



THE UNIVERSITY OF QUEENSLAND
A U S T R A L I A

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	9
MAXIMUM LIKELIHOOD ESTIMATES	10
METHODS OF EVALUATING ESTIMATORS	12
SUFFICIENCY AND UNBIASEDNESS	14
REFERENCES	15

SYMBOLS AND NOTATION

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

<i>Syntax</i>	<i>Meaning</i>
\triangleq	An equality which acts as a statement
$ \mathbf{A} $	The determinate of a matrix.
$\mathbf{x}^\top, \mathbf{X}^\top$	The transpose operator.
$\mathbf{x}^*, \mathbf{X}^*$	The hermitian operator.
$\mathbf{a}.*\mathbf{b}$ or $\mathbf{A}.*\mathbf{B}$	Element-wise vector (matrix) multiplication, similar to Matlab.
\propto	Proportional to.
∇ or $\nabla_{\mathbf{f}}$	The partial derivative (with respect to \mathbf{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.

$\mathbf{A}_{(:,)}$	Index slicing to extract a submatrix from the elements of $\mathbf{A} \in \mathbb{R}^{n \times m}$, 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 $\mathbf{A}_{(1:3,j:j')}$ is the submatrix $\mathbb{R}^{3 \times (j'-j+1)}$ matrix containing the first three rows of \mathbf{A} and columns j to j' . As another example, $\mathbf{A}_{(:,j)}$ is the j^{th} column of \mathbf{A} .
\mathbf{A}^\dagger	Denotes the unique psuedo inverse or Moore-Penore inverse of \mathbf{A} .
\mathbb{C}	The complex numbers.
$\text{diag}(\mathbf{w})$	Vector argument, a diagonal matrix containing the elements of vector \mathbf{w} .
$\text{diag}(\mathbf{W})$	Matrix argument, a vector containing the diagonal elements of the matrix \mathbf{W} .
\mathbb{E} or $\mathbb{E}_{q(x)}[z(x)]$	Expectation, or expectation of $z(x)$ where $x \sim q(x)$.
\mathbb{R}	The real numbers.
$\text{tr}(\mathbf{A})$	The trace of a matrix.
\mathbb{V} or $\mathbb{V}_{q(x)}[z(x)]$	Variance, the variance of $z(x)$ when $x \sim q(x)$.
\mathbb{Z}	The integers, $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$.
Ω	The sample space.

REVIEW

Theorems and definitions 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 $\frac{1}{p} + \frac{1}{q} = 1$, 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$(12) \quad \mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

$$(\text{Bayes' Theorem}) \quad \mathbb{P}(A | B) = \frac{\mathbb{P}(B | A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

$$(13) \quad \mathbb{P}(A_1, \dots, A_n) = \mathbb{P}(A_1) \mathbb{P}(A_2 | A_1) \mathbb{P}(A_3 | A_1, A_2) \cdots \mathbb{P}(A_n | A_1, A_2, \dots, A_{n-1})$$

$$(14)$$

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

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

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

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

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

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

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

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

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

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

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

Probabilistic Identities. If $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Ber}(p)$ then

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

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

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

If $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma)$ then

$$(17) \quad \sum_{i=1}^n X_i^2 = \chi_n^2.$$

POINT ESTIMATION

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 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 arithmetic average of the values in a random sample. It is usually denoted by

$$(18) \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

[Cas01, page 212].

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

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

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

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

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

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

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

[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, θ , 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 θ in this restriction) and $\max_i x_i \leq \theta$. If we define $T(\mathbf{x}) = \max_i x_i = x_{(n)}$,

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

and

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

It is easily verified that $f(\mathbf{x} \mid \theta) = g(T(\mathbf{x}) \mid \theta)$ for all \mathbf{x} and θ . Thus, according to Theorem 7, the largest order statistic, $T(\mathbf{X}) = X_{(n)}$, 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}_θ be the natural numbers up to and including θ . Then the joint pmf of X_1, \dots, X_n is

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

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

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

thus providing the factorization

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

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

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

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

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

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

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

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

[Kro13, page 165].

Definition 10 (Exponential Family). *In the case of p -dimensional observation $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{C}^p$, a d -dimensional parameter vector $\boldsymbol{\theta} \in \mathbb{C}^d$, and a q -dimensional sufficient statistic $T(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{C}^q$, 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***

$$L(\boldsymbol{\theta}) = b(\mathbf{x}_1, \dots, \mathbf{x}_n) \exp \{ c(\boldsymbol{\theta})^\top T(\mathbf{x}_1, \dots, \mathbf{x}_n) \} / a(\boldsymbol{\theta})$$

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

Theorem 11. Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be iid observations from a pdf or pmf $f(\mathbf{x} \mid \theta)$ that belongs to an exponential family as seen in Definition 10, then

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

is a sufficient statistic for θ [Cas01, page 279].

Definition 12 (Minimal Sufficient Statistic). A sufficient statistic $T(\mathbf{X})$ is called a **minimal sufficient statistic** if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ [Cas01, page 280].

Theorem 13. Let $f(\mathbf{x} \mid \theta)$ be the pdf of a sample \mathbf{X} . Suppose there exists a function $T(\mathbf{x})$ such that, for every two sample points \mathbf{x} and \mathbf{y} , the ratio $f(\mathbf{x} \mid \theta) / f(\mathbf{y} \mid \theta)$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is a minimal sufficient statistic [Cas01, page 281].

Example 14 (Normal Minimal Sufficient Statistic). Example taken from [Cas01, page 281]. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 unknown. Let \mathbf{x} and \mathbf{y} denote two sample points, and let (\bar{x}, s_x^2) and (\bar{y}, s_y^2) be the sample means and variances corresponding to the \mathbf{x} and \mathbf{y} samples, respectively. Then, the ratio of the densities becomes

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

This ratio will be constant as a function of μ and σ^2 if and only if $\bar{x} = \bar{y}$ and $s_x^2 = s_y^2$. Thus by Theorem 13, (\bar{X}, S^2) is a minimal sufficient statistic for (μ, σ^2) .

Definition 15 (Ancillary Statistic). A statistic $S(\mathbf{X})$ whose distribution does not depend on the parameter θ is called an **ancillary statistic** [Cas01, page 282].

Definition 16 (Complete Distributions and Statistics). Let $f(t \mid \theta)$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$. The family of probability distributions is called **complete** if $\mathbb{E}_\theta g(T) = 0$ for all θ implies $\mathbb{P}(g(T) = 0) = 1$ for all θ . Equivalently, $T(\mathbf{X})$ is called a **complete statistic** [Cas01, page 285].

Example 17 (Binomial Complete Statistic). Example taken from [Cas01, page 285]. Suppose that T has a $\text{Bin}(n, p)$ distribution, $0 < p < 1$. Let g be a function such that $\mathbb{E}_p g(T) = 0$. Then

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

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 18 (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].

Method of Moments.

Definition 19 (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 equating the first k sample moments to the corresponding k population moments, and solving the resulting system of simultaneous equations. More precisely, define

$$\begin{aligned} m_1 &= \frac{1}{n} \sum_{i=1}^n X_i^1, & \mu'_1 &= \mathbb{E}X^1 \\ m_2 &= \frac{1}{n} \sum_{i=1}^n X_i^2, & \mu'_2 &= \mathbb{E}X^2 \\ &\vdots \\ m_k &= \frac{1}{n} \sum_{i=1}^n X_i^k, & \mu'_k &= \mathbb{E}X^k. \end{aligned}$$

The population moment μ'_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 20 (Normal Methods of Moments). Example taken from [Cas01, page 313]. Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathbf{N}(\theta, \sigma^2)$. In the preceding notation, $\theta_1 = \theta$ and $\theta_2 = \sigma^2$. We have $m_1 = \bar{X}$, $m_2 = (1/n) \sum X_i^2$, $\mu'_1 = \theta$, $\mu'_2 = \theta^2 + \sigma^2$, and hence we must solve

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

Solving for θ and σ^2 yields the methods of moments estimators

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

Maximum Likelihood Estimates.

Definition 21 (Maximum Likelihood Estimator). For each sample point \mathbf{x} , let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta | \mathbf{x})$ attains its maximum as a function of θ , with \mathbf{x} held fixed. A **maximum likelihood estimator (MLE)** of the parameter θ based on a sample \mathbf{X} is $\hat{\theta}(\mathbf{X})$ [Cas01, page 316].

Example 22 (Normal Likelihood). Example taken from [Cas01, page 316]. Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$, and let $L(\theta | \mathbf{x})$ denote the likelihood function. Then

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

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

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

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

$$\frac{d^2}{d\theta^2} L(\theta | \mathbf{x})|_{\theta=\bar{x}} < 0.$$

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

Theorem 23. If $\hat{\theta}$ is the MLE of θ , then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$ [Cas01, page 320].

Example 24 (Normal MLE, μ and σ unknown). Example taken from [Cas01, page 321]. Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$ with both μ and σ^2 unknown. Then

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

and

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

The partial derivatives, with respect to θ and σ^2 are

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

and

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

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

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

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

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

Methods of Evaluating Estimators.

Definition 25 (Mean Square Error). The **mean square error** (MSE) of an estimator W of a parameter θ is the function θ defined by $\mathbb{E}_\theta(W - \theta)^2$ [Cas01, page 330].

Definition 26 (Bias). The **bias** of an estimator W of a parameter θ is the difference between the expected value of W and θ ; that is $\text{Bias}_\theta W = \mathbb{E}_\theta W - \theta$. An estimator whose bias is identically (in θ) equal to 0 is called an **unbiased estimator** and satisfies $\mathbb{E}_\theta W = \theta$ for all θ [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 27 (Normal MSE). Example taken from [Cas01, page 331]. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. The statistics \bar{X} and S^2 are both unbiased estimators since

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

The MSEs of these estimators are given by

$$\begin{aligned} \mathbb{E}(\bar{X} - \mu)^2 &= \text{Var}\bar{X} = \frac{\sigma^2}{n} \\ \mathbb{E}(S^2 - \sigma^2)^2 &= \text{Var}S^2 = \frac{2\sigma^4}{n-1}. \end{aligned}$$

The MSE of \bar{X} remains σ^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 σ^2 is the MLE $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S^2$. 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 σ^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

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

Thus we have

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

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

Definition 28 (Best Unbiased Estimator). An estimator W^* is a **best unbiased estimator** of $\tau(\theta)$ if it satisfies $\mathbb{E}W^* = \tau(\theta)$ for all θ and, for any other estimator W with $\mathbb{E}W = \tau(\theta)$, we have $\text{Var}_\theta W^* \leq \text{Var}_\theta W$ for all θ . W^* is also called a **uniform minimum variance unbiased estimator** (UMVUE) of $\tau(\theta)$ [Cas01, page 334].

Theorem 29 (Cramer-Rao Inequality). Let X_1, \dots, X_n be a sample with pdf $f(\mathbf{x} \mid \theta)$, and let $W(\mathbf{X}) = W(X_1, \dots, X_n)$ be any estimator satisfying

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

and

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

Then

$$\text{Var}_\theta(W(\mathbf{X})) \geq \frac{\left(\frac{d}{d\theta} \mathbb{E}_\theta W(\mathbf{X})\right)^2}{\mathbb{E}_\theta \left(\left(\frac{\partial}{\partial \theta} \ln f(\mathbf{X} | \theta)\right)^2\right)}$$

[Cas01, page 335].

Corollary 30 (Cramer-Rao Inequality, iid Case). *If the assumptions of Theorem 29 are satisfied and, additionally, if X_1, \dots, X_n are iid with pdf $f(x | \theta)$, then*

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

[Cas01, page 337].

Lemma 31. *If $f(x | \theta)$ satisfies*

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

(true for the exponential family), then

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

[Cas01, page 338].

Example 32 (Poisson Unbiased Estimate). Example taken from [Cas01, page 338]. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Poi}(\lambda)$, and let \bar{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 λ . We have

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

so both \bar{X} and S^2 are unbiased estimators of λ . To determine the better estimator, \bar{X} or S^2 , we should now compare the variances. We have $\text{Var}_\lambda \bar{X} = \lambda/n$, but $\text{Var}_\lambda S^2$ is quite a lengthy calculation. Not only this, 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 , $\mathbb{E}_\lambda W_a = \lambda$, so now we have infinitely many unbiased estimators of λ . Instead, let us show that \bar{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 31 gives us

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

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

Hence for any unbiased estimator, W , of λ , from Corollary 30 we must have

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

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

Corollary 33 (Attainment). *Let X_1, \dots, X_n be a sample with pdf $f(\mathbf{x} | \theta)$, where $f(\mathbf{x} | \theta)$ satisfies the conditions of the Cramer-Rao Theorem. $L(\theta | \mathbf{x}) = \prod_{i=1}^n f(x_i | \theta)$ denote the likelihood function. If $W(\mathbf{X}) = W(X_1, \dots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramer-Rao Lower Bound if and only if*

$$a(\theta) [W(\mathbf{x}) - \tau(\theta)] = \frac{\partial}{\partial\theta} \ln L(\theta | \mathbf{x})$$

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

Example 34 (Continuation of Example 24). Example taken from [Cas01, page 341]. Here we know

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

and hence

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

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

Sufficiency and Unbiasedness.

Theorem 35 (Rao-Blackwell). *Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient statistic for θ . Define $\phi(T) = \mathbb{E}(W | T)$. Then $\mathbb{E}_\theta \phi(T) = \tau(\theta)$ and $\text{Var}_\theta \phi(T) \leq \text{Var}_\theta W$ for all θ ; that is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$ [Cas01, page 342].*

Theorem 36. *If W is the best unbiased estimator of $\tau(\theta)$, then W is unique [Cas01, page 343].*

Theorem 37. *Let T be a complete sufficient statistic for a parameter θ , 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].*

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).