] \\
&= 0 - \mathrm{E}_\theta\left[\left(\frac{\frac{\partial}{\partial\theta}f(X\mid\theta)}{f(X\mid\theta)}\right)^2\right] \\
&= -\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right)^2\right]
\end{aligned}
$$

□

# Correlation review

- $\mathrm{E}[X] = \mu_X, \mathrm{E}[Y] = \mu_Y, \mathrm{Var}[X] = \sigma_X^2, \mathrm{Var}[Y] = \sigma_Y^2$
- Assume $0 < \sigma_X^2 < \infty$ and $0 < \sigma_Y^2 < \infty$

### Definition 1.7 (Correlation coefficient).

The correlation of $X$ and $Y$ is the number defined by

$$\rho_{XY} = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \sigma_Y}$$

The value $\rho_{XY}$ is also called the correlation coefficient.

### Theorem 1.8 (Bounds of $\rho_{X,Y}$).

*For any random variables $X$ and $Y$,*

(a) $-1 \leq \rho_{XY} \leq 1$

(b) $|\rho_{XY}| = 1$ *if and only if there exists numbers $a \neq 0$ and $b$ such that $P(Y = aX + b) = 1$. If $\rho_{XY} = 1$ then $a > 0$, and if $\rho_{XY} = -1$ then $a < 0$*

# Now we are ready to prove Theorem 1.5 (1/3)

## Proof.

- Let $D_i = \frac{\partial}{\partial\theta}\log f(X_i \mid \theta)$ so that

  $$D = \frac{\partial}{\partial\theta}\left\{\log\prod_{i=1}^{n} f(X_i \mid \theta)\right\} = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i \mid \theta) = \sum_{i=1}^{n} D_i$$

  Since Theorem 1.8 implies $\{\mathrm{Cov}[W(\mathbf{X}), D]\}^2 \leq \mathrm{Var}[W(\mathbf{X})]\,\mathrm{Var}[D]$ it follows that

  $$\mathrm{Var}[W(\mathbf{X})] \geq \frac{\{\mathrm{Cov}[W(\mathbf{X}), D]\}^2}{\mathrm{Var}[D]}$$

- Since $E[D] = \sum_{i=1}^{n} E\left[D_i\right] \overset{\mathit{Thm.}\ 1.6}{=} 0$, we have

  $$\mathrm{Cov}[W(\mathbf{X}), D] = \mathrm{E}[W(\mathbf{X})D]$$

- Note that we can write $D = \sum_{i=1}^{n}\frac{\frac{\partial}{\partial\theta}f(X_i|\theta)}{f(X_i|\theta)}$.

# Proof of Theorem 1.5 (2/3)

$$k(\theta) = \int \cdots \int W(\boldsymbol{x}) \prod_{i=1}^{n} f(x_i \mid \theta)\, dx_1 \cdots dx_n$$

Differentiating with respect to $\theta$, we obtain

$$\begin{aligned}
k'(\theta) &= \int \cdots \int W(\boldsymbol{x}) \frac{\partial}{\partial\theta} \prod_{i=1}^{n} f(x_i \mid \theta)\, dx_1 \cdots dx_n \\
&= \int \cdots \int W(\boldsymbol{x}) \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial\theta} f(x_i \mid \theta) \prod_{j \neq i} f(x_j \mid \theta) \right\} dx_1 \cdots dx_n \\
&= \int \cdots \int W(\boldsymbol{x}) \underbrace{\sum_{i=1}^{n} \left\{ \frac{\frac{\partial}{\partial\theta} f(x_i \mid \theta)}{f(x_i \mid \theta)} \right\}}_{D} f(\boldsymbol{x} \mid \theta)\, dx_1 \cdots dx_n \\
&= E[W(\mathbf{X}) D]
\end{aligned}$$

# Proof of Theorem 1.5 (3/3)

Furthermore, we have

$$
\begin{aligned}
&\mathrm{Var}[D] = \mathrm{E}\left[D^2\right] = \mathrm{E}\left[\left(\sum_{i=1}^n D_i\right)^2\right] \\
&= \mathrm{E}\left[\sum_i D_i \sum_j D_j\right] = \mathrm{E}\left[\sum_i \sum_j D_i D_j\right] \\
&= \sum_{i=1}^n \sum_{j=1}^n \mathrm{E}\left[D_i D_j\right] \\
&= \sum_{i=1}^n \mathrm{E}\left[D_i^2\right] + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \mathrm{E}\left[D_i D_j\right] \\
&= \sum_{i=1}^n \mathrm{E}\left[\left(\frac{\partial}{\partial\theta} \log f(x \mid \theta)\right)^2\right] + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \mathrm{E}\left[D_i D_j\right] \\
&= \sum_{i=1}^n I(\theta) + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \mathrm{E}\left[D_i\right] E\left[D_j\right] = nI(\theta) + 0.
\end{aligned}
$$

So, we have $\mathrm{Var}[W(\mathbf{X})] \geq \frac{\{\mathrm{Cov}[W(\mathbf{X}), D]\}^2}{\mathrm{Var}[D]} = \frac{\left\{k'(\theta)\right\}^2}{nI(\theta)}$. □

# A useful corollary

**Corollary 1.9.**

*Under the assumptions of Theorem 1.5, if $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$ is an unbiased estimator of $\theta$, so that $k(\theta) = \theta$, then the Rao-Cramér inequality becomes*

$$Var(W(\boldsymbol{X})) \geq \frac{1}{nI(\theta)}.$$

# Poisson example revisited (again) I

## Example 2.

*Let $X_1, \ldots, X_n$ be iid Poisson $(\lambda)$. Find the Cramér-Rao lower bound on the variance of unbiased estimators of $\lambda$. Also, find the MLE and show that it attains the Cramér-Rao lower bound. Since $\frac{\partial^2}{\partial\lambda^2} \log f(x \mid \lambda) =$*

$$\frac{\partial^2}{\partial\lambda^2}\left[\log\left\{\lambda^x e^{-\lambda}(x!)^{-1}\right\}\right] = \frac{\partial^2}{\partial\lambda^2}[x\log\lambda - \lambda - \log(x!)] = -\frac{x}{\lambda^2}$$

*we have*

$$\mathrm{E}\left[\frac{\partial^2}{\partial\lambda^2}\log f(X \mid \lambda)\right] = \mathrm{E}\left[-\frac{1}{\lambda^2}X\right] = -\frac{1}{\lambda^2}\mathrm{E}[X] = -\frac{1}{\lambda^2}\lambda = -\frac{1}{\lambda}$$

*By Theorem 1.6,*

$$\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\lambda^2}\log f(X \mid \lambda)\right] = \frac{1}{\lambda}$$

# Poisson example revisited (again) II

**Example 2.**

*So the Cramér-Rao lower bound for an unbiased estimator in the iid case is*

$$\frac{\left(\frac{d}{d\theta}\mathrm{E}_\theta[W(\boldsymbol{X})]\right)^2}{n\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right]} = \frac{1}{n\left(\frac{1}{\lambda}\right)} = \frac{\lambda}{n}$$

*The MLE of* $\lambda$ *is* $\hat{\lambda} = \bar{X}$ *and* $\mathrm{Var}[\bar{X}] = \frac{\mathrm{Var}[X_1]}{n} = \frac{\lambda}{n}$ *so it attains the CRLB.*

# CRLB for Normal Distribution Variance Estimator I

## Example 3.

*Let $X_1, \ldots, X_n$ be iid Normal $\left(\mu, \sigma^2\right)$ random variables. Find the Cramér-Rao lower bound on unbiased estimators of $\sigma^2$. Does $S^2$ satisfy the CRLB?*

$$
\begin{aligned}
\frac{\partial^2}{\partial\left(\sigma^2\right)^2} \log f\left(x \mid \mu, \sigma^2\right) &= \frac{\partial^2}{\partial\left(\sigma^2\right)^2}\left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}(x-\mu)^2\right] \\
&= \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}
\end{aligned}
$$

*Theorem 1.6 implies that*

$$
\begin{aligned}
\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f\left(X \mid \mu, \sigma^2\right)\right)^2\right] &= -\mathrm{E}\left[\frac{\partial^2}{\partial(\sigma^2)^2}\log f\left(X \mid \mu, \sigma^2\right)\right] \\
&= -\mathrm{E}\left[\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\right]
\end{aligned}
$$

# CRLB for Normal Distribution Variance Estimator II

**Example 3.**

$$= -\mathrm{E}\left[\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\right]$$

$$= -\frac{1}{2\sigma^4} + \frac{\mathrm{E}\left[(X-\mu)^2\right]}{\sigma^6}$$

$$= -\frac{1}{2\sigma^4} + \frac{\sigma^2}{\sigma^6} = \frac{1}{2\sigma^4}$$

*Thus, the CRLB is*

$$\frac{1}{n\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right]} = \frac{2\sigma^4}{n}.$$

*So, $S^2$ does not satisfy the CRLB since*

$$\mathrm{Var}\left[S^2\right] = \frac{2\sigma^4}{n-1} = \frac{n}{n-1}\left(\frac{2\sigma^4}{n}\right) > \frac{2\sigma^4}{n} = \mathit{CRLB}$$

- In the previous example we are left with an incomplete answer; that is, is there a better unbiased estimator of $\sigma^2$ than $S^2$, or is the CRLB unattainable?

- In the previous example we are left with an incomplete answer; that is, is there a better unbiased estimator of $\sigma^2$ than $S^2$, or is the CRLB unattainable?
- The conditions for attainment of the CRLB are actually quite simple.
- Recall that the bound follows from an application of the Cauchy-Schwarz Inequality, so conditions for attainment of the bound are the conditions for equality in the Cauchy-Schwarz Inequality.
- The following Corollary is a useful tool because it gives us a way of finding a best unbiased estimator

- In the previous example we are left with an incomplete answer; that is, is there a better unbiased estimator of $\sigma^2$ than $S^2$, or is the CRLB unattainable?
- The conditions for attainment of the CRLB are actually quite simple.
- Recall that the bound follows from an application of the Cauchy-Schwarz Inequality, so conditions for attainment of the bound are the conditions for equality in the Cauchy-Schwarz Inequality.
- The following Corollary is a useful tool because it gives us a way of finding a best unbiased estimator

### Corollary 1.10 (Attainment).

*Let* $X_1, \cdots, X_n$ *be iid with pdf* $/\mathrm{pmf} f_X(x \mid \theta)$, *where* $f_X(x \mid \theta)$ *satisfies the assumptions of the Cramer-Rao Theorem. Let* $L(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$ *denote the likelihood function. If* $W(\mathbf{X})$ *is unbiased for* $\tau(\theta)$, *then* $W(\mathbf{X})$ *attains the Cramer-Rao lower bound if and only if*

$$\frac{\partial}{\partial \theta} \log L(\theta \mid \mathbf{x}) = S_n(\mathbf{x} \mid \theta) = a(\theta)[W(\mathbf{X}) - \tau(\theta)]$$

*for some function* $a(\theta)$.

# Proof of Corollary 1.10 (1/2)

Proof.

We used Cauchy-Schwarz inequality to prove that

$$\left[\operatorname{Cov}\left\{W(\mathbf{X}), \frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X} \mid \theta)\right\}\right]^2 \le \operatorname{Var}[W(\mathbf{X})] \operatorname{Var}\left[\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X} \mid \theta)\right]$$

In Cauchy-Schwarz inequality, the equality satisfies if and only if there is a linear relationship between the two variables (see Theorem 1.8), that is

$$\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \frac{\partial}{\partial\theta}\log L(\theta \mid \mathbf{x}) = a(\theta) W(\mathbf{x}) + b(\theta)$$

# Proof of Corollary 1.10 (2/2)

$$\begin{aligned}
E\left[\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X}\mid\theta)\right] &= E\left[S_n(\mathbf{X}\mid\theta)\right] = 0\\
E[a(\theta)\,W(\mathbf{X}) + b(\theta)] &= 0\\
a(\theta)E[W(\mathbf{X})] + b(\theta) &= 0\\
a(\theta)\tau(\theta) + b(\theta) &= 0\\
b(\theta) &= -a(\theta)\tau(\theta)\\
\frac{\partial}{\partial\theta}\log L(\theta\mid\mathbf{x}) &= a(\theta)\,W(\mathbf{x}) - a(\theta)\tau(\theta) = a(\theta)[W(\mathbf{x}) - \tau(\theta)]
\end{aligned}$$

□

# Continuation of Example 3

Is CRLB for $\sigma^2$ attainable?

$$
\begin{aligned}
L\left(\sigma^2 \mid \mathbf{x}\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right] \\
\log L\left(\sigma^2 \mid \mathbf{x}\right) &= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n} \frac{\left(x_i-\mu\right)^2}{2\sigma^2} \\
\frac{\partial \log L\left(\sigma^2 \mid \mathbf{x}\right)}{\partial\sigma^2} &= -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} + \sum_{i=1}^{n} \frac{\left(x_i-\mu\right)^2}{2\left(\sigma^2\right)^2} \\
&= -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{\left(x_i-\mu\right)^2}{2\sigma^4} \\
&= \frac{n}{2\sigma^4}\left(\frac{\sum_{i=1}^{n}\left(x_i-\mu\right)^2}{n} - \sigma^2\right) \\
&= a\left(\sigma^2\right)\left(W(\mathbf{x}) - \sigma^2\right)
\end{aligned}
$$

# Continuation of Example 3

Therefore,

1. If $\mu$ is known, the best unbiased estimator for $\sigma^2$ is $\sum_{i=1}^{n} \left(x_i - \mu\right)^2 / n$, and it attains the Cramer-Rao lower bound, i.e.

$$\operatorname{Var}\left[\frac{\sum_{i=1}^{n} \left(X_i - \mu\right)^2}{n}\right] = \frac{2\sigma^4}{n}$$

2. If $\mu$ is not known, the Cramer-Rao lower-bound cannot be attained.

# Bernoulli example I

# Bernoulli example II

## Example 4 (Bernoulli).

*Let* $X_1, \cdots, X_n \overset{i.i.d.}{\sim}$ *Bernoulli* $(p)$. *Is* $\bar{X}$ *the best unbiased estimator of* $p$? *Does it attain the Cramer-Rao lower bound?*

$$
\begin{aligned}
L(p \mid \mathbf{x}) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
\log L(p \mid \mathbf{x}) &= \log \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
&= \sum_{i=1}^{n} \log\left[p^{x_i}(1-p)^{1-x_i}\right] \\
&= \sum_{i=1}^{n} \left[x_i \log p + (1-x_i)\log(1-p)\right] \\
&= \log p \sum_{i=1}^{n} x_i + \log(1-p)\left(n - \sum_{i=1}^{n} x_i\right)
\end{aligned}
$$

# Bernoulli example III

**Example 4 (Bernoulli).**

$$
\begin{aligned}
\frac{\partial}{\partial p} \log L(p \mid \mathbf{x}) &= \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} \\
&= \frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} \\
&= \frac{(1-p)n\bar{x} - np(1-\bar{x})}{p(1-p)} \\
&= \frac{n(\bar{x}-p)}{p(1-p)} \\
&= a(p)[W(\mathbf{x}) - \tau(p)]
\end{aligned}
$$

*where* $a(p) = \dfrac{n}{p(1-p)}$, $W(\mathbf{x}) = \bar{x}, \tau(p) = p$. *Therefore,* $\bar{X}$ *is the best unbiased estimator for* $p$ *and attains the Cramer-Rao lower bound.*

Recap

- Point Estimation
- Maximum Likelihood Estimator
- MLE example

Methods of Evaluating Estimators

- Uniformly Minimum Variance Unbiased Estimator
- Challenges in finding the UMVUE
- Cramér–Rao Lower Bound (CRLB)
- Fisher Information
- Correlation review
- CRLB proof
- Attainment of CRLB

Efficiency and Asymptotic Normality of MLE

- Efficiency

# Efficient Estimator and Efficiency

**Definition 1.11 (Efficient Estimator).**

Let $W(\mathbf{X})$ be an unbiased estimator of a parameter $\theta$ in the case of point estimation. The statistic $W(\mathbf{X})$ is called an **efficient estimator** of $\theta$ if and only if the variance of $W(\mathbf{X})$ attains the Rao-Cramér lower bound.

# Efficient Estimator and Efficiency

## Definition 1.11 (Efficient Estimator).

Let $W(\mathbf{X})$ be an unbiased estimator of a parameter $\theta$ in the case of point estimation. The statistic $W(\mathbf{X})$ is called an **efficient estimator** of $\theta$ if and only if the variance of $W(\mathbf{X})$ attains the Rao-Cramér lower bound.

## Definition 1.12 (Efficiency).

In cases in which we can differentiate with respect to a parameter under an integral or summation symbol, the ratio of the Rao–Cramér lower bound to the actual variance of any unbiased estimator of a parameter is called the **efficiency** of that estimator.

# Poisson example revisited (again again)

**Example 5.**

- *Recall from Example 2 where we showed that* $W(\boldsymbol{X}) = \bar{X}$ *is an MLE of* $\lambda$
- *We also showed that the CRLB is* $\lambda/n$
- *The variance of* $W(\boldsymbol{X})$ *is equal to* $\lambda/n$
- *Therefore, by definition,* $W(\boldsymbol{X}) = \bar{X}$ *is an efficient estimator of* $\lambda$

# Poisson example revisited (again again)

**Example 5.**

- *Recall from Example 2 where we showed that* $W(\boldsymbol{X}) = \bar{X}$ *is an MLE of* $\lambda$
- *We also showed that the CRLB is* $\lambda/n$
- *The variance of* $W(\boldsymbol{X})$ *is equal to* $\lambda/n$
- *Therefore, by definition,* $W(\boldsymbol{X}) = \bar{X}$ *is an efficient estimator of* $\lambda$

In the above example, we were able to obtain the MLE in closed form along with their distributions and, hence, moments. This is often not the case. Maximum likelihood estimators, however, have an asymptotic normal distribution. In fact, MLEs are asymptotically efficient as we will now show, i.e.., MLE achieves the lowest possible variance → the CRLB.

# Asymptotic Normality

**Theorem 1.13 (Asymptotic normality of MLE).**

*Assume $X_1, \ldots, X_n$ are iid with pdf $f(x|\theta_0)$ for $\theta_0 \in \Omega$ and some regularity conditions are satisfied. Suppose further that the Fisher information satisfies $0 < I(\theta_0) < \infty$. Then any consistent sequence of solutions of the MLE equations satisfies*

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{D} N\left(0, \frac{1}{I(\theta_0)}\right)$$

As we can see, the asymptotic variance/dispersion of the estimate around true parameter will be smaller when Fisher information is larger.

## Proof.

Define the normalized log-likelihood function and its first and second derivatives with respect to $\theta$ as

$$L_n(\theta) = \frac{1}{n} \log f_X(x|\theta) \tag{6}$$

$$L_n'(\theta) = \frac{\partial}{\partial\theta}\left(\frac{1}{n} \log f_X(x|\theta)\right) \tag{7}$$

$$L_n''(\theta) = \frac{\partial^2}{\partial\theta^2}\left(\frac{1}{n} \log f_X(x \mid \theta)\right). \tag{8}$$

Since MLE $\hat{\theta}$ is maximizer of $L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i \mid \theta)$, we have $L_n'(\hat{\theta}) = 0$. Let us use the Mean Value Theorem

$$\frac{f(a) - f(b)}{a - b} = f'(c) \text{ or } f(a) = f(b) + f'(c)(a - b) \text{ for } c \in [a, b]$$

with $f(\theta) = L_n'(\theta)$, $a = \hat{\theta}$ and $b = \theta_0$. Then we can write,

$$0 = L_n'(\hat{\theta}) = L_n'(\theta_0) + L_n''\left(\hat{\theta}_1\right)\left(\hat{\theta} - \theta_0\right)$$

for some $\hat{\theta}_1 \in \left[\hat{\theta}, \theta_0\right]$.

From here we get that

$$\hat{\theta} - \theta_0 = -\frac{L_n'(\theta_0)}{L_n''\left(\hat{\theta}_1\right)} \rightarrow \sqrt{n}\left(\hat{\theta} - \theta_0\right) = -\frac{\sqrt{n}L_n'(\theta_0)}{L_n''\left(\hat{\theta}_1\right)} \tag{9}$$

The numerator in (9) :

$$\begin{aligned}
\sqrt{n}L_n'(\theta_0) &= \sqrt{n}\left(\frac{1}{n}\left[\frac{\partial}{\partial\theta}\log f_X(X \mid \theta_0)\right]\right) \\
&= \sqrt{n}\left(\frac{1}{n}\left[\frac{\partial}{\partial\theta}\log\prod_{i=1}^{n} f_X(X_i \mid \theta_0)\right]\right) \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}\log f_X(X_i \mid \theta_0)\right]\right) \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}\log f_X(X_i \mid \theta_0)\right] - \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f_X(X_1 \mid \theta_0)\right]\right) \\
&\rightarrow^d \mathcal{N}\left(0, \mathrm{Var}\left[\frac{\partial}{\partial\theta}\log f_X(X_1 \mid \theta_0)\right]\right)
\end{aligned}$$

converges in distribution by Central Limit Theorem.

Next, let us consider the denominator in (9).

$$
\begin{aligned}
L_n''(\theta) &= \frac{1}{n}\left(\frac{\partial^2}{\partial\theta^2}\log f_X(X \mid \theta)\right) \\
&= \frac{1}{n}\left(\frac{\partial^2}{\partial\theta^2}\log\prod_{i=1}^{n} f_X(X_i \mid \theta)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2}{\partial\theta^2}\log f_X(X_i \mid \theta)\right) \\
&\to^p \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_X(X_1 \mid \theta)\right].
\end{aligned}
$$

In the last step we invoke WLLN without loss of generality on $X_1$. Since $\hat{\theta}_1 \in \left[\hat{\theta}, \theta_0\right]$ by constructions and we assume consistency $\hat{\theta} \to \theta_0$, we have $\hat{\theta}_1 \to \theta_0$. Taken together we have

$$
L_n''(\theta) \to^p \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_X(X_1 \mid \theta)\right] = -I(\theta_0)
$$

To summarize, we have shown that

$$\sqrt{n}L_n'\left(\theta_0\right) \rightarrow^d \mathcal{N}\left(0, I\left(\theta_0\right)\right)$$

and

$$L_n''(\tilde{\theta}) \rightarrow^p -I\left(\theta_0\right)$$

We invoke Slutsky's theorem, and we're done:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \rightarrow^d \mathcal{N}\left(\frac{1}{I(\theta_0)}\right)$$

As discussed in the introduction, asymptotic normality immediately implies

$$\hat{\theta}_n \rightarrow^d \mathcal{N}\left(\theta_0, I_n\left(\theta_0\right)^{-1}\right)$$

As our finite sample size $n$ increases, the MLE becomes more concentrated or its variance becomes smaller and smaller. In the limit, MLE achieves the lowest possible variance, the Cramér-Rao lower bound. □

# Generalization of efficiency in the asymptotic case

## Definition 1.14.

Let $X_1, \ldots, X_n$ be independent and identically distributed with probability density function $f(x;\theta)$. Suppose $\hat{\theta}_{1n} = \hat{\theta}_{1n}(X_1, \ldots, X_n)$ is an estimator of $\theta_0$ such that $\sqrt{n}\left(\hat{\theta}_{1n} - \theta_0\right) \xrightarrow{D} N\left(0, \sigma^2_{\hat{\theta}_{1n}}\right)$. Then

(a) The asymptotic efficiency of $\hat{\theta}_{1n}$ is defined to be

$$e\left(\hat{\theta}_{1n}\right) = \frac{1/I(\theta_0)}{\sigma^2_{\hat{\theta}_{1n}}}$$

(b) The estimator $\hat{\theta}_{1n}$ is said to be asymptotically efficient if the ratio in part (a) is 1.

(c) Let $\hat{\theta}_{2n}$ be another estimator such that $\sqrt{n}\left(\hat{\theta}_{2n} - \theta_0\right) \xrightarrow{D} N\left(0, \sigma^2_{\hat{\theta}_{2n}}\right)$. Then the asymptotic relative efficiency $(ARE)$ of $\hat{\theta}_{1n}$ to $\hat{\theta}_{2n}$ is the reciprocal of the ratio of their respective asymptotic variances; i.e.,

$$e\left(\hat{\theta}_{1n}, \hat{\theta}_{2n}\right) = \frac{\sigma^2_{\hat{\theta}_{2n}}}{\sigma^2_{\hat{\theta}_{1n}}}$$

# Poisson example revisited (again again again)

- Assume we observe i.i.d. samples $X_1, \ldots, X_n$ drawn from a Poisson $(\lambda)$ distribution with true parameter $\lambda = \lambda_0$. The loglikelihood is:

  $$\ell\left(\lambda; X_1, \ldots, X_n\right) = \sum_{i=1}^{n} -\lambda + X_i \log(\lambda) + \log\left(X_i!\right)$$

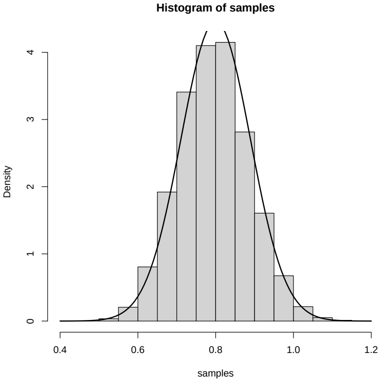
  Taking the derivative with respect to $\lambda$, setting it equal to zero, and solving for $\lambda$ gives the mle as the sample mean, $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The Fisher information is:

  $$I_n(\lambda) = E_\lambda\left[-\frac{d^2}{d\lambda^2}\ell(\lambda)\right] = \sum_{i=1}^{n} E\left[X_i/\lambda^2\right] = \frac{n}{\lambda}$$

  The main result says that (for large $n$ ), $\hat{\lambda}$ is approximately $N\left(\lambda, \frac{1}{n\lambda}\right)$. We illustrate this by simulation:

# Poisson

```r
num.iterations          <- 7000
lambda.truth            <- 0.8
num.samples.per.iter    <- 100
samples                 <- numeric(num.iterations)
for(iter in seq_len(num.iterations)) {
  samples[iter] <- mean(rpois(num.samples.per.iter, lambda.truth))
}
hist(samples, freq=F)
curve(dnorm(x, mean=lambda.truth,sd=sqrt(lambda.truth/num.samples.per.iter) ),
   0.4, 1.2, lwd=2, xlab = "", ylab = "", add = T)
```


Histogram of samples
Density
samples