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
$\lvert\boldsymbol{A}\rvert$	The determinate of a matrix.
$\boldsymbol{x}^\intercal, \boldsymbol{X}^\intercal$	The transpose operator.
$\boldsymbol{x}^*, \boldsymbol{X}^*$	The hermitian operator.
$\boldsymbol{a} .* \boldsymbol{b}$ or $\boldsymbol{A} .* \boldsymbol{B}$	Element-wise vector (matrix) multiplication, similar to Matlab.
$\propto$	Proportional to.
∇ 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.
$\mathrm{diag}\,(\boldsymbol{w})$	Vector argument, a diagonal matrix containing the elements of vector $\boldsymbol{w}$.
$\mathrm{diag}\,(\boldsymbol{W})$	Matrix argument, a vector containing the diagonal elements of the matrix $\boldsymbol{W}$.
$\mathbb{E}$ or $\mathbb{E}_{q(x)}\left[z(x)\right]$	Expectation, or expectation of $z(x)$ where $x \sim q(x)$.
$\mathbb{R}$	The real numbers.
$\mathrm{tr}\,(\boldsymbol{A})$	The trace of a matrix.
$\mathbb{V}$ or $\mathbb{V}_{q(x)}\left[z(x)\right]$	Variance, the variance of $z(x)$ when $x \sim q(x)$.
$\mathbb{Z}$	The integers, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$.
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	$\mathsf{U}[a,b]$	$[a,b]$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$
Beta	$\mathsf{Beta}(\alpha,\beta)$	$[0,1]$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Exponential	$\mathsf{Exp}(\lambda)$	$\mathbb{R}^+$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
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	$2n$
Dirichlet	$\mathsf{Dir}(\boldsymbol{\alpha})$	$x_i\in(0,1),\sum_i x_i=1$	$\frac{1}{B(\boldsymbol{\alpha})}\prod_{i=1}^K x_i^{\alpha_i-1}$	OM [1]	OM
White-Noise	$\mathsf{WN}(\mu,\sigma^2)$	NA	NA	μ	σ^2

[1] NA means not applicable, OM means omitted.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + n(\mu - \bar{X})^2 \tag{14}$$

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

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

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

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 $\varepsilon > 0$

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

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

(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 $cX \sim \mathsf{Gam}(k, c\lambda)$.

Point Estimation

Methods of Finding Estimates Introduction.

Definition 2 (Statistic)**.** *Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. The the random variable or random vector $Y = T(X_1, \ldots, X_n)$ 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 $q(t \mid \theta)$ 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 $X_1, \ldots, X_n$ be iid observations from the discrete uniform distribution on $1, \ldots, \theta$. That is, the unknown parameter, θ, is a positive integer and the pmf of X_i is

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

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

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

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.

Theorem 9 (Basu's Theorem)**.** *Let T be a complete sufficient statistic for $\boldsymbol{\theta}$. If the distribution of some statistic V $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ does not depend on $\boldsymbol{\theta}$, then V is an ancillary statistic (Definition 17). An ancillary statistic will be distributed independently of T* [Cas01, page 287].

Definition 10 (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 11. *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 12 (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 13. *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 12, 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 14 (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 15. *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 16 (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 $(\overline{x}, s^2_{\boldsymbol{x}})$ and $(\overline{y}, s^2_{\boldsymbol{y}})$ 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 $\overline{x} = \overline{y}$ and $s^2_{\boldsymbol{x}} = s^2_{\boldsymbol{y}}$. Thus by Theorem 15, $(\overline{X}, S^2)$ is a minimal sufficient statistic for (μ, σ^2).

Definition 17 (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 18 (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 19. *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 12, 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 20. *If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic* [Cas01, page 289].

Theorem 21. *A complete, sufficient statistic is always minimal* [*Background Notes, page 25*].

Example 22 (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

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

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 $t = 0, 1, \ldots n$. 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 23 (Sum of iid Bernoulli RVs). Example taken from [Tutorial Sheet 2, Q6]. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$. The likelihood function for θ is given by

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

where

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

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 10) 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 46) 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 18, V is not complete. Furthermore, as T is a complete, sufficient statistic, it is a minimal sufficient statistic (Theorem 21) for θ. It is a function of every other sufficient statistic (Definition 14) 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 14). Now lets consider the variance between two estimators of θ, $T = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $W(V) = \mathbb{E}[X_1 \mid V]$. We saw that T is UMVU and its variance attains MVB. Its variance can be computed as

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

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

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

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

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

Definition 24 (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 25 (Fisher Information Matrix)**.** *For the model $\boldsymbol{X} \sim f(\cdot; \boldsymbol{\theta})$, let $S(\boldsymbol{\theta})$ be the score function (see Definition 10) 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 $\mathbb{E}[S(\theta)] = 0$ *(see Theorem 11) so that*

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

[Kro13, page 168].

Definition 26 (Observed Information)**.** *For the model* $\boldsymbol{X} \sim f(\cdot; \boldsymbol{\theta})$*, let* $S(\boldsymbol{\theta})$ *be the score function (see Definition 10) 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 27. *Under regularity conditions, the following equality holds*

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

[Kro13, page 169].

Theorem 28 (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 29. *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 30 (Method of Moments)**.** *Let $X_1, \ldots, X_n$ be a random sample of size n from a population with pf $f(x \mid \theta_1, \ldots, \theta_k)$. 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 $\theta_1, \ldots, \theta_k$, say $\mu_j'(\theta_1, \ldots, \theta_k)$. The method of moments estimator $(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ of $(\theta_1, \ldots, \theta_k)$ is obtained by solving the following system of equations for $(\theta_1, \ldots, \theta_k)$ in terms of $(m_1, \ldots, m_k)$

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

[Cas01, page 312].

Example 31 (Normal Methods of Moments). Example taken from [Cas01, page 313]. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2)$. 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 32 (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 33 (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 34. *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 35 (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 36 (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 37. 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 34, $g(\hat{\boldsymbol{\theta}}) = \hat{\mu} + \hat{\sigma}\Phi^{-1}(\alpha)$ 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 38 (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 39 (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

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

Example 40 (Normal MSE). Example taken from [Cas01, page 331]. Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu,\sigma^2)$. The statistics $\overline{X}$ and S^2 are both unbiased estimators since

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

The MSEs of these estimators are given by

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

The MSE of $\overline{X}$ remains 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

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

and hence, its MSE is given by

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

Thus we have

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

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

Definition 41 (Best Unbiased Estimator)**.** *An estimator W^* is a* **best unbiased estimator** *of $\tau(\theta)$ if it satisfies $\mathbb{E}_\theta W^* = \tau(\theta)$ for all θ and, for any other estimator W with $\mathbb{E}_\theta W = \tau(\theta)$, we have $\text{Var}_\theta W^* \le \text{Var}_\theta W$ for all θ. W^* is also called a* **uniform minimum variance unbiased estimator** *(UMVUE) of $\tau(\theta)$* [Cas01, page 334].

Theorem 42 (Cramer-Rao Inequality)**.** *Let $X_1,\ldots,X_n$ be a sample with pdf $f(\boldsymbol{x} \mid \theta)$, and let $W(\boldsymbol{X}) = W(X_1,\ldots,X_n)$ 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 43 (Cramer-Rao Inequality, iid Case)**.** *If the assumptions of Theorem 42 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 44. *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 45 (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 44 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 43 we must have

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

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

Corollary 46 (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 47. 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 42) for the variance of an unbiased estimator of θ, we can show that

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

where $\mathcal{J}(\theta)$ is the Fisher information matrix (Definition 25). To start, since $b(\theta) = \mathbb{E}[T] - \theta$ 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,

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

Now

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

Example 48 (Continuation of Example 35). 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 49 (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 50. *If W is the best unbiased estimator of $\tau(\theta)$, then W is unique* [Cas01, page 343].

Theorem 51. *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 52 (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 53. *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

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

On the region of integration in the above,

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

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

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

Thus

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

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

Example 54 (Continuation of Example 40). 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 $\mathrm{Var}\left(\hat{\sigma}^2\right) \to 0$ as $n \to \infty$. Therefore, by Theorem 53, $\hat{\sigma}^2$ is a consistent estimator of σ^2.

Example 55 (Conclusion of Example 8). Example taken from [Cas01, page 338] and can also be found on tutorial sheet 3. We saw from Example 8 that the likelihood was

$$\begin{aligned} L(\theta) &= \theta^{-n}\mathbb{1}_{N_\theta}(x_{(n)})\left(\prod_{i=1}^{n}\mathbb{1}_N(x_i)\right) \\ &= \theta^{-n}\mathbb{1}_{[0,\theta]}(x_{(n)})\left(\prod_{i=1}^{n}\mathbb{1}_{[0,\infty)}(x_i)\right). \end{aligned}$$

Thus $L(\theta) = \theta^{-n}$ when $\theta \geq x_{(n)}$ and 0 otherwise. This implies the MLE of θ is $\theta_{(n)}$. Since $\frac{\partial}{\partial\theta}\ln f(x \mid \theta) = -\frac{1}{\theta}$, we have

$$\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\ln f(x \mid \theta)\right)^2\right) = \frac{1}{\theta^2}.$$

The Cramer-Rao Theorem (see Theorem 42) suggests that if W is any unbiased estimator of θ then,

$$\mathbb{V}_\theta W \geq \frac{\theta^2}{n}.$$

We would like to find an unbiased estimator with small variance. As a first guess, consider the sufficient statistic $Y = X_{(n)}$, the largest order statistic and MLE. The pdf of Y is $f_Y(y \mid \theta) = ny^{n-1}/\theta^n,\ 0 < y < \theta$, so

$$\mathbb{E}_\theta Y = \int_0^\theta \frac{ny^{n-1}}{\theta^n}\,dy = \frac{n}{n+1}\theta,$$

showing that $\frac{n+1}{n}Y$ is an unbiased estimator of θ. We next calculate

$$\begin{aligned} \mathbb{V}_\theta\left(\frac{n+1}{n}Y\right) &= \left(\frac{n+1}{n}\right)^2\mathbb{V}_\theta(Y) \\ &= \left(\frac{n+1}{n}\right)^2\left(\mathbb{E}_\theta Y^2 - \left(\frac{n}{n+1}\theta\right)^2\right) \\ &= \left(\frac{n+1}{n}\right)^2\left(\frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2\right) \\ &= \frac{1}{n(n+2)}\theta^2, \end{aligned}$$

which is uniformly smaller than $\frac{\theta^2}{n}$. This indicates that the Cramer-Rao Theorem is not applicable to this pdf. To see this is so, we can use the Leibnitz's Rule which states that if $f(x,\theta), a(\theta)$ and $b(\theta)$ are differentiable with respect to θ, then

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)dx = f(b(\theta),\theta)\frac{d}{d\theta}b(\theta) - f(a(\theta),\theta)\frac{d}{d\theta}a(\theta) + \int_{a(\theta)}^{b(\theta)}\frac{\partial}{\partial\theta}f(x,\theta)dx$$

we can calculate

$$\begin{aligned} \frac{d}{d\theta}\int_0^\theta h(x)f(x \mid \theta)\,dx &= \frac{d}{d\theta}\int_0^\theta h(x)\frac{1}{\theta}\,dx \\ &= \frac{h(\theta)}{\theta} + \int_0^\theta h(x)\frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\right)\,dx \end{aligned}$$

$$\neq \int_0^\theta h(x) \frac{\partial}{\partial \theta} f(x \mid \theta)\, dx,$$

unless $h(\theta)/\theta = 0$ for all θ. Hence the Cramer-Rao Theorem does not apply. In general, if the range of the pdf depends on the parameter, the theorem will no be applicable.

Large-Sample Comparisons of Estimators.

Theorem 56 (Asymptotic Distribution of the MLE)**.** *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 $\mathsf{N}\left(\mathbf{0}, \mathring{\mathcal{J}}^{-1}(\boldsymbol{\theta})\right)$ distributed random vector as $n \to \infty$. In other words,*

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

Theorem 57 (Multivariate Central Limit Theorem)**.** *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} \mathsf{WN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For large n, the random vector $\sum_i \boldsymbol{X}_i$ approximately has a $\mathsf{N}(n\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution.*

Theorem 58 (Delta Method)**.** *Let $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots$ be a sequence of random vectors such that $\sqrt{n}\left(\boldsymbol{Z}_n - \boldsymbol{\mu}\right) \to \boldsymbol{K} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$. Then for any continuously differentiable function g of $\boldsymbol{Z}_n$,*

$$\sqrt{n}\left(g(\boldsymbol{Z}_n) - g(\boldsymbol{\mu})\right) \to \boldsymbol{R} \sim \mathsf{N}\left(\mathbf{0}, \boldsymbol{J}\boldsymbol{\Sigma}\boldsymbol{J}^\intercal\right),$$

where $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\mu}) = \left(\frac{\partial g_i(\boldsymbol{\mu})}{\partial x_j}\right)$ is the Jacobian of the matrix g evaluated at $\boldsymbol{\mu}$ [Kro13, page 92].

Definition 59 (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 60 (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 25) for p, note that regularity conditions hold so that

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

where

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

This means that

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

Theorem 56 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 $\boldsymbol{\Psi} = (\Psi_1, \Psi_2, \ldots, \Psi_d)^\intercal$ 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 61 ($E-$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 62 ($M-$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

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

changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood value $L(\boldsymbol{\Psi}^{(k)})$. Another way of expressing Definition 61 is to say that $\boldsymbol{\Psi}^{(k+1)}$ belongs to

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

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

Example 63 (A Multinomial Example). Example taken from [McL08, page 10] [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 $0 \leq \Psi \leq 1$. 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

$$y_{11}^{(0)} = \frac{1}{2}y_1/\left(\frac{1}{2}+\frac{1}{4}\Psi^{(0)}\right).$$

We also have

$$\begin{aligned} y_{12}^{(0)} &= \frac{1}{4}y_1 - y_{11}^{(0)} \\ &= \frac{1}{4}y_1\Psi^{(0)}/\left(\frac{1}{2}+\frac{1}{4}\Psi^{(0)}\right). \end{aligned}$$

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

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

Example 64 (Bin Grouped Data (Normal)). Example taken from class notes. We consider now the application of the EM algorithm to continuous data that are grouped into intervals. More specifically, let W be a random variable with pdf $f(w;\theta)$ specified up to a vector θ of unknown parameters. Suppose that the sample space $\mathcal{W}$ of W is partitioned into v mutually exclusive intervals $W_j,\ (j = 1, \ldots, v)$. Independent observations are made on W , but only the number n_j falling in $W_j,\ (j = 1, \ldots, r)$ is recorded where $r \leq v$. That is, individual observations are not recorded but only the class intervals W_j in which they fall are recorded; further, even such observations are made only if the W value falls in one of the intervals $W_j, (j = 1, \ldots, r)$. We can solve this problem within the EM framework by introducing the vectors

$$\boldsymbol{u} = (n_{r+1}, \ldots, n_v)^T, \quad \boldsymbol{w}_j = \left(w_{j1}, \ldots, w_{j,n_j}\right)^T \quad (j = 1, \ldots, v)$$

as the missing data. The vector $\boldsymbol{u}$ contains the unobservable frequencies in the case of truncation $(r < v)$, while $\boldsymbol{w}_j$ contains the n_j unobservable individual observations in the jth interval $W_j,\ (j = 1, \ldots, v)$. The complete-data vector $\boldsymbol{x}$ corresponding to the missing data is

$$\boldsymbol{x} = \left(\boldsymbol{y}^T, \boldsymbol{u}^T, \boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_v^T\right)^T.$$

The complete-data log likelihood function for $\boldsymbol{\theta},\ \ln L_c(\boldsymbol{\theta})$ is

$$\log L_c(\boldsymbol{\theta}) = \sum_{j=1}^{v} \sum_{l=1}^{n_j} \log f\left(w_{jl}; \boldsymbol{\theta}\right).$$

To perform the $E-$step, Q can be computed to be

$$\begin{aligned}
Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right) &= \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\log L_c(\boldsymbol{\theta}) \mid \boldsymbol{y}\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\sum_{j=1}^{v} \sum_{l=1}^{n_j} \log f\left(W_{jl}; \boldsymbol{\theta}\right) \mid \boldsymbol{y}\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\sum_{j=1}^{v} \sum_{l=1}^{n_j} \log f\left(W_{jl}; \boldsymbol{\theta}\right) \mid \boldsymbol{y}, u\right]\right] \\
&= \sum_{j=1}^{v} \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[n_j \mid \boldsymbol{y}\right] \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\log f\left(W; \boldsymbol{\theta}\right) \mid W \in \mathcal{W}_j\right] \\
&= \sum_{j=1}^{v} n_j^{(k)} Q_j(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})
\end{aligned}$$

where

$$Q_j(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\log f(W; \boldsymbol{\theta}) \mid W \in \mathcal{W}_j\right]$$

and

$$n_j^{(k)} = n_j, \quad 1 \leq j \leq r$$

$$= \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[n_j \mid \boldsymbol{y}\right] = \frac{nP_j(\boldsymbol{\theta}^{(k)})}{P(\boldsymbol{\theta}^{(k)})}, j = v.$$

To perform the $M-$step,

$$\partial Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right)/\partial\boldsymbol{\theta} = \sum_{j=1}^{v} n_j^{(k)} \partial Q_j\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right)/\partial\boldsymbol{\theta}$$

where

$$\partial Q_j\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right)/\partial\boldsymbol{\theta} = E_{\boldsymbol{\theta}^{(k)}}\left\{\partial \log f(W;\boldsymbol{\theta})/\partial\boldsymbol{\theta} \mid W \in \mathcal{W}_j\right\}$$

on interchanging the operations of differentiation and expectation. In this case

$$\begin{aligned} Q_j(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\log\left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{W-\mu}{\sigma}\right)^2\right)\right) \mid W \in \mathcal{W}_j\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[-\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(\sigma^2\right) - \frac{1}{2}\left(\frac{W-\mu}{\sigma}\right)^2 \mid W \in \mathcal{W}_j\right] \\ &= -\frac{1}{2}\left(\log\left(2\pi\right) + \log\left(\sigma^2\right)\right) - \frac{\sigma^{-2}}{2}\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[(W-\mu)^2 \mid W \in \mathcal{W}_j\right] \end{aligned}$$

so that

$$\begin{aligned} Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right) &= \sum_{j=1}^{v} n_j^{(k)} Q_j(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) \\ &= \sum_{j=1}^{v} n_j^{(k)}\left[-\frac{1}{2}\left(\log\left(2\pi\right) + \log\left(\sigma^2\right)\right) - \frac{\sigma^{-2}}{2}\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[(W-\mu)^2 \mid W \in \mathcal{W}_j\right]\right] \\ &= -\frac{1}{2}\sum_{j=1}^{v} n_j^{(k)}\left(\log\left(2\pi\right) + \log\left(\sigma^2\right)\right) - \frac{\sigma^{-2}}{2}\sum_{j=1}^{v} n_j^{(k)}\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[(W-\mu)^2 \mid W \in \mathcal{W}_j\right]. \end{aligned}$$

To complete the $M-$step, from the above we have

$$\begin{aligned} \partial Q_j\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right)/\partial\mu &= \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left\{\partial \log f(W;\boldsymbol{\theta})/\partial\mu \mid W \in \mathcal{W}_j\right\} \\ &= \frac{\sigma^{-2}}{2}\sum_{j=1}^{v} n_j^{(k)}\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[(W-\mu) \mid W \in \mathcal{W}_j\right] \end{aligned}$$

so that

$$\begin{aligned} &\frac{\sigma^{-2}}{2}\sum_{j=1}^{v} n_j^{(k)}\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\left(W-\mu^{(k+1)}\right) \mid W \in \mathcal{W}_j\right] = 0 \\ &\sum_{j=1}^{v} n_j^{(k)}\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\left(W-\mu^{(k+1)}\right) \mid W \in \mathcal{W}_j\right] = 0 \\ &\sum_{j=1}^{v} n_j^{(k)}\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[W \mid W \in \mathcal{W}_j\right] - \mu^{(k+1)}\sum_{j=1}^{v} n_j^{(k)} = 0 \\ &\mu^{(k+1)} = \sum_{j=1}^{v} n_j^{(k)}\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[W \mid W \in \mathcal{W}_j\right]\left(\sum_{j=1}^{v} n_j^{(k)}\right)^{-1}. \end{aligned}$$

Do something similar for $\sigma^{2(k+1)}$ provides an iterate of

$$\sigma^{(k+1)^2} = \sum_{j=1}^{v} n_j^{(k)} \left[\mathbb{E}_{\theta^{(k)}} \left\{ \left(W - \mu^{(k+1)} \right)^2 \mid W \in \mathcal{W}_j \right\} \right] \left(\sum_{j=1}^{v} n_j^{(k)} \right)^{-1}.$$

Methods of Finding Tests.

Definition 65 (Hypothesis)**.** *A* **hypothesis** *is a statement about a population parameter* [Cas01, page 373].

Definition 66 (Null and Alternative Hypothesis)**.** *The complementary hypotheses in a hypothesis testing problem are called the* **null hypothesis** *and the* **alternative hypothesis***. They are denoted by H_0 and H_1 respectively* [Cas01, page 373].

Definition 67 (Hypothesis Test)**.** *A* **hypothesis testing procedure** *or* **hypothesis test** *is a rule that specifies*

- *For which sample values the decision is made to accept H_0 as true.*
- *For which sample values H_0 is rejected and H_1 is accepted as true.*

[Cas01, page 374].

Definition 68 (Acceptance and Rejection Regions)**.** *The subset of the sample space for which H_0 will be rejected is called the* **rejection region** *or* **critical region***. The complement of the rejection region is called the* **acceptance region** [Cas01, page 374].

Definition 69 (Acceptance and Rejection Regions)**.** *The subset of the sample space for which H_0 will be rejected is called the* **rejection region** *or* **critical region***. The complement of the rejection region is called the* **acceptance region** [Cas01, page 374].

Definition 70 (Likelihood Ratio Test (LRT))**.** *The* **likelihood ratio test statistic** *for tetsing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is*

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\Theta_0} L\left(\theta \mid \boldsymbol{x}\right)}{\sup_{\Theta} L\left(\theta \mid \boldsymbol{x}\right)}.$$

A **likelihood ratio test** *(LRT) is any test that has a rejection region of the form $\{\boldsymbol{x} : \lambda\left(\boldsymbol{x}\right) \leq c\}$, where c is any number satisfying $0 \leq c \leq 1$* [Cas01, page 375].

Example 71 (Normal LRT). Example taken from [Cas01, page 375]. Let $X_1, X_2, \ldots, X_n$ be a random sample from a $\mathsf{N}\left(\theta, 1\right)$ population. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Here θ_0 is a number fixed by the experimenter prior the experiment. Since there is only one value of θ specified by H_0, the numerator of $\lambda\left(\boldsymbol{x}\right)$ is $L\left(\theta_0 \mid \boldsymbol{x}\right)$. In Example 33, the (unrestricted MLE) we found to be $\overline{X}$, the sample mean. Thus the denominator of $\lambda(\boldsymbol{x})$ is $L(\overline{x} \mid \boldsymbol{x})$. So the LRT statistic is

$$\begin{aligned}\lambda(\boldsymbol{x}) &= \frac{(2\pi)^{-n/2} \exp\left[-\sum_{i=1}^{n}(x_i - \theta_0)^2/2\right]}{(2\pi)^{-n/2} \exp\left[-\sum_{i=1}^{n}(x_i - \overline{x})^2/2\right]} \\ &= \exp\left[\left(-\sum_{i=1}^{n}(x_i - \theta_0)^2 + \sum_{i=1}^{n}(x_i - \overline{x})^2\right)/2\right].\end{aligned}$$

The expression for $\lambda(\boldsymbol{x})$ can be simplified by nothing that

$$\sum_{i=1}^{n}(x_i - \theta_0)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 n(\overline{x} - \theta_0)^2.$$

Thus the LRT statistic is $\lambda(\boldsymbol{x}) = \exp\left[-n\left(\overline{x} - \theta_0\right)^2/2\right]$.

Theorem 72. *If $T(\boldsymbol{X})$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(x)$ are the LRT statistics based on T and $\boldsymbol{X}$, respectively, then $\lambda^*(T(\boldsymbol{X})) = \lambda(\boldsymbol{x})$ for every $\boldsymbol{x}$ in the sample space* [Cas01, page 376].

Definition 73 (Type I and Type II Errors)**.** *Suppose R denotes the rejection of rejection for a test. then for $\theta \in \Theta_0$, the test will make a mistake if $\boldsymbol{x} \in R$, so the probability of a* **Type I Error** *is $\mathbb{P}_\theta(\boldsymbol{X} \in R)$. For $\theta \in \Theta_0^c$, the probability of a* **Type II Error** *is $\mathbb{P}(\boldsymbol{X} \in R^c)$* [Cas01, page 383].

Definition 74 (Power Function)**.** *The* **power function** *of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = \mathbb{P}_\theta(\boldsymbol{X} \in R)$* [Cas01, page 383].

Example 75 (Normal Power Function). Example taken from [Cas01, page 375]. Let $X_1, X_2, \ldots, X_n$ be a random sample from a $\mathsf{N}(\theta, \sigma^2)$ population. AN LRT of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ is a test that rejects H_0 is $\left(\overline{X} - \theta_0\right) / (\sigma/\sqrt{n}) > c$. That constant c can be any positive number. The power function of this test is

$$\begin{aligned}\beta(\theta) &= \mathbb{P}\left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} > c\right) \\ &= \mathbb{P}\left(\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\ &= \mathbb{P}\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right),\end{aligned}$$

where Z is a standard normal random variable, since $\left(\overline{X} - \theta\right) / (\sigma/\sqrt{n}) \sim \mathsf{N}(0, 1)$. As θ increases from $-\infty$ to ∞, it is easy to see that this normal probability increases from 0 to 1. Therefore, it follows that $\beta(\theta)$ is an increasing function of θ, with

$$\lim_{\theta \to -\infty} \beta(\theta) = 0, \ \lim_{\theta \to \infty} \beta(\theta) = 1, \text{ and } \beta(\theta_0) = \alpha \text{ if } \mathbb{P}(Z > c) = \alpha.$$

Suppose the experimenter wishes to have a maximum Type I Error probability of 0.1. Suppose, in addition, the experimenter wishes to have a maximum Type II Error probability of 0.2 if $\theta \geq \theta_0 + \sigma$. We now show how to choose c and n to achieve these goals, using a test that rejects $H_0 : \theta \leq \theta_0$ if $\left(\overline{X} - \theta_0\right) / (\sigma/\sqrt{n}) > c$. As noted above, the power function of such a test is

$$\beta(\theta) = \mathbb{P}\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right).$$

Because $\beta(\theta)$ is increasing in θ, the requirements will be met if

$$\beta(\theta_0) = 0.1 \quad \text{and} \quad \beta(\theta_0 + \sigma) = 0.8.$$

By choosing $c = 1.28$, we achieve $\beta(\theta_0) = \mathbb{P}(Z > 1.28) = 0.1$, regardless of n. Now we wish to choose n so that $\beta(\theta_0 + \sigma) = \mathbb{P}(Z > 1.28 - \sqrt{n}) = 0.8$. But, $\mathbb{P}(Z > -0.84) = 0.8$. So setting $1.28 - \sqrt{n} = -0.84$ and solving for n yields $n = 4.49$. Of course n must be an integer. So choosing $c = 1.28$ and $n = 5$ yields a test with error probabilities controlled as a specified by the experimenter.

Definition 76 (Size α Test)**.** *For $0 \leq \alpha \leq 1$, a test with a power function $\beta(\theta)$ is a* **size α test** *if $\sup_{\theta \in \Theta_0} = \alpha$* [Cas01, page 385].

Definition 77 (Level α Test)**.** *For $0 \leq \alpha \leq 1$, a test with a power function $\beta(\theta)$ is a* **level α test** *if $\sup_{\theta \in \Theta_0} \leq \alpha$* [Cas01, page 385].

Definition 78 (Unbiased Tests)**.** *A test with a power function $\beta(\theta)$ is* **unbiased** *if $\beta(\theta') \geq \beta(\theta'')$ for every $\theta' \in \Theta_0^c$ and $\theta'' \in \Theta_0$* [Cas01, page 387].

Uniformly Most Powerful Tests.

Definition 79 (Uniformly Most Powerful Tests)**.** *Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test in class $\mathcal{C}$, with power function $\beta(\theta)$, is a* **uniformly most powerful** *(UMP) class $\mathcal{C}$ test if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta_0^c$ and every $\beta'(\theta)$ that is a power function of a test in class $\mathcal{C}$* [Cas01, page 388].

Theorem 80 (Neyman-Pearson Lemma)**.** *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the pdf or pmf corressponding to θ_i is $f(\boldsymbol{x} \mid \theta_i),\ i = 0, 1$, using a test with rejection region R that satisfies*

$$\boldsymbol{x} \in R \quad \text{if} \quad f(\boldsymbol{x} \mid \theta_1) > k f(\boldsymbol{x} \mid \theta_0) \tag{20}$$

and

$$\boldsymbol{x} \in R^c \quad \text{if} \quad f(\boldsymbol{x} \mid \theta_1) < k f(\boldsymbol{x} \mid \theta_0) \tag{21}$$

for some $k \geq 0$, and

$$\alpha = \mathbb{P}_{\theta_0}(\boldsymbol{X} \in R). \tag{22}$$

Then

- *(Sufficiency) Any test that satisfies 20, 21 and 22 is a UMP α test.*
- *(Necessity) If there exists a test satisfying 20, 21 and 22 with $k > 0$, then every UMP level α test is a size α test (satisfies 22) and every UMP level α test satisfies 20 and 21 except perhaps on a set A satisfying $\mathbb{P}_{\theta_0}(\boldsymbol{X} \in A) = \mathbb{P}_{\theta_1}(\boldsymbol{X} \in A) = 0$*

[Cas01, page 388].

Corollary 81. *Consider the hypothesis problem posed in 80. Suppose $T(\boldsymbol{X})$ is a sufficient statistic (see 5) for θ and $g(t \mid \theta_i)$ is the pdf or pmf of T corressponding to $\theta_i = 0, 1$. Then any test based of T with rejection region S (a subset of the sample space of T) is a UMP level α test if it satisfies*

$$t \in S \quad \text{if} \quad g(t \mid \theta_1) > k g(t \mid \theta_0) \tag{23}$$

and

$$t \in S^c \quad \text{if} \quad g(t \mid \theta_1) < k g(t \mid \theta_0) \tag{24}$$

for some $k \geq 0$, where

$$\alpha = \mathbb{P}_{\theta_0}(T \in S) \tag{25}$$

[Cas01, page 389].

Definition 82 (Monotone Likelihood Ratio)**.** *A family of pdfs or pmfs $\{g(t \mid \theta) : \theta \in \Theta\}$ for a univariate random variable T with real-valued parameter θ has a* **monotone likelihood ratio** *(MLR) if, for every $\theta_2 > \theta_1$, $g(t \mid \theta_2)/g(t \mid \theta_1)$ is a monotone (nonincreasing or nondecreasing) function of t on $\{t : g(t \mid \theta_1) > 0 \text{ or } g(t \mid \theta_2) > 0\}$. Note that $c/0$ is defind as ∞ if $0 < c$* [Cas01, page 391].

Theorem 83 (Karlin-Rubin)**.** *Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose that T is a sufficient statistic for θ and the family of pdfs or pmfs $\{g(t \mid \theta) : \theta \in \Theta\}$ of T has a MLR. Then for any t_0, the test that rejects H_0 if and only if $T > t_0$ is a UMP level α test, where $\alpha = \mathbb{P}_{\theta_0}(T > t_0)$* [Cas01, page 391].

Example 84 (Karlin-Rubin Applied to Beta Distribution). Example taken from [Cas01, page 405]. Suppose X is one observation from a population with $\text{Beta}(\theta, 1)$ pdf. Furthermore, consider testing $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$ where the null hypothesis is rejected if $X > \frac{1}{2}$. The power function is simply

$$\begin{aligned}\beta(\theta) &= \mathbb{P}_\theta\left(X > \frac{1}{2}\right) \\ &= \int_{1/2}^{1} \frac{\Gamma(\theta+1)}{\Gamma(\theta)\Gamma(1)} x^{\theta-1}(1-x)^{1-1}\, dx \\ &= \left[\theta\frac{1}{\theta}x^\theta\right]_{1/2}^{1} = 1 - \frac{1}{2}^\theta.\end{aligned}$$

This size is $\sup_{\theta\in H_0}\beta(\theta) = \sup_{\theta\leq 1}\left(1-\frac{1}{2}^\theta\right) = 1 - 1/2 = 1/2$.

Now suppose we would like to find the powerful level α test for testing $H_0 : \theta = 1$ versus $H_1 : \theta = 2$. By Theorem 80, the most powerful test of $H_0 : \theta = 1$ versus $H_1 : \theta = 2$ is given by rejecting H_0 if $f(x \mid 2)/f(x \mid 1) > k$ for some $k > 0$. Substituting the beta pdf gives

$$\frac{f(x \mid 2)}{f(x \mid 1)} = \frac{\frac{1}{\Gamma(1)\Gamma(2)/\Gamma(3)} x^{2-1}(1-x)^{1-1}}{\frac{1}{\Gamma(1)\Gamma(1)/\Gamma(2)} x^{1-1}(1-x)^{1-1}} = \frac{\Gamma(3)}{\Gamma(2)\Gamma(1)} x = 2x.$$

Thus the MP test will reject H_0 if $2x > k$. We can now use the α to determine a suitable k. We have

$$\alpha = \sup_{\theta\in\Theta_0}\beta(\theta) = \beta(1) = \int_{k/2}^{1} f_X(x \mid 1)\, dx = \int_{k/2}^{1} \frac{1}{\Gamma(1)\Gamma(1)/\Gamma(2)} x^{1-1}(1-x)^{1-1}\, dx = 1 - \frac{k}{2}.$$

Thus $1 - k/2 = \alpha$, so the most powerful α level test is to reject H_0 if $X > 1 - \alpha$.

Now consider the existence of a UMP test for testing $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$. For $\theta_2 > \theta_1$ we have $\frac{f(x|2)}{f(x|1)} = (\theta_2/\theta_1)x^{\theta_2-\theta_1}$, an increasing function of x (since $\theta_2 > \theta_1$). So this family has MLR (see Definition 82). By Theorem 83, the test that rejects H_0 if $X > t$ is the UMP test of this size.

p-Values.

Definition 85 (p-Values)**.** *A* **p-value** $p(\boldsymbol{X})$ *is a test statistic satisfying* $0 \leq p(\boldsymbol{x}) \leq 1$ *for every sample point* $\boldsymbol{x}$*. Small value of* $p(\boldsymbol{X})$ *give evidence that* H_1 *is true. A p-value is* **valid** *if, for every* $\theta \in \Theta_0$ *and every* $0 \leq \alpha 1$*,*

$$\mathbb{P}_\theta \left(p(\boldsymbol{X}) \leq \alpha \right) \leq \alpha$$

[Cas01, page 397].

Theorem 86. *Let* $W(\boldsymbol{X})$ *be a test statistic such that large values of* W *give evidence that* H_1 *is true. For each sample point* $\boldsymbol{x}$*, define*

$$p(\boldsymbol{x}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left(W(\boldsymbol{X}) \geq W(\boldsymbol{x}) \right).$$

Then, $p(\boldsymbol{X})$ *is a valid p-value* [Cas01, page 397].

Monte Carlo Sampling.

Definition 87 (Empirical distribution)**.** *Let $x_1, \ldots, x_n$ be an iid real-valued sample from a cdf F. The function*

$$F_n(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{x_i \leq x\}}(x), \quad x \in \mathbb{R}$$

is called the **empirical cdf** *of the data* [Kro13, page 196].

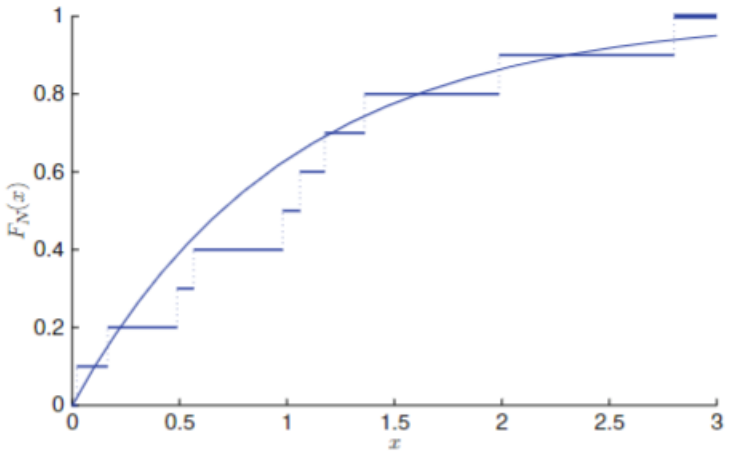


Figure 1. The empirical cdf 87 for a sample size 10 from a $\mathsf{Exp}(0.2)$ distribution as well as the true cdf. Image from [Kro13, page 197].

Note that for the ordered sample $x_{(1)} < x_{(2)} < \cdots < x_{(N)}$,

$$F_N\left(x_{(i)}\right) = \frac{i}{N}.$$

assuming for the sake of simplicity that all $\{x_i\}$ take on different values. If instead of deterministic $\{x_i\}$ we take random X_i, then $F_N(x)$ becomes random as well. To distinguish between the deterministic and the random case, let us denote the random empirical cdf by $\hat{F}_N(x)$. We now have

$$\mathbb{P}\left[\hat{F}_N(x) = \frac{i}{N}\right] = \mathbb{P}\left[X_{(i)} \leq x, X_{(i+1)} > x\right] = \binom{N}{i} \left(F(x)\right)^i \left(1 - F(x)\right)^{N-i}$$

[Kro13, page 198]. The above equation can be summarized as $N\hat{F}_N(x) \sim \mathsf{Bin}(N, F(x))$. Consequently,

$$\begin{aligned} \mathbb{E}\left[\hat{F}_N(x)\right] &= F(x) \\ \mathrm{Var}\left[\hat{F}_N(x)\right] &= F(x)\left(1 - F(x)\right). \end{aligned}$$

Moreover, by the law of large numbers and the central limit theorem, we have

$$\mathbb{P}\left[\lim_{N\to\infty} \hat{F}_N(x) = F(x)\right] = 1,$$

and

$$\lim_{N\to\infty} \mathbb{P}\left[\frac{\hat{F}_N(x) - F(x)}{\sqrt{F(x)(1-F(x)/N)}}\right] = \Phi(z).$$

Definition 88 (Confidence Intervals)**.** *Let $X_1, \ldots, X_n$ be random variables with a joint distribution depdending on a parameter $\theta \in \Theta$. Let $T_1 < T_2$ be functions of the data but not of θ. The random interval (T_1, T_2) is called a* **stochastic confidence interval** *for θ with confidence $1-\alpha$ if*

$$\mathbb{P}_\theta\left[T_1 < \theta < T_2\right] \geq 1-\alpha \quad \textit{for all} \quad \theta \in \Theta.$$

If t_1 and t_2 are the observed values of T_1 and T_2, then the interval (t_1, t_2) is called the **numerical confidence interval** *for θ with confidence $1-\alpha$* [Kro13, page 128].

An approximate $1-\alpha$ confidence interval for $F(x)$ is

$$F_N(x) \pm z_{1-\alpha/2}\sqrt{\frac{F_N(x)(1-F_N(x))}{N}}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Moreover for the ordered sample $x_{(1)} < x_{(2)} < \cdots < x_{(N)}$, an approximate $1-\alpha$ confidence interval for $F(x_{(i)})$ is

$$\frac{i}{N} \pm z_{1-\alpha/2}\sqrt{\frac{i(1-i/n)}{N^2}}$$

[Kro13, page 198].

Simultaneous Confidence Bands.

Definition 89 (Kolmogorov–Smirnov statistic)**.** *For a continuous F, the statistic of*

$$D_n = \sup_{x\in\mathbb{R}} \left|\hat{F}_N(x) - F(x)\right|$$

is called the **Kolmogorov–Smirnov statistic** [Kro13, page 200].

Note that this statistic does not depend on F and can be used to test whether iid samples $X_1, \ldots, X_n$ come from a specified distribution as well as constructing a simultaneous confidence band for F. The empirical cdf can be used to estimate any cdf, both discrete and continuous, but it is always 'jumpy' which may be undesirable, for example when the underlying distribution is continuous. When we have continuous data, we may want a continuous density estimate instead. This leads to our next topic of density estimation.

Kernel Density Estimation.

Definition 90 (Kernel Density Estimator)**.** *Let $X_1, \ldots X_n \overset{iid}{\sim} f$ from a continuous distribution with cdf F. A* **Kernel density estimator** *(KDE), also known as a sliding window estimator, with bandwidth h is given by*

$$\hat{f}(x;h) = \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right).$$

The function $K(\cdot)$ is the kernel function ("window shape") and h is the window ("window size").

There are many choice of K, but we shall just focus on the Gaussian kernel.

Definition 91 (Gaussian Kernel)**.** *The* **Gaussian Kernel** *uses a kernel of*

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

so that

$$\hat{f}(x;h) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i - x}{h}\right)^2\right)$$

[Kro13, page 201].

Essentially, the Gaussian kernel is just an equally weighted $\frac{1}{N}$ normal mixture model. It has N Gaussian "humps", centered at each observation and has standard deviation h. As it turns out, the choice of the kernel is not that crucial, but the choice of the bandwidth is very important. An often used *rule of thumb* is to take

$$h_{\text{Rot}} = \left(\frac{4S^5}{3N}\right)^{4/5}$$

called the Silverman's rule of thumb where S^2 is the variance. This is based on a theoretical analysis of the discrepancy between $f_n(x;h)$ and the true pdf, as $n \to \infty$. This rule of thumb is best suited when the underlying data is roughly normal-looking (i.e unimodel and symmeteric). A more recent bandwidth selection methods [ZIBaJFGaDPK10], called the theta KDE, adaptively changes the bandwidth.

Bootstrap Method. A key concept in statistics is the idea of a sampling distribution (i.e. the distribution of the sample quantity, if we could repeat the random experiment over and over again). Unfortunately, we typically only get one realization of the random process, from which we compute sample quantity we want (for example the sample mean or medium). So, how can we get a feel about the sampling distribution if we have only one realization?

Example 92 (Bootstrap Median). Suppose we have the following sample

$$6.45, 3.52, 3.92, 0.64, 3.73, 2.00, 2.73, 4.20, 20.58, 6.80, 3.22, 7.11, 1.08, 5.94.$$

From this data, how do we construct a confidence interval for the true population medium M of the distribution F that generated the data? Ideally we sample from F over and over again and each time we compute the sample medium and then we look at the distribution of the computed sample medians. Since we don't have access to the distribution F, we can instead employ the bootstrap method.

- To start we take iid sample of size n, x^*, from the empirical cdf F_n with replacement.
- For each sample, we compute the sample medium m^* of the resampled data.
- Then we look at the 2.5 percentile and the 97.5 percentile of the resampled values of the mediums.
- This percentiles together provide an approximate 95 percent confidence interval for the population median M.

We can use the bootstrap method to estimate any property of the sample distribution, for example, the variance of the sample mean, medium or IQR. In general we may be interested in some property h (eg. bias, variance, MSE) of some statistic $H(\boldsymbol{x})$ (eg. median, $\frac{1}{\bar{X}}$) we can apply the bootstrap resampling approach to estimate these properties.

Algorithm 1: Bootstrap Method

input : Observations $\boldsymbol{x}$.

output: Estimate of $\mathbb{E}\left[h(H(\boldsymbol{x}))\right]$.

for $i \leftarrow 1, \ldots, K$ **do**
  Resample data $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_n^*$ from the original data.
  For each sample, calculate $h(H_k^*) = h(H(\boldsymbol{x}_k^*))$
end
Estimate $\mathbb{E}\left[h(H(\boldsymbol{x}))\right]$ using $\frac{1}{K}\sum_{i=1}^{K} h(H_k^*)$

For example to compute the variance of the sample medium we can compute

$$\mathbb{V}(\text{medium}) = \mathbb{E}\left[(m - \mathbb{E}(m))\right]^2 \simeq \sum_{i=1}^{K}(m_i^* - \overline{m}^*)^2$$

where m_i^* is the sample median of the ith resampled dataset and $\overline{m}^*$ is the mean of the resampled median values. If we replace F by the empirical cdf, F_n, this is non-parameteric bootstrap. On this otherhand if replace F by $F_{\hat{\theta}}$, this is parameteric bootstrap. As an example, the mean of the bootstrap estimate of the expectation of H is

$$\widehat{\mathbb{E}H} = \bar{H}^* = \frac{1}{K}\sum_{i=1}^{K} H_i^*,$$

which is simply the sample mean of $\{H_i^*\}$. Similarly, the bootstrap estimate for $\mathbb{V}(H)$ is the sample variance

$$\widehat{\mathbb{V}(H)} = \frac{1}{K}\sum_{i=1}^{K}\left(H_i^* - \bar{H}^*\right)^2$$

[Kro13, page 206].

Markov Chain Monte Carlo. All of the sampling methods so far require us to know the explicit form of the distribution we are sampling from. For example, we know how to sample from $\mathsf{U}[0,1], \mathsf{N}(\mu, \sigma^2), \ldots, F_n$. But often we want to sample from a distribution F for which we don't know the exact form. In a special cases, we may know F up to a normalizing constant, or related to some other known function.

Example 93. Consider sampling from the distribution

$$f(x) \propto x^2 \exp\left(-x^2 + \sin(x)\right).$$

Here, the normalizing constant c has no explicit form. From STAT3004, we can approximately generate samples from a "target" distribution by instead sampling from a Markov Chain whose limiting distribution is the target distribution.

This is called Markov Chain Monte Carlo (MCMC). We *could* try and integrate $f(x)$ numerically. However, for this amount of computation we could have already obtained an MCMC sample already. Moreover, after we find c, obtaining the quantity such as quantiles is not easy (even if we know c). This is again very expensive for quantile.

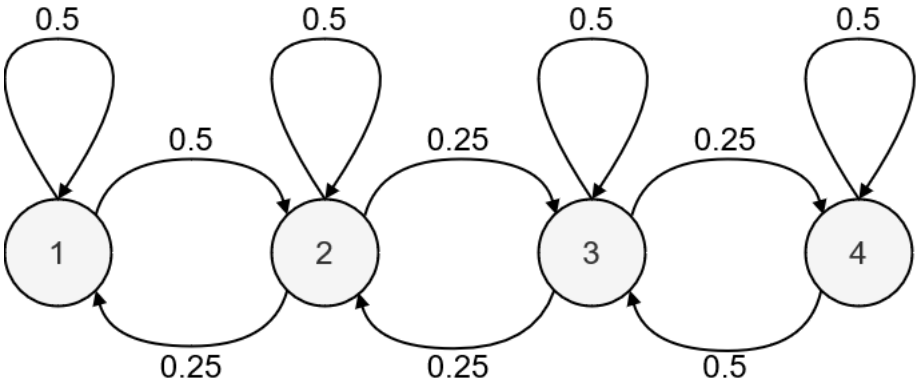
Definition 94 (Markov Chain)**.** *A* **Markov Chain** *is a collection of random variables* $\{X_t : t = 0, 1, \ldots\}$ *indexed by* t *such that* $(X_{t+1} \mid X_t, X_{t-1} \mid X_0) = (X_{t+1} \mid X_t)$ *for all* t*. This is called the* **Markov Property**.

To specify a Markov Chain we need to define a few things:

- The distribution of X_0, called the initial distribution of a MC
- The probability of getting from one state to another. That is probability of getting from state y when we are currently at the state x, $q(y \mid x)$ (the transition probability).

Both of these together will complete determine the distribution of the Markov Chain.

Example 95. Consider the 4 states shown in the diagram below. Suppose the probability of staying in the current state is 0.5 and the probability of moving to a neighbouring state is equally likely, as depicted below.



The first question we might ask is what are the transition probabilities? We also might want to ask, what is the long-term relative frequency of visiting each of the four states? We can compute probabilities of the first question analytically as

$$
\begin{aligned}
q(1 \mid 1) &= q(2 \mid 1) = 0.5 \\
q(3 \mid 1) &= q(4 \mid 1) = 0 \\
q(2 \mid 2) &= 0.5 \\
q(3 \mid 2) &= q(1 \mid 2) = 0.25 \\
q(4 \mid 2) &= 0 \\
&\vdots
\end{aligned}
$$

which can be written in matrix form as

$$
\boldsymbol{Q} = \begin{pmatrix}
0.5 & 0.25 & 0 & 0 \\
0.5 & 0.5 & 0.25 & 0 \\
0 & 0.25 & 0.5 & 0.5 \\
0 & 0 & 0.25 & 0.5
\end{pmatrix}
$$

were $\boldsymbol{Q}_{ij} = q(i \mid j)$. We will see soon how to answer the second question.

Definition 96 (Ergodic Markov Chain)**.** *A Markov Chain is* **Ergodic** *if the probability of X_t converges to some fixed probability function $f(x)$ as $t \to \infty$.*

Definition 97 (Global Balance Equation)**.** *For an ergodic MC, the limiting probability function $f(x)$ must satisfy a set of recursive relationships, called the* **global balance equations***.*

$$f(x) = \sum_y f(y) q\left(x \mid y\right)$$

in the discrete case and

$$f(x) = \int_y f(y) q(x \mid y)\, dy$$

in the continuous case. The function $f(x)$ is called the long term frequency at state x.

Example 98 (Continuation of 95). Hence we can find a limiting distribution $\boldsymbol{f}$ using the global balance equations. Again we have the transition matrix

$$\boldsymbol{Q} = \begin{pmatrix} 0.5 & 0.25 & 0 & 0 \\ 0.5 & 0.5 & 0.25 & 0 \\ 0 & 0.25 & 0.5 & 0.5 \\ 0 & 0 & 0.25 & 0.5 \end{pmatrix}.$$

Let's calculate f using the global balance equations.

$$\begin{aligned} f(1) &= \sum_{i=1}^{4} f(i) q(1 \mid i) = 0.75 \\ f(2) &= \sum_{i=1}^{4} f(i) q(2 \mid i) = 1.25 \\ f(3) &= \sum_{i=1}^{4} f(i) q(3 \mid i) = 1.25 \\ f(4) &= \sum_{i=1}^{4} f(i) q(4 \mid i) = 0.75. \end{aligned}$$

We can use computer sotware to solve this. For the discrete case, the global balance equations can be written in matrix form as $\boldsymbol{Q}\boldsymbol{f} = \boldsymbol{f}$, where $\boldsymbol{Q}$ is the one-step transition matrix and $\boldsymbol{f}$ the *row vector of the limiting probabilities*. This leads to solving the linear equation $(\mathbb{1} - \boldsymbol{Q})\boldsymbol{f} = \boldsymbol{0}$, where $\mathbb{1}$ is the identity matrix (with appropriate size). We also know that since $\boldsymbol{f}$ is a distribution, meaning must also impose the constraint $\sum_i \boldsymbol{f}(i) = 1$. This can be carried out using the following lines of `MatLab` code.

```
>> Q = [0.5,0.25,0,0;0.5,0.5,0.25,0;0,0.25,0.5,0.5;0,0,0.25,0.5];
>> f = null(eye(4) - Q);
>> f = f/sum(f);
>> f
f =

    0.1667
```

```
0.3333
0.3333
0.1667
```

Thus the limiting probabilities are $\approx [0.166, 0.333, 0.333, 0.166]^\intercal$.

Definition 99 (Time Reversible (Markov Chain))**.** *A MC is called* **time-reversible** *if it is a MC when run backwards in time. Note, not all MC are reversible but all ergodic MCs are reversible.*

Definition 100 (Reverse Transition Probabilities)**.** *The* **reverse transition probability** *is*

$$\tilde{q}(y \mid x) = \frac{f(y)q(x \mid y)}{f(x)}.$$

The influx into state y from state x in the reverse chain is $\tilde{q}(y \mid x)f(x)$ while the influx into state x from state y in the forward chain is $q(y \mid x)f(y)$. At the limit state, the forward MC and backward chain should look the same, that is

$$f(x)q(y \mid x) = f(y)q(x \mid y)$$

for all pairs x and y. This set of equations are known as the **local (or detailed) balance equations**.

Example 101 (Continuation of 98). This suggests we can also solve $\boldsymbol{f}$ from Example 98 using the local balance equations. For $x = y$

$$\boldsymbol{f}(x)q(x \mid x) = \boldsymbol{f}(x)q(x \mid x).$$

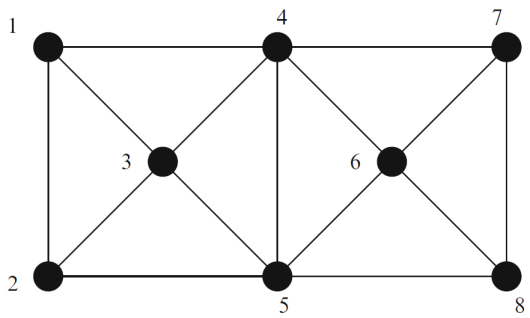
For $x = 1, y = 2$

$$\begin{aligned}\boldsymbol{f}(1)q(2 \mid 1) &= \boldsymbol{f}(2)q(1 \mid 2)\\ 2\boldsymbol{f}(1) &= \boldsymbol{f}(2).\end{aligned}$$

Similarly for $x = 4, y = 3$ we have $2\boldsymbol{f}(4) = \boldsymbol{f}(1)$ and by symmetry for $x = 2, y = 3$ we have $\boldsymbol{f}(2) = \boldsymbol{f}(3)$. This means

$$\begin{aligned}&[\boldsymbol{f}(1), \boldsymbol{f}(2), \boldsymbol{f}(3), \boldsymbol{f}(4)] \propto [1, 2, 2, 1]\\ \Rightarrow\ &[\boldsymbol{f}(1), \boldsymbol{f}(2), \boldsymbol{f}(3), \boldsymbol{f}(4)] = [1/6, 2/6, 2/6, 1/6]\,.\end{aligned}$$

Example 102. Example taken from [Kro13, page 213]. Consider a Markov chain that performs a random walk on the graph in the below diagram, at each step jumping from the current vertex (node) to one of the adjacent vertices, with equal probability.



Clearly this Markov chain is reversible. It is also irreducible and aperiodic. Let $f(x)$ denote the limiting probability that the chain is in vertex x. By symmetry, $f(1) = f(2) = f(7) = f(8), f(4) = f(5)$, and $f(3) = f(6)$. Moreover by the detail balance equations $f(4)/5 = f(1)/3$, and $f(3)/4 = f(1)/3$. This means that f, $\sum_i f(i) = 4f(1) + 2 \cdot 5/3f(1) + 2 \cdot 4/3f(1) = 10f(1) = 1$ so that $f(1) = 1/10, f(2) = 2/15$, and $f(4) = 1/6$.

Metropolis Hastings Algorithm.

Example 103. Suppose we would like draw samples from

$$f(x) \propto x^2 \exp\left(-x^2 + \sin(x)\right).$$

Instead of sampling directly, we can build a MC whose limiting distribution is will be our target function $f(x)$. After a long burn, 0 to T, the random variables will form an approximate but dependent sample from $f(x)$. The idea of Metropolis and Hastings has a two phases.

- Proposal: Given we are currently at state x, we generate a proposal state Y, from some transition probability $q(\cdot \mid x)$.
- Accept/Reject: Accept Y with some probability $\alpha(x, Y)$, and reject with probability $1 - \alpha(x, Y)$.

We can choose $\alpha(x, y)$ such that we get $f(x)$ as our limiting distribution. To do this we set

$$\alpha(x, y) = \min\left\{\frac{f(y)q(x \mid y)}{f(x)q(y \mid x)}, 1\right\}.$$

If the ratio id greater that 1, we will always make the move. Otherwise we only sometimes make the move. We only need to know $f(x)$ up to a normalizing constant.

Proof. Check the local balance equations, for this MHMC. The one-step transition probability is

$$q_{MH}(y \mid x) = \begin{cases} q(y \mid x)\alpha(x, y), & \text{if } x \neq y \text{ (Case 1)} \\ 1 - \sum_{z \neq x} q(z \mid x)\alpha(x, z), & \text{if } x = y \text{ (Case 2)} \end{cases}.$$

Let us check the local balance equation to find the limiting distribution. For Case 1 ($y \neq x$), the left hand side of the local balance equation is

$$\begin{aligned} LHS &= f(x)q_{MH}(y \mid x) \\ &= f(x)q(y \mid x)\alpha(x, y) \\ &= f(x)q(y \mid x) \min\left\{\frac{f(y)q(x \mid y)}{f(x)q(y \mid x)}, 1\right\} \end{aligned}$$

if $\frac{f(y)q(x|y)}{f(x)q(y|x)} \geq 1$ then this becomes $f(x)q(y \mid x)$. On the other hand if $\frac{f(y)q(x|y)}{f(x)q(y|x)} < 1$ or $\frac{f(x)q(y|x)}{f(y)q(x|y)} > 1$, then

$$\begin{aligned} & f(x)q(y \mid x)\alpha(x, y) \\ &= f(x)q(y \mid x)\frac{f(y)q(x \mid y)}{f(x)q(y \mid x)} \\ &= f(y)q(x \mid y). \end{aligned}$$

The right hand side of the local balance equation is

$$\begin{aligned} RHS &= f(y)q_{MH}(x \mid y) \\ &= f(y)q(x \mid y)\min\left\{\frac{f(y)q(x \mid y)}{f(x)q(y \mid x)}, 1\right\} \\ &= \begin{cases} f(y)q(x \mid y), & \text{if ratio} \geq 1 \\ f(x)q(y \mid x), & \text{if ratio} < 1 \end{cases}. \end{aligned}$$

For Case 2

$$LHS = f(x)q_{MH}(y \mid x) = RHS$$

so $f(x)$ is indeed the limiting distribution if the MH samples. □

Example 104 (Example 7.12 form Dirk). At the current state x, propose a new state $Y = x + Z$, where Z has a symmeteric distribution around 0 (eg Z comes from a standard normal distribution). In this case

$$\alpha(x, y) = \min\left\{\frac{f(y)q(x \mid y)}{f(x)q(y \mid x)}, 1\right\} = \min\left\{\frac{f(y)}{f(x)}, 1\right\}.$$

Example 105. Choose Y from $g(y \mid x)$ for some density $g(\cdot)$ that does not depend on the current state x. Accept this proposal to Y with probability

$$\alpha(x, y) = \min\left\{\frac{f(y)}{f(x)} \cdot \frac{g(x)}{g(y)}, 1\right\}$$

where $\frac{f(y)}{f(x)}$ is proportional to our target density f and $\frac{g(x)}{g(y)}$ is inversely proportional.

When might an independent sampler be better over a random walk sampler? It is usually better when there are both local and global features of our target f or when we can easily sample from our proposed g, that somehow resembles our target f.

Gibbs Samplings. Suppose we want to sample from a joint pdf $f(x)$. Direct sampling of joints pdfs are usually very difficult, especially if the dimension p is very high. In particular, numerical integration in high dimensions is very computationally expensive. What if instead we sample from each conditional pdf instead. This leads to the Gibbs Sampling Algorithm seen below.

Algorithm 2: Gibbs Sampling

input : A distribution f.
output: Samples from f.

Given the current state $X_t = x$, draw a vector $Y = y$ using sequnetial sampling from the conditional pdfs
$Y_1 \sim f(y_1 \mid x_2, x_3, \ldots, x_p)$
$Y_2 \sim f(y_2 \mid Y_1, x_3, \ldots, x_p)$
$Y_3 \sim f(y_3 \mid Y_1, Y_2, x_4, \ldots, x_p)$
$\vdots$
$Y_p \sim f(y_p \mid Y_1, Y_2, \ldots, Y_{p-1})$
Set $X_{t+1} = Y = (Y_1, Y_2, \ldots, Y_p)^\intercal$.
When this is run long enough ($> T$ simulations), then $X_t \sim f(x)$ for $t \geq T$.

This is actually special case of MH with transition probability from state x to y as

$$q_{1\to p}(y \mid x) = f_1(y_1 \mid x_2, x_3, \ldots x_p) \cdot f_2(y_2 \mid y_1, x_3, \ldots x_p) \cdot f_p(y_p \mid y_1, y_2, \ldots, y_{p-1}).$$

We want to check whether the $X_1, X_2, \ldots$ indeed has a limiting distribution f we want. Note that the global balance equation do not hold in general in this case. This is because of the one-step transition requiring p small sequential steps that are not reversible

$$f(x)q_{1\to p}(y \mid x) \neq f(y)q_{1\to p}(x \mid y).$$

But we do have a modified local balance equation

$$f(x)q_{1\to p}(y \mid x) = f(y)q_{p\to 1}(x \mid y)$$

where $f(x)q_{1\to p}(y \mid x)$ is the influx from state x into y in our forward chain and $f(y)q_{p\to 1}(x \mid y)$ is the influx from state y into x in our backward sequence. We can use the above expression to show that the global balance equations hold

$$\begin{aligned} RHS &= \int f(x)q_{1\to p}(y \mid x)\, dx \\ &= \int f(y)q_{p\to 1}(x \mid y)\, dx \\ &= f(y) \int q_{p\to 1}(x \mid y)\, dx \\ &= f(y) \cdot 1 = f(y) = LHS \end{aligned}$$

Example 106 (Example 7.14 form Dirk). Suppose we want to sample $\boldsymbol{X} = (X_1, X_2)^\intercal$ from the following bivariate pdf.

$$f(x_1, x_2) \propto \exp\left(-x_1x_2 - x_1 - x_2\right)$$

for $x_1 \geq 0, x_2 \geq 0$. The normalizing constant is not given meaning we cannot sample directly. What are the conditional pdfs?

$$f_1(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f(x_2)} \propto f(x_1, x_2)$$

as a function of x_1 only so that

$$f_1(x_1 \mid x_2) \propto f(x_1, x_2) \propto \exp\left(-x_1 x_2 - x_1 - x_2\right) \propto \exp\left(-x_1 x_2 - x_1\right) \sim \mathsf{Exp}(\lambda = x_2 + 1).$$

Also

$$f_2(x_2 \mid x_1) \propto \exp\left(-x_1 x_2 - x_1 - x_2\right) \sim \mathsf{Exp}(\lambda = x_1 + 1).$$

Given any starting vector $X^{(0)} = \left(X_1^{(0)}, X_2^{(0)} > \mathbf{0}\right)$, the Gibbs sampling procedure is

- $X_1^{(1)} \sim \mathsf{Exp}(\lambda = x_2^{(0)} + 1)$
- $X_2^{(1)} \sim \mathsf{Exp}(\lambda = x_1^{(1)} + 1)$

and we repeat this for as many iterations as we are allowed. After a burn-in period T, the sequence $\boldsymbol{X}_{T+1}, \boldsymbol{X}_{T+2}, \ldots$ forms an approximately (dependent) sample from f.

Bayesian Statistics.

Definition 107 (Prior, Liklihood and Posterior)**.** *Let $\boldsymbol{x}$ and $\boldsymbol{\theta}$ denote the data and the parameters in a Bayesian model and let $f(\boldsymbol{z})$ be the distribution of $\boldsymbol{z}$ (for some random variable $\boldsymbol{z}$)*

- *The pdf of $\boldsymbol{\theta}$ is called the* **prior** *pdf.*
- *The conditional pdf $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ is called the* **likelihood** *function.*
- *The central object of interest is the* **posterior** *pdf $f(\boldsymbol{\theta} \mid \boldsymbol{x})$ which, Baye's theorem, is proportional to the product of the prior and the likelihood:*

$$f(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}).$$

[Kro13, page 228].

There is also not much distinction between random variables and their realizations, both are written in lower case. Bayesian analysis usually consists of three main steps.

- Specify a prior belief or prior distribution $f(\boldsymbol{\theta})$ on the parameter $\boldsymbol{\theta}$. This represents our prior belief about $\boldsymbol{\theta}$ before looking at the data.
- The likelihood (model) $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ characterises how the distribution of $\boldsymbol{x}$ depends on $\boldsymbol{\theta}$.
- The posterior distribution of $\boldsymbol{\theta}$ is given by

$$f(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})$$

  where $f(\boldsymbol{\theta} \mid \boldsymbol{x})$ is called the **posterior**, $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ is called the likelihood and $f(\boldsymbol{\theta})$, the **prior**.

Recall Baye's rule

$$f(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x}, \boldsymbol{\theta})}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\boldsymbol{x})} \propto f(\boldsymbol{x} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})$$

Example 108 (Coin tosses). Suppose we tossed a coin 10 times, and observe S heads. What is the probability of θ of heads for this coin? To start we can specify a prior belief on this coin so that $\theta \sim \mathsf{U}[0,1]$, that is $f(\theta) = 1$ (all possible values of $\theta \in [0,1]$ have equal chance of occur). Next, suppose we observe $S = 3$ heads out of 10 tosses. The likelihood of the data is

$$f(x \mid \theta) = \binom{10}{3} \theta^3 (1-\theta)^7 \propto \theta^3 (1-\theta)^7.$$

The find the posterior we compute

$$\begin{aligned} f(\theta \mid \boldsymbol{x}) &\propto f(\boldsymbol{x} \mid \theta) f(\theta) \\ &= \theta^3 (1-\theta)^7 \cdot 1 \text{ for } \theta \in [0,1] \\ &\sim \mathsf{Beta}(\alpha = 4, \beta = 8). \end{aligned}$$

Suppose we started with a different prior and instead suspect θ to be around 0.5 where are change in the prior is made so that $f(\theta) \sim \mathsf{Beta}(3,3)$. Using a more general form of the likelihood

$$f(x \mid \theta) \propto \theta^S (1-\theta)^{n-S}.$$

The posterior is now

$$f(\theta \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \theta) f(\theta)$$

$$
\begin{aligned}
&= \theta^S(1-\theta)^{n-S} \cdot \theta^2(1-\theta)^2 \\
&= \theta^{S+2}(1-\theta)^{n-S+2} \\
&\sim \mathsf{Beta}(\alpha = S+3, \beta = n-S+3).
\end{aligned}
$$

Example 109 (Year 2021 Final, Q4). Suppose $x_1, x_2, \ldots, x_n \sim \mathsf{Geo}(p)$ where $p \in [0,1]$ is unknown. Note that in this parametrization, x represents the number of failures before the First success. Consider a prior distribution on p given by $p \sim \mathsf{Beta}(\alpha, \beta)$ with pdf

$$
f(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad p \in [0,1].
$$

First note that the likelihood is

$$
f(\boldsymbol{x} \mid p) = (1-p)^{\sum_i x_i} p^n
$$

where the MLE of p can be computed as

$$
\begin{aligned}
\ln f(\boldsymbol{x} \mid p) &= \sum_i \ln(1-p) + n \ln p \\
\frac{\partial}{\partial p} \ln f(\boldsymbol{x} \mid p) &= -\frac{\sum_i x_i}{1-p} + \frac{n}{p} \\
0 &= -\frac{\sum_i x_i}{1-\hat{p}} + \frac{n}{\hat{p}} \\
0 &= (-\sum_i x_i)\hat{p} + n(1-\hat{p}) \\
n &= (\sum_i x_i + n)\hat{p} \\
\hat{p} &= \frac{n}{n+n\bar{x}}.
\end{aligned}
$$

The posterior can then be computed as

$$
\begin{aligned}
f(p \mid \boldsymbol{x}) &\propto f(p) f(\boldsymbol{x} \mid p) \\
&= p^{n+\alpha-1}(1-p)^{\sum_i x_i + \beta - 1}
\end{aligned}
$$

so that $p \mid \boldsymbol{x} \sim \mathsf{Beta}\left(n+\alpha, \sum_i x_i + \beta\right)$. Since the posterior distribution and prior distribution belong to the same class of distributions (both beta distributions), $p \sim \mathsf{Beta}(\alpha, \beta)$ is indeed conjugate for this problem. The mean of the posterior is then

$$
\mathbb{E}\left[p \mid \boldsymbol{x}\right] = \frac{n+\alpha}{n+\alpha+\sum_i x_i + \beta}.
$$

As $n \to \infty$

$$
\lim_{n\to\infty} \frac{n+\alpha}{n+\alpha+\sum_i x_i + \beta} = \frac{1}{1+\bar{x}} = \frac{n}{n+n\bar{x}}.
$$

Thus as $n \to \infty$, the posterior mean will approach the MLE of p. In other words, our prior information becomes less significant as $n \to \infty$. When $\alpha, \beta \to \infty$

$$
\lim_{(\alpha,\beta)\to(0,0)} \frac{n+\alpha}{n+\alpha+\sum_i x_i + \beta} = \frac{n}{n+n\bar{x}}
$$

so, again, we find the posterior mean will approach the MLE of p. The effect of the prior parameter α and β on the posterior mean is like observing α experiments with β total failures prior to our current experiment.

Example 110 (Bayesian Inference for Normal Data). Suppose $x_1, x_2, \ldots, x_n \sim \mathsf{N}\left(\mu, \sigma^2\right)$ where σ^2 is known so only μ is unknown. Let's try $\mathsf{N}\left(\mu_0, \sigma_0^2\right)$ for some hyperparamters μ_0 and σ_0^2. To specify the likelihood of the data, from the question we know that

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

so for the posterior distribution

$$\begin{aligned} f(\mu \mid x) &\propto f(x \mid \mu) f(\mu) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{\sigma_0^2}(\mu-\mu_0)^2\right]\right). \end{aligned}$$

Isolating the exponent

$$\begin{aligned} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{\sigma_0^2}(\mu-\mu_0)^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\mu + \mu^2\right) + \frac{1}{\sigma_0^2}\left(\mu^2 - 2\mu\mu_0 + \mu^2\right) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i^2 - 2n\bar{x}\mu + n\mu^2\right) + \frac{1}{\sigma_0^2}\left(\mu^2 - 2\mu\mu_0 + \mu^2\right) \\ &= \mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right) - 2\mu\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) + c_1 \\ &= \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left[\mu^2 - 2\mu\frac{\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}\right] + c_1 \end{aligned}$$

This means that $f(\mu \mid x)$ is a normal distribution with parameters $\mathsf{N}\left(\mathbb{E}(\mu \mid x), \mathrm{Var}(\mu \mid x)\right)$ where

$$\begin{aligned} \mathbb{E}(\mu \mid x) &= \frac{\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)} \\ &= \frac{n\bar{x}\sigma_0^2 + \sigma^2\mu_0}{n\sigma_0^2+\sigma^2} \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(\mu \mid x) &= \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1} \\ &= \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}. \end{aligned}$$

So the mean/mode is a weighted average between the sample and $\bar{x}$ and our prior mean μ_0, with the weight being $n\sigma_0^2$ and σ^2, respectively. Observe

$$\begin{aligned}\mathbb{E}(\mu \mid x) &= \frac{n\bar{x}\sigma_0^2 + \sigma^2\mu_0}{n\sigma_0^2 + \sigma^2} \\ &= \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \\ &= (1-w)\,\bar{x} + w\mu_0\end{aligned}$$

and for our variance

$$\begin{aligned}\mathrm{Var}(\mu \mid x) &= \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} \\ &= w^2\sigma_0^2 + (1-w)^2\frac{\sigma^2}{n} \\ &= w^2\mathrm{Var}(\mu) + (1-w)^2\mathrm{Var}(\bar{x})\end{aligned}$$

What is the effect of the prior on the posterior? The information in the prior of $\mu \sim \mathsf{N}\left(\mu_0, \sigma_0^2\right)$ is roughly the same as it would be provided is a sample with sample mean $\bar{x} = \mu_0$ and standard error being σ_0.

Example 111 (Bayesian Inference for Normal Data with Unknown σ^2). Suppose $x_1, x_2, \ldots, x_n \sim \mathsf{N}\left(\mu, \sigma^2\right)$ where μ is known so only σ^2 is known. We can use a prior of a Inverse-Gamma distribution. Recall $\mathsf{InvGamma}(\alpha_0, \beta_0)$ has density

$$f(x) = \beta_0^{\alpha_0}\,(x)^{-\alpha_0-1}\exp\left(-\beta_0/x\right)$$

for $x > 0$. If $\sigma^2 \sim \mathsf{InvGamma}(\alpha_0, \beta_0)$, then $1/\sigma^2 \sim \mathsf{Gam}(\alpha_0, \beta_0)$. The data likelihood has a distribution of

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

meaning that posterior can be computed as

$$\begin{aligned}f(\sigma^2 \mid x) &= f(x \mid \sigma^2)f(\sigma^2) \\ &= \left(\frac{1}{\sigma}\right)^{n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2\right)(\sigma)^{-\alpha_0-1}\exp\left(-\frac{\beta_0}{\sigma^2}\right) \\ &= (\sigma^2)^{-(\frac{n}{2}+\alpha_0+1)}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 + \beta_0\right) \\ &\sim \mathsf{InvGamma}\left(\frac{n}{2} + \alpha_0, \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 + \beta_0\right).\end{aligned}$$

The posterior is then

$$\frac{\frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 + \beta_0}{\frac{n}{2} + \alpha_0 + 1} = \frac{2\beta_0 + \sum_{i=1}^{n}\left(x_i - \mu\right)^2}{n + \alpha_0 + 2}.$$

It is a weighted average between the prior mode and the sample variance. The Inverse-Gamma is a popular choice of prior since it gives a closed form distribution, and posterior in the same family as the prior.

Definition 112 (Conjugacy)**.** *Whenever the posterior distribution has the same form as the prior, this property is called* **conjugacy**. *The corresponding prior is called the* **conjugate prior**.

Example 113 (Continuation of Example 110). What is the effect of the hyperparameters α_0 and β_0 on the posterior mode on the above example? Recall, the mode was found to be

$$\frac{2\beta_0 + \sum_{i=1}^{n} (x_i - \mu)^2}{n + \alpha_0 + 2}.$$

This is "like" observing $2\beta_0$ sum of the squares from $2\alpha_0 + 2$ samples prior to our current experiment. Now what if we take $\alpha_0 \to 0$ and $\beta_0 \to 0$? The mode $\frac{\beta_0}{\alpha_0+1} \to 0$ meaning the posterior distribution is $\sigma^2 \mid x \sim \mathsf{InvGamma}(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2)$ which is a valid distribution. But, what about the prior? In this case $f(\sigma^2) \propto \frac{1}{\sigma^2}$ for $\sigma^2 > 0$, this is impossible to normalize since $\int_0^\infty \frac{1}{\sigma^2}\, d\sigma^2 = \infty$, so this is a *improper* prior!

Example 114 (Normal Data with Unknown μ and σ^2). Suppose $x_1, x_2, \ldots, x_n \sim \mathsf{N}\left(\mu, \sigma^2\right)$ where both μ and σ^2 is unknown. Let's consider independent priors for μ and σ^2 where

$$\begin{aligned} \mu &\sim \mathsf{N}\left(\mu_0, \sigma_0^2\right) \\ \sigma^2 &\sim \mathsf{InvGamma}(\alpha_0, \beta_0) \end{aligned}$$

so that the joint prior becomes

$$\begin{aligned} f(\mu, \sigma^2) &= f(\mu) f(\sigma^2) \\ &= \beta_0^{\alpha_0} \left(\sigma^2\right)^{-\alpha_0 - 1} \exp\left(-\beta_0/\sigma^2\right) \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left(-\frac{1}{2\sigma^2}\left(\mu - \mu_0\right)^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(\mu - \mu_0\right)^2\right) \left(\sigma^2\right)^{-\alpha_0 - 1} \exp\left(-\beta_0/\sigma^2\right). \end{aligned}$$

The likelihood is then

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

and finally, the posterior is

$$\begin{aligned} f(\mu, \sigma^2 \mid \boldsymbol{x}) &\propto f(\boldsymbol{x} \mid \mu, \sigma^2) \cdot f(\mu, \sigma^2) \\ &\propto \left(\sigma^2\right)^{-n/2 + \alpha_0 + 1} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{1}{2\sigma^2}\left(\mu - \mu_0\right)^2 - \beta_0/\sigma^2\right). \end{aligned}$$

Unfortunately, this joint distribution does not correspond to any common distribution. To sample from this distributions we can employ the Gibbs sampling methods. We find

$$\begin{aligned} f(\mu, \sigma^2 \mid x) &\propto f(\mu \mid \sigma^2, x) \\ &= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{1}{2\sigma^2}\left(\mu - \mu_0\right)^2\right) \\ &\sim \mathsf{N}\left(\frac{\sigma_0 n\bar{x} + \mu\sigma^2}{n\sigma_0^2 + \sigma^2}, \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right) \end{aligned}$$

and

$$\begin{aligned} f(\sigma^2 \mid \mu, x) &\propto f(\mu, \sigma \mid x) \\ &\propto \left(\sigma^2\right)^{-n/2+\alpha_0+1} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\sigma^2}\left(\mu-\mu_0\right)^2 - \beta_0/\sigma^2\right) \\ &\propto \left(\sigma^2\right)^{-n/2+\alpha_0+1} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \beta_0/\sigma^2\right) \\ &\sim \mathsf{InvGamma}\left(\alpha_0+\frac{n}{2}, \beta_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right). \end{aligned}$$

So now, we can use Gibbs sampling to generate (approximate) samples from the joint posterior $f(\mu, \sigma^2 \mid x)$. Firts we start with initial values $\mu^{(0)}$ and $\sigma^{(0)} > 0$. Then we repeat the following K times: where on the k^{th} we compute

$$\begin{aligned} \mu^{(k+1)} &\sim \mathsf{N}\left(\frac{\sigma_0 n\bar{x} + \mu\sigma^{2(k)}}{n\sigma_0^2 + \sigma^{2(k)}}, \frac{\sigma^{2(k)}\sigma_0^2}{n\sigma_0^2+\sigma^{2(k)}}\right) \\ \sigma^{2(k+1)} &\sim \mathsf{InvGamma}\left(\alpha_0+\frac{n}{2}, \beta_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu^{(k+1)})^2\right). \end{aligned}$$

For large $k \geq T$ (burn in period), we have $\left\{\mu^{(k)}, \sigma^{2(k)}_{K\geq T} \sim f(\mu, \sigma^2 \mid x)\right\}$ approximately.

Example 115 (Un informative prior for μ and σ^2). Suppose $x_1, x_2, \ldots, x_n \sim \mathsf{N}\left(\mu, \sigma^2\right)$ where both μ and σ^2 is unknown. Let's consider a joint prior $f(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$ for $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. To start we have

$$\begin{aligned} f(\mu, \sigma^2) &\propto \frac{1}{\sigma^2} \\ f(x \mid \mu, \sigma^2) &\propto \left(\sigma^2\right)^{-n/2+\alpha_0+1} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\sigma^2}\left(\mu-\mu_0\right)^2\right) \end{aligned}$$

and

$$\begin{aligned} f(\mu, \sigma^2 \mid x) &\propto f(x \mid \mu, \sigma^2) f(\mu, \sigma^2) \\ &= \left(\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\sigma^2}\left(\mu-\mu_0\right)^2\right). \end{aligned}$$

Let's start with the marginal posterior,

$$\begin{aligned} f(\sigma^2 \mid x) &= \int_{-\infty}^{\infty} f(\mu, \sigma^2 \mid x) d\,\mu \\ &= (\sigma^2)^{\frac{n+1}{2}} \exp\left(-(\frac{n-1}{2\sigma^2})S^2\right) \int_{-\infty}^{\infty} \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right) d\,\mu \end{aligned}$$

where $\exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right)$ is the kernel of $\mathsf{N}\left(\bar{x}, \frac{\sigma^2}{n}\right)$ so that

$$= (\sigma^2)^{\frac{n+1}{2}} \exp\left(-(\frac{n-1}{2\sigma^2})S^2\right) \int_{-\infty}^{\infty} \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right) d\,\mu$$

$$
\begin{aligned}
&= (\sigma^2)^{\frac{n+1}{2}} \exp\left(-(\frac{n-1}{2\sigma^2})S^2\right)\left(\frac{\sigma^2}{n}\right)^{\frac{1}{2}}(2\pi)^{\frac{1}{2}} \\
&\propto (\sigma^2)^{\frac{n+1}{2}} \exp\left(-(\frac{n-1}{2\sigma^2})S^2\right) \\
&\propto \mathsf{InvGamma}\left(\frac{n-1}{2}, \left(\frac{n-1}{2}S^2\right)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
f(\mu \mid x) &= \int_0^\infty f(\mu, \sigma^2 \mid x)\, d\sigma^2 \\
&= \int_0^\infty (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right)\, d\sigma^2
\end{aligned}
$$

where $(\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 - \beta_0/\sigma^2\right)$ is the kernel of $\mathsf{InvGamma}\left(n/2, \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right)$ meaning

$$
\begin{aligned}
&= \int_0^\infty (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right)\, d\sigma^2 \\
&= \Gamma(n/2)\left[\frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2\right]^{-n/2} \\
&\propto \left[\sum_{i=1}^n (x_i-\mu)^2\right]^{-n/2} \\
&\propto \left[(n-1)S^2 + n(\bar{x}-\mu)^2\right]^{-n/2} \\
&\propto \left[1 + \frac{1}{n+1}\left(\frac{\bar{x}-\mu}{S/\sqrt{n}}\right)\right]^{-n/2}
\end{aligned}
$$

This is the kernel of the non-standard t-distribution with $n-1$ degrees of freedom. Recall that

$$\frac{\bar{x}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

so $\frac{\mu-\bar{x}}{S/\sqrt{n}} \sim t_{n-1}$ and we write $\mu \mid x \sim t_{n-1}\left(\bar{x}, S^2/n\right)$. Observe, that the joint prior posterior can be written as

$$\sigma^2 \mid x \sim \mathsf{InvGamma}\left(\frac{n-1}{2}, \left(\frac{n-1}{2}S^2\right)\right)$$

and

$$\frac{\mu-\bar{x}}{S/\sqrt{n}} \mid x \sim t_{n-1}\left(\bar{x}, S^2/n\right)$$

independently.

The Bayesian normal model can be generalised to a Bayesian linear regression model:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i,$$

for $1 \leq i \leq n$ where $\varepsilon_i \sim \mathsf{N}\left(0, \sigma^2\right)$ and $\boldsymbol{x}_i \in \mathbb{R}$ (the vectors of covariates) and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression coefficients. What are the parameters in this model? We have two, $\boldsymbol{\beta}$ and σ^2. We can set the prior distribution on our model parameters

$$\begin{aligned}\boldsymbol{\beta} &\sim \mathsf{N}\left(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0\right) \\ \sigma^2 &\sim \mathsf{InvGamma}\left(\alpha_0, \beta_0\right)\end{aligned}$$

and $\boldsymbol{\beta}$ and σ^2 are independent. The likelihood of the data is set to be

$$f(y \mid x, \boldsymbol{\beta}, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^{\intercal} \boldsymbol{\beta}\right)^2\right)$$

which means the posterior turns out to be

$$\begin{aligned} f\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{x}, y\right) &\propto f\left(y \mid \boldsymbol{\beta}, \sigma^2 \boldsymbol{x}\right) \\ &= \left(\sigma^2\right)^{-(\alpha_0 + n/2 - 1)} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^{\intercal} \boldsymbol{\beta}\right)^2 - \frac{1}{2} \left(\boldsymbol{\beta} - \boldsymbol{\beta_0}\right)^{\intercal} \boldsymbol{\Sigma}_0 \left(\boldsymbol{\beta} - \boldsymbol{\beta_0}\right) - \frac{\boldsymbol{\beta}_0}{2}\right). \end{aligned}$$

We can sample from this distribution using Gibbs sampling method, so let's find the conditional posteriors:

$$\begin{aligned} f(\sigma^2 \mid \boldsymbol{x}, y, \boldsymbol{\beta}) &\propto \left(\sigma^2\right)^{-(\alpha_0 + n/2 - 1)} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^{\intercal} \boldsymbol{\beta}\right)^2 - \frac{\beta_0}{\sigma^2}\right) \\ &\sim \mathsf{InvGamma}\left(\alpha_0 + n/2, \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^{\intercal} \boldsymbol{\beta}\right)^2 + \beta_0\right) \end{aligned}$$

and

$$\begin{aligned} f(\mu \mid \sigma^2, \boldsymbol{x}, y, \boldsymbol{\beta}) &\propto f(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{x}, y) \\ &\vdots \\ &\sim \mathsf{N}\left(\mathbb{E}\left(\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{x}, y\right), \mathbb{V}\left(\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{x}, y\right)\right). \end{aligned}$$

So we can now use Gibbs sampling to obtain update for σ^2 and $\boldsymbol{\beta}$ to give approximate samples from our joint posterior.

Definition 116 (Bayeisan Credible Interval)**.** *A $100(1-\alpha)$ credible interval for θ is an (L, U) such that*

$$1 - \alpha = \mathbb{P}\left(L \leq \theta \leq U \mid x\right) = \begin{cases} \int_L^U f(\theta \mid x)\, d\theta & \textit{continuous} \\ \sum_{\theta=L}^{U} f(\theta \mid x) & \textit{discrete} \end{cases}$$

Example 117 (Credible Interval Example). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}\left(\mu, \sigma^2\right)$. Suppose $\sigma^2 = 4$ and a prior for μ is $\mu \sim \mathsf{N}\left(0, 1\right)$. To find $100(1-\alpha)\%$ credible interval for μ, from a previous lecture

$$\mu \mid x \sim \mathsf{N}\left(\frac{\bar{x}}{1 + \frac{4}{n}}, \frac{1}{1 + \frac{4}{n}}\right).$$

We need to find (L, U) such that

$$1 - \alpha \ = \mathbb{P}\left(L \leq \mu \leq U \mid x\right)$$

$$= \mathbb{P}\left(-z_{\alpha/2} \leq \frac{\mu - \frac{\bar{x}}{1+4/n}}{\sqrt{\frac{1}{1+4/n}}} \leq z_{\alpha/2} \mid x \right)$$
$$= \mathbb{P}\left(\frac{\bar{x}}{1+4/n} - z_{\alpha/2}\frac{1}{\sqrt{1+4/n}} \leq \mu \leq \frac{\bar{x}}{1+4/n} + z_{\alpha/2}\frac{1}{\sqrt{1+4/n}} \mid x \right).$$

Thus the $100(1-\alpha)\%$ credible interval for μ is $\frac{\bar{x}}{1+4/n} \pm z_{\alpha/2}\frac{1}{\sqrt{1+4/n}}$. If $f(\theta \mid x)$ is not easy to work with we can use MCMC sampling to generate an approximate sample for it, and then find the appropriate quantiles from the sample.

Bayesian Multinomial Model. Let's look at a model for categorical data, in general if we have $K \geq 2$ classes, then we have multinomial data. In this case, the data $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)$, which is a vector of counts of each category out of sample of $n = \sum_i X_i$. What is the probability function of $X \sim \mathsf{Multi}_K(n, \boldsymbol{p})$

$$f(x_1, \ldots, x_k) = \mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} \prod_i p_i^{x_i} \propto \prod_i p_i^{x_i}.$$

Note that $\sum_i x_i = n$ and $\sum_i p_i = 1$, which means x_j or p_j could be written as a linear combination of the other x_is or p_is. What would be a good prior for distribution for $\boldsymbol{p}$. Dirichlet distribution is typically used as a prior here.

Definition 118 (Dirichlet Distribution)**.** *We say that a random vector $\boldsymbol{p} = (p_1, \ldots, p_k)$ has a Dirichlet distribution with shape parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ if its density is in the form*

$$f(\boldsymbol{p} \mid \boldsymbol{\alpha}) \propto \prod_{i=1}^{k} p_i^{\alpha_i - 1}, \quad \textit{for } 0 \leq p_i \leq 1 \textit{ and } \sum_i p_i = 1.$$

Alternatively, for a vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_m)$ is said to follow a Dirichlet distribution with shape parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{m+1})$ if its density is

$$f(\boldsymbol{z} \mid \alpha) \propto \left(\prod_{i=1}^{m} z_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{m} z_i \right)^{\alpha_{m+1} - 1}$$

for $0 \leq z_i \leq 1$ and $\sum_{i=1}^{m} z_i \leq 1$. We write $\boldsymbol{z} \sim \mathsf{Dir}\,(\alpha_1, \alpha_2, \ldots, \alpha_{m+1})$.

What's s special about the Dirichlet distribution? Let's look at the special case of $k = 2$, then

$$f(\boldsymbol{p} \mid \boldsymbol{\alpha}) \propto p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} = p_1^{\alpha_1 - 1}(1 - p_1)^{\alpha_2 - 1}$$

for $0 \leq p_1 \leq 1$. This is the $\mathsf{Beta}(\alpha_1, \alpha_2)$ distribution and so the Dirichlet distribution is a generalization of the Beta distribution. Since $p_k = 1 - \sum_{i=1}^{k-1} p_i$ we can infact specify a Dirichlet distribution by replacing the last component p_k with $1 - \sum_{i=1}^{k-1} p_i$. A special case of the Dirichlet distribution with parameters $\alpha_i = 1,\ 0 \leq i \leq k$ is a uniform distribution over the set of all multinomial probabilities of K categories, i.e., $f(p) \propto 1$. This is not the same as setting $\alpha_i = 1/K,\ 0 \leq i \leq k$ since this is just one distribution on K categories.

Bayesian Inference for the Multinomial Model. If the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_K) \sim \mathsf{Multi}_K(n, \boldsymbol{p})$ with prior $p \sim \mathsf{Dir}\,(\alpha_1, \alpha_2, \ldots, \alpha_K)$, then our posterior is

$$\begin{aligned} f(p_1, p_2, \ldots, p_k \mid \boldsymbol{x}) &\propto f(\boldsymbol{x} \mid \boldsymbol{p}) f(\boldsymbol{p}) \\ &\propto \prod_{i=1}^{K} p_i^{x_i} \prod_{i=1}^{K} p_i^{\alpha_i - 1} \\ &\propto \prod_{i=1}^{K} p_i^{x_i + \alpha_i - 1} \\ &\sim \mathsf{Dir}\,(\alpha_1 + x_1, \ldots, \alpha_K + x_K) \end{aligned}$$

From the form of the posterior distribution, what is the effect of our prior like? In terms of the posterior mode:

$$\left(\frac{\alpha_1 + x_1 - 1}{\sum_i \alpha_i + x_i - 1}, \frac{\alpha_2 + x_2 - 1}{\sum_i \alpha_i + x_i - 1}, \ldots, \frac{\alpha_K + x_K - 1}{\sum_i \alpha_i + x_i - 1} \right)$$

The effect of our priors parameters, α_i, is like observing $(\alpha_i - 1)$ counts in each category prior to our current count. In particular, for our non-informative prior $(\alpha_i = 1)$, then in terms of the posterior mode, it is like observing 0 counts in each category prior to our current experiment.

Sampling from the Dirichlet Distribution. How do we sample from the Dirichlet distribution?

Theorem 119 (Sampling from a Dirichlet distribution)**.** *Let $Y_i \sim \Gamma(\alpha_i, 1)$. Define*

$$Z_j = \frac{Y_j}{\sum_{i=1}^{K} Y_i}$$

for $1 \leq j \leq K$. Then $\boldsymbol{Z} = (Z_1, \ldots, Z_k) \sim \mathsf{Dir}\,(\alpha_1, \ldots, \alpha_k)$. Using the textbook notation $\boldsymbol{Z} = (Z_1, \ldots, Z_{k-1}) \sim \mathsf{Dir}\,(\alpha_1, \alpha_2, \ldots, \alpha_k)$.

Example 120 (Credible Interval Example). Taken from Example 8.3 of Dirk. Suppose we sample $n = 501$ people randomly from the same population and ask them if they dislike, and neutral or like anti-smoking campaigns. Gender is also recorded. For male counts $53, 57$ and 147 were against, neutral and encouraging of these campaigns respectively. For female counts $93, 38$ and 113 were against, neutral and encouraging of these campaigns respectively. Let $\boldsymbol{p} = (p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23})$ be the underlying probabilities of each genders position on the campaigns. For our prior beliefs we can set $\boldsymbol{p} \sim \mathsf{Dir}(1, 1, 1, 1, 1, 1)$, i.e., the uniform belief over all the probability vectors of length 6. The likelihood for the data x can be given as $x \mid p \sim \mathsf{Multi}_6(n = 501, \boldsymbol{p})$ so that $f(\boldsymbol{x} \mid \boldsymbol{p}) \propto p_{11}^{53} p_{12}^{57} p_{13}^{147} p_{21}^{93} p_{22}^{38} p_{23}^{113}$. Our posterior belief is then

$$f(\boldsymbol{p} \mid \boldsymbol{x}) \propto \text{likelihood} \times \text{prior}$$

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.

[ZIBaJFGaDPK10] Z. I. Botev and J. F. Grotowski and D. P. Kroese, *Kernel density estimation via diffusion*, Vol. 38, Institute of Mathematical Statistics, 2010.