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

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

Syntax	*Meaning*
$\triangleq$	An equality which acts as a statement
$\|\boldsymbol{A}\|$	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.
∇ 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 $X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$
$\mathbf{0}$ 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.
$\mathrm{diag}\left(\boldsymbol{w}\right)$	Vector argument, a diagonal matrix containing the elements of vector $\boldsymbol{w}$.
$\mathrm{diag}\left(\boldsymbol{W}\right)$	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.
$\mathrm{tr}\left(\boldsymbol{A}\right)$	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, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$.
Ω	The sample space.

Review

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

Useful Formulae and Theorems.

(Combination)
$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

(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	$\mathsf{Poi}(\lambda)$	$\mathbb{N}_0$	$\frac{\lambda^x}{x!}e^{-\lambda}$	λ	λ
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	$\mathsf{N}(\mu,\sigma^2)$	$\mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$	μ	σ^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	χ^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	μ	σ^2

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

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

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

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

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

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 \tag{4}$$

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

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

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

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

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

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

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

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

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

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

(14)

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

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

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

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

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

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 $\varepsilon > 0$

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

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 $x \in \mathbb{R}$

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

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

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

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

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

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

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

If $X \sim \chi_2^2$, then $X \sim \mathsf{Exp}(1/2)$.

If $X \sim \mathsf{U}(0, 1)$, then $-2\ln(X)\chi_2^2$.

If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Exp}(\lambda)$ then $\sum_i X_i \sim \mathsf{Gam}(n, \lambda)$.

If $X \sim \mathsf{Gam}(k, \lambda)$ then for any $c > 0$ we have $cX \sim \mathsf{Gam}(k, c\lambda)$.

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 θ if the conditional distribution of the sample $\boldsymbol{X}$ given the value of $T(\boldsymbol{X})$ does not depend on θ* [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 θ 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 θ* [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 θ, 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 θ,*

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

[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, 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 $\max_i x_i \le \theta$. 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 θ. 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 θ. 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 θ. 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 θ 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 θ, 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 θ,

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

On differentiating both sides of the above with respect to θ 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 θ 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 μ and σ^2 unknown. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote two sample points, and let $(\bar{x}, s^2_{\boldsymbol{x}})$ and $(\bar{y}, s^2_{\boldsymbol{y}})$ be the sample means and variances corresponding to the $\boldsymbol{x}$ and $\boldsymbol{y}$ samples, respectively. Then, the ratio of the densities becomes

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

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

Definition 16 (Ancillary Statistic)**.** *A statistic $S(\boldsymbol{X})$ whose distribution does not depend on the parameter θ 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 some function g, for all θ implies $\mathbb{P}(g(T) = 0) = 1$ for all θ. Equivalently, $T(\boldsymbol{X})$ is called a* **complete statistic** [Cas01, page 285]*.*

Theorem 18. *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 the statistic*

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

is complete as long as the parameter space is non-meager [Cas01, page 288]*.*

Theorem 19. *If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic* [Cas01, page 289]*.*

Theorem 20. *A complete, sufficient statistic is always minimal [Background Notes, page 25].*

Example 21 (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 $t = 0, 1, \ldots n$. Since T takes on the values $0, 1, \ldots n$ 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.

Example 22 (Sum of iid Bernoulli RVs)*.* Example taken from [Tutorial Sheet 2, Q6]. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$. The likelihood function for θ is given by

$$\begin{aligned} L(\theta) &= \prod_{j=1}^{n} \binom{n}{x_j} \theta^{x_j} (1-\theta)^{1-x_j} \\ &= \left[\prod_{j=1}^{n} \binom{n}{x_j} \right] \theta^t (1-\theta)^{n-t} \\ &= \left[\prod_{j=1}^{n} \binom{n}{x_j} \right] \exp\left[c(\theta) t\right] (1-\theta)^n \end{aligned}$$

$$= b(\boldsymbol{x}) \exp \left[c(\theta) t \right] / a(\theta)$$

where

$$t(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$$
$$c(\theta) = \ln \frac{\theta}{1-\theta}$$
$$a(\theta) = (1-\theta)^{-n}$$
$$b(\boldsymbol{x}) = \prod_{j=1}^{n} \binom{n}{x_j}.$$

Clearly, the likelihood belongs to the regular exponential family with canonical parameter $c(\theta)$ and complete sufficient statistic $T = t(\boldsymbol{X})$. Also, the score statistic (Definition 9) is given by

$$S(\theta) = \frac{\partial}{\partial \theta} \ln L(\theta) = \frac{n}{\theta(1-\theta)} \left(\frac{t}{n} - \theta \right)$$

showing that the estimator T attains the Cramer-Rao lower bound is estimating θ. Hence, it attains the MVB (Corollary 45) and is therefore also a UMVU estimator of θ. On the other hand, the estimator

$$V = (X_n, T_{n-1})^\intercal$$

where $T_{n-1} = \sum_{j=1}^{n-1} X_j$, while sufficient (with canonical parameter $c(\theta) = (\ln \frac{\theta}{1-\theta}, \ln \frac{\theta}{1-\theta})^\intercal$), is not complete. To demonstrate that V is not complete, we have that

$$\mathbb{E}\left[X_n - \frac{1}{n-1} T_{n-1} \right] = 0$$

however, consider

$$\mathbb{P}\left[X_n - \frac{1}{n-1} T_{n-1} = 0 \right].$$

Since, $X_n \sim \mathsf{Ber}(\theta)$, $T_{n-1} \sim \mathsf{Bin}(n-1, \theta)$ and X_i are iid

$$\begin{aligned} \mathbb{P}\left[X_n - \frac{1}{n-1} T_{n-1} = 0 \right] &= \mathbb{P}\left[T_{n-1} = 0 \mid X_n = 0 \right] \cdot \mathbb{P}\left[X_n = 0 \right] + \mathbb{P}\left[T_{n-1} = n-1 \mid X_n = 0 \right] \cdot \mathbb{P}\left[X_n = 1 \right] \\ &= (1-\theta)^n + \theta^n \neq 1 \end{aligned}$$

for $0 < \theta < 1$. So by Definition 17, V is not complete. Furthermore, as T is a complete, sufficient statistic, it is a minimal sufficient statistic (Theorem 20) for θ. It is a function of every other sufficient statistic (Definition 13) and here we can see it is a function of V with

$$T = (V)_1 + (V)_2 = X_n + T_{n-1}.$$

This also shows that V is not a (sufficient) minimal statistic (again by Definition 13). Now lets consider the variance between two estimators of θ, $T = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $W(V) = \mathbb{E}\left[X_1 \mid V \right]$. We saw that T is UMVU and its variance attains MVB. Its variance can be computed as

$$\operatorname{Var}(T) = \frac{1}{n^2}(n\theta(1-\theta)) = \frac{1}{n}\theta(1-\theta).$$

Now let us try and find an explicit espression for $W(V(\boldsymbol{x}))$. We have

$$\begin{aligned}
W(V(\boldsymbol{x})) &= \mathbb{E}\left[X_1 \mid X_n = x_n, \sum_{i=1}^{n-1} X_i = t_{n-1}\right] \\
&= \sum_{x_1=0}^{1} x_1 \cdot \mathbb{P}\left[X_1 = x_1 \mid X_n = x_n, \sum_{i=1}^{n-1} X_i = t_{n-1}\right] \\
&= \mathbb{P}\left[X_1 = 1 \mid X_n = x_n, \sum_{i=1}^{n-1} X_i = t_{n-1}\right] \\
&= \frac{\mathbb{P}\left[X_1 = 1, X_n = x_n, \sum_{i=1}^{n-1} X_i = t_{n-1}\right]}{\mathbb{P}\left[X_n = x_n, \sum_{i=1}^{n-1} X_i = t_{n-1}\right]} \\
&= \frac{\mathbb{P}\left[X_1 = 1, X_n = x_n, \sum_{i=2}^{n-1} X_i = t_{n-1} - 1\right]}{\mathbb{P}\left[X_n = x_n, \sum_{i=1}^{n-1} X_i = t_{n-1}\right]} \\
&= \frac{\mathbb{P}\left[X_1 = 1\right]\mathbb{P}\left[X_n = x_n\right]\mathbb{P}\left[\sum_{i=2}^{n-1} X_i = t_{n-1} - 1\right]}{\mathbb{P}\left[X_n = x_n\right]\mathbb{P}\left[\sum_{i=1}^{n-1} X_i = t_{n-1}\right]}.
\end{aligned}$$

Since $X_1 \sim \mathsf{Ber}(\theta)$, $\sum_{i=1}^{n-1} X_i \sim \mathsf{Bin}(n-1, \theta)$ and $\sum_{i=2}^{n-1} X_i \sim \mathsf{Bin}(n-2, \theta)$, we have

$$\begin{aligned}
W(V(\boldsymbol{x})) &= \frac{\theta\binom{n-2}{t_{n-1}-1}\theta^{t_{n-1}-1}(1-\theta)^{(n-2)-(t_{n-1}-1)}}{\binom{n-1}{t_{n-1}}\theta^{t_{n-1}}(1-\theta)^{(n-1)-t_{n-1}}} \\
&= t_{n-1}/(n-1)
\end{aligned}$$

where $t_{n-1} = \sum_{i=1}^{n-1} x_i$. This means $W(V(\boldsymbol{X})) = \frac{1}{n-1}\sum_{i=1}^{n-1} X_i$ and

$$\operatorname{Var}(W(V)) = \frac{(n-1)}{(n-1)^2}\theta(1-\theta) = \frac{1}{(n-1)}\theta(1-\theta) < \frac{1}{n}\theta(1-\theta).$$

Definition 23 (Point Estimator)**.** *A* **point estimator** *is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic (see Definition 2) is a point estimator* [Cas01, page 311].

Definition 24 (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

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

$$= \mathrm{Var}_\theta\left(S(\theta)\right)$$

[Kro13, page 168].

Definition 25 (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 26. *Under regularity conditions, the following equality holds*

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

[Kro13, page 169].

Theorem 27 (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 28. *If the* $L(\theta)$ *belongs to the regular exponential family, then the likelihood equation*

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

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 29 (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 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 30 (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 θ and σ^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 31 (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 θ, with $\boldsymbol{x}$ held fixed. A* **maximum likelihood estimator (MLE)** *of the parameter θ based on a sample $\boldsymbol{X}$ is $\hat{\theta}(\boldsymbol{X})$* [Cas01, page 316].

Example 32 (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 33. *If $\hat{\theta}$ is the MLE of θ, the for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$* [Cas01, page 320].

Example 34 (Normal MLE, μ and σ unknown)*.* Example taken from [Cas01, page 321]. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2)$ with both μ and σ^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 θ and σ^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 σ^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.

Definition 35 (Quantile of Order α)**.** *The quantile of order α, q_α, is the value of the random variable X such that*

$$\mathbb{P}\left[X \leq q_\alpha\right] = \alpha.$$

Example 36. Let $X_1, \ldots, X_n$ denote a random sample from a $N(\mu, \sigma^2)$ distribution, where both μ and σ are unknown. Let us consider a way to find the maximum likelihood for quantile of order α. Take $X \sim N\left(\mu, \sigma^2\right)$. As $Z = (X - \mu)/\sigma$ has a standard normal distribution with the function $\Phi(z)$, we can express the left hand for the expression of the quantile of order α as

$$\mathbb{P}\left[X \leq q_\alpha\right] = \mathbb{P}\left[\frac{X-\mu}{\sigma} \leq \frac{q_\alpha - \mu}{\sigma}\right] = \mathbb{P}\left[Z \leq \frac{q_\alpha - \mu}{\sigma}\right].$$

This means

$$\begin{aligned} \frac{q_\alpha - \mu}{\sigma} &= \Phi^{-1}(\alpha) \\ q_\alpha &= \mu + \sigma\Phi^{-1}(\alpha). \end{aligned}$$

Hence, as a consequence of Theorem 33, $g(\hat{\boldsymbol{\theta}}) = \hat{\mu} + \hat{\sigma}\Phi^{-1}(\alpha)$ is the maximum likelihood estimate of q_α. Note, however, that this is not an unbiased estimate of q_α by virtue of the fact that $\mathbb{E}[\hat{\sigma}] \neq \sigma$. If we adjusted this estimate by multiplying by some constant k_n, that is,

$$\mathbb{E}[\hat{\sigma}] = k_n\sigma$$

then it would be an unbiased estimator of q_α. To compute such a k_n, we have that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$ that is, $\frac{1}{2}n\hat{\sigma}^2/\sigma^2 \sim \gamma(m/2)$ where $m = (n-1)/2$. This is equivalent to saying $\hat{\sigma} = \sigma\sqrt{2/n}\sqrt{Y}$ where $Y \sim \gamma(m)$. Thus $\mathbb{E}[\hat{\sigma}] = k_n\sigma$, where $k_n = \sqrt{2/n}\mathbb{E}[\sqrt{Y}]$ and where

$$\begin{aligned} \mathbb{E}[\sqrt{Y}] &= \frac{\int_0^\infty y^{1/2}\exp(-y)y^{m-1}\,dy}{\Gamma(m)} \\ &= \frac{\int_0^\infty \exp(-y)y^{m+\frac{1}{2}-1}\,dy}{\Gamma(m)} \\ &= \frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)} \\ &= \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \end{aligned}$$

Upon substituting the result for $\mathbb{E}[\sqrt{Y}]$ into the right-hand side of of expression for k_n we obtain

$$k_n = \sqrt{\frac{2}{n}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}.$$

Methods of Evaluating Estimators.

Definition 37 (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 38 (Bias)**.** *The* **bias** *of an estimator W of a parameter θ is the difference between the expected value of W and θ; that is* $\mathrm{Bias}_\theta W = \mathbb{E}_\theta W - \theta$*. 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

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

Example 39 (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

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

The MSE of $\overline{X}$ 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 $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^n\left(X_i-\overline{X}\right)^2 = \frac{n-1}{n}S^2$. 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 σ^2. The variance of $\hat{\sigma}^2$ can also be calculated as

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

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

Theorem 41 (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

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

which is commonly refereed to as the **minimum variance bound** *(MVB). If $W(\boldsymbol{X})$ attains the MVB (for all values of θ), it is said to be a MVB estimator* [Cas01, page 335].

Corollary 42 (Cramer-Rao Inequality, iid Case)**.** *If the assumptions of Theorem 41 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 43. *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 44 (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 λ. 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 λ. 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 λ. 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 43 gives us

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

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

Hence for any unbiased estimator, W, of λ, from Corollary 42 we must have

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

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

Corollary 45 (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 46*.* Example taken from Tutorial Sheet 2 Q5. Let T be an estimator of the parameter θ, having bias $b(\theta)$. Assuming that the usual regularity conditions and using the Cramer-Rao lower bound (Theorem 41) for the variance of an unbiased estimator of θ, we can show that

$$
\mathrm{MSE}(T) \geq \left[1 + \frac{\partial}{\partial\theta}b(\theta)\right]^2 \cdot \mathcal{J}^{-1}(\theta) + [b(\theta)]^2
$$

where $\mathcal{J}(\theta)$ is the Fisher information matrix (Definition 24). To start, since $b(\theta) = \mathbb{E}[\theta] - \theta$ we have

$$
\mathbb{E}[T] = \theta + b(\theta) \triangleq g(\theta)
$$

so that T is an unbiased estimate for $g(\theta)$. By the Cramer-Rao lower bound,

$$
\mathrm{Var}(T) \geq [g'(\theta)]^2 \cdot \mathcal{J}^{-1}(\theta) = \left[1 + \frac{\partial}{\partial\theta}b(\theta)\right]^2 \cdot \mathcal{J}^{-1}(\theta).
$$

Now

$$
\begin{aligned}
\mathrm{MSE}(T) &= \mathrm{Var}(T) + [\mathrm{Bias}(T)]^2\\
&= \mathrm{Var}(T) + [b(\theta)]^2\\
&\geq \left[1 + \frac{\partial}{\partial\theta}b(\theta)\right]^2 \cdot \mathcal{J}^{-1}(\theta) + [b(\theta)]^2.
\end{aligned}
$$

Example 47 (Continuation of Example 34). 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 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 48 (Rao-Blackwell)**.** *Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient statistic for θ. Define $\phi(T) = \mathbb{E}(W \mid T)$. Then $\mathbb{E}_\theta\phi(T) = \tau(\theta)$ and $\mathrm{Var}_\theta\phi(T) \leq \mathrm{Var}_\theta W$ for all θ; that is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$* [Cas01, page 342].

Theorem 49. *If W is the best unbiased estimator of $\tau(\theta)$, then W is unique* [Cas01, page 343].

Theorem 50. *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].

Consistency.

Definition 51 (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

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

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

Theorem 52. *If $\mathrm{Var}(T_n) \to 0$ and $\mathrm{Bias}(T_n) \to 0$, 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

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

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

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

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

Example 53 (Continuation of Example 39). 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 σ^2. This is easily seen with s^2 since

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

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

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

so that

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

and

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

Hence

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

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 52, $\hat{\sigma}^2$ is a consistent estimator of σ^2.

Large-Sample Comparisons of Estimators.

Theorem 54 (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 55 (Asymptotic Relative Efficiency)**.** *Suppose that $\hat{\theta}_{n_1}$ and $\hat{\theta}_{n_2}$ are two single variable estimates such that*

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

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

$$\mathrm{ARE}(\hat{\theta}_{n_2}) = \tau_1^2/\tau_2^2$$

[*Background Notes, page 46*].

Example 56 (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 24) for p, note that regularity conditions hold so that

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

where

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

This means that

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

Theorem 54 states that
$$\hat{p}_n \overset{\text{approx}}{\sim} \mathsf{N}\left(p, \frac{p(1-p)}{n}\right).$$

Expectation Maximization Algorithm. The expectation-maximization (EM) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood (ML) estimates, useful in a variety of incomplete data problems, where algorithms such as the Newton-Raphson method may turn out to be more complicated. On each iteration of the EM algorithm, there are two steps called the Expectation step or the $E-$step and the Maximization step or the $M-$step. Because of this, the algorithm is called the EM algorithm.

Formulation of the EM Algorithm. We let $\boldsymbol{Y}$ be the random vector corresponding to the observed data $\boldsymbol{y}$ having p.d.f. postulated as $g(\boldsymbol{y};\boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = (\Psi_1, \Psi_2, \dots, \Psi_d)^\intercal$ is a vector of unknown parameters with parameter space Ω. We let $g_c(\boldsymbol{x};\boldsymbol{\Psi})$ denote the p.d.f. of the random vector $\boldsymbol{X}$ corresponding to the complete data vector $\boldsymbol{x}$. Then the complete-data log likelihood function that could be formed for $\boldsymbol{\Psi}$ is $\boldsymbol{x}$ were fully observable is given by

$$\ln L_c(\boldsymbol{\Psi}) = \ln g_c(\boldsymbol{x};\boldsymbol{\Psi}).$$

Formally, we have two sample space X and Y and a many-to-one mapping X to Y. Instead of observing the complete-data vector $\boldsymbol{x} \in X$, we observe the incomplete-data vector $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}) \in Y$. It follows that

$$g(\boldsymbol{y};\boldsymbol{\Psi}) = \int_{X(\boldsymbol{y})} g_c(\boldsymbol{x};\boldsymbol{\Psi})\, d\boldsymbol{x}$$

where $X(\boldsymbol{y})$ is the subset of X determined by the equation $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$. The EM algorithm approaches the problem of solving the incomplete-data likelihood equation

$$\nabla_{\boldsymbol{\Psi}} \ln L(\boldsymbol{\Psi}) = 0$$

indirectly by proceeding iteratively in terms of the complete-data log likelihood function, $\ln L_c(\boldsymbol{\Psi})$. As it is unobservable, it is replaced by its conditional expectation given $\boldsymbol{y}$, using the current fit for $\boldsymbol{y}$. More specifically, let $\boldsymbol{\Psi}^{(0)}$ be some initial value for $\boldsymbol{\Psi}$. Then on the first iteration, the $E-$step requires the calculation of

$$Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(0)}\right) = \mathbb{E}_{\boldsymbol{\Psi}^{(0)}}\left[\ln L_c\left(\boldsymbol{\Psi}\right) \mid \boldsymbol{y}\right].$$

The $M-$step requires the maximization of $Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(0)}\right)$ with respect to $\boldsymbol{\Psi}$ over the parameter space Ω. That is, we choose $\boldsymbol{\Psi}^{(1)}$ such that

$$Q\left(\boldsymbol{\Psi}^{(1)};\boldsymbol{\Psi}^{(0)}\right) \geq Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(0)}\right)$$

for all $\boldsymbol{\Psi} \in \Omega$. The $E-$ and $M-$steps are then carried out again, but this time with $\boldsymbol{\Psi}^{(0)}$ replaced by the current fit $\boldsymbol{\Psi}^{(1)}$. On the $(k+1)^{th}$ iteration, the $E-$ and $M-$steps are defined as follows:

Definition 57 ($E-$step)**.** *Calculate* $Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}\right)$ *as*

$$Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}\right) = \mathbb{E}_{\boldsymbol{\Psi}^{(k)}}\left[\ln L_c\left(\boldsymbol{\Psi}\right) \mid \boldsymbol{y}\right].$$

Definition 58 ($M-$step)**.** *Choose* $\boldsymbol{\Psi}^{(k+1)}$ *to be any value of* $\boldsymbol{\Psi} \in \Omega$ *that maximises* $Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}\right)$*, that is,*

$$Q\left(\boldsymbol{\Psi}^{(k+1)};\boldsymbol{\Psi}^{(k)}\right) \geq Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}\right)$$

for all $\boldsymbol{\Psi} \in \Omega$.

The $E-$ and $M-$ steps are alternated repeatedly until the difference

$$L(\mathbf{\Psi}^{(k+1)}) - L(\mathbf{\Psi}^{(k)})$$

changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood value $L(\mathbf{\Psi}^{(k)})$. Another way of expressing Definition 57 is to say that $\mathbf{\Psi}^{(k+1)}$ belongs to

$$\mathcal{M}\left(\mathbf{\Psi}^{(k)}\right) = \underset{\mathbf{\Psi}}{\operatorname{argmax}}\, Q\left(\mathbf{\Psi}; \mathbf{\Psi}^{(k)}\right),$$

which is the set of points that maximise $Q\left(\mathbf{\Psi}; \mathbf{\Psi}^{(k)}\right)$.

Example 59 (A Multinomial Example). Example taken from [McL08, page] [Kro13, page 185]. We consider first the multinomial example that DLR used to introduce the EM algorithm and that has been subsequently used many times in the literature to illustrate various modifications and extensions of this algorithm. The data relates to a problem of estimation of linkage in genetics where an observed data vector of frequencies

$$\boldsymbol{y} = (y_1, y_2, y_3, y_4)^\intercal$$

is a postulated to arise from a multinomial distribution with four cells with cell probabilities

$$\frac{1}{2} + \frac{1}{4}\Psi, \frac{1}{4}(1-\Psi), \frac{1}{4}(1-\Psi), \text{ and } \frac{1}{4}\Psi$$

with $0 \le \Psi \le 1$. The parameter Ψ is to be estimated on the basis of the observed information $\boldsymbol{y}$. The probability of the observed data $\boldsymbol{y}$ is given by

$$g(\boldsymbol{y}; \Psi) = \frac{n!}{y_1! y_2! y_3! y_4!} \left(\frac{1}{2} + \frac{1}{4}\Psi\right)^{y_1} \left(\frac{1}{4} - \frac{1}{4}\Psi\right)^{y_2} \left(\frac{1}{4} - \frac{1}{4}\Psi\right)^{y_3} \left(\frac{1}{4}\Psi\right)^{y_4}.$$

Suppose now that the first of the original four multinomial cells, which has an associated probability of $\frac{1}{2} + \frac{1}{4}\Psi$, could be split into two subcells have probabilities $\frac{1}{2}$ and $\frac{1}{4}$, respectively, and let y_{11} and y_{12} be the corresponding split of y_1, where

$$y_1 = y_{11} + y_{12}.$$

Thus, the observed vector of frequencies y is viewed as being incomplete and the complete-data vector is taken to be

$$\boldsymbol{x} = (y_{11}, y_{12}, y_2, y_3, y_4)^\intercal.$$

The cell frequencies in $\boldsymbol{x}$ are assumed to arise from a multinomial distribution having five cells with probabilities

$$\frac{1}{2}, \frac{1}{4}\Psi, \frac{1}{4}(1-\Psi), \frac{1}{4}(1-\Psi), \text{ and } \frac{1}{4}\Psi.$$

In this framework, y_{11} and y_{12} are regraded as the unobservable or missing data since we only get their sum y_1. The complete-data log likelihood is then

$$g_c(\boldsymbol{y}; \Psi) = C(\boldsymbol{x}) \left(\frac{1}{2}\right)^{y_{11}} \left(\frac{1}{4}\Psi\right)^{y_{12}} \left(\frac{1}{4} - \frac{1}{4}\Psi\right)^{y_2} \left(\frac{1}{4} - \frac{1}{4}\Psi\right)^{y_3} \left(\frac{1}{4}\Psi\right)^{y_4}$$

Thus, the complete-data log likelihood is, therefore

$$\ln L_c(\Psi) = (y_{12} + y_4) \ln \Psi + (y_2 + y_3) \ln(1-\Psi) + c.$$

for some constant c not involving Ψ. Let $\Psi^{(0)}$ be the value specified initially for Ψ. Then on the first iteration of the EM algorithm, the $E-$step requires the computation of the conditional expectation of $L_c(\Psi)$ given $\boldsymbol{y}$, using $\Psi^{(0)}$, which can be written as

$$Q\left(\Psi; \Psi^{(0)}\right) = \mathbb{E}_{\Psi^{(0)}}\left[\ln L_c(\Psi) \mid \boldsymbol{y}\right].$$

As $\ln L_c(\Psi)$ is a linear function of the unobservable data y_{11} and y_{12} for this problem, the $E-$step is effected simply by replacing y_{11} and y_{12} by their current conditional expectations given the observed data $\boldsymbol{y}$. Considering the random variable Y_{11} corresponding to y_{11}, it is easy to verify that conditional on $\boldsymbol{y}$, effectively y_1, Y_{11} has a binomial distribution with sample size y_1 and probability parameter

$$\frac{1}{2}/\left(\frac{1}{2}+\frac{1}{4}\Psi^{(0)}\right),$$

where $\Psi^{(0)}$ is used in place of the unknown parameter Ψ. Thus the initial conditional expectation of Y_{11} given y_1 is

$$\mathbb{E}_{\Psi^{(0)}}\left[Y_{11} \mid y_1\right] = y_{11}^{(0)},$$

where

$$\begin{aligned} y_{11}^{(0)} &= \frac{1}{2}y_1 - y_{11}^{(0)} \\ &= \frac{1}{4}y_1\Psi^{(0)}/\left(\frac{1}{2}+\frac{1}{4}\Psi^{(0)}\right) \end{aligned}$$

The M-step is undertaken on the first iteration by choosing $\Psi^{(1)}$ to be the value of Q that maximizes $Q\left(\Psi; \Psi^{(0)}\right)$ with respect to Ψ. Since this $Q-$function is given simply by replacing the unobservable frequencies y_{11} and y_{12} with their current conditional expectations $y_{11}^{(0)}$ and $y_{12}^{(0)}$ in the complete-data log likelihood, $\Psi^{(1)}$ is obtained taking the derivative of $\ln L_c(\Psi)$ to find its maximising value, that is,

$$\begin{aligned} \frac{\partial}{\partial\Psi}\ln L_c(\Psi) = 0 \ &= \left(y_{12}^{(0)} + y_4\right)\frac{1}{\Psi} - (y_2+y_3)\frac{1}{1-\Psi} \\ \Psi\left(y_2+y_3\right) &= (1-\Psi)\left(y_{12}^{(0)}+y_4\right) \\ \Psi\left(y_{12}^{(0)}+y_2+y_3+y_4\right) &= \left(y_{12}^{(0)}+y_4\right) \\ \Psi &= \frac{y_{12}^{(0)}+y_4}{y_{12}^{(0)}+y_2+y_3+y_4} \\ &= \frac{y_{12}^{(0)}+y_4}{n-y_{11}^{(0)}}. \end{aligned}$$

It follows on so alternating the $E-$ and $M-$steps on the $(k+1)^{th}$ iteration of the EM algorithm that

$$\Psi^{(k+1)} = \frac{y_{12}^{(k)}+y_4}{n-y_{11}^{(k)}}$$

where

$$\begin{aligned} y_{11}^{(k)} &= \frac{1}{2}y_1/\left(\frac{1}{2}+\frac{1}{4}\Psi^{(k)}\right) \\ y_{12}^{(k)} &= y_1 - y_{11}^{(k)} \end{aligned}$$

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).

[McL08] Geoffrey John and Krishnan McLachlan T. (Thriyambakam) and McLachlan, *The EM algorithm and extensions / Geoffrey J. McLachlan, Thriyambakam Krishnan.*, Wiley series in probability and statistics, Wiley-Interscience, 2008.