SCIENTIA AC LABORE

THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

# Course Notes for STAT3001 Mathematical Statistics

**Contributors:**

Michael Ciccotosto-Camp

NAME2

The University of Queensland

School of Mathematics and Physics

## Contents

Symbols and Notation . . . . . iii
Review . . . . . 1
Useful Formulae and Theorems . . . . . 1
Common Distributions . . . . . 2
Common Probabilistic Properties and Identities . . . . . 3
Probabilistic Properties . . . . . 3
Probabilistic Identities . . . . . 4
Point Estimation . . . . . 5
Methods of Finding Estimates Introduction . . . . . 5
Method of Moments . . . . . 10
Maximum Likelihood Estimates . . . . . 11
Methods of Evaluating Estimators . . . . . 13
Sufficiency and Unbiasedness . . . . . 15
Consistency . . . . . 16
Large-Sample Comparisons of Estimators . . . . . 17
References . . . . . 19

## Symbols and Notation

Matrices are capitalized bold face letters while vectors are lowercase bold face letters.

| *Syntax* | *Meaning* |
|---|---|
| $\triangleq$ | An equality which acts as a statement |
| $\lvert\boldsymbol{A}\rvert$ | The determinate of a matrix. |
| $\boldsymbol{x}^\intercal, \boldsymbol{X}^\intercal$ | The transpose operator. |
| $\boldsymbol{x}^*, \boldsymbol{X}^*$ | The hermitian operator. |
| $\boldsymbol{a}. * \boldsymbol{b}$ or $\boldsymbol{A}. * \boldsymbol{B}$ | Element-wise vector (matrix) multiplication, similar to Matlab. |
| $\propto$ | Proportional to. |
| $\nabla$ or $\nabla_{\boldsymbol{f}}$ | The partial derivative (with respect to $\boldsymbol{f}$). |
| $\nabla\nabla$ or $H(f)$ | The Hessian. |
| $\sim$ | Distributed according to, example $X \sim \mathcal{N}(0,1)$ |
| $\overset{\text{iid}}{\sim}$ | Identically and independently distributed according to, example $X_1, X_2, \dots X_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ |
| $\mathbf{0}$ or $\mathbf{0}_n$ or $\mathbf{0}_{n\times m}$ | The zero vector (matrix) of appropriate length (size) or the zero vector of length $n$ or the zero matrix with dimensions $n \times m$. |
| $\mathbf{1}$ or $\mathbf{1}_n$ or $\mathbf{1}_{n\times m}$ | The one vector (matrix) of appropriate length (size) or the one vector of length $n$ or the one matrix with dimensions $n \times m$. |
| $\mathbb{1}_A(x)$ | The indicator function. $\mathbb{1}_A(x) = 1$ if $x \in A$, 0 otherwise. |

$\boldsymbol{A}_{(\cdot,\cdot)}$ Index slicing to extract a submatrix from the elements of $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, similar to indexing slicing from the python and Matlab programming languages. Each parameter can receive a single value or a ’slice’ consisting of a start and an end value separated by a semicolon. The first and second parameter describe what row and columns should be selected, respectively. A single value means that only values from the single specified row/column should be selected. A slice tells us that all rows/columns between the provided range should be selected. Additionally if now start and end values are specified in the slice then all rows/columns should be selected. For example, the slice $\boldsymbol{A}_{(1:3,j:j')}$ is the submatrix $\mathbb{R}^{3\times(j'-j+1)}$ matrix containing the first three rows of $\boldsymbol{A}$ and columns $j$ to $j'$. As another example, $\boldsymbol{A}_{(:,j)}$ is the $j^{th}$ column of $\boldsymbol{A}$.

$\boldsymbol{A}^{\dagger}$ Denotes the unique psuedo inverse or Moore-Penore inverse of $\boldsymbol{A}$.

$\mathbb{C}$ The complex numbers.

$\operatorname{diag}(\boldsymbol{w})$ Vector argument, a diagonal matrix containing the elements of vector $\boldsymbol{w}$.

$\operatorname{diag}(\boldsymbol{W})$ Matrix argument, a vector containing the diagonal elements of the matrix $\boldsymbol{W}$.

$\mathbb{E}$ or $\mathbb{E}_{q(x)}\left[z(x)\right]$ Expectation, or expectation of $z(x)$ where $x \sim q(x)$.

$\mathbb{R}$ The real numbers.

$\operatorname{tr}(\boldsymbol{A})$ The trace of a matrix.

$\mathbb{V}$ or $\mathbb{V}_{q(x)}\left[z(x)\right]$ Variance, the variance of $z(x)$ when $x \sim q(x)$.

$\mathbb{Z}$ The integers, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$.

$\Omega$ The sample space.

## Review

Theorems and defintions here are mostly concepts seen before from other courses.

**Useful Formulae and Theorems.**

(Geometric Series)

$$\sum_{k=0}^{n-1} r^k = \left(\frac{1-r^n}{1-r}\right)$$

or

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r} \quad \text{with} \quad |r| < 1$$

(Euler's formula) $$e^{ix} = \cos x + i \sin x$$

(Newton's Binomial formula) $$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$$

**Theorem 1** (Young's inequality for products)**.** *If $a \geq 0$ and $b \geq 0$ are nonnegative real numbers and if $p > 1$ and $q > 1$ are real numbers such that $\frac{1}{p} + \frac{1}{p} = 1$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

*Equality holds iff $a^p = b^q$.*

**Common Distributions.** Common distributions seen from prior courses. Notations mostly borrowed from STAT2003.

| *Name* | *Notation* | *Support* | *pf* | *Expectation* | *Variance* |
|---|---|---|---|---|---|
| Bernoulli | $\mathsf{Ber}(p)$ | $\{0,1\}$ | $p^k(1-p)^{1-k}$ | $p$ | $p(1-p)$ |
| Binomial | $\mathsf{Bin}(n,p)$ | $\{0,\dots,n\}$ | $\binom{n}{k}p^k(1-p)^{n-k}$ | $np$ | $np(1-p)$ |
| Negative-Binomial | $\mathsf{NB}(r,p)$ | $\mathbb{N}_0$ | $\binom{x+r-1}{x}p^x(1-p)^r$ | $\frac{rp}{1-p}$ | $\frac{rp}{(1-p)^2}$ |
| Geometric | $\mathsf{Geo}(n,p)$ | $\mathbb{N}_0$ | $(1-p)^k p$ | $\frac{1-p}{p}$ | $\frac{1-p}{p^2}$ |
| Poisson | $\mathsf{Poi}(\lambda)$ | $\mathbb{N}_0$ | $\frac{\lambda^x}{x!}e^{-\lambda}$ | $\lambda$ | $\lambda$ |
| Uniform | $\mathsf{U}[a,b]$ | $[a,b]$ | $\frac{1}{b-a}$ | $\frac{a+b}{2}$ | $\frac{(a-b)^2}{12}$ |
| Exponential | $\mathsf{Exp}(\lambda)$ | $\mathbb{R}^+$ | $\lambda e^{-\lambda x}$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda}$ |
| Normal | $\mathsf{N}(\mu,\sigma^2)$ | $\mathbb{R}$ | $\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$ | $\mu$ | $\sigma^2$ |
| Gamma | $\mathsf{Gam}(\alpha,\lambda)$ | $\mathbb{R}^+$ | $\frac{\lambda^\alpha x^{\alpha-1}\exp(-\lambda x)}{\Gamma(\alpha)}$ | $\frac{\alpha}{\lambda}$ | $\frac{\alpha}{\lambda^2}$ |
| Chi-Squared | $\chi^2_n$ | $\mathbb{R}^+$ | $\frac{x^{\frac{n}{2}-1}\exp(-\frac{1}{2}x)}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}$ | $n$ | $2n$ |
| White-Noise | $\mathsf{WN}(\mu,\sigma^2)$ | NA | NA | $\mu$ | $\sigma^2$ |

**Common Probabilistic Properties and Identities.** Common probabilistic properties seen from prior courses.

*Probabilistic Properties.* For any random variables, the following hold.

(1) $$\mathbb{E}(X) = \int_0^\infty (1 - F(X))\, dx$$

(2) $$\mathbb{E}\left(aX + b\right) = a\mathbb{E}X + b$$

(3) $$\mathbb{E}\left(g(X) + h(X)\right) = \mathbb{E}g(X) + \mathbb{E}h(X)$$

(4) $$\text{Var}(X) = \mathbb{E}X^2 - \left(\mathbb{E}X\right)^2$$

(5) $$\text{Var}(aX + b) = a^2\text{Var}(X)$$

(6) $$\text{Cov}(X, Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y$$

(7) $$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

(8) $$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$$

(9) $$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$$

(10) $$|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

(11) $$|\text{Cov}(XY)|^2 \leq \text{Var}(X)\text{Var}(Y)$$

(12) $$\mathbb{P}\left(A \mid B\right) = \frac{\mathbb{P}\left(A \cap B\right)}{\mathbb{P}\left(B\right)}$$

(Bayes' Theorem) $$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

(13) $$\mathbb{P}\left(A_1, \ldots, A_n\right) = \mathbb{P}\left(A_1\right)\mathbb{P}\left(A_2 \mid A_1\right)\mathbb{P}\left(A_3 \mid A_1, A_2\right) \cdots \mathbb{P}\left(A_n \mid A_1, A_2, \ldots, A_{n-1}\right)$$

(14)

Let $\Omega = \bigcup\limits_{i=1}^{n} B_i$ (that is $B_i$ partitions the sample space) then

(TLoP) $$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$$

(TLoE) $$\mathbb{E}(A) = \sum_{i=1}^{n} \mathbb{E}(A \mid B_i)\mathbb{P}(B_i)$$

which, when TLoP used in conjunction with Bayes' Rule gives

(15) $$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{n} \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}.$$

If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{WN}(\mu, \sigma^2)$ and $S_n = \sum_{i=1}^{n} X_i$, then for all $\varepsilon > 0$

(Weak Law of Large Numbers) $$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0.$$

If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{WN}(\mu, \sigma^2)$ and $S_n = \sum_{i=1}^n X_i$, then for all $x \in \mathbb{R}$

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}}\right) \leq x = \Phi(x). \tag{CLT}$$

If $X$ is a random variable and $h$ is a convex function then

$$h(\mathbb{E}(X)) \leq \mathbb{E}(h(X)). \tag{Jensens Inequality}$$

*Probabilistic Identities.* If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(p)$ then

$$\sum_{i=1}^{n} X_i \sim \mathsf{Bin}(n, p). \tag{16}$$

If $X \sim \mathsf{Bin}(n, p)$ and $Y \sim \mathsf{Bin}(m, p)$, then $X + Y \sim \mathsf{Bin}(n + m, p)$.

If $X \sim \mathsf{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$, then $X + Y \sim \mathsf{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

If $X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ then

$$\sum_{i=1}^{n} X_i^2 = \chi_n^2. \tag{17}$$

**Methods of Finding Estimates Introduction.**

**Definition 2** (Statistic)**.** *Let $X_1, \dots, X_n$ be a random sample of size $n$ from a population and let $T(x_1, \dots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \dots, X_n)$. The the random variable or random vector $Y = T(X_1, \dots, X_n)$ is called a* **statistic***. The probability distribution of a statistic $Y$ is called the* **sampling distribution** *of $Y$* [Cas01, page 211].

**Definition 3** (Sample Mean)**.** *The* **sample mean** *is the arthicmetic average of the values in a random sample. It is usually denoted by*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{18}$$

[Cas01, page 212].

**Definition 4** (Sample Variance and Standard Deviation)**.** *The* **sample variance** *is the statistic defined by*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{19}$$

*The* **sample standard deviation** *is the statistic defined by $S = \sqrt{S^2}$* [Cas01, page 212].

**Definition 5** (Sufficient Statistic)**.** *A statistic $T(\boldsymbol{X})$ is a* **sufficient statistic** *for $\theta$ if the conditional distribution of the sample $\boldsymbol{X}$ given the value of $T(\boldsymbol{X})$ does not depend on $\theta$* [Cas01, page 272].

**Theorem 6.** *If $p(\boldsymbol{x} \mid \theta)$ is the joint pdf or pmf of $\boldsymbol{X}$ and $q(\theta \mid \theta)$ is the pdf or pmf of $T(\boldsymbol{X})$, then $T(\boldsymbol{X})$ is a sufficient statistic for $\theta$ if, for every $\boldsymbol{x}$ in the sample space, the ratio $p(\boldsymbol{x} \mid \theta)/q(T(\boldsymbol{x}) \mid \theta)$ is a constant function of $\theta$* [Cas01, page 274].

**Theorem 7** (Factorization Theorem)**.** *Let $f(\boldsymbol{x} \mid \theta)$ denote the joint pdf or pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for $\theta$, if and only if there exist function $g(t \mid \theta)$ and $h(\boldsymbol{x})$ such that, for all sample points $\boldsymbol{x}$ and all parameter points $\theta$,*

$$f(\boldsymbol{x} \mid \theta) = g(T(\boldsymbol{x}) \mid \theta)h(\boldsymbol{x})$$

[Cas01, page 276].

*Example* 8 (Uniform Sufficient Statistic)*.* Example taken from [Cas01, page 277] and can also be found on tutorial sheet 3. Let $X_1, \dots, X_n$ be iid observations from the discrete uniform distribution on $1, \dots, \theta$. That is, the unknown parameter, $\theta$, is a positive integer and the pmf of $X_i$ is

$$f(x \mid \theta) = \begin{cases} \frac{1}{\theta}, & x = 1, 2, \dots \theta \\ 0, & \text{otherwise} \end{cases}.$$

The restriction $x_i \in \{1, \dots, \theta\}$ for $i = 1, \dots, n$ can be re-expressed as $x_i \in \{1, 2, \dots\}$ for $i = 1, \dots n$ (note that there is no $\theta$ in this restriction) and $\max_i x_i \le \theta$. If we define $T(\boldsymbol{x}) = \max_i x_i = x_{(n)}$,

$$h(x) = \begin{cases} 1, & x_i \in \{1, \dots, \theta\} \text{ for } i = 1, \dots, n \\ 0, & \text{otherwise} \end{cases}$$

and

$$g(t \mid \theta) = \begin{cases} \theta^{-n} & t \leq \theta \\ 0, & \text{otherwise} \end{cases}.$$

It is easily verified that $f(\boldsymbol{x} \mid \theta) = g(T(\boldsymbol{x}) \mid \theta)$ for all $\boldsymbol{x}$ and $\theta$. Thus, according to Theorem 7, the largest order statistic, $T(\boldsymbol{X}) = X_{(n)}$, is a sufficient statistic in this problem. This type of analysis can sometimes be carried out more clearly and concisely using indicator function. Let $\mathbb{N}$ be the set of natural numbers (discluding $0$) and $\mathbb{N}_\theta$ be the natural numbers up to and including $\theta$. Then the joint pmf of $X_1, \ldots, X_n$ is

$$f(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} \theta^{-1} \mathbb{1}_{N_\theta}(x_i) = \theta^{-n} \prod_{i=1}^{n} \mathbb{1}_{N_\theta}(x_i).$$

Defining $T(\boldsymbol{x}) = x_{(n)}$, we see that

$$\prod_{i=1}^{n} \mathbb{1}_{N_\theta}(x_i) = \left( \prod_{i=1}^{n} \mathbb{1}_{N}(x_i) \right) \mathbb{1}_{N_\theta}(T(x))$$

thus providing the factorization

$$f(\boldsymbol{x} \mid \theta) = \theta^{-n} \mathbb{1}_{N_\theta}(T(x)) \left( \prod_{i=1}^{n} \mathbb{1}_{N}(x_i) \right).$$

The first factor depends on $x_1, \ldots, x_n$ only through the value of $T(\boldsymbol{x}) = x_{(n)}$, and the second factor does not depend on $\theta$. Again, according to Theorem 7, $T(\boldsymbol{X}) = X_{(n)}$, is a sufficient statistic in this problem.

**Definition 9** (Likelihood, Log-Likelihood and Score Function)**.** *Let $f(\boldsymbol{x} \mid \theta)$ denote the joint pdf or pmf of the sample $\boldsymbol{X} = (X_1, \ldots, X_2)$. Then, given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function of $\theta$ defined by*

$$L(\theta \mid \boldsymbol{x}) = f(\boldsymbol{x} \mid \theta)$$

*is called the* **likelihood function** [Cas01, page 290]. *For a given outcome $\boldsymbol{x}$ of $\boldsymbol{X}$, the* **log-likelihood function**, *denoted $l$, is the natural logarithm of the likelihood function*

$$l(\theta \mid \boldsymbol{x}) = \ln L(\theta \mid \boldsymbol{x}) = \ln f(\boldsymbol{x} \mid \theta).$$

*It's gradient with respect to $\theta$, denoted $S$, is called the* **score function**

$$S(\theta \mid \boldsymbol{x}) = \nabla_\theta \, l(\theta \mid \boldsymbol{x}) \frac{\nabla_\theta f(\boldsymbol{x} \mid \theta)}{f(\boldsymbol{x} \mid \theta)}$$

[Kro13, page 165].

**Theorem 10.** *Under regularity conditions*

$$\mathbb{E}\left[S(\theta \mid \boldsymbol{x})\right] = 0$$

[*Background Notes, page 10*].

*Proof.* Since $L(\theta)$ is a density when viewed as a function of the observed data $x_1, \ldots, x_n$ we have the following identity in $\theta$,

$$\int \cdots \int L(\theta) \, dx_1 \, \ldots \, dx_n = 1.$$

On differentiating both sides of the above with respect to $\theta$ gives

$$\int \cdots \int \left[ \frac{\partial L(\theta)}{\partial \theta} \right] \, dx_1 \, \ldots \, dx_n = 0.$$

Apply the chain rule to $\frac{\partial \ln L(\theta)}{\partial \theta}$ we find

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{\partial \ln L(\theta)}{\partial L(\theta)} \cdot \frac{\partial L(\theta)}{\partial \theta} = \frac{1}{L(\theta)} \frac{\partial L(\theta)}{\partial \theta}$$

meaning

$$\frac{\partial \ln L(\theta)}{\partial \theta} L(\theta) = \frac{\partial L(\theta)}{\partial \theta}$$

so that

$$\begin{aligned} \int \cdots \int \left[ \frac{\partial L(\theta)}{\partial \theta} \right] \, dx_1 \, \ldots \, dx_n &= 0 \\ \int \cdots \int \left[ \frac{\partial \ln L(\theta)}{\partial \theta} \right] L(\theta) \, dx_1 \, \ldots \, dx_n &= 0 \\ \mathbb{E}\left[ S(\theta) \right] &= 0 \end{aligned}$$

as wanted. □

**Definition 11** (Expotential Family)**.** *In the case of $p-$dimensional observation $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{C}^p$, a $d-$dimensional parameter vector $\boldsymbol{\theta} \in \mathbb{C}^d$, and a $q-$dimensional sufficient statistic $T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{C}^q$, the likelihood function $L(\boldsymbol{\theta})$ for the $d-$parameter vector $\boldsymbol{\theta}$ has the following form if it belongs to the $d-$parameter* **exponential family**

$$L(\boldsymbol{\theta}) = b(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \exp \left\{ c(\boldsymbol{\theta})^\intercal T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \right\} / a(\boldsymbol{\theta})$$

*where $c(\boldsymbol{\theta}) \in \mathbb{C}^q$ and $b(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $a(\boldsymbol{\theta})$ are scalar functions* [Cas01, page 279].

**Theorem 12.** *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be iid observations from a pdf or pmf $f(x \mid \boldsymbol{\theta})$ that belongs to an exponential family as seen in Definition 11, then*

$$T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \left( \sum_{j=1}^{n} t_1(\boldsymbol{X}_j), \ldots, \sum_{j=1}^{n} t_k(\boldsymbol{X}_j) \right)$$

*is a sufficient statistic for $\boldsymbol{\theta}$* [Cas01, page 279].

**Definition 13** (Minimal Sufficient Statistic)**.** *A sufficient statistic $T(\boldsymbol{X})$ is called a* **minimal sufficient statistic** *if, for any other sufficient statistic $T'(\boldsymbol{X})$, $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$* [Cas01, page 280].

**Theorem 14.** *Let $f(\boldsymbol{x} \mid \theta)$ be the pd of a sample $\boldsymbol{X}$. Suppose there exists a function $T(\boldsymbol{x})$ such that, for every two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the ratio $f(\boldsymbol{x} \mid \theta)/f(\boldsymbol{y} \mid \theta)$ is constant as a function of $\theta$ if and only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then $T(\boldsymbol{X})$ is a minimal sufficient statistic* [Cas01, page 281].

*Example* 15 (Normal Minimal Sufficient Statistic)*.* Example taken from [Cas01, page 281]. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$, where both $\mu$ and $\sigma^2$ unknown. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote two sample points, and let $(\bar{x}, s_{\boldsymbol{x}}^2)$ and $(\bar{y}, s_{\boldsymbol{y}}^2)$ be the sample means and variances corresponding to the $\boldsymbol{x}$ and $\boldsymbol{y}$ samples, respectively. Then, the ratio of the densities becomes

$$\begin{aligned} \frac{f\left(\mathbf{x} \mid \mu, \sigma^2\right)}{f\left(\mathbf{y} \mid \mu, \sigma^2\right)} &= \frac{\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\left[n(\bar{x}-\mu)^2 + (n-1)s_{\mathbf{x}}^2\right] / \left(2\sigma^2\right)\right)}{\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\left[n(\bar{y}-\mu)^2 + (n-1)s_{\mathbf{y}}^2\right] / \left(2\sigma^2\right)\right)} \\ &= \exp\left(\left[-n\left(\bar{x}^2 - \bar{y}^2\right) + 2n\mu(\bar{x}-\bar{y}) - (n-1)\left(s_{\mathbf{x}}^2 - s_{\mathbf{y}}^2\right)\right] / \left(2\sigma^2\right)\right). \end{aligned}$$

This ratio will be constant as a function of $\mu$ and $\sigma^2$ if and only if $\bar{x} = \bar{y}$ and $s_{\boldsymbol{x}}^2 = s_{\boldsymbol{y}}^2$. Thus by Theorem 14, $(\overline{X}, S^2)$ is a minimal sufficient statistic for $(\mu, \sigma^2)$.

**Definition 16** (Ancillary Statistic)**.** *A statistic $S(\boldsymbol{X})$ whose distribution does not depend on the parameter $\theta$ is called an ancillary statistic* [Cas01, page 282].

**Definition 17** (Complete Distributions and Statistics)**.** *Let $f(t \mid \theta)$ be a family of pdfs or pmfs for a statistic $T(\boldsymbol{X})$. The family of probability distributions is called* **complete** *if $\mathbb{E}_\theta g(T) = 0$ for all $\theta$ implies $\mathbb{P}(g(T) = 0) = 1$ for all $\theta$. Equivalently, $T(\boldsymbol{X})$ is called a* **complete statistic** [Cas01, page 285].

*Example* 18 (Binomial Complete Statistic)*.* Example taken from [Cas01, page 285]. Suppose that $T$ has a $\mathsf{Bin}(n, p)$ distribution, $0 < p < 1$. Let $g$ be a function such that $\mathbb{E}_p g(T) = 0$. Then

$$\begin{aligned} 0 = \mathbb{E}_p g(T) &= \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-1} \\ &= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t \end{aligned}$$

for all $p$, $0 < p < 1$. The factor $(1-p)^n$ is not 0 for any $p$ in this range. Thus it must be that

$$0 = \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \sum_{t=0}^{n} g(t) \binom{n}{t} r^t$$

for all, $0 < r < \infty$. But the last expression is a polynomial of degree $n$ in $r$, where the coefficient of $r^t$ is $g(t)\binom{n}{t}$. For the polynomial to be 0 for all $r$, each coefficient must be 0. Since none of the $\binom{n}{t}$ terms is 0, this implies that $g(t) = 0$ for $t = 0, 1, \dots n$. Since $T$ takes on the values $0, 1, \dots n$ with probability 1, this means that $\mathbb{P}_p(g(T) = 0) = 1$ for all $p$, the desired conclusion. Hence, $T$ is a complete statistic.

**Definition 19** (Point Estimator)**.** *A* **point estimator** *is any function $W(X_1, \dots, X_n)$ of a sample; that is, any statistic (see Definition 2) is a point estimator* [Cas01, page 311].

**Definition 20** (Fisher Information Matrix)**.** *For the model $\boldsymbol{X} \sim f(\cdot; \boldsymbol{\theta})$, let $S(\boldsymbol{\theta})$ be the score function (see Definition 9) of $\boldsymbol{\theta}$. The covariance matrix of the random vector $S(\boldsymbol{\theta})$, denoted by $\mathcal{J}(\theta)$, is called the* **Fisher Information Matrix** *where*

$$\mathcal{J}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[ S(\boldsymbol{\theta}) S(\boldsymbol{\theta})^\intercal \right]$$

*in the multivariate case and*

$$\mathcal{J}(\theta) = \mathbb{E}_\theta \left( \frac{d}{d\theta} \ln f(\boldsymbol{X}; \theta) \right)^2$$

*in the one-dimensional case. Note that under regularity conditions $\mathbb{E}\left[S(\theta)\right] = 0$ (see Theorem 10) so that*

$$\begin{aligned} \mathcal{J}(\theta) &= \mathbb{E}_\theta \left[ \frac{d}{d\theta} \ln f(\boldsymbol{X}; \theta) \right]^2 \\ &= \mathrm{Var}_\theta \left( \frac{d}{d\theta} \ln f(\boldsymbol{X}; \theta) \right) + \left( \mathbb{E}_\theta \left[ \frac{d}{d\theta} \ln f(\boldsymbol{X}; \theta) \right] \right)^2 \\ &= \mathrm{Var}_\theta \left( S(\theta) \right) + \left( \mathbb{E}_\theta \left[ S(\theta) \right] \right)^2 \\ &= \mathrm{Var}_\theta \left( S(\theta) \right) \end{aligned}$$

[Kro13, page 168].

**Definition 21** (Observed Information)**.** *For the model $\boldsymbol{X} \sim f(\cdot; \boldsymbol{\theta})$, let $S(\boldsymbol{\theta})$ be the score function (see Definition 9) of $\boldsymbol{\theta}$. The negative of the Hessian of the random vector $S(\boldsymbol{\theta})$, denoted by $I(\theta)$, is called the* **Observed Information** *where*

$$I(\boldsymbol{\theta}) = -\nabla\nabla S(\boldsymbol{\theta})$$

*in the multivariate case and*

$$I(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\theta^2} \ln f(\boldsymbol{X}; \theta)$$

*in the one-dimensional case [Background Notes, page 8].*

**Theorem 22.** *Under regularity conditions, the following equality holds*

$$\mathcal{J}(\boldsymbol{\theta}) = \mathbb{E}\left[I(\boldsymbol{\theta})\right]$$

[Kro13, page 169].

**Theorem 23** (Fisher Information Matrix for iid Data)**.** *Let $\boldsymbol{X} = (X_1, \ldots, X_n) \overset{iid}{\sim} \mathring{f}(x; \boldsymbol{\theta})$, and let $\mathring{\mathcal{J}}(\boldsymbol{\theta})$ be the information matrix corresponding to $X \sim \mathring{f}(x; \boldsymbol{\theta})$. Then the information matrix for $\boldsymbol{X}$ is given by*

$$\mathcal{J}(\boldsymbol{\theta}) = n\mathring{\mathcal{J}}(\boldsymbol{\theta})$$

[Kro13, page 170].

**Theorem 24.** *If the $L(\theta)$ belongs to the regular exponential family, then the likelihood equation*

$$\frac{d}{d\boldsymbol{\theta}} \ln L(\boldsymbol{\theta}) = \boldsymbol{0},$$

*can be expressed as*

$$T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \mathbb{E}\left[T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)\right]$$

*[Lecture Notes 1, page 8].*

**Method of Moments.**

**Definition 25** (Method of Moments)**.** *Let $X_1, \dots, X_n$ be a random sample of size $n$ from a population with pf $f(x \mid \theta_1, \dots, \theta_k)$. Method of moments estimators are found by equation the first $k$ sample moments to the corresponding $k$ population moments, and solving the resulting system of simultaneous equations. More precisely, define*

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad \mu_1' = \mathbb{E}X^1 \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \mu_2' = \mathbb{E}X^2 \\ &\vdots \\ m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \mu_k' = \mathbb{E}X^k. \end{aligned}$$

*The population moment $\mu_j'$ will typically be a function of $\theta_1, \dots, \theta_k$, say $\mu_j'(\theta_1, \dots, \theta_k)$. The method of moments estimator $(\tilde{\theta}_1, \dots, \tilde{\theta}_k)$ of $(\theta_1, \dots, \theta_k)$ is obtained by solving the following system of equations for $(\theta_1, \dots, \theta_k)$ in terms of $(m_1, \dots, m_k)$*

$$\begin{aligned} m_1 &= \mu_1'(\theta_1, \dots, \theta_k) \\ m_2 &= \mu_2'(\theta_1, \dots, \theta_k) \\ &\vdots \\ m_k &= \mu_k'(\theta_1, \dots, \theta_k) \end{aligned}$$

[Cas01, page 312].

*Example* 26 (Normal Methods of Moments)*.* Example taken from [Cas01, page 313]. Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2)$. In the preceding notation, $\theta_1 = \theta$ and $\theta_2 = \sigma^2$. We have $m_1 = \overline{X},\ m_s = (1/n)\sum X_i^2,\ \mu_1' = \theta,\ \mu_2' = \theta^2 + \sigma^2$, and hence we must solve

$$\overline{X} = \theta, \quad \frac{1}{n}\sum X_i^2 = \theta^2 + \sigma^2.$$

Solving for $\theta$ and $\sigma^2$ yields the methods of moments estimators

$$\tilde{\theta} = \overline{X} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \overline{X}^2 = \frac{1}{n}\sum (X_i^2 - \overline{X}^2).$$

**Maximum Likelihood Estimates.**

**Definition 27** (Maximum Likelihood Estimator)**.** *For each sample point $\boldsymbol{x}$, let $\hat{\theta}(\boldsymbol{x})$ be a parameter value at which $L(\theta \mid \boldsymbol{x})$ attains its maximum as a function of $\theta$, with $\boldsymbol{x}$ held fixed. A* **maximum likelihood estimator (MLE)** *of the parameter $\theta$ based on a sample $\boldsymbol{X}$ is $\hat{\theta}(\boldsymbol{X})$* [Cas01, page 316].

*Example* 28 (Normal Likelihood)*.* Example taken from [Cas01, page 316]. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$, and let $L(\theta \mid \boldsymbol{x})$ denote the likelihood function. Then

$$L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} \exp\left(-(1/2)(x_i - \theta)^2\right) = \frac{1}{(2\pi)^{1/2}} \exp\left(-(1/2)\sum_{i=1}^{n}(x_i - \theta)^2\right).$$

The equation $(d/d\theta)L(\theta \mid \boldsymbol{x}) = 0$ reduces to

$$\sum_{i=1}^{n}(x_i - \theta) = 0,$$

which has the solution $\hat{\theta} = \overline{x}$. Hence, $\overline{x}$ is a candidate for the MLE. To verify that $\overline{x}$ is, in fact, a global maximim of the likelihood function, we can use the following argument. First, note that $\hat{\theta} = \overline{x}$ is the only solution to $\sum_{i=1}^{n}(x_i - \theta) = 0$; hence $\overline{x}$ is the only zero of the first derivative. Second, verify that

$$\frac{d^2}{d\theta^2} L(\theta \mid \boldsymbol{x})|_{\theta=\overline{x}} < 0.$$

Thus, $\overline{x}$ is the only extreme point in the interior and it is a maximum. To finally verify that $\overline{x}$ is a global maximum, we must check the boundaries at $\pm\infty$. So $\tilde{\theta} = \overline{x}$ is a global maximum and hence $\overline{X}$ is the MLE.

**Theorem 29.** *If $\hat{\theta}$ is the MLE of $\theta$, the for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$* [Cas01, page 320].

*Example* 30 (Normal MLE, $\mu$ and $\sigma$ unknown)*.* Example taken from [Cas01, page 321]. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2)$ with both $\mu$ and $\sigma^2$ unknown. Then

$$L(\theta \mid \boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-(1/2)\sum_{i=1}^{n}(x_i - \theta)^2/\sigma^2\right)$$

and

$$\ln L(\theta \mid \boldsymbol{x}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2/\sigma^2.$$

The partial derivatives, with respect to $\theta$ and $\sigma^2$ are

$$\frac{\partial}{\partial\theta}\ln L(\theta \mid \boldsymbol{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \theta)$$

and

$$\frac{\partial}{\partial\sigma^2}\ln L(\sigma^2 \mid \boldsymbol{x}) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \theta).$$

Setting the partial derivatives equal to 0 and solving for the solution $\hat{\theta} = \overline{x}$, $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})$. To verify that this solution is, in fact, a global maximum, recall first that if $\theta \neq \overline{x}$, then $\sum(x_i - \theta)^2 >$

$\sum(x_i - \overline{x})^2$. Hence, for any value of $\sigma^2$,

$$\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-(1/2)\sum_{i=1}^{n}(x_i - \overline{x})^2/\sigma^2\right) \geq \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-(1/2)\sum_{i=1}^{n}(x_i - \theta)^2/\sigma^2\right).$$

Therefore, verifying that we have found the maximum likelihood estimators is reduced to a one-dimensional problem, verifying that $(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2}\sum(x_i - \overline{x})^2/\sigma^2\right)$ achieves its global maximum at $\sigma^2 = n^{-1}\sum(x_i - \overline{x})^2$. This is straightforward to do using univariate calculus and, in fact, the estimators $\left(\overline{X}, n^{-1}\sum\left(X_i - \overline{X}\right)^2\right)$ are the MLEs.

**Methods of Evaluating Estimators.**

**Definition 31** (Mean Square Error)**.** *The* **mean square error** (*MSE*) *of an estimator $W$ of a parameter $\theta$ is the function $\theta$ defined by* $\mathbb{E}_\theta(W-\theta)^2$ [Cas01, page 330].

**Definition 32** (Bias)**.** *The* **bias** *of an estimator $W$ of a parameter $\theta$ is the difference between the expected value of $W$ and $\theta$; that is* $\text{Bias}_\theta W = \mathbb{E}_\theta W - \theta$*. An estimator whose bias is identically (in $\theta$) equal to 0 is called an* **unbiased estimator** *and satisfies $\mathbb{E}_\theta W = \theta$ for all $\theta$* [Cas01, page 330].

It is important to note that

$$\mathbb{E}_\theta (W-\theta)^2 = \text{Var}_\theta + (\mathbb{E}_\theta W - \theta)^2 = \text{Var}_\theta W + (\text{Bias}_\theta W)^2 .$$

*Example* 33 (Normal MSE). Example taken from [Cas01, page 331]. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$. The statistics $\overline{X}$ and $S^2$ are both unbiased estimators since

$$\mathbb{E}\overline{X} = \mu, \quad \mathbb{E}S^2 = \sigma^2, \text{ for all } \mu \text{ and } \sigma^2.$$

The MSEs of these estimators are given by

$$\mathbb{E}\left(\overline{X}-\mu\right)^2 = \text{Var}\overline{X} = \frac{\sigma^2}{n}$$
$$\mathbb{E}\left(S^2-\sigma^2\right)^2 = \text{Var}S^2 = \frac{2\sigma^4}{n-1}.$$

The MSE of $\overline{X}$ remains $\sigma^2/n$ even if the normality assumption is dropped. However, the above expression for the MSE of $S^2$ does not remain the same if the normality assumption is relaxed. An alternative estimator for $\sigma^2$ is the MLE $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 = \frac{n-1}{n}S^2$. It is straightforward to calculate

$$\mathbb{E}\hat{\sigma}^2 = \mathbb{E}\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2,$$

so that $\hat{\sigma}^2$ is a biased estimator of $\sigma^2$. The variance of $\hat{\sigma}^2$ can also be calculated as

$$\text{Var }\hat{\sigma}^2 = \text{Var}\left(\frac{n-1}{n}S^2\right) = \left(\frac{n-1}{n}\right)^2 \text{Var}S^2 = \frac{2(n-1)\sigma^4}{n^2},$$

and hence, its MSE is given by

$$\mathbb{E}\left(\hat{\sigma}^2-\sigma^2\right) = \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4.$$

Thus we have

$$\mathbb{E}\left(\hat{\sigma}^2-\sigma^2\right)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4 < \left(\frac{2}{n-1}\right)\sigma^4 = \mathbb{E}\left(\hat{\sigma}^2-\sigma^2\right)^2,$$

showing that $\hat{\sigma}^2$ has a smaller MSE than $S^2$. Thus, by trading off variance for bias, the MSE is improved.

**Definition 34** (Best Unbiased Estimator)**.** *An estimator $W^*$ is a* **best unbiased estimator** *of $\tau(\theta)$ if it satisfies $EE_\theta W^* = \tau(\theta)$ for all $\theta$ and, for any other estimator $W$ with $\mathbb{E}_\theta W = \tau(\theta)$, we have $\text{Var}_\theta W^* \leq \text{Var}_\theta W$ for all $\theta$. $W^*$ is also called a* **uniform minimum variance unbiased estimator** (*UMVUE*) *of $\tau(\theta)$* [Cas01, page 334].

**Theorem 35** (Cramer-Rao Inequality)**.** *Let $X_1, \dots, X_n$ be a sample with pdf $f(\boldsymbol{x} \mid \theta)$, and let $W(\boldsymbol{X}) = W(X_1, \dots, X_n)$ be any estimator satisfying*

$$\frac{d}{d\theta}\mathbb{E}_\theta W(\boldsymbol{X}) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\left[W(\boldsymbol{x}) f(\boldsymbol{x} \mid \theta)\right]$$

*and*

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

*Then*

$$\mathrm{Var}_\theta(W(\boldsymbol{X})) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(\boldsymbol{X})\right)^2}{\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\ln f(\boldsymbol{X} \mid \theta)\right)^2\right)} = \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(\boldsymbol{X})\right)^2}{\mathcal{J}(\theta)}$$

[Cas01, page 335].

**Corollary 36** (Cramer-Rao Inequality, iid Case)**.** *If the assumptions of Theorem 35 are satisfied and, additionally, if $X_1, \ldots, X_n$ are iid with pdf $f(\boldsymbol{x} \mid \theta)$, then*

$$\mathrm{Var}_\theta(W(\boldsymbol{X})) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(\boldsymbol{X})\right)^2}{n\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\ln f(X \mid \theta)\right)^2\right)}$$

[Cas01, page 337].

**Lemma 37.** *If $f(\boldsymbol{x} \mid \theta)$ satisfies*

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

(*true for the exponential family*), *then*

$$\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\ln f(X \mid \theta)\right)^2\right) = -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\ln f(X \mid \theta)\right)$$

[Cas01, page 338].

*Example* 38 (Poisson Unbiased Estimate). Example taken from [Cas01, page 338]. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Poi}(\lambda)$, and let $\overline{X}$ and $S^2$ be the sample mean and variance, respectively. Recall that for the Poisson pmf both the mean and variance are equal to $\lambda$. We have

$$\begin{aligned}\mathbb{E}_\lambda\overline{X} &= \lambda, \quad \text{for all } \lambda,\\ \mathbb{E}_\lambda S^2 &= \lambda, \quad \text{for all } \lambda,\end{aligned}$$

so both $\overline{X}$ and $S^2$ are unbiased estimators of $\lambda$. To determine the better estimator, $\overline{X}$ or $S^2$, we should now compare the variances. We have $\mathrm{Var}_\lambda\overline{X} = \lambda/n$, but $\mathrm{Var}_\lambda S^2$ is quiet a lengthy calculation. Not only this, even if we can establish that $\overline{X}$ is better than $S^2$, consider the class of estimators

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

For every constant $a$, $\mathbb{E}_\lambda W_a = \lambda$, so now we have infinitely many unbiased estimators of $\lambda$. Instead, let us show that $\overline{X}$ is the best estimator directly using the Cramer-Rao inequality. Here we are estimating $\tau(\lambda) = \lambda$, so that $\tau'(\lambda) = 1$. Also, since we have an exponential family, using Lemma 37 gives us

$$\begin{aligned}\mathbb{E}_\lambda\left(\left(\frac{\partial}{\partial\lambda}\ln f(X \mid \lambda)\right)^2\right) &= -n\mathbb{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\ln f(X \mid \lambda)\right)\\ &= -n\mathbb{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\ln\left(\frac{e^{-\lambda}\lambda^X}{X!}\right)\right)\\ &= -n\mathbb{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\left(-\lambda + X\ln\lambda - \ln X!\right)\right)\end{aligned}$$

$$= -n\mathbb{E}_\lambda\left(-\frac{X}{\lambda^2}\right)$$
$$= \frac{n}{\lambda}.$$

Hence for any unbiased estimator, $W$, of $\lambda$, from Corollary 36 we must have

$$\begin{aligned}\operatorname{Var}_\theta(W(\boldsymbol{X})) &\geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(\boldsymbol{X})\right)^2}{n\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\ln f(X \mid \theta)\right)^2\right)} \\ &= \frac{(1)^2}{\left(\frac{n}{\lambda}\right)} \\ &= \frac{\lambda}{n}.\end{aligned}$$

Since $\operatorname{Var}_\lambda\overline{X} = \lambda/n$, $\overline{X}$ must be the best unbiased estimator.

**Corollary 39** (Attainment)**.** *Let $X_1, \ldots, X_n$ be a sample with pdf $f(\boldsymbol{x} \mid \theta)$, where $f(x \mid \theta)$ satisfies the conditions of the Cramer-Rao Theorem. $L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^n f(x_1 \mid \theta)$ denote the likelihood function. If $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(\boldsymbol{X})$ attains the Cramer-Rao Lower Bound if and only if*

$$a(\theta)\left[W(\boldsymbol{x}) - \tau(\theta)\right] = \frac{\partial}{\partial\theta}\ln L(\theta \mid \boldsymbol{x})$$

*for some function $a(\theta)$* [Cas01, page 341].

*Example* 40 (Continuation of Example 30). Example taken from [Cas01, page 341]. Here we know

$$L(\mu, \sigma^2 \mid \boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-(1/2)\sum_{i=1}^n (x_i - \mu)^2/\sigma^2\right),$$

and hence

$$\frac{\partial}{\partial\sigma^2}\ln L(\mu, \sigma^2 \mid \boldsymbol{x}) = \frac{n}{2\sigma^4}\left(\sum_{i=1}^n \frac{(x_i - \mu)^2}{n} - \sigma^2\right).$$

Thus, taking $a(\sigma^2) = n/(2\sigma^4)$ shows that the best unbiased estimator of $\sigma^2$ is $\frac{(x_i-\mu)^2}{n}$, which is calculable only if $\mu$ is known. If $\mu$ is not known, the bound *cannot* be attained.

*Sufficiency and Unbiasedness.*

**Theorem 41** (Rao-Blackwell)**.** *Let $W$ be any unbiased estimator of $\tau(\theta)$, and let $T$ be a sufficient statistic for $\theta$. Define $\phi(T) = \mathbb{E}(W \mid T)$. Then $\mathbb{E}_\theta\phi(T) = \tau(\theta)$ and $\operatorname{Var}_\theta\phi(T) \leq \operatorname{Var}_\theta W$ for all $\theta$; that is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$* [Cas01, page 342].

**Theorem 42.** *If $W$ is the best unbiased estimator of $\tau(\theta)$, then $W$ is unique* [Cas01, page 343].

**Theorem 43.** *Let $T$ be a complete sufficient statistic for a parameter $\theta$, and let $\phi(T)$ be any estimator based only on $T$. Then $\phi(T)$ is the best unbiased estimator of its expected value* [Cas01, page 347].

*Consistency.*

**Definition 44** (Consistency)**.** *A sequence of estimators $T_n$ of $g(\boldsymbol{\theta})$ is said to be* **consistent** *if for every $\boldsymbol{\theta} \in \Omega$,*

$$T_n \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}}} g(\boldsymbol{\theta}), \quad \text{as } n \to \infty$$

*that is, given any $\varepsilon > 0$, then*

$$\mathbb{P}\left[\left|T_n\left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right) - g(\boldsymbol{\theta})\right| \geq \varepsilon\right] \to 0, \quad \text{as } n \to \infty$$

[Kro13, page 176] [*Background Notes, page 44*].

**Theorem 45.** *If* $\operatorname{Var}(T_n) \to 0$ *and* $\operatorname{Bias}(T_n) \to 0$, *as $n \to \infty$, then the sequence of estimates $T_n$ is consistent for estimating $g(\boldsymbol{\theta})$* [*Background Notes, page 44*].

*Proof.* Let $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta})$ denote the joint pdf of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Then we have

$$\mathbb{P}\left[|T_n - g(\boldsymbol{\theta})| \geq \varepsilon\right] = \int \ldots \int_{|T_n - g(\boldsymbol{\theta})| \geq \varepsilon} f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta})\, d\boldsymbol{x}_1 \ \ldots \ d\boldsymbol{x}_n.$$

On the region of integration in the above,

$$\begin{gathered} |T_n - g(\boldsymbol{\theta})| \geq \varepsilon \\ (T_n - g(\boldsymbol{\theta}))^2 \geq \varepsilon^2 \\ \frac{(T_n - g(\boldsymbol{\theta}))^2}{\varepsilon^2} \geq 1, \end{gathered}$$

and since $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta})$ is non-negative,

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}) \leq f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}) \frac{(T_n - g(\boldsymbol{\theta}))^2}{\varepsilon^2}.$$

Thus

$$\begin{aligned} &\mathbb{P}\left[|T_n - g(\boldsymbol{\theta})| \geq \varepsilon\right] \\ &= \int \ldots \int_{|T_n - g(\boldsymbol{\theta})| \geq \varepsilon} f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta})\, d\boldsymbol{x}_1 \ \ldots \ d\boldsymbol{x}_n \\ &\leq \int \ldots \int_{|T_n - g(\boldsymbol{\theta})| \geq \varepsilon} f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}) \frac{(T_n - g(\boldsymbol{\theta}))^2}{\varepsilon^2}\, d\boldsymbol{x}_1 \ \ldots \ d\boldsymbol{x}_n \\ &\leq \frac{1}{\varepsilon^2} \int \ldots \int f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}) \left(T_n - g(\boldsymbol{\theta})\right)^2 \ d\boldsymbol{x}_1 \ \ldots \ d\boldsymbol{x}_n \\ &\leq \frac{1}{\varepsilon^2} \operatorname{MSE}(T_n) \\ &\leq \frac{1}{\varepsilon^2} \left[\operatorname{Var}(T_n) + (\operatorname{Bias}(T_n))^2\right] \end{aligned}$$

which tends to 0 as $n \to \infty$, since $\operatorname{Var}(T_n)$ and $\operatorname{Bias}(T_n)$ both tend to 0. □

*Example* 46 (Continuation of Example 33). Example taken from Background Notes, page 44. The estimators

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

are both consistent for $\sigma^2$. This is easily seen with $s^2$ since

$$\mathbb{E}[s^2] = \sigma \quad \text{and} \quad \operatorname{Var}(s^2) \to 0$$

as $n \to \infty$. To show that $\hat{\sigma}^2$ is also a consistent estimator, note that

(see 17)
$$\frac{\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2}{\sigma^2} \sim \chi^2_{n-1}$$

meaning

$$\mathbb{E}\left[\frac{\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2}{\sigma^2}\right] = n - 1$$

and

$$\operatorname{Var}\left[\frac{\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2}{\sigma^2}\right] = 2(n - 1)$$

so that

$$\mathbb{E}\left[\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2\right] = (n-1)\sigma^2$$

and

$$\operatorname{Var}\left[\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2\right] = 2(n-1)\sigma^4.$$

Hence

$$\begin{aligned} \mathbb{E}\left(\hat{\sigma}^2\right) &= \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{\sigma^2}{n} \\ \operatorname{Var}\left(\hat{\sigma}^2\right) &= \frac{2(n-1)\sigma^4}{n^2} \end{aligned}$$

meaning $\mathbb{E}\left(\hat{\sigma}^2\right) \to 0$ and $\operatorname{Var}\left(\hat{\sigma}^2\right) \to 0$ as $n \to \infty$. Therefore, by Theorem 45, $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2$.

*Large-Sample Comparisons of Estimators.*

**Theorem 47** (Information Matrix for iid Data)**.** *Suppose that $\hat{\boldsymbol{\theta}}_n$ is a sequence of consistent ML estimates for $\boldsymbol{\theta}$. Then $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)$ converges in distribution to a $\mathsf{N}\left(\boldsymbol{\theta}, \mathring{\mathcal{J}}^{-1}(\boldsymbol{\theta})\right)$ distributed random vector as $n \to \infty$. In other words,*

$$\hat{\boldsymbol{\theta}}_n \overset{approx}{\sim} \mathsf{N}\left(\boldsymbol{\theta}, \mathring{\mathcal{J}}^{-1}(\boldsymbol{\theta})/n\right).$$

**Definition 48** (Asymptotic Relative Efficiency)**.** *Suppose that $\hat{\theta}_{n_1}$ and $\hat{\theta}_{n_2}$ are two single variable estimates such that*

$$\begin{aligned} \hat{\theta}_{n_1} &\overset{approx}{\sim} \mathsf{N}\left(\theta, \tau_1^2/n\right) \\ \hat{\theta}_{n_2} &\overset{approx}{\sim} \mathsf{N}\left(\theta, \tau_2^2/n\right). \end{aligned}$$

*The* **Asymptotic Relative Efficiency** *(ARE) of $\hat{\theta}_{n_2}$ with respect to $\hat{\theta}_{n_1}$ is given by*

$$\operatorname{ARE}(\hat{\theta}_{n_2}) = \hat{\theta}_{n_1}/\hat{\theta}_{n_2}$$

[*Background Notes, page 46*].

*Example* 49 (Asymptotic Distribution of Bernoulli MLE). Example taken from [Kro13, page 177]. For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(p)$, the MLE for $p$ is

$$\hat{p}_n = \overline{x} = \frac{1}{n} \sum_i x_i.$$

To compute the information number (see Definition 20) for $p$, note that regularity conditions hold so that

$$\mathring{\mathcal{J}}(p) = \mathrm{Var}_p \left( S(p) \right)$$

where

$$\begin{aligned} S(p) &= \frac{d}{dp} \ln \left( p^x (1-p)^{1-x} \right) \\ &= \frac{d}{dp} \left[ x \cdot \ln(p) + (1-x) \ln(1-p) \right] \\ &= \frac{x}{p} - \frac{(1-x)}{1-p} = \frac{x}{p(1-p)}. \end{aligned}$$

This means that

$$\begin{aligned} \mathring{\mathcal{J}}(p) &= \mathrm{Var}_p \left( S(p) \right) \\ &= \mathrm{Var}_p \left( \frac{X}{p(1-p)} \right) \\ &= \frac{p(1-p)}{p^2(1-p)^2} = \frac{1}{p(1-p)}. \end{aligned}$$

Theorem 47 states that

$$\hat{p}_n \overset{\text{approx}}{\sim} \mathsf{N} \left( p, \frac{p(1-p)}{n} \right).$$

## References

[Cas01] George and Berger Casella Roger, *Statistical Inference*, Cengage, Mason, OH, 2001 (eng).

[Kro13] Dirk P and C.C. Chan Kroese Joshua, *Statistical Modeling and Computation*, Springer New York, New York, NY, 2013 (eng).