University of California Davis

STA131B: Mathematical Statistics

Lecturer: Christiana Drake

Scribe: Ramneek Narayan

Last Revision: May 31, 2020

Table of Contents

About/Usage

About/Usage

About this Book

These are some notes Christiana Drake took about STA131B typeset by Ramneek Narayan after he completed the course. They aim to be easy to read and provide more precision in content. If there are any typos, let the writers know, we appreciate it.

How to use this Book

This book was written with the student in mind and comes with colored environments to make reading easier. In addition, at the end of each environment are symbols used conclude the environment (show that it is completed); they are there for organization and for your ease of reading. We list the environments below for clarity:

Example = Red Violet concludes with '♥'

Remark = Teal concludes with '♦'

Definition = Lime Green concludes with '♣'

Theorem = Royal Purple concludes with '□'

Proposition = Mulberry concludes with '□'

Lemma = Goldenrod concludes with '□'

Note = Orange concludes with '∞'

Corollary = Melon concludes with '□'

Emphasis = Royal Blue concludes with '☕'

Extra = Gray has no symbol to conclude

Read at your own pace and if anything doesn't make sense, argue with the instructor (this is how we make more knowledge)! It makes sense at the end if nothing comes to the mind immediately. We hope you enjoy reading it!

Overview of STA131B

> *"Statistics is the art and science of gathering, modeling and making inference from data."*

What you need to know to succeed in this course

In order to do well in this course, it will help if you are familiar with...

1. Basic probability
   - if $A =$ event, then $0 \leq P(A) \leq 1$ for every $A \subset S$
   - For $A_1,\ A_2$ **disjoint** events: $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ and $P(A_1 \cap A_2) = \varnothing$
   - $P(S) = 1$ and $P(\varnothing) = 0$
2. Conditional probability (including Bayes' Theorem)
   - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
   - $P(A \cap B) = P(A)P(B)$ or $P(A|B) = P(A)$
3. Random variables and probability distributions...
   - Discrete and continuous random variables
   - Probability mass function and density function
   - Cumulative distribution function
   - Univariate and bivariate distributions
   - Marginal distributions and bivariate distributions
   - Functions of one or more random variables
4. Mathematical expectation...

- Mean, variance and other moments
- Marginal and conditional moments

5. (Strong/Weak) Law of Large numbers and Central Limit Theorem
   - **Weak Law of Large Numbers:** If $X_1, X_2, ..., X_n$ is a sequence of i.i.d random variables, then $\overline{X} \xrightarrow{P} \mu$ as $n \to \infty$ or $\lim_{n\to\infty} P(|\overline{X}_n - \mu| \geq \epsilon) = 0$.
6. Special distributions...
   - Bernoulli, Binomial and Hypergeometric distribution
   - Poisson and Negative Binomial distribution
   - Normal distribution
   - Gamma distribution (including exponential)
   - Beta distributions (including uniform)
   - Bivariate Normal and Multinomial distribution
7. Transformation of variables
   - Univariate transformation and probability integral transformation
     - **Probability Integral Transform:** If $X \sim F$, then $Y = \phi(X) \sim \mathcal{U}$
   - Bivariate transformation of variables

What you will know by the end of this course

The focus in this course will be on principles of *estimation* and *hypothesis testing.*

1. General concepts in estimation...
   - Bayes' Estimators
   - Method of Maximum Likelihood (MLE's)
   - Sufficient Statistics
2. Distributional properties of estimators
   - The sampling distribution of a statistic
   - the t-distribution and χ^2 distribution in estimation
   - Calculating and interpreting confidence intervals

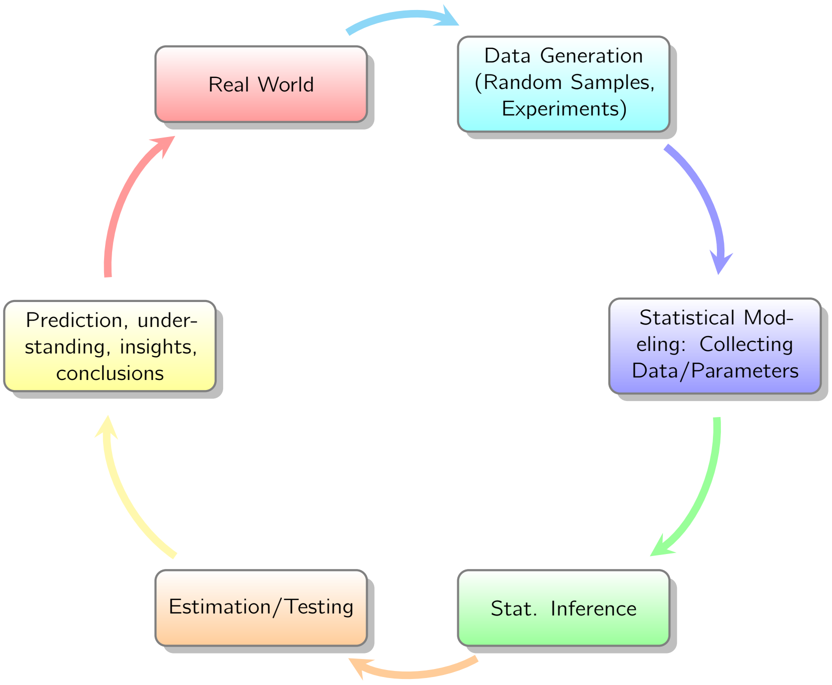
   - Unbiased estimation and Fisher information
   - Bayesian inference

3. Hypothesis testing
   - Simple hypothesis; type I and type II error and most powerful tests
     - **Simple Hypothesis:** $H_0 : \mu = \mu_0$ vs. $H_A : \mu = \mu_1$
   - Composite hypothesis and uniformly most powerful tests
   - Likelihood ratio test
   - p-values
   - t-tests
   - F-tests
   - Relationship between a hypothesis test and a confidence interval
   - Bayes' procedures in hypothesis testing

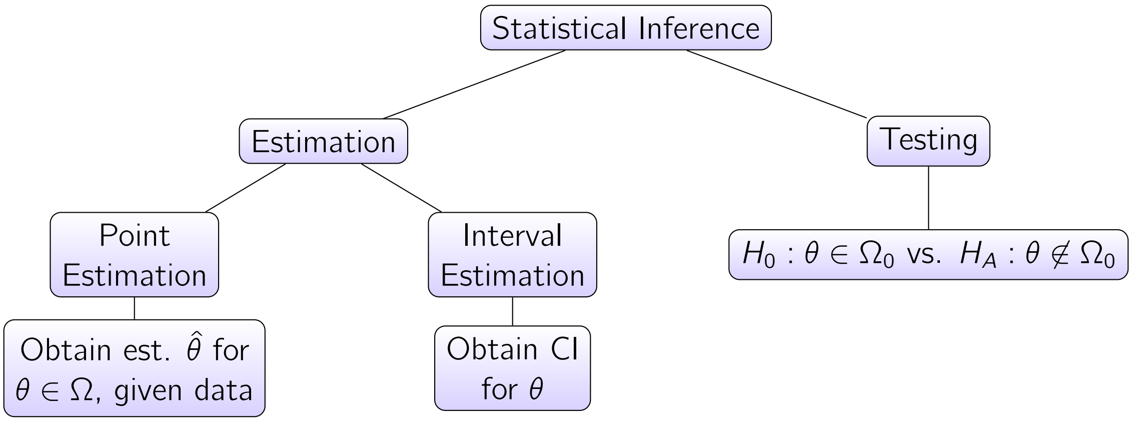
☕

Note: For this class, the tools we will be using are mathematics and probability (131A and its prerequisites are required for 131B). Pioneers of modern statistics include K. Pearson, R.A. Fisher, J. Neyman. ∞

Below is a visual depiction of the general statistical framework:


Real World
Data Generation (Random Samples, Experiments)
Statistical Modeling: Collecting Data/Parameters
Stat. Inference
Estimation/Testing
Prediction, understanding, insights, conclusions

In this course we outline statistics as follows:


Statistical Inference
Estimation
Testing
Point Estimation
Interval Estimation
$H_0 : \theta \in \Omega_0$ vs. $H_A : \theta \notin \Omega_0$
Obtain est. $\hat{\theta}$ for $\theta \in \Omega$, given data
Obtain CI for θ

Outline of statistics

Chapter 1 — Inference: Estimation

1.1 Probability vs. Statistics

We begin by discussing the difference between probability and statistics. Both are different, though related, fields and require different ways of interpretation. We use a simple coin-flip example to show probability and its aims. We will record the set of outcomes as $S = \{H, T\}$ for clarity.

Probabilist View

1. If the coin is fair, then
$$P(H) = P(T) = \frac{1}{2}$$
2. If we set $X = \#$ of heads in 10 flips, then
$$X \sim \text{Bin}(n = 10, p = 1/2)$$
3. If we set $Y = \#$ of flips until (before) first heads, then
$$Y \sim \text{Geometric}(p = 1/2)$$
i.e.
$$\text{p.m.f. } P(Y = y | p = 1/2) = \left(\frac{1}{2}\right)^{y+1}$$
4. If we set $T = \#$ of tails until (before) 10 heads, then
$$T \sim \text{NegBin}(r = 10, p = 1/2)$$
and its p.m.f is described by
$$P(T = t | r = 10, p = 1/2) = \begin{cases} \binom{10+t-1}{t}\left(\frac{1}{2}\right)^{10+t} & \forall t \in \mathbb{N} \\ 0 & \text{o.w.} \end{cases}$$

The probabilist view uses these probability models to build probabilities of interest.

Now be discuss the methods of a **statistical model:**

1.1.1 Statistical Model

In a statistical model, we set

$$X = \text{outcome of some experiment } E$$

Now E could be flipping a coin once, a fixed number of times, or until the first occasion of some event of interest and the question is: **is the coin fair?** The statistician asks whether the coin is fair and how to reasonably answer that question using probabilisitic methods.

To do this, the statistician creates a model/experiment to answer the question about the coin. For the above, there are several approaches to obtaining data to answer that question and each approach has a somewhat different **probability model**. We give now the steps toward making a statistical model:

Steps to Make a Statistical Model

1. Identify the random variables of interest: X, Y, Z, &etc... from before
2. Specify the joint distribution or family of joint distributions
3. Identify the parameters of interest (both known and unknown)
4. (Possibly) Set $\theta =$ unknown parameter as **fixed but unknown** or **random** and create a probability distribution for θ. If either case, we consider the distribution of X as conditional on θ or $X|\theta$ that describes behavior of phenomenon

Example 1.1.1 (Statistical Modeling)**.**

1. *Suppose* $X_1, X_2, ..., X_n \overset{iid}{\sim} Exp(\lambda)$ *and specify the times until failures of tires. We know*

$$f(x_i) = \lambda e^{-\lambda x_i}$$

then the joint distribution becomes

$$f(x_1, x_2, ..., x_n|\lambda) = \prod_i f(x_i) = \lambda^n e^{\lambda \sum_i x_i}$$

2. *Suppose $X_1, ..., X_n = \#$ of heads in 10 flips of a coin, then*

$$X_i \sim Bin(n = 10, p)$$

and the joint distribution is

$$f(x_1, ..., x_n|p) = \prod_i f(x_i) = \prod_i \binom{10}{x_i} (p)^{x_i} (1-p)^{10-x_i}$$

3. *Suppose $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$ are unknown. Then, the joint distribution is*

$$f(x_1, ..., x_n|\mu, \sigma^2) = \prod_i f(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\}$$

♥

Generally we denote a model parameter by θ as seen in the example above and the distribution by P and the random observable (or possibly hypothetically observable) by X. Before moving on, we give one definition of the set of values a parameter θ can take.

Definition 1.1.1 (Parameter Space)**.** *A* **parameter space** Ω *is the set of all possible values a parameter θ can take.*

♣

In summary then, the components to a statistical model are:

1. $X =$ random quantity
2. $P =$ probability model (distribution) indexed by parameter θ
3. $\Omega =$ parameter space of possible parameter values

More formally and succinctly, we can specify a statistical model as

Definition 1.1.2 (Statistical Model)**.** *A* **statistical model** *consists of random variables following some distribution under a parameter belonging to some set, i.e.*

$$X_1, ..., X_n \overset{iid}{\sim} (P_\theta : \theta \in \Omega)$$

♣

These can be seen already in the examples we have given:

Example 1.1.2 (Identifying Statistical Models)**.**

1. *Exponential Model*
   - $P =$ *exponential distribution*
   - $\theta = \lambda$ *(usually generic parameter rate)*
   - $\Omega = (0, \infty)$
2. *Binomial Model*
   - $P =$ *binomial distribution*
   - $\theta = p$
   - $\Omega = [0, 1]$
3. *Normal Model*
   - $P =$ *family of normal distributions*
   - $\theta = (\mu, \sigma)$ *or* (μ, σ^2)
   - $\mu \in (-\infty, \infty)$ *and* $\sigma^2 \in (0, \infty)$ *hence,*

$$\Omega = (-\infty, \infty) \times (0, \infty)$$

♥

1.1.2 Statistical Inference

Now that we know about models, we can start answering some of the questions they are about. We call these **statistical inferences**. More formally,

Definition 1.1.3 (Statistical Inference)**.** *A* **statistical inference** *is a procedure that results in a probabilistic statement about some or all parts of a statistical model.*

♣

Note: In this class, we will be concerned with **parametric inference**, i.e. a model with a known distribution family P_θ indexed by an unknown parameter θ. We can also assume the family is unknown but this would be a non-parametric setting. ∞

So, what are some problems statistical inference can solve? We give some below:

Classification of Problems

- Prediction: Machine Learning
- Statistical Decision Problems: Estimation and Testing
- Experimental Design: note that the book talks about this in classical design situation but you can also think of it in cases of observational data as well
- Other Types...

☕

Remark 1.1.1. *We will study MLE first, then bayesian inference. This is 7.5-6 first in the book, then 7.2-4 later.*

♦

1.2 Method of Maximum Likelihood Estimation (MLE)

Before discussing the MLE method, we give a precise definition of a statistic. You may have heard of them in previous classes as *anything that can be computed given the data*. We use this idea to create a concise definition of a statistic:

Definition 1.2.1 (Statistic)**.** *A* **statistic** *is a real-valued function of the data $X_1, \ldots, X_n$ noted as*

$$T = \underbrace{\varphi(X_1, \ldots, X_n)}_{\text{observable}}$$

♣

Example 1.2.1 (Common Statistics)**.** *Some statistics that are common in statistics are:*

1. *Sample mean:* $\bar{X}$
2. *Sample maximum:* $X_{(n)} = \max\{X_1, \ldots, X_n\}$
3. *Sample variance:* $\hat{\sigma}^2$
4. *(Just for fun) Arbitrary function:* $\varphi(X_1, \ldots, X_n) = 3$

Basically we can think of a statistic as a function of sample values that have no unknown parameters.

♥

With this definition in mind, we can now study the MLE since it is (as you will see) a statistic as well. In practice, we rarely know the distribution parameter θ in advance, so it is useful to make a "best guess". To do this, we begin with the concept of a *likelihood function* which measures how likely a given value of a parameter θ is given observed our data.

Definition 1.2.2 (Likelihood Function)**.** *Let $X_1, ..., X_n \overset{iid}{\sim} P_\theta$ with joint p.d.f (or p.m.f)*

$$f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

where we assume θ is unknown. When viewed as a function of θ for a fixed set of data points, the joint p.d.f/p.m.f is known as the **likelihood function**:

$$\mathcal{L}(\theta|x_1, ..., x_n) \sim_F \prod_{i=1}^{n} f(x_i|\theta) \qquad (\textit{Likelihood Function})$$

where $\sim_F$ means "has the same form as". Note, however, $f(x_1, ..., x_n|\theta)$ and $\mathcal{L}(\theta|x_1, ..., x_n)$ are different objects. One describes the chance of observing the data for a given θ but the other describes the chance of observing θ given that the data is already observed.

♣

Note: Although

$$\int_\Omega \mathcal{L}(\theta|x_1, ..., x_n)d\theta = 1$$

not always, we can regularize $\mathcal{L}$ to give $\mathcal{L}'$ such that

$$\mathcal{L}'(\theta|x_1, ..., x_n) = \frac{\mathcal{L}(\theta|x_1, ..., x_n)}{\int_\Omega \mathcal{L}(\theta|x_1, ..., x_n)d\theta}$$

note that this is a monotone transform (mapping). ∞

Example 1.2.2 (Exponential Likelihood)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} Exp(\lambda)$ then the joint distribution of the data is written as*

$$f(x_1, ..., x_n|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda)$$

Now, $f(x_i|\lambda) = \lambda e^{-\lambda x_i}$ and this makes the joint p.d.f as

$$f(x_1, ..., x_n|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}$$

and hence

$$\mathcal{L}(\lambda|x_1, ..., x_n) = \lambda^n e^{-\lambda \sum x_i}$$

♥

Now that we know what the likelihood is, we can define the MLE or $\hat{\theta}_{MLE}$:

Definition 1.2.3 (MLE)**.** *The* **MLE** *or* $\hat{\theta}_{MLE}$ *is the value of* θ *such that the likelihood* $\mathcal{L}(\theta|x_1, ..., x_n)$ *is a maximum. Note, this is the mode of the regularized likelihood function* $\mathcal{L}'$*. We can write this as*

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Omega} \mathcal{L}(\theta|x_1, ..., x_n)$$

♣

In practice, it is intractable to work with the likelihood function itself, so we take the natural logarithm of it and call it the **log-likelihood** that is

$$\log(\mathcal{L}(\theta|x_1, ..., x_n)) = \ell(\theta|x_1, ..., x_n)$$

Note: Since the logarithm is a monotone transform (mapping), the maximum argument of $\mathcal{L}(\theta|x_1, ..., x_n)$ is the same maximum argument as that of $\ell(\theta|x_1, ..., x_n)$. Hence,

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Omega} \mathcal{L}(\theta|x_1, ..., x_n) = \arg\max_{\theta \in \Omega} \ell(\theta|x_1, ..., x_n)$$

∞

The MLE can be thought of as an estimator in the sense that it has the potential or use of estimating the true value of the parameter $\theta \in \Omega$. In general, we define an estimator as

Definition 1.2.4 (Estimator)**.** *An* **estimator** $\delta(X_1, ..., X_n)$ *is any random statistic contained in the parameter space. In other words if*

$$\delta(X_1, ..., X_n) \in \Omega \qquad\qquad (\Omega = \textit{parameter space})$$

then, $\delta(X_1, ..., X_n)$ *is an estimator of* θ*. Note that* δ *is a function such that* $\delta : X_1, ..., X_n \to \Omega$*.*

♣

We then obtain a proposition about the MLE:

Proposition 1.2.1 (MLE is an Estimator)**.** *The MLE is an estimator of* θ*.*

Proof. Note that for $\mathcal{L}$ and ℓ the domain is Ω by definition. It then follows (if we assume the data is unobserved) that

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max_{\theta\in\Omega} \mathcal{L}(\theta|X_1, ..., X_n)\\ &= \arg\max_{\theta\in\Omega} \ell(\theta|X_1, ..., X_n)\\ &= \delta(X_1, ..., X_n)\end{aligned}$$

as we sought to show. □

Concept Note: The unobserved MLE is an estimator. Once, we collect the data, the observed MLE is called an **estimate**. In symbols,

$$\begin{aligned}\hat{\theta}_{MLE}(X_1, ..., X_n) &= \text{estimator}\\ \hat{\theta}_{MLE}(x_1, ..., x_n) &= \text{estimate}\end{aligned}$$

∞

How to find MLE

1. Brute force: use a search algorithm to find the maximum of the function. We could also use a graphical procedure too.
2. Often we can use calculus

Example 1.2.3 (MLE of Exponential)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} Exp(\lambda)$. Then, we know from Exponential Likelihood the form of the likelihood function is*

$$\mathcal{L}(\lambda|x_1, ..., x_n) = \lambda^n e^{-\lambda \sum x_i}$$

We can make a log-likelihood out of this

$$\ell(\lambda|x_1, ..., x_n) = n\log(\lambda) - \lambda\sum_{i=1}^{n} x_i$$

Since this is a smooth function, we use calculus to find the maximum via derivatives

$$\begin{aligned}\frac{d\ell(\lambda)}{d\lambda} &= \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0\\ \Rightarrow \frac{n}{\lambda} &= \sum_{i=1}^{n} x_i\\ \Rightarrow \hat{\lambda}_{MLE} &= \frac{n}{\sum_{i=1}^{n} x_i} = 1/\bar{x}\end{aligned}$$

♥

Note: We can verify an extreme value as an MLE by taking the 2nd derivative and performing the second derivative test. ∞

We now give two more examples of computing the MLE using the Bernoulli and Normal Distributions.

Example 1.2.4 (Sampling from Bernoulli)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim}$ *Bernoulli*(p)*. Then,*

$$P(X = x) = p^x(1-p)^{1-x}$$

and the likelihood becomes

$$\begin{aligned}\mathcal{L}(p|x_1, ..., x_n) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}\\ &= p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}\\ &= \left(\frac{p}{1-p}\right)^{\sum_i x_i}(1-p)^n\end{aligned}$$

It then follows that the log-likelihood is

$$\ell(p) = \left(\sum_{i=1}^{n} x_i\right)(\log(p) - \log(1-p)) + n\log(1-p)$$

Using calculus we find

$$\begin{aligned}
\ell'(p) &= \frac{\sum_i x_i}{p} - \frac{n - \sum_i x_i}{1-p} = 0 \\
&\implies \frac{\sum_i x_i}{p} = \frac{n - \sum_i x_i}{1-p} \\
&\implies \left(\sum_{i=1}^{n} x_i\right)(1-p) = np - p\sum_{i=1}^{n} x_i \\
&\implies \hat{p}_{MLE} = \frac{\sum_i x_i}{n} = \bar{x} = \textit{sample mean (proportion)}
\end{aligned}$$

♥

Example 1.2.5 (Sampling from Normal)**.** *Suppose* $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. *Then, the likelihood function is*

$$\mathcal{L}(\mu, \sigma^2 | x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\}$$

which implies that the log-likelihood is

$$\ell(\mu, \sigma^2 | x_1, \ldots, x_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

From here there are three cases for the information given about the parameters:

Case 1: *(σ^2 known, μ unknown) In this case we only have one variable to maximize, μ so the maximum is found via calculus:*

$$\begin{aligned}
\frac{d\ell}{d\mu} &= \frac{1}{2\sigma^2} \cdot 2\left(\sum_{i=1}^{n}(x_i - \mu)\right) = 0 \\
&\implies \sum_{i=1}^{n}(x_i - \mu) = 0 \\
&\implies \hat{\mu}_{MLE} = \bar{x}
\end{aligned}$$

just as we would have expected

Case 2: *(σ^2 unknown, μ known) We again proceed as we did with μ:*

$$\begin{aligned}\frac{d\ell}{d\sigma} &= \frac{d}{d\sigma}\left(-n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\\ &= -\frac{n}{\sigma} + \frac{2}{2\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0\\ \implies & -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0\\ \implies & \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}(x_i-\mu)^2\end{aligned}$$

as would have guessed

Case 3: *(μ, σ^2 unknown) In this case we must optimize with respect to both parameters. We thus have to take derivatives with respect to both μ and σ and obtain the* **likelihood equations** *(same as the ones we already computed):*

$$\frac{\partial\ell}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \tag{1.2.1}$$

$$\frac{\partial\ell}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \tag{1.2.2}$$

From (1.2.1) we get as before

$$\hat{\mu}_{MLE} = \bar{x} \qquad \textit{(this doesn't involve } \sigma^2 \textit{at all)}$$

and from (1.2.2) we also get

$$\begin{aligned}\hat{\sigma}_{MLE} &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\end{aligned}$$

♥

1.2.1 Bivariate Normal: Review

In a bivariate normal distribution, we assume to observed a pair of observations (X_1, X_2) such that $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ and an association between X_1 and X_2 can exist or $\mathrm{Cov}(X_1, X_2) \neq 0$ possibly. The context of the bivariate distribution has to do with linear regression[1]. If we consider this context, then the joint distribution between X_1 and X_2, we can be written it as

$$(X_1, X_2) \sim \text{BVN}(\vec{\mu}, \Sigma)$$

where

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

and

$$\begin{aligned} \Sigma &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} && \text{(covariance matrix)} \\ &= \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \end{aligned}$$

The joint density function is then given as

$$f(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})\right\} \qquad \left(\vec{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right)$$

This is fine by itself, but we can expand it for calculations. First we begin with $|\Sigma|$:

$$\begin{aligned} |\Sigma| &= \sigma_1^2\sigma_2^2 - \sigma_{12}^2 \\ &= \sigma_1^2\sigma_2^2 - (\rho\sigma_1\sigma_2)^2 \\ &= \sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2 \\ &= \sigma_1^2\sigma_2^2(1 - \rho^2) \end{aligned}$$

The normalizing constant $1/2\pi|\Sigma|^{1/2}$ is then

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}}$$

Expanding the exponent we see the vector-matrix calculation is expanded as

$$\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}$$

and Σ^{-1} can be written as

[1] See the Appendix for more information

$$
\begin{aligned}
\Sigma^{-1} &= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix}\\
&= \frac{1}{1-\rho^2}\begin{pmatrix}1/\sigma_1^2 & -\rho/\sigma_1\sigma_2\\ -\rho/\sigma_1\sigma_2 & 1/\sigma_2^2\end{pmatrix}
\end{aligned}
$$

after some algebra and simplification, we get the final result for the exponent:

$$
-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]
$$

This makes the alternate form of the bivariate normal distribution as

Bivariate Normal p.d.f

$$
\begin{aligned}
f(x_1,x_2) = \frac{1}{2\pi\sqrt{(1-\rho^2)}\sigma_1\sigma_2}\exp\Bigg\{-\frac{1}{2(1-\rho^2)}\Bigg[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\\
+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\Bigg]\Bigg\}
\end{aligned}
$$

Now we can see what the likelihood estimates for this distribution are.

Example 1.2.6 (Bivariate Normal Likelihood)**.** *The form of the likelihood for this distribution is*

$$
\begin{aligned}
\mathcal{L} = \prod_{i=1}^{n} f(x_{1i},x_{2i}) = \prod_{i=1}^{n}\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\Bigg\{-\frac{1}{2(1-\rho^2)}\Bigg[\left(\frac{x_{1i}-\mu_1}{\sigma_1}\right)^2\\
-2\rho\left(\frac{x_{1i}-\mu_1}{\sigma_1}\right)\left(\frac{x_{2i}-\mu_2}{\sigma_2}\right)\\
+\left(\frac{x_{2i}-\mu_2}{\sigma_2}\right)^2\Bigg]\Bigg\}
\end{aligned}
$$

Here, there are 5 unknowns $\mu_1, \mu_2, \sigma_1, \sigma_2, \sigma_{12}$ and if we solve the simultaneous likelihood *equations we arrive at*

$$\begin{aligned}
\hat{\mu}_1 &= \bar{x}_1 \\
\hat{\mu}_2 &= \bar{x}_2 \\
\hat{\sigma}_1^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \\
\hat{\sigma}_2^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2 \\
\hat{\sigma}_{12} &= \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)
\end{aligned}$$

♥

1.2.2 Summary about Bivariate Normal

We now give a summary about the properties of the bivariate normal distribution.

Facts about the Bivariate Normal Distribution

1. Each marginal distribution is normal, i.e. $X_i \sim N(\mu_i, \sigma_i^2)$ for all $i \in \{1, 2\}$
2. Each linear combination of X_1, X_2 is also normal, that is if

$$\begin{aligned}
Z_1 &= a_1X_1 + a_2X_2 + a_3 \\
Z_2 &= b_1X_1 + b_2X_2 + b_3
\end{aligned}$$

   then Z_1, Z_2 also follow a bivariate normal distribution
3. The conditional distribution of X_1 given X_2 is normal with

$$\begin{aligned}
E(X_1|X_2) &= \mu_1 + \rho\sigma_1\left(\frac{X_2 - \mu_2}{\sigma_2}\right) \\
V(X_1|X_2) &= (1 - \rho^2)\sigma_1^2
\end{aligned}$$

   and

$$\begin{aligned}
E(X_2|X_1) &= \mu_2 + \rho\sigma_2\left(\frac{X_1 - \mu_1}{\sigma_1}\right) \\
V(X_2|X_1) &= (1 - \rho^2)\sigma_2^2
\end{aligned}$$

Proof. To find the conditional distribution we proceed as follows

$$\begin{aligned} f(x_2|x_1) &= \frac{f(x_1,x_2)}{f(x_1)dx_1} \\ &= \frac{\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\{A\}}{\frac{1}{\sqrt{2\pi}\sigma_1}\exp\{B\}} \end{aligned}$$

where

$$A = \left\{ -\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

$$B = -\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2$$

This results in the simplification

$$\begin{aligned} f(x_2|x_1) =& \frac{1}{\sqrt{2\pi\sigma_2^2}\sqrt{1-\rho^2}} \times \exp\left\{ -\frac{1}{(1-\rho^2)}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \right. \\ & \frac{2\rho}{2(1-\rho^2)}\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} - \frac{1}{(1-\rho^2)}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \\ & \left. \frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 \right\} \end{aligned}$$

We then collect the terms and simplify (there is a perfect square in the exponent above), we arrive at a form for $f(x_2|x_1)$ that is of a normal distribution with mean and variance given in fact #2. This concludes the proof. □

4. If there is no association between X_1, X_2, then the joint distribution is a product of the marginal distributions (independence between normal variates).

   Proof. We begin with the simplified p.d.f for the bivariate normal and factorize it accordingly:

$$\begin{aligned} f(x_1,x_2) &= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right\} \times \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\} \\ &= f(x_1) \times f(x_2) \end{aligned}$$

which shows what we sought. □

1.2.3 Non-Calculus MLE

We now give some examples of how other methods of calculating the MLE are better used rather than calculus. These are discrete in nature.

Example 1.2.7 (Screening Test)**.** *Suppose a test for a disease has the property:*

- *Positive if disease presented with probability* 0.9 *(true positive)*
- *Positive if no disease presented with probability* 0.1 *(false positive)*

A test result X could be a binary variable where

$$X = \begin{cases} 1 & \textit{test positive} \\ 0 & \textit{test negative} \end{cases}$$

We want to estimate or determine the chance if a subject has the disease given the result of our test is positive. In this case, the parameter space (values of probabilities we choose) has two values

$$\Omega = \{0.1, 0.9\}$$

We set

$$\begin{cases} \theta = 0.1 & \textit{when person tested does not have disease} \\ \theta = 0.9 & \textit{when person tested has disease} \end{cases}$$

then, $P(X = x|\theta) = \theta^x(1-\theta)^{1-x}$ and as a function of θ, this is a likelihood function. It returns a probability of true positive/false positive under the true outcome of the person. Now, if $X = 0$ (test negative), then

$$P(X = 0|\theta) = \begin{cases} 0.9 & \textit{if } \theta = 0.1 \textit{ (error)} \\ 0.1 & \textit{if } \theta = 0.9 \textit{ (no error)} \end{cases}$$

so the MLE is the value with the highest chance, or $\theta = 0.1$ if the test is negative. Similarly,

$$P(X = 1|\theta) = \begin{cases} 0.9 & \textit{if } \theta = 0.9 \textit{ (no error)} \\ 0.1 & \textit{if } \theta = 0.1 \textit{ (error)} \end{cases}$$

so the MLE is $\theta = 0.9$ if $X = 1$ is observed. Thus, the values of the MLE are

$$\hat{\theta}_{MLE} = \begin{cases} 0.1 & \textit{if } X = 0 \\ 0.9 & \textit{if } X = 1 \end{cases}$$

♥

Example 1.2.8 (Sampling from Uniform)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} U(0, \theta)$*, that is*

$$f(x_i|\theta) = \begin{cases} 1/\theta & x_i \in [0, \theta] \\ 0 & else \end{cases}$$

The joint p.d.f is given by

$$f(x_1, ..., x_n|\theta) = \begin{cases} 1/\theta^n & \forall x_i \in [0, \theta] \\ 0 & o.w. \end{cases}$$

Now, any estimate of θ *must meet or exceed* x_i *for all* $i \in \{1, ..., n\}$*. Since* $1/\theta^n$ *is decreasing in* θ*,*

- *the larger* θ*, the smaller* $1/\theta^n$
- $\hat{\theta}_{MLE}$ *is the smallest value that* θ *can be while still leaving the likelihood non-zero. This occurs when* $\theta = \max_i\{x_i\}$*.*

Hence,

$$\hat{\theta}_{MLE} = \max_{1 \le i \le n}\{x_i\}$$

♥

Remark 1.2.1.

Some facts about MLE

1. *They do not always exist*
2. *They are not always unique (multiple can exist)*

♦

The following examples will help illustrate these ideas.

Example 1.2.9 (Non-Existent MLE)**.** *Consider Sampling from Uniform but with a slight modification:*

$$f(x|\theta) = \begin{cases} 1/\theta & x \in (0, \theta) \\ 0 & o.w. \end{cases}$$

Since $x \in (0, \theta)$ *we can not have* $\hat{\theta}_{MLE} = \max\{x_1, ..., x_n\}$ *because* $\theta \neq x_i$ *for all* $i \in \{1, ..., n\}$*. We can, however, choose* $\hat{\theta}_{MLE}$ *to be arbitarily close to* $\max\{x_1, ..., x_n\}$ *but can never equal* $\max\{x_1, ..., x_n\}$*. Hence, the MLE cannot exist.*

♥

Example 1.2.10 (Non-Unique MLE). *Consider $X_1, ..., X_n \overset{iid}{\sim} U[\theta, \theta+1]$ for all $\theta \in \mathbb{R}$. Then, the p.d.f for each X_i is*

$$f(x_i|\theta) = \begin{cases} 1 & x_i \in [\theta, \theta+1] \\ 0 & o.w. \end{cases}$$

This makes the joint p.d.f as

$$f(x_1, ..., x_n|\theta) = \begin{cases} 1 & x_i \in [\theta, \theta+1] \quad \forall i \in \{1, ..., n\} \\ 0 & o.w. \end{cases}$$

Now, we know by the support of this function that $x_i \geq \theta$ for all $i \in \{1, ..., n\}$ implies that even the least of these x_i is at or above θ, or $\min\{x_1, ..., x_n\} \geq \theta$. Also, when $x_i \leq \theta + 1$ for all $i \in \{1, ..., n\}$ we know that $\max\{x_1, ..., x_n\} \leq \theta + 1$ or $\max\{x_1, ..., x_n\} - 1 \leq \theta$. Hence,

$$x_i \in [\theta, \theta+1] \quad \forall i \in \{1, ..., n\} \implies \theta \in [\max\{x_1, ..., x_n\} - 1, \min\{x_1, ..., x_n\}]$$

and the likelihood function is

$$\mathcal{L}(\theta|x_1, ..., x_n) = \begin{cases} 1 & \theta \in [\max\{x_1, ..., x_n\} - 1, \min\{x_1, ..., x_n\}] \\ 0 & else \end{cases}$$

therefore any value for θ in the interval

$$[\max\{x_1, ..., x_n\} - 1, \min\{x_1, ..., x_n\}]$$

can be used as an MLE[2].

Note: *in this problem we have a uniform over an interval of length 1 but we cannot specifiy where this interval is located.*

♥

1.2.4 More Properties of MLE

We give some more information about MLE's and their properties that we may use in a statistical setting.

Properties of MLE's

1. Invariance

   Theorem 1.2.1 (MLE Invariance). *If $\hat{\theta}_{MLE}$ is MLE of θ and g is a one-to-one function*

[2] $\max\{x_1, ..., x_n\} - 1 \leq \min\{x_1, ..., x_n\}$ since at most minimum and maximum differ by at most one.

of θ, then $\widehat{g(\theta)}_{MLE} = g(\hat{\theta}_{MLE})$.

Proof. Let $\Omega =$ parameter space of original parametrization. Then when we transform θ get get a new parameter space: $g(\Omega) = \Gamma$. If we set $\psi = g(\theta)$ and $g^{-1} = h$, then it follows that

$$\theta = h(\psi)$$

Notice, the original p.d.f is $f(x|\theta)$ and this is equivalent to $f(x|h(\psi))$. When we transform θ, the likelihoods per $g(\theta)$ stay the same for all $\theta \in \Omega$ since g is bijective. Thus,

$$\mathcal{L}(\psi|x_1, ..., x_n) = \mathcal{L}(\theta|x_1, ..., x_n) = f(x_1, ..., x_n|h(\psi))$$

and we look for $\hat{\psi}_{MLE} = \hat{\psi}$ such that $\hat{\psi}$ maximizes $\mathcal{L}(\psi)$. Now, by definition of MLE, $\hat{\theta}_{MLE}$ maximizes $\mathcal{L}(\theta|x_1, ..., x_n)$ and since $\theta = h(\psi)$ we must have $\hat{\theta}_{MLE} = h(\hat{\psi}_{MLE})$ as the likelihoods between parametizations are the same. Since $g^{-1} = h$, we have $g(\hat{\theta}_{MLE}) = \hat{\psi}_{MLE}$ as we sought to show. □

2. Defining the MLE of a function $g(\theta)$

Definition 1.2.5 (MLE of function $g(\theta)$)**.** *If $g(\theta)$ is an arbitrary function (not always one-to-one) such that*

$$g : \Omega \rightarrow \underbrace{G}_{\text{image of } \Omega}$$

Then, we can define a set $G_t \subset \Omega$ where $G_t = \{\theta : g(\theta) = t\}$. If we further define $L^*(t) = \max_{\theta \in G_t} \ell(\theta|x_1, ..., x_n)$, that is, L^* is the maximum of the log-likelihood function such that the likelihood is maximized over all $\hat{\theta}$ that map into t. With this, we define the MLE of $g(\theta)$ to be the value $\hat{t}$ such that

$$L^*(\hat{t}) = \max_{t \in G_t} L^*(t)$$

so we have possibly several θ such that $g(\theta) = t$ mapping into the same t.

Note: *$L^*(t)$ is the value of log-likelihood where the log-likelihood is maximized over the values of θ that map into t, this takes care of any mappings that are surjective as well. We then find the value of t that maximizes L^*; this value is $\hat{t}_{MLE} = \widehat{g(\theta)}_{MLE}$.*

3. Invariance (generalized)

Theorem 1.2.2 (Generalized Invariance)**.** *If $\hat{\theta} =$ MLE of θ and $g(\theta)$ is* **any** *function of θ, then*

$$\left[\widehat{g(\theta)}\right]_{MLE} = g(\hat{\theta}_{MLE})$$

Proof. We need to show that $\hat{t} = g(\hat{\theta}_{MLE})$ satisfies

$$\max_{t \in G} L^*(t) = L^*(\hat{t})$$

Note that $L^*(t)$ is the maximum of $\ell(\theta|x_1, ..., x_n)$ over one subset of Ω because we choose that value of θ such that $L^*(t)$ is a maximum. Now, $\ell(\hat{\theta}_{MLE}|x_1, ..., x_n)$ is a maximum over all θ. This implies

$$L^*(t) \leq \ell(\hat{\theta}_{MLE}|x_1, ..., x_n) \qquad \forall t \in G$$

Notice that if I maximize over a larger set of parameters (i.e. more than just t if there are 2 or more parameters in the setting), the maximum cannot get smaller. As reminder, all we have left to show is

$$L^*(\hat{t}) = \underbrace{\ell(\hat{\theta}_{MLE}|x_1, ..., x_n)}_{\text{max likelihood parametrized by } \theta}$$

Now, by definition $\hat{\theta}_{MLE} \in G_{\hat{t}}$ and furthermore $\hat{\theta}_{MLE}$ maximizes $\ell(\theta|x_1, ..., x_n)$ over all θ and since $G_t \subset \Omega$, we know $\hat{\theta}_{MLE}$ also maximizes over all $\theta \in G_t$. This implies that

$$L^*(\hat{t}) = \ell(\hat{\theta}_{MLE}|x_1, ..., x_n)$$

this means that $\hat{t} = g(\hat{\theta}_{MLE})$ is the MLE of $g(\theta)$. $\square$

We now give an example that illustrates the property of invariance:

Example 1.2.11 (Invariance for Variance)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ *where* μ, σ^2 *are both unknown. By Sampling from Normal we know*

$$\begin{aligned} \hat{\mu}_{MLE} &= \bar{x} \\ \hat{\sigma}^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{aligned}$$

and we want the MLE of σ *(std. dev.) and* $E(X^2)$*. We can set* $g(x) = \sqrt{x}$ *and arrive at (by invariance)*

$$\left[\widehat{g(\sigma^2)}\right]_{MLE} = g(\hat{\sigma}^2_{MLE})$$

which can be simplified into

$$\hat{\sigma}_{MLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Note also that

$$\begin{aligned} &\sigma^2 = E(X^2) - (E(X))^2 \\ &\implies E(X^2) = \sigma^2 + (E(X))^2 \\ &\implies E(X^2) = h(\sigma^2, E(X)) \qquad\qquad (h(x, y) = x + y^2) \end{aligned}$$

and again by invariance (note multiple parameters $\sigma^2, E(X)$) we have

$$\begin{aligned}\left[\widehat{E(X^2)}\right]_{MLE} &= h\left(\hat{\sigma}^2_{MLE}, \left[\widehat{E(X)}\right]_{MLE}\right)\\ &= h(\hat{\sigma}^2_{MLE}, \bar{x})\\ &= \hat{\sigma}^2_{MLE} + \bar{x}^2\end{aligned}$$

♥

One other property of the MLE is **consistency**.

Consistency of MLE

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$ and for some sample size $n \geq n_0$, there is a unique MLE $\underbrace{\hat{\theta}_{[n,MLE]}}_{\text{depends on } n}$.

Then, under some conditions satisfied under certain scenarios:

$$\hat{\theta}_{[n,MLE]} \xrightarrow{P} \theta \qquad (\text{as } n \to \infty)$$

☕

1.2.5 Computing MLE's (Caveats)

Computing MLE's can be a challenge sometimes as there is not always an analytical solution to the likelihood equation. We give two examples to show this.

Example 1.2.12 (Gamma Distribution MLE)**.** *Suppose $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\alpha)$ where*

$$f(x|\alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} \qquad \forall x > 0, \alpha > 0$$

then we can show:

$$\mathcal{L}(\alpha|x_1, \ldots, x_n) = \frac{1}{\Gamma^n(\alpha)} \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \times \exp\left(-\sum_{i=1}^{n} x_i\right)$$

The log-likelihood equation is then

$$\frac{\partial \ell(\alpha|x_1, \ldots, x_n)}{\partial \alpha} = 0$$

which results in

$$\underbrace{\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}}_{\textit{digamma function}} = \frac{1}{n} \sum \log x_i$$

Notice that the digamma function is tabulated meaning that there is no analytical solution for α. This means that we can only use numerical approximations.

♥

Example 1.2.13 (Cauchy Distribution MLE)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} f(x|\theta)$ where*

$$f(x|\theta) = \frac{1}{\pi(1+(x-\theta)^2)} \qquad -\infty < x < \infty$$

Then, the likelihood function is

$$\mathcal{L}(\theta|x_1, ..., x_n) = \frac{1}{\pi^n \prod_{i=1}^n [1+(x-\theta)^2]}$$

and again we need a numerical technique to find a solution.

♥

One technique to extract MLE's is **Newton-Raphson** since we wish to solve the equation $\mathcal{L}'(\theta|x_1, ..., x_n) = f(\theta) = 0$. In this method we use a Taylor Approximation. If we set θ_0 as an initial guess, then we use

$$\theta_1 = \theta_0 - \frac{f(\theta_0)}{f'(\theta_0)}$$

then we keep updating until there is no (noticeable) change in updated value of θ.

1.3 Method of Moments (MM)

In this method, we assume $X_1, ..., X_n \overset{iid}{\sim} P_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ and $E(X^k) < \infty$ (finite moments). We set

$$\mu_j = E(X^j|\theta)$$

and assume $\mu(\boldsymbol{\theta}) = (\mu_1(\theta), ..., \mu_k(\theta))$ is one-to-one (i.e. there is an inverse function M such that $\boldsymbol{\theta} = M(\mu_1(\theta), ..., \mu_k(\theta))$). When we observe the data, we let

$$m_j = \frac{1}{n}\sum_{i=1}^n x_i^j \qquad j = 1, ..., k$$

be the sample moments. The **method of moments estimator** is then

$$\hat{\theta}_{MM} = M(m_1, ..., m_k)$$

This implies $M(m_1, ..., m_k) \approx \boldsymbol{\theta}$ and further per parameter: $m_j \approx \mu_j$. Using this information, we can give a procedure for finding the MM estimates:

Steps to Find MM Estimates

1. Empirically find $\hat{\mu}^j$
2. Find the parametric forms for μ^j
3. Equate each $\hat{\mu}_i^j$ with μ_i^j respectively
4. Solve the resulting system of equations for the parameters

We give an example to illustrate this process

Example 1.3.1 (Gamma Distribution MM)**.** *Suppose we sample $X_1, \ldots, X_n$ from a Gamma distribution and want to estimate unknown parameters α and β. Note that the pdf of a Gamma distributed variable X_i is:*

$$f(x_i|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i}.$$

This makes the likelihood function:

$$L(x_1, \ldots, x_n|\alpha,\beta) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n \left(\prod_{i=1}^n x_i\right)^{\alpha-1} e^{-\beta\sum_{i=1}^n x_i}$$

which is very difficult to optimize due to the gamma function. Since MLE is not possible easily we will ustilize MOM to generate point estimates of α and β. As there are 2 parameters unknown we only need to use the first 2 sample and population moments to compute our point estimates. Equating the respective moments gives us:

$$E(X) = \frac{\alpha}{\beta} = \frac{\sum_{i=1}^n x_i}{n} = \overline{x} \tag{1.3.1}$$

$$E(X^2) = Var(X) + (E(X))^2 = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2} = \frac{\sum_{i=1}^n x_i^2}{n}. \tag{1.3.2}$$

Now we can solve with substitution the point estimates for our parameters, rewriting the value of α in terms of β:

$$\alpha = \beta\overline{x}$$

and substituting the value in (1.3.2) gives:

$$\frac{\overline{x}}{\beta} + (\overline{x})^2 = \frac{1}{n}\sum_{i=1}^n x_i^2$$

and solving for β results in:

$$\hat{\beta}_{MM} = \frac{n\overline{x}}{\sum_{i=1}^n x_i^2 - n(\overline{x})^2} = \frac{n\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

Similarly, we can solve for α:

$$\beta = \frac{\alpha}{\overline{x}}$$

and rewrite (1.3.2):

$$\frac{(\overline{x})^2}{\alpha} + (\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \implies \frac{(\overline{x})^2}{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2\right) \implies \frac{1}{\alpha} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n(\overline{x})^2} \implies$$

$$\implies \hat{\alpha}_{MM} = \frac{n(\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

♥

So, why do we need this process? How does it relate to the MLE? We give 3 reasons:

- MM estimates do not require optimizing a likelihood function, so they make be less computationally extensive
- Often MM estimates are also MLE's
- MM estimates are **consistent** by the law of large numbers. So,

$$\hat{\theta}_{MM} \xrightarrow{P} \theta \qquad (\text{as } n \to \infty)$$

We give one more example that relates this to the MLE

Example 1.3.2 (Normal Distribution MM). *Suppose $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where μ, σ^2 are unknown. Then, the observed moments are:*

$$\hat{\mu}_1 = m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\mu}_2 = m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Then, $\hat{\mu}_{MM} = \hat{\mu}_{MLE} = \bar{x}$ and

$$\begin{aligned} & m_2 \approx \sigma^2 + \mu^2 && (\sigma^2 = E(X^2) - \mu^2) \\ & \implies m_2 - \mu^2 \approx \sigma^2 \\ & \implies \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \hat{\sigma}^2_{MM} = \hat{\sigma}^2_{MLE} \end{aligned}$$

Hence, both MM estimates are of the same form as the MLE estimates.

♥

1.4 Appendix (Inference: Estimation)

We now give an account of the bivariate distribution based on simple linear regression. The results as we already derived are the same only this way gives an idea of how one might create the form of the distribution. We begin with the regression setup. We assume $X_1 \sim N(\mu_{X_1}, \sigma^2_{X_1})$ and $X_2 \sim N(\mu_{X_2}, \sigma^2_{X_2})$ on the condition that $X_2 = \alpha + \beta X_1 + \epsilon$ for parameters α, β. If we set the error per the true regression line as $\epsilon = Z = X_2 - \alpha - \beta X_1$, then we have $E(Z) = 0$ and $V(Z) = \sigma_{X_2|X_1}$ since the variation of the error is the sole source of the variation of the conditional distribution of $X_2|X_1$.

Now, how would we know the true form of α and β? One solution is to consider the covariance between X_1 and X_2. We use the form $\text{Cov}(X_1, X_2) = E(X_1X_2) - \mu_{X_1}\mu_{X_2}$. We compute $E(X_1X_2)$ as follows:

$$\begin{aligned}
E(X_1X_2) &= \int_{x_1} E(X_1X_2|X_1 = x_1)f(x_1)dx_1 \\
&= \int_{x_1} x_1E(X_2|X_1 = x_1)f(x_1)dx_1 \\
&= \int_{x_1} x_1E(Z)f(x_1)dx_1 \\
&= \int_{x_1} x_1(\alpha + \beta x_1)f(x_1)dx_1 \\
&= \alpha\mu_{X_1} + \beta \int_{x_1} x_1^2 f(x_1)dx_1 \\
&= \alpha\mu_{X_1} + \beta(\sigma^2_{X_1} + \mu^2_{X_1}) \\
&= \alpha\mu_{X_1} + \beta\sigma^2_{X_1} + \beta\mu^2_{X_1}
\end{aligned}$$

Note that $\mu_{X_2} = \alpha + \beta\mu_{X_1}$ since $X_2 = \alpha + \beta X_1 + \epsilon$ where $\epsilon = Z \sim N(0, \sigma_{X_2|X_1})$. This makes $\mu_{X_1}\mu_{X_2} = \alpha\mu_{X_1} + \beta\mu^2_{X_1}$. This all implies

$$\begin{aligned}
&\text{Cov}(X_1, X_2) = \beta\sigma^2_{X_1} \\
&\implies \rho = \beta\frac{\sigma_{X_1}}{\sigma_{X_2}} \\
&\implies \beta = \rho\frac{\sigma_{X_2}}{\sigma_{X_1}}
\end{aligned}$$

The corresponding form of α is then

$$\alpha = \mu_{X_2} - \rho\frac{\sigma_{X_2}}{\sigma_{X_1}}\mu_{X_1}$$

We next compute σ_Z^2. Since $Z = X_2 - \alpha - \beta X_1$ it follows that

$$\begin{aligned}
\sigma_Z^2 &= V(X_2 - \alpha - \beta X_1) \\
&= \sigma_{X_2}^2 + \beta^2\sigma_{X_1}^2 - 2\beta(\beta\sigma_{X_1}^2) && (\text{Cov}(X_2, -\beta X_1) = -\beta\text{Cov}(X_2, X_1)) \\
&= \sigma_{X_2}^2 - \beta^2\sigma_{X_1}^2 \\
&= \sigma_{X_2}^2(1-\rho^2) && \left(\beta^2 = \rho^2\frac{\sigma_{X_2}^2}{\sigma_{X_1}^2}\right)
\end{aligned}$$

We are now ready to compute the distribution of Z. By our assumption, $Z \sim N(\mu_Z, \sigma_Z^2)$ where $\mu_Z = 0$ and $\sigma_Z^2 = \sigma_{X_2}^2(1-\rho^2)$. It now follows that

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma_Z}\exp\left\{-\frac{1}{2}\left(\frac{z}{\sigma_Z}\right)^2\right\}$$

Since $f_Z(z) = f_{X_2|X_1=x_1}(x_2)$ since variation is only from the error assuming X_1 is observed, some substitution gives:

$$\begin{aligned}
f_{X_2|X_1=x_1}(x_2) &= \frac{1}{\sqrt{2\pi}\sigma_{X_2}\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}\left(\frac{x_2-\alpha-\beta x_1}{\sigma_{X_2}\sqrt{1-\rho^2}}\right)^2\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma_{X_2}\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x_2-\alpha-\beta x_1}{\sigma_{X_2}}\right)^2\right\}
\end{aligned}$$

Which shows that $(X_2|X_1 = x_1) \sim N(\alpha + \beta x_1, \sigma_{X_2}^2(1-\rho^2))$. More precisely, we have

$$\begin{aligned}
E(X_2|X_1 = x_1) &= \alpha + \beta x_1 \\
&= \mu_{X_2} - \rho\frac{\sigma_{X_2}}{\sigma_{X_1}}\mu_{X_1} + \rho\frac{\sigma_{X_2}}{\sigma_{X_1}}x_1 \\
&= \mu_{X_2} + \rho\frac{\sigma_{X_2}}{\sigma_{X_1}}(x_1 - \mu_{X_1})
\end{aligned}$$

which is the same form as we gave in point #3 in BVN Facts.

Note, the quantity

$$\left(\frac{x_2-\alpha-\beta x_1}{\sigma_{X_2}}\right)^2$$

can be simplified into (with the addition and subtraction of $\alpha - \beta\mu_{X_1}$) into

$$\left(\frac{(x_2-\mu_{X_2}) - \beta(x_1-\mu_{X_1})}{\sigma_{X_2}}\right)^2$$

and expanding this square gives us

$$
\begin{aligned}
&\left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2 - \frac{2\beta(x_2-\mu_{X_2})(x_1-\mu_{X_1})}{\sigma^2_{X_2}} + \beta^2\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_2}}\right)^2 \\
&= \left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2 - \frac{2\rho(x_2-\mu_{X_2})(x_1-\mu_{X_1})}{\sigma_{X_2}\sigma_{X_1}} + \rho^2\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2 \quad \text{since } \left(\beta^2 = \rho^2\frac{\sigma^2_{X_2}}{\sigma^2_{X_1}}\right)
\end{aligned}
$$

since the marginal distribution of X_1 is

$$
f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_{X_1}} \exp\left\{-\frac{1}{2}\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2\right\}
$$

it follows that $f(x_2, x_1) = f(x_2|x_1)f(x_1)$ has the form

$$
\begin{aligned}
f(x_1, x_2) = \frac{1}{2\pi\sqrt{(1-\rho^2)}\sigma_{X_1}\sigma_{X_2}} \exp\Bigg\{ -\frac{1}{2(1-\rho^2)}\Bigg[&\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2 - 2\rho\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)\left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right) \\
&+ \left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2\Bigg]\Bigg\}
\end{aligned}
$$

which is the same form as the distribution for the bivariate normal. This argument is axis invariant since we kept both X_1 and X_2 random and can be repeated for $X_1|X_2$ and give similar results (the 2's are permuted with 1's).

Chapter 2 — Bayesian Inference

Bayesian inference is a different type of statistical method. This theory has its roots in **Bayes' Theorem** and thinks of probabilities as **degrees of belief** rather than **ratios observed by experiment**. To begin, we recall the methods of frequentist inference. In a frequentist statistical model we have

1. $X =$ random variable of interest
2. $X \sim P_\theta =$ distribution variable follows (the model)
3. $\theta \in \Omega =$ the parameter space

In this method of inference, we usually consider θ a fixed but unknown number and the objective of statistical activity is to estimate θ. What makes the bayesian viewpoint different is that θ **is NOT fixed**. Instead, θ is a **random variable** representing out degree of belief over the possible values that parameter can take. This makes θ have a probability distribution much like our sampled variates X_i.

2.1 Prior and Posterior Distribution

In a Bayesian setting, we sample variates as we do with frequentist methods only we assume a random value of θ as opposed to a fixed on. So, our model is then

$$X_1, \ldots, X_n | \theta \sim P_\theta \qquad (\theta \text{ random})$$

where $\theta \sim \xi(\theta)$. Here $\xi(\theta)$ is called the **prior distribution**. We give some facts about it below

Prior Distribution Facts

- It is determined for θ **prior** to observing the data (or performing the experiment)
- The range (support for θ) is the parameter space Ω
- It is a probability distribution (integrates to 1) except in the case of the **improper prior**

The aim of Bayesian inference to use the data we have collected to *update* the distribution of θ so as to reflect what we have observed. This is known as **Bayesian Updating**. The result of bayesian updating is to generate a **posterior distribution** $\xi(\theta|x_1, ..., x_n)$.

2.1.1 Posterior Distribution

The posterior distribution is one conditional on the data. To find it we first consider the probability in the discrete case, $P(\theta|\text{Data})$ by Bayes' Theorem this can we written as

$$\begin{aligned}
P(\theta|\text{Data}) &= \frac{P(\text{Data}|\theta)P(\theta)}{P(\text{Data},\theta)} \\
&= \frac{P(\text{Data}|\theta)P(\theta)}{\sum\limits_{\Omega} P(\text{Data}|\theta)P(\theta)} && \text{(Law of Total Prob.)} \\
&= \frac{P(x_1, ..., x_n|\theta)P(\theta)}{\sum\limits_{\Omega} P(x_1, ..., x_n|\theta)P(\theta)} && (\text{Data} = x_1, ..., x_n) \\
&= \frac{f(x_1, ..., x_n|\theta)P(\theta)}{\sum\limits_{\Omega} f(x_1, ..., x_n|\theta)P(\theta)} && \text{(by definition of joint mass function)}
\end{aligned}$$

Note that the denominator of this fraction is called the **normalizing constant**. If we keep this form, this is how we will compute the posterior probability $P(\theta|\text{Data})$ in the discrete case. In the continuous case, we set $P(\theta|\text{Data}) = \xi(\theta|\text{Data})d\theta$ and $P(\theta) = \xi(\theta)d\theta$. The summation over the parameter space then also becomes and integral to adjust for the continuous nature of θ. We then have the continuous form as

$$\begin{aligned}
[\xi(\theta|\text{Data})]d\theta &= \left[\frac{f(\text{Data}|\theta)\xi(\theta)}{\displaystyle\int_{\Omega} f(\text{Data}|\theta)\xi(\theta)d\theta}\right] d\theta = \\
\xi(\theta|x_1, ..., x_n) &= \frac{f(x_1, ..., x_n|\theta)\xi(\theta)}{\underbrace{\displaystyle\int_{\Omega} f(x_1, ..., x_n|\theta)\xi(\theta)d\theta}_{g(x_1,...,x_n)}} && (d\theta\text{'s cancel for densities})
\end{aligned}$$

where we give a new name for the normalizing constant $\displaystyle\int_{\Omega} f(x_1, ..., x_n|\theta)\xi(\theta)d\theta$ as $g(x_1, ..., x_n)$ for brevity and to reflect that the resulting integration marginalizes out the variable θ. We formalize this result by a theorem.

Theorem 2.1.1 (Posterior Distribution Form)**.** *The form for the posterior distribution is given by*

$$\xi(\theta|x_1,...,x_n) = \frac{f(x_1,...,x_n|\theta)\xi(\theta)}{g(x_1,...,x_n)}$$

where $X_1,...,X_n|\theta \overset{iid}{\sim} P_\theta$

Proof. By conditional independence of the data, we have the form of the joint p.d.f given a value of θ as

$$f(x_1,...,x_n|\theta) = f(x_1|\theta) \times \cdots \times f(x_n|\theta)$$

Further, we can create a joint density for $x_1,...,x_n,\theta$ noting it with the function z. To make notation easier, we set $x = (x_1,...,x_n)$. By property of joint p.d.f's, we can write

$$z(\theta,x) = z(x,\theta)$$

and by conditional probability for continuous functions (each factor belongs to a different probability distribution as we speak of different events) we have

$$\begin{aligned}
\xi(\theta|x)g(x) &= f(x|\theta)\xi(\theta) = \\
\xi(\theta|x) &= \frac{f(x|\theta)\xi(\theta)}{g(x)} = \\
\xi(\theta|x) &= \frac{f(x|\theta)\xi(\theta)}{\displaystyle\int_\Omega f(x|\theta)\xi(\theta)d\theta} && \text{(def. of marginal distn.)}
\end{aligned}$$

which is the same form as we intuitively found out previously. □

Example 2.1.1 (Fluorescent Lamps)**.** *Suppose the lifetimes for fluorescent lamps is given by* $X_1,...,X_n \overset{iid}{\sim} \exp(\theta)$. *This means that*

$$f(x_i|\theta) = \theta e^{-\theta x} \qquad x > 0, \theta > 0$$

The aim is to find a posterior distribution for θ *given our data* $x_1,...,x_n$. *The joint distribution* $f(x_1,...,x_n|\theta)$ *is given by*

$$\begin{aligned}
\prod_{i=1}^n f(x_i|\theta) &= \prod_{i=1}^n \theta e^{-\theta x_i} \\
&= \theta^n e^{-\theta \sum x_i} \\
&= \theta^n e^{-\theta y} && \left(\text{let } y = \sum x_i\right)
\end{aligned}$$

where $\theta \in (0,\infty)$. *Now, how can we find a prior before we observe the lifetimes? A distribution with support* $\theta \in (0,\infty)$ *is given by the family of gamma distributions*[1]. *The* *form of the gamma prior is*

$$\xi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \qquad \theta > 0,\ \alpha > 0,\ \beta > 0$$

for some given α, β. Note that $\frac{\beta^\alpha}{\Gamma(\alpha)}$ is the reciprocal of the area under $\theta^{\alpha-1}e^{-\beta\theta}$ over the support. Some facts about this prior are:

$$E(\theta^k) = \frac{\Gamma(\alpha+\beta)}{\beta^k\Gamma(\alpha)} \qquad E(\theta) = \frac{\alpha}{\beta} \qquad V(\theta) = \frac{\alpha}{\beta^2}$$

What values of α, β can be a 'good' prior (reflects population closely)? Suppose we assume the prior mean and variance are

$$E(\theta) = \frac{\alpha_0}{\beta_0} = 0.0002 \qquad V(\theta) = \frac{\alpha_0}{\beta_0^2} = 0.0001$$

Solving for α_0 and β_0 we get

$$\alpha_0 = 4 \qquad \beta_0 = 20,000$$

Hence, the prior is

$$\begin{aligned}\xi(\theta) &= \frac{(20,000)^4}{\Gamma(4)}\theta^3 e^{-20,000} \\ &= \frac{(20,000)^4}{3!}\theta^3 e^{-20,000} \qquad (\Gamma(\alpha_0) = (\alpha_0 - 1)! = 3!)\end{aligned}$$

The posterior is then given by

$$\begin{aligned}\xi(\theta|y) &= \frac{\theta^n e^{-\theta y} \times \frac{(20,000)^4}{3!}\theta^3 e^{-20,000\theta}}{\int_0^\infty \theta^n e^{-\theta y} \times \frac{(20,000)^4}{3!}\theta^3 e^{-20,000\theta} d\theta} \\ &= \frac{\theta^{n+3}e^{-(y+20,000)\theta}}{\int_0^\infty \theta^{n+3}e^{-(y+20,000)\theta} d\theta} \qquad (\textit{const. cancel})\end{aligned}$$

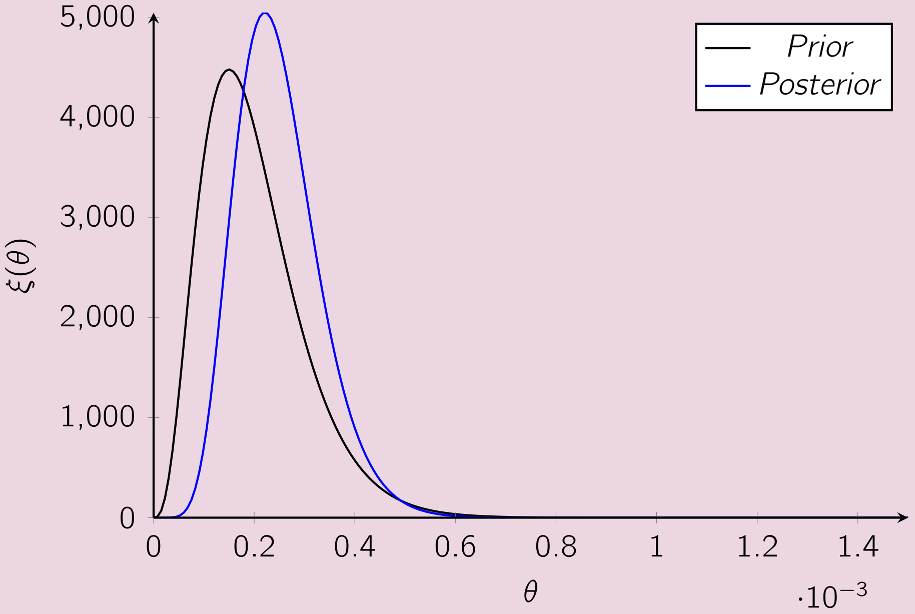
We know that

$$\int_0^\infty \theta^{n+3}e^{-(y+20000)\theta} d\theta = \frac{\Gamma(n+4)}{(y+20000)^{n+4}}$$

so

$$\begin{aligned}\xi(\theta|y) &= \frac{-(y+20000)^{n+4}}{\Gamma(n+4)}\theta^{n+3}e^{-(y+20000)\theta} \\ &\sim \textit{Gamma}(\alpha = n+4, \beta = (y+20000))\end{aligned}$$

Plots of the prior and posterior distribution are given below. Note that in the posterior *distribution plot $n = 5$ and $y = 16,178$.*



♥

Remark 2.1.1. *Since $g(x_1, \ldots, x_n)$ is constant, we can write the posterior as*

$$\xi(\theta | x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\theta) \times \xi(\theta)}{\text{constant}}$$

this would then imply

$$\begin{aligned}\xi(\theta | x_1, \ldots, x_n) &\propto f(x_1, \ldots, x_n|\theta)\xi(\theta) = \prod_i f(x_i|\theta)\xi(\theta) \\ &\propto (\text{likelihood}) \times (\text{prior})\end{aligned}$$

so the density of the posterior is proportional to the product of the densities of the likelihood and prior.

♦

2.1.2 Conjugate Prior

Suppose we have a data distribution $f(x_1, \ldots, x_n|\theta)$ and a prior from some family of distributions such that the posterior is from the same family. Then, we call the family the prior and posterior are from the **conjugate family** and the prior the **conjugate prior**. We have seen a conjugate prior and family in the previous example. The conjugate prior is distributed

[1] **Fact:** the exponential distribution is a special case of the gamma distribution with $\alpha = 1$.

Gamma(α_0, β_0), the posterior is Gamma$(\alpha_0 + n, \sum x_i + \beta_0)$, and the conjugate family is the set of all Gamma distributions. We give more examples of conjugate priors below

Example 2.1.2 (Beta is a Conjugate Prior)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim}$ Bernoulli(θ). Then*

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

which implies that the joint distribution is

$$f(x_1, ..., x_n|\theta) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} = \theta^y(1-\theta)^{n-y}$$

if we set $y = \sum_i x_i$. For the prior $\xi(\theta)$ we need a distribution with support $\Omega = [0, 1]$. A good candidate is the beta distribution, so we set

$$\xi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad 0 \leq \theta \leq 1 \quad \alpha, \beta > 0$$

This results in

$$f(x_1, ..., x_n|\theta)\xi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$$

Also, notice that

$$\int_0^1 \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}d\theta = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}$$

which then implies

$$\begin{aligned}\xi(\theta|x_1, ..., x_n) &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &\sim \text{Beta}(y+\alpha, n-y+\beta)\end{aligned}$$

So, the posterior is a beta distribution as well which makes the beta distribution a conjugate prior assuming the data is Bernoulli.

♥

Example 2.1.3 (Poisson is Conjugate Prior)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim}$ Poisson(θ) where $\theta \sim$ Gamma(α, β). Then,*

$$\begin{aligned}f(x_1, ..., x_n|\theta)\xi(\theta) &= \frac{1}{x_1! \ldots x_n!}\theta^{\sum_i x_i}e^{-n\theta}\left[\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\right] \\ &= \frac{\beta^\alpha}{x_1! \ldots x_n! \times \Gamma(\alpha)}\left[\theta^{y+\alpha-1}e^{-\theta(n+\beta)}\right] \qquad (y = \textstyle\sum_i x_i)\end{aligned}$$

with some calculations and simplifications of the constants (anything observed and without the quantity θ cancel out in the form of the posterior), we can show that the posterior also *follows a Gamma distribution too.*

♥

Example 2.1.4 (Normal is a Conjugate Prior)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ *where* μ *is unknown and* σ^2 *is known. Since we don't know the value of* μ*, we set* $\mu = \theta$ *and have*

$$\theta \sim N(\mu_0, \nu_0^2)$$

with some calculations[2] *we can show*

$$(\theta | x_1, ..., x_n) \sim N(\mu_1, \nu_1^2)$$

where

$$\mu_1 = \frac{\sigma^2\mu_0 + n\nu_0^2\bar{x}_n}{\sigma^2 + n\nu_0^2} \qquad \text{and} \qquad \nu_1^2 = \frac{\sigma^2\nu_0^2}{\sigma^2 + n\nu_0^2}$$

♥

2.2 Bayes' Estimator

We now will use the theory of Bayesian Statistics to create point estimations for population parameters as we did with likelihood functions in the estimation chapter. First we review the setup. Suppose we have $X_1, ..., X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$ where $X_i | \theta \sim P_\theta$ and

$$\xi(\theta) = \text{prior distribution}$$
$$\xi(\theta | x_1, ..., x_n) = \text{posterior distribution}$$

The question is: **how can we use this information to find an actual estimate of θ?** Recall that the definition of an **estimator** $\delta(X_1, ..., X_n)$ of a parameter θ is a real valued function such that

$$\delta : \mathbb{R}^n \to \Omega$$

Observed values of the estimator are called *estimates* $\delta(x_1, ..., x_n)$. In order to create a Bayesian estimate, we can employ what is called at **Loss Function.**

2.2.1 Loss Function

To motivate the idea behind the loss function, we give the following dialogue:

[2] see Appendix for a derivation

Q: What to we want from an estimator?
A: It should yield a value close to the true quantity.

Q: How can we measure the performance of an estimator?
A: Through a Loss Function

We define a loss function as follows:

Definition 2.2.1 (Loss Function)**.** *A loss function $L(\theta, a)$ is a function of 2 variables θ and a whose output can be described by a probability distribution. If θ is a true parameter and an estimate of it is $\hat{\theta} = a$, then*

$$L(\theta, \hat{\theta})$$

is the loss. In other words, $L(\theta, \hat{\theta})$ measures the discrepancy between θ and its estimate $\hat{\theta}$. ♣

In Bayesian inference, θ is a random variable with support Ω and so then is the loss $L(\theta, a)$ which is a function of it. While we cannot then find a deterministic value for the loss, we can compute the *expected loss*. This would be

$$\underbrace{E\left[L(\theta, a)\right]}_{\text{average loss}} = \int_{\Omega} L(\theta, a)\xi(\theta)d\theta$$

assuming we have not conducted the experiment yet and only use prior information. The loss over all possible values of θ is then a function of a.

Note: We have the value of a as

$$a = \text{specific value for a specific sample}$$

and will have to create some condition to extract it. ∞

When we observe the data, it is best not to use the prior for computing the expected loss. Instead, we use the **expected posterior loss** which is given by

$$E\left[L(\theta, a)|\, x_1, ..., x_n\right] = \int_{\Omega} L(\theta, a)\xi(\theta|x_1, ..., x_n)d\theta$$

Some examples of typical loss functions are

1. **Squared Error Loss:** $L(\theta, a) = (\theta - a)^2$
2. **Absolute Error Loss:** $L(\theta, a) = |\theta - a|$

Since we are looking at how far away the estimate a is from θ, we use differences.

2.2.2 Defining the Bayes' Estimator

The specific estimator that is "best" for the data we observe relative to the expected posterior loss is known as the **Bayes' Estimator** which we define below:

Definition 2.2.2 (Bayes' Estimator)**.** *A Bayes' Estimator $\delta^*(X_1, ..., X_n)$ is an estimator such that its observed value $\delta^*(x_1, ..., x_n)$ (the Bayes' estimate) minimizes the expected posterior loss. In other words, $\delta^*(x_1, ..., x_n)$ is value such that*

$$\delta^*(x_1, ..., x_n) = E\left[L(\theta, \delta^*(x_1, ..., x_n))|\, x_1, ..., x_n\right] \leq \underbrace{E\left[L(\theta, a)|\, x_1, ..., x_n\right]}_{\text{for any other "a"}}$$

or, equivalently it is a value such that

$$E\left[L(\theta, \delta^*(x_1, ..., x_n))|\, x_1, ..., x_n\right] = \min_{a\in\Omega}\{E\left[L(\theta, a)|\, x_1, ..., x_n\right]\}$$

♣

To summarize the definition above:

1. a is some number/value $\in \Omega$
2. $L(\theta, a)$ is the loss for estimating θ with a
3. $E[L(\theta, a)]$ is the expected loss over the distribution for θ
4. $\delta^*(x_1, ..., x_n)$ is the estimate (i.e. specific value) for θ
5. $E\left[L(\theta, \delta^*(x_1, ..., x_n))|\, x_1, ..., x_n\right]$ is the expected loss if $a = \delta^*(x_1, ..., x_n)$ for the posterior distribution

When the expected posterior loss is minimized, the resulting estimate that did so is called the **Bayes' Estimate**. If we can find a rule $\delta^*(X_1, ..., X_n)$ that can calculate an estimate for each sample we might have, then we get a **Bayes' Estimator**.

2.2.3 Examples

We now give some examples using the method Bayesian estimation with respect to minimizing the expected loss.

Example 2.2.1 (Bayes' Estimate for Bernoulli)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim}$ Bernoulli(θ) and*

$$\theta \sim Beta(\alpha, \beta)$$
$$L(\theta, a) = (\theta - a)^2$$

i.e. for a specific example we will have $\delta^*(x_1, ..., x_n)$ as our estimate.

Q: What is the form of this estimate?

If we set $y = \sum_i x_i$, then by our previous calculations (see Beta is a Conjugate Prior), we know

$$(\theta | x_1, ..., x_n) \sim Beta(\underbrace{\alpha + y}_{\alpha_1}, \underbrace{\beta + n - y}_{\beta_1})$$

For a squared error loss, any fixed sample $x_1, ..., x_n$ gives the minimum expected square loss as the mean of the posterior distribution[3]*. In other words,*

$$E(\theta | x_1, ..., x_n) = \frac{\alpha_1}{\alpha_1 + \beta_1} = \frac{\alpha + y}{\alpha + \beta + n}$$

is $\delta^*(x_1, ..., x_n)$. The Bayes' estimator is then given by

$$\delta^*(X_1, ..., X_n) = \frac{\alpha + \sum_{i=1}^n X_i}{\alpha + \beta + n}$$

i.e. the value to calculate is a modified sample proportion. Note the similarity to the MLE estimate for this type of data which is $\hat{\theta}_{MLE} = \sum_i X_i / n$ which too is a sample proportion.

♥

Example 2.2.2 (Bayes' Estimate for Normal)**.** *Suppose that $X_1, ..., X_n | \theta \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $\theta \sim N(\mu_0, \nu_0^2)$ where $\theta = \mu =$ unknown and $\sigma^2 =$ known. Then, since θ and $X_1, ..., X_n$ are normally distributed, the posterior distribution $\theta | x_1, ..., x_n$ is also normally distributed with*

$$\mu_1 = \frac{\mu_0 \sigma^2 + n \nu_0^2 \bar{x}}{n \nu_0^2 + \sigma^2}$$
$$\nu_1^2 = \frac{\sigma^2 \nu_0^2}{n \nu_0^2 + \sigma^2}$$

this makes the Bayes' estimate $\delta^*(x_1, ..., x_n) = \dfrac{\mu_0 \sigma^2 + n \nu_0^2 \bar{x}}{n \nu_0^2 + \sigma^2}$ as it is the mean of the posterior distribution.

♥

2.2.4 Absolute Error Loss

If we choose the absolute error loss where $L(\theta, a) = |\theta - a|$, we let $\delta^*(x_1, ..., x_n)$ is an estimate that minimizes this loss. The quantity that minimizes the absolute error loss is the median

[3] See the Appendix for a proof

of the posterior distribution[4]. We define the median as the point $\tilde{\theta}$ such that

$$P(X \geq \tilde{\theta}) = P(X \leq \tilde{\theta}) = \frac{1}{2}$$

for a random sample X. It is typically complicated to calculate and thus hard to find a closed form expression for. However, for a symmetric distribution (like the Normal Distribution) we know median = mean. This means for a normal data with a normal prior and normal posterior the Bayes' estimator for the squared error loss is the same value as the absolute error loss.

Note: Both the choice of the loss function and prior has an effect on the Bayes' estimator. However, with increasing sample size, the effect of the prior distribution choice diminishes. ∞

2.2.5 Consistency

Recall that an estimator $\hat{\theta}_n$ is called consistent if $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$. We know that under fairly general conditions, the Bayes' estimator is consistent. This means

$$\delta^*(X_1, ..., X_n) \xrightarrow{P} \theta$$

for large n. This is true for a wide class of loss functions, but does depend on the loss function. We give some examples to show this.

Example 2.2.3 (Bernoulli Distribution)**.** *We know for a Bernoulli distribution*

$$\begin{aligned}\delta^*(X_1, ..., X_n) &= \frac{\alpha + \sum X_i}{\alpha + \beta + n} \\ &= \frac{\alpha}{\alpha + \beta + n} + \frac{\sum X_i}{\alpha + \beta + n}\end{aligned}$$

Notice for $n \to \infty$:

$$\frac{\alpha}{\alpha + \beta + n} \to 0 \quad \text{and} \quad \frac{\sum X_i}{\alpha + \beta + n} \to \bar{X}$$

and since $\bar{X} \xrightarrow{P} \theta$ eventually, we have $\delta^*(X_1, ..., X_n) \xrightarrow{P} \theta$.

♥

Example 2.2.4 (Normal Distribution)**.** *We have a similar process, recall*

$$\delta^*(X_1, ..., X_n) = \frac{\sigma^2 \mu_0}{\sigma^2 + n\nu_0^2} + \frac{n\nu_0 \bar{X}}{\sigma^2 + n\nu_0}$$

[4] See Appendix for a proof

As $n \to \infty$

$$\frac{\sigma^2\mu_0}{\sigma^2 + n\nu_0^2} \to 0 \quad \textit{and} \quad \frac{n\nu_0\bar{X}}{\sigma^2 + n\nu_0} \to \bar{X}$$

Since $\bar{X} \xrightarrow{P} \mu$ eventually, we have $\delta^*(X_1, \ldots, X_n) \xrightarrow{P} \mu$.

♥

Note: Bayes estimation can also be used for vector valued parameters, i.e. $\theta = (\theta_1, \ldots, \theta_n)$. In this case, we need a **joint prior** $\xi(\theta)$ and computations can be complex. In addition, we might also be interested in estimating some image of a function of a parameter $\psi = h(\theta)$ rather than the parameter itself, these are Bayes' estimators for a function. ∞

2.2.6 MLE's and Bayes' Estimation

In practice, we find that $\hat{\theta}_{MLE}$ and $\delta^*(X_1, \ldots, X_n)$ are often both similar for large samples and are based on functions of likelihood.

2.3 Appendix (Bayesian Inference)

2.3.1 Derivation of Posterior for Normal Prior

In this section we derive the form of the posterior of a normal prior as it is given in the example. Recall that $X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $\theta = \mu =$ unknown and $\sigma^2 =$ known. We choose $\theta \sim N(\mu_0, \nu_0^2)$ as the prior and need to find the form of the posterior. We make the observation that any constants that make up $f(x_1, \ldots, x_n|\theta)\xi(\theta)$ will cancel, so we only need to see what the posterior is proportional to to see what the posterior could look like. First,

$$f(x_1, \ldots, x_n|\theta) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}$$

while

$$\xi(\theta) \propto \exp\left\{-\frac{1}{2\nu_0^2}(\theta-\mu_0)^2\right\}$$

hence,

$$\xi(\theta|x_1, \ldots, x_n) \propto \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\nu_0^2}\right]\right\}$$

Notice, that with the addition and subtraction of $\bar{x}$, we can write $\sum(x_i-\theta)^2 = n(\theta-\bar{x})^2 + \sum(x_i-\bar{x})^2$. This then makes

$$\begin{aligned}\xi(\theta|x_1, \ldots, x_n) &\propto \exp\left\{-\frac{1}{2}\left[\frac{n(\theta-\bar{x})^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\nu_0^2}\right]\right\} \times \exp\left\{-\frac{1}{2}\frac{\sum(x_i-\bar{x})^2}{\sigma^2}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{n(\theta-\bar{x})^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\nu_0^2}\right]\right\} \qquad \text{(right side does not depend on } \theta)\end{aligned}$$

So, all the information we need to find the form of the posterior is given in

$$\exp\left\{-\frac{1}{2}\left[\frac{n(\theta-\bar{x})^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\nu_0^2}\right]\right\}$$

To do this, we will collect the terms (after squaring) and complete the square. We focus only on the terms inside the exponent now for easier math.

$$\begin{aligned}
-\frac{1}{2}\left[\frac{n(\theta-\bar{x})^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\nu_0^2}\right] &= -\frac{1}{2}\left[\frac{n}{\sigma^2}(\theta^2+\bar{x}^2-2\theta\bar{x})+\frac{\theta^2+\mu_0-2\theta\mu_0}{\nu_0^2}\right] \\
&= -\frac{1}{2}\left[\frac{n\nu_0^2}{\nu_0^2\sigma^2}(\theta^2+\bar{x}^2-2\theta\bar{x})+\frac{\sigma^2}{\sigma^2\nu_0^2}(\theta^2+\mu_0-2\theta\mu_0)\right] \\
&= -\frac{1}{2}\left[\frac{1}{\nu_0^2\sigma^2}\Big(\theta^2(n\nu_0^2+\sigma^2)-2\theta(\mu_0\sigma^2+n\nu_0^2\bar{x})+(n\nu_0^2\bar{x}^2+\sigma^2\mu_0^2)\Big)\right] \\
&= -\frac{1}{2}\frac{n\nu_0^2+\sigma^2}{\sigma^2\nu_0^2}\left(\theta^2-2\theta\left[\frac{\mu_0\sigma^2+n\nu_0^2\bar{x}}{n\nu_0^2+\sigma^2}\right]+\frac{n\nu_0^2\bar{x}^2+\sigma^2\mu_0^2}{n\nu_0^2+\sigma^2}\right) \\
&\propto -\frac{1}{2}\frac{n\nu_0^2+\sigma^2}{\sigma^2\nu_0^2}\left(\theta-\left[\frac{\mu_0\sigma^2+n\nu_0^2\bar{x}}{n\nu_0^2+\sigma^2}\right]\right)^2
\end{aligned}$$

From this form we find

$$\begin{aligned}
\mu_1 &= \frac{\mu_0\sigma^2+n\nu_0^2\bar{x}}{n\nu_0^2+\sigma^2} \\
\nu_1^2 &= \frac{\sigma^2\nu_0^2}{n\nu_0^2+\sigma^2}
\end{aligned}$$

and futher that the posterior has the form of a normal distribution with mean μ_1 and variance ν_1^2. Hence, the posterior is a conjugate prior of the normal conjugate family.

2.3.2 Minimum Expected Square Error Loss

In this part we will show that the mean of the posterior distribution is the minimum of the expectation of $L(\theta, a) = (\theta - a)^2$.

Theorem 2.3.1 (Minimum Square Error Loss)**.** *The minimum of the expected square error loss $E[L(\theta, a)|x_1, ..., x_n]$ where $L(\theta, a) = (\theta - a)^2$ is the mean of the posterior distribution or $E[\theta|x_1, ..., x_n]$, assuming it is finite.*

Proof. We prove this directly by following the definition of expected values:

$$\begin{aligned}
E[L(\theta, a)|x_1, ..., x_n] &= \int_\Omega L(\theta, a)\xi(\theta|x_1, ..., x_n)d\theta \\
&= \int_\Omega (\theta-a)^2\xi(\theta|x_1, ..., x_n)d\theta
\end{aligned}$$

We want to minimize this function with respect to a. Since it is continuous, we take the derivative check for extrema. We arrive at

$$\begin{aligned}\frac{\partial E[L(\theta,a)|x_1,...,x_n]}{\partial a} &= \int_\Omega 2(a-\theta)\xi(\theta|x_1,...,x_n)d\theta = 0\\ &= a - \int_\Omega \theta\xi(\theta|x_1,...,x_n)d\theta = 0\\ &\implies a = E(\theta|x_1,...,x_n)\end{aligned}$$

since

$$\frac{\partial^2 E[L(\theta,a)|x_1,...,x_n]}{\partial a^2} = 2 > 0$$

we know that this is a minimum. Hence, $\min_a\{E[L(\theta,a)|x_1,...,x_n)]\} = E(\theta|x_1,...,x_n)$ when the loss is the square error loss, as we sought to show. Notice, the minimum expected loss is the variance of the posterior distribution or $V(\theta|x_1,...,x_n)$. A similar argument can be made for the discrete case. □

2.3.3 Minimum Expected Absolute Error Loss

In this part we will show that the mean of the posterior distribution is the minimum of the expectation of $L(\theta,a) = |\theta - a|$.

Theorem 2.3.2 (Minimum Absolute Error Loss)**.** *The minimum of the expected absolute error loss or $E[L(\theta,a)|x_1,...,x_n]$ where $L(\theta,a) = |\theta - a|$ is the median of the posterior distribution or median$(\theta|x_1,...,x_n) = \tilde{\theta}$.*

Proof. We work by definition:

$$E[L(\theta,a)|x_1,...,x_n] = \int_\Omega |\theta - a|\xi(\theta|x_1,...,x_n)d\theta$$

Since we are optimizing with respect to a we can see

$$\frac{\partial E[L(\theta,a)|x_1,...,x_n]}{\partial a} = \int_{\Omega'} \xi(\theta|x_1,...,x_n)d\theta - \int_{\Omega''} \xi(\theta|x_1,...,x_n)d\theta = 0$$

under the condition $\Omega' \cup \Omega'' = \Omega$ since

$$\frac{\partial|\theta - a|}{\partial a} = \begin{cases} 1 & a \in (\theta,\infty) = \Omega' \\ -1 & a \in (-\infty,\theta) = \Omega'' \end{cases}$$

Notice this constraint implies that

$$\int_{\Omega'} \xi(\theta|x_1, ..., x_n) d\theta = \int_{\Omega''} \xi(\theta|x_1, ..., x_n) d\theta$$

and since $\int_{\Omega'} \xi(\theta|x_1, ..., x_n) d\theta \in [0, 1]$ we must have

$$\int_{\Omega'} \xi(\theta|x_1, ..., x_n) d\theta = \int_{\Omega''} \xi(\theta|x_1, ..., x_n) d\theta = \frac{1}{2}$$

This can only happen when $a = \tilde{\theta} = \theta_{\text{median}}$. Also, when we choose $a < \tilde{\theta}$, $\frac{\partial E[L(\theta, a)|x_1, ..., x_n]}{\partial a} < 0$ and when we choose $a > \tilde{\theta}$, $\frac{\partial E[L(\theta, a)|x_1, ..., x_n]}{\partial a} > 0$, so indeed $a = \tilde{\theta}$ is a minimum of the expected absolute loss which is what we sought to show. A similar argument can be made for the discrete case. □

Chapter 3 — Sufficiency

3.1 Sufficient Statistics

Sometimes a Bayes' Estimate/MLE are not available, may not exist, or are hard to calculate. In this case, we need to find other options. One possibility is using what is called a **sufficient statistic**. Informally, a sufficient statistic $T = r(X_1, ..., X_n)$ is a summary of the data that is not reliant on the actual data values $X_1, ..., X_n$ themselves.[1] Notice when we know a population parameter's true value, the joint distribution depends only on the data we collect. The same idea occurs with a sufficient statistic, so it is a good candidate for that population parameter's point estimate.

Example 3.1.1 (Normal Distribution)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ *where* μ *is unknown. We wish to estimate* μ *as we have done before but cannot use the MLE nor the Bayes' estimator to do so. Using the method of sufficient statistics, we can show that to estimate* μ *it is enough to know*

$$S = \sum_i^n X_i$$

rather than the data values $X_1, ..., X_n$*. Such a statistic* S*, is called sufficient.*

♥

We now give some motivating questions to consider when considering how to calculate sufficient statistics.

Q: How can we find (compute) sufficient statistics?

Q: What exactly do they mean?

To help answer these questions, we give some more information about sufficient statistics below.

[1] we can simulate the data closely with this single statistic

Sufficient Statistics Facts

1. If $T(X_1, ..., X_n)$ is sufficient, then all we need to know about the data $X_1, ..., X_n$ is contained in T. Therefore, there is no information about θ in the data conditional (given) T.
2. A statistic is sufficient for the parameter we are interested in, to find *joint sufficiency*, we need to take both parameters into account. In other words, $T(X_1, ..., X_n)$ may be sufficient for θ_1, but not for θ_2 in a two parameter problem.
3. If $T = r(X_1, ..., X_n)$ is sufficient, then

   $$f(x_1, ..., x_n | t = r(x_1, ..., x_n)) = \frac{f(x_1, ..., x_n)}{g(t)}$$

   does not depend on θ. Further, if $X_1', ..., X_n'$ are simulated from a distribution such that for all θ, the distribution of $X_1', ..., X_n'$ is the about the same as that of $X_1, ..., X_n$, then T is sufficient.

☕

3.1.1 Formal Definition of Sufficiency

Formally, we define what a sufficient statistic is below.

Definition 3.1.1 (Sufficient Statistic). *Suppose $X_1, ..., X_n \overset{iid}{\sim} P_\theta$ for some θ. We then define a statistic $T = r(X_1, ..., X_n)$ such that given $T = t$, the distribution of $x_1, ..., x_n$ given $T = t$ does not depend on θ. Such a statistic $T = r(X_1, ..., X_n)$ is a sufficient statistic for θ. Notice, the symbolic form of this definition is*

$$\begin{aligned} f(x_1, ..., x_n | t = r(x_1, ..., x_n)) &= \frac{h(t = r(x_1, ..., x_n) | x_1, ..., x_n) f(x_1, ..., x_n)}{g(t)} \\ &= \frac{f(x_1, ..., x_n)}{g(t)} \end{aligned}$$

for the continuous case. For the last line in the equation, we used a Dirac measure for the event $E = (t = r(x_1, ..., x_n) | x_1, ..., x_n)$, so $\delta_{\mathbb{R}}(E) = 1$ since the event is always true.

♣

With this definition, we can learn how to find a sufficient statistic from the data we are given. It is through a method known as the **factorization criterion** that allows us to identify sufficient statistics.

Theorem 3.1.1 (Likelihood Factorization Criterion)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} P_\theta$ *for some* $\theta \in \Omega$. *Then, a statistic* $T = r(X_1, ..., X_n)$ *is sufficient for* θ **if an only if** *the following holds.*

$$\mathcal{L}(\theta|x_1, ..., x_n) = f(x_1, ..., x_n|\theta) = u(x_1, ..., x_n) \times v(r(x_1, ..., x_n), \theta)$$

Notice, u *can depend on* $X_1, ..., X_n$, *but not* θ *and* v *can only depend on* $X_1, ..., X_n$ *through* $r(x_1, ..., x_n)$.

Proof. We will prove this only for the discrete data only; the proof for continuous data is left for more advanced statistics courses. In a discrete setting,

$$f(x_1, ..., x_n|\theta) = P(X_1 = x_1, ..., X_n = x_n|\theta)$$

We will proceed in the proof by proving each side of the biconditional seperately.

1. ($\Longrightarrow$) Suppose f can be factored for any data we observe and any true value of the parameter θ, that is

   $$\forall x_1, ..., x_n \in \mathbb{R}^n \qquad \text{and} \qquad \theta \in \Omega$$

   we have

   $$f(x_1, ..., x_n|\theta) = u(x_1, ..., x_n) \times v(r(x_1, ..., x_n), \theta)$$

   then,

   $$\begin{aligned} P(X_1 = x_1, ..., X_n = x_n|T = t, \theta) &= \frac{P(X_1 = x_1, ..., X_n = x_n|\theta)}{P(T = t|\theta)} \quad \text{(Bayes' Thm.)} \\ &= \frac{P(X_1 = x_1, ..., X_n = x_n|\theta)}{\sum\limits_{y_1,...,y_n \in A(t)} P(X_1 = y_1, ..., X_n = y_n|\theta)} \end{aligned}$$

   where

   $$\begin{aligned} A(t) &= \{y_1, ..., y_n : r(y_1, ..., y_n) = t\} \\ &= \text{set of all samples where } T = t \end{aligned}$$

   Since $r(y_1, ..., y_n) = t$ for all $y_1, ..., y_n \in A(t)$ including $r(x_1, ..., x_n)$, we have (using $f(x_1, ..., x_n|\theta) = u(x_1, ..., x_n) \times v(r(x_1, ..., x_n), \theta)$)

   $$P(X_1 = x_1, ..., X_n = x_n|T = t, \theta) = \frac{u(x_1, ..., x_n)}{\sum\limits_{y_1,...,y_n \in A(t)} u(y_1, ..., y_n)} = \text{constant}$$

   where

   $$u(\alpha_1, ..., \alpha_n|\theta) = \text{function purely of } \alpha's$$

Notice if $x_1, ..., x_n$ is observed such that $x_1, ..., x_n \notin A(t)$, then

$$P(X_1 = x_1, ..., X_n = x_n | T = t, \theta) = 0 = \text{constant}$$

and does not depend on θ. For these samples, then, T is a sufficient statistic too.

2. ($\Longleftarrow$) Suppose $T = r(x_1, ..., x_n)$ is sufficient, then by definition

$$P(X_1 = x_1, ..., X_n = x_n | T = t, \theta) = u(x_1, ..., x_n)$$

Further, in a similar fashion as with the forward direction of this proof,

$$P(X_1 = x_1, ..., X_n = x_n | \theta) = P(X_1 = x_1, ..., X_n = x_n | T = t, \theta) P(T = t | \theta)$$

From here, we note again $u(x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n | T = t, \theta)$ by definition and set $v(r(x_1, ..., x_n), \theta) = P(T = t | \theta)$ to arrive at

$$P(X_1 = x_1, ..., X_n = x_n | \theta) = u(x_1, ..., x_n) \times v(r(x_1, ..., x_n), \theta)$$

which shows that a sufficient statistic can factorize the joint conditional distribution.

We have proven both sides of the biconditional; this concludes the proof. □

Note:

1. $T = r(x_1, ..., x_n)$ is sufficient if and only if the likelihood $\mathcal{L}(\theta | x_1, ..., x_n) = f(x_1, ..., x_n | \theta) = u(x_1, ..., x_n)$ (as a function of θ) is proportional to a function that depends on the data only through $r(x_1, ..., x_n)$ or

$$\mathcal{L}(\theta | x_1, ..., x_n) \propto \underbrace{v(r(x_1, ..., x_n), \theta)}_{\text{depends only on } r(x_1, ..., x_n)} \times u(x_1, ..., x_n)$$

2. When using the likelihood for finding a posterior distribution, any factor not depending on θ can be removed from the likelihood without affecting the calculation of the posterior.

∞

Point 2 of the note above gives a corollary.

Corollary 3.1.1 (Bayesian Sufficiency)**.** *$T = r(x_1, ..., x_n)$ is a sufficient statistic if and only if the posterior depends on the data only through T for any prior we might choose, in other words $\xi(\theta | x_1, ..., x_n) = \xi(\theta | r(x_1, ..., x_n))$.*

Proof. We prove this for each side of the biconditional.

1. ($\Longrightarrow$) Suppose the posterior depends on the data only by $T = r(x_1, ..., x_n)$. Then,

$$\xi(\theta | r(x_1, ..., x_n)) = \frac{h(r(x_1, ..., x_n), \theta) \times \xi(\theta)}{\displaystyle\int_\Omega h(r(x_1, ..., x_n), \theta) \times \xi(\theta) d\theta}$$

   if we assume the likelihood is factorable, then

$$\begin{aligned}\xi(\theta | r(x_1, ..., x_n)) &= \frac{u(x_1, ..., x_n) h(r(x_1, ..., x_n), \theta) \times \xi(\theta)}{\displaystyle\int_\Omega u(x_1, ..., x_n) h(r(x_1, ..., x_n), \theta) \times \xi(\theta) d\theta} \\ &= \frac{\mathcal{L}(\theta | x_1, ..., x_n) \xi(\theta)}{\displaystyle\int_\Omega \mathcal{L}(\theta | x_1, ..., x_n) \xi(\theta) d\theta} \\ &= \xi(\theta | x_1, ..., x_n)\end{aligned}$$

   and by Likelihood Fact. Crit. we know T is a sufficient statistic.

2. ($\Longleftarrow$) Suppose T is a sufficient statistic, then by the Likelihood Factorization Criterion, $\mathcal{L}(\theta | x_1, ..., x_n) = u(x_1, ..., x_n) v(r(x_1, ..., x_n), \theta)$. When computing the posterior, we have

$$\begin{aligned}\xi(\theta | x_1, ..., x_n) &= \frac{\mathcal{L}(\theta | x_1, ..., x_n) \xi(\theta)}{\displaystyle\int_\Omega \mathcal{L}(\theta | x_1, ..., x_n) \xi(\theta) d\theta} \\ &= \frac{u(x_1, ..., x_n) v(r(x_1, ..., x_n), \theta) \xi(\theta)}{\displaystyle\int_\Omega u(x_1, ..., x_n) v(r(x_1, ..., x_n), \theta) \xi(\theta) d\theta} \\ &= \frac{v(r(x_1, ..., x_n), \theta) \xi(\theta)}{\displaystyle\int_\Omega v(r(x_1, ..., x_n), \theta) \xi(\theta) d\theta} \\ &= \xi(\theta | r(x_1, ..., x_n))\end{aligned}$$

   Hence, the posterior only depends on the sufficient statistic and is a function of θ.

Since both sides of the biconditional are proven, this concludes the proof. □

To find a sufficient statistic, we use the following steps:

Steps to Find a Sufficient Statistic

1. Find the likelihood function of the data
2. Factorize the likelihood function into $u(x_1, ..., x_n) \times v(r(x_1, ..., x_n), \theta)$
3. The function that depends only on $r(x_1, ..., x_n)$ is the sufficient statistic by factorization criterion

3.1.2 Examples

We now give some examples using the idea of sufficient statistics.

Example 3.1.2 (Poisson Sufficient Statistic)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim}$ *Poisson*(θ) *where* $\theta > 0$ *only. We want to find the sufficient statistic for* θ*. We note that the p.d.f for each* X_i *is*

$$f(x_i|\theta) = \frac{\theta^{x_i} e^{-\theta}}{x_i!}$$

then, the joint p.d.f is

$$\begin{aligned} f(x_1, ..., x_n|\theta) &= \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} \\ &= \underbrace{\prod_{i=1}^{n} \frac{1}{x_i!}}_{u(x_1,...,x_n)} \times \underbrace{\exp\left(-\sum_{i=1}^{n} x_i\right) \exp(-n\theta)}_{v(r(x_1,...,x_n),\theta)} \end{aligned}$$

by the likelihood factorization theorem, we know that $T = r(X_1, ..., X_n) = \sum_{i=1}^{n} X_i$ *is sufficient for* θ*.*

♥

Example 3.1.3 (Continuous Data)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} f$ *where the p.d.f is*

$$f(x_i|\theta) = \begin{cases} \theta x^{\theta-1} & 0 < x < 1 \\ 0 & o.w. \end{cases}$$

The likelihood is then

$$\begin{aligned} f(x_1, ..., x_n|\theta) &= \prod_{i=1}^{n} \theta x^{\theta-1} \\ &= \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1} \\ &= \underbrace{1}_{u(x_1,...,x_n)} \times \underbrace{\theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1}}_{v(r(x_1,...,x_n),\theta)} \end{aligned}$$

notice that $u(x_1, ..., x_n)$ *does not have to depend solely on the data; it can be a constant as well. Hence, the statistic sufficient for* θ *is* $T = r(X_1, ..., X_n) = \prod_{i=1}^{n} X_i$.

♥

Example 3.1.4 (Normal Sufficient Statistic)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where μ is unknown and σ^2 is known. Then, the joint p.d.f can be written as*

$$\begin{aligned} f(x_1, ..., x_n|\theta) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2}\frac{\sum(x_i-\mu)^2}{\sigma^2}\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum x_i^2\right\} \exp\left\{\frac{\mu}{\sigma^2}\sum x_i - \frac{n\mu^2}{2\sigma^2}\right\} \end{aligned}$$

then the first term is $u(x_1, ..., x_n)$ since it does not depend on μ and the second term is $v(r(x_1, ..., x_n), \theta)$ and depends on the data only through

$$T = r(X_1, ..., X_n) = \sum X_i$$

Hence, $\sum X_i$ is sufficient for θ by the likelihood factorization criterion.

♥

Example 3.1.5 (Uniform Distribution)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} U[0, \theta]$. We know the individual p.d.f is*

$$f(x_i|\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & o.w. \end{cases}$$

Then, the joint p.d.f can be written as

$$f(x_1, ..., x_n|\theta) = \begin{cases} 1/\theta^n & 0 \le x_i \le \theta \ \forall i \\ 0 & o.w. \end{cases}$$

Notice if $0 \le x_i \le \theta$ then we know $\max\{x_1, ..., x_n\}$ is in this interval as well. So

$$0 \le x_i \le \theta \implies \max\{x_1, ..., x_n\} \le \theta$$

We can then write the p.d.f as

$$f(x_1, ..., x_n|\theta) = \begin{cases} 1/\theta^n & \max\{x_1, ..., x_n\} \le \theta \\ 0 & \max\{x_1, ..., x_n\} \not\le \theta \end{cases}$$

this suggests that $T = r(X_1, ..., X_n) = \max\{X_1, ..., X_n\}$ is the sufficient statistic and

$$v(r(x_1, ..., x_n), \theta) = \begin{cases} 1/\theta^n & t \le \theta \\ 0 & t \not\le \theta \end{cases}$$

and $u(x_1, ..., x_n) = 1$ *leaving* $f(x_1, ..., x_n|\theta) = u(x_1, ..., x_n)v(r(x_1, ..., x_n), \theta) = v(r(x_1, ..., x_n), \theta)$. *So, for a uniform distribution* $T = r(x_1, ..., x_n) = \max\{X_1, ..., X_n\}$ *is the sufficient statistic.*

♥

3.1.3 Jointly Sufficient Statistics

We now turn our attention to the case where two or more parameters are unknown. How would calculate the jointly sufficient statistics then? To get started, suppose $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. We are looking for statistics T_1 and T_2 such that

$$T_1 = r_1(X_1, ..., X_n)$$
$$T_2 = r_2(X_1, ..., X_n)$$

If $f(x_1, ..., x_n|T_1, T_2)$ does not depend on $(\mu, \sigma^2) = \theta$, then we say T_1 and T_2 are jointly sufficient for $\theta = (\mu, \sigma^2)$.

More generally, if $X_1, ..., X_n \overset{iid}{\sim} P_\theta$ where $\theta \in \Omega \subset \mathbb{R}^k$ and $T = (T_1, ..., T_n)$ are statistics for θ. Then we say that they are jointly sufficient statistics for θ when $f(x_1, ..., x_n|T_1, ..., T_n)$ does not depend on $\theta = (\theta_1, ..., \theta_n)$.

To find a sufficient statistic, we use what is known as the **Multivariate Likelihood Factorization Theorem**.

Theorem 3.1.2 (Multivariate Likelihood Factorization Theorem (MLFT))**.** *Suppose* $r_1, ..., r_k$ *are functions of n real variables. In other words,*

$$T_i = r_i(X_1, ..., X_n) \qquad \forall i = 1, ..., k$$

These statistics are jointly sufficient statistics for θ *if and only if*

$$f(x_1, ..., x_n|\theta) = u(x_1, ..., x_n) \times v(r_1, ..., r_k, \theta)$$

as before $u(x_1, ..., x_n)$ *can depend on the data* $x_1, ..., x_n$ *but not on the parameters* $\theta = (\theta_1, ..., \theta_k)$ *and* $v(r_1, ..., r_k, \theta)$ *only depends on* $r_1, ..., r_k$ *and* θ.

Proof. We prove the biconditonal separately (only for discrete case).

1. ($\Longrightarrow$) [Contrapositive Method] Suppose $T = (T_1, ..., T_k)$ are not sufficient statistics. Then we know $P(X_1 = x_1, ..., X_n = x_n|T, \theta) \neq u(x_1, ..., x_n)$ and if we set $P(T = t|\theta) =$

$v(r_1, ..., r_k, \theta)$ we see that

$$\begin{aligned} P(X_1 = x_1, ..., X_n = x_n|\theta) &= P(X_1 = x_1, ..., X_n = x_n|T, \theta)P(T = t|\theta) \\ &\neq u(x_1, ..., x_n) \times v(r_1, ..., r_k, \theta) \end{aligned}$$

is not possible as no joint sufficient statistics exist. Hence, the factorization we are after does not exist.

2. ($\Longleftarrow$) Suppose $T = (T_1, ..., T_k)$ are sufficient statistics. Then we know $P(X_1 = x_1, ..., X_n = x_n|T, \theta) = u(x_1, ..., x_n)$ and if we set $P(T = t|\theta) = v(r_1, ..., r_k, \theta)$ we see that

$$\begin{aligned} P(X_1 = x_1, ..., X_n = x_n|\theta) &= P(X_1 = x_1, ..., X_n = x_n|T, \theta)P(T = t|\theta) \\ &= u(x_1, ..., x_n) \times v(r_1, ..., r_k, \theta) \end{aligned}$$

which is precisely the factorization criterion we are after.

We have just proven both sides of the implication, we have proven the bi-implication. This concludes the proof. □

We now give some examples illustrating this method.

Example 3.1.6 (Normal Multi. Sufficient Statistics)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ *where both parameters are unknown. The joint distribution is*

$$\begin{aligned} f(x_1, ..., x_n|\mu, \sigma^2) &= \prod_{x=1x}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i + \mu^2\right)\right) \\ &= v(r_1, r_2, \mu, \sigma^2) \end{aligned}$$

This means for the factorization, we have $u(x_1, ..., x_n) = 1$ *and the sufficient statistics are*

$$\begin{aligned} T_1 &= \sum X_i \implies \hat{\mu} = \bar{x} = \frac{1}{n}T_1 \\ T_2 &= \sum X_i^2 \implies \hat{\sigma}^2 = \frac{1}{n}T_2 - \left(\frac{1}{n}T_1\right)^2 = s(T_1, T_2) \end{aligned}$$

The basic sufficient statistics are T_1, T_2 *but we can use functions of them to estimate the unknown parameters* μ, σ^2.

♥

Example 3.1.7 (Uniform Multi. Sufficient Statistics)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} U[a, b]$ *where*

$a < b$. We know the p.d.f for each variate is

$$f(x|a,b) = \frac{1}{b-a} \qquad a \leq x \leq b$$

This makes the joint p.d.f as

$$f(x_1,...,x_n|a,b) = \left(\frac{1}{b-a}\right)^n \qquad a \leq x_i \leq b$$

Notice, if $x_i \geq a$ for each observation, then $\min\{X_1,...,X_n\} \geq a$ too. Also if $x_i \leq b$, then $\max\{X_1,...,X_n\} \leq b$ too. This lets us reformulate the likelihood as

$$f(x_1,...,x_n|a,b) = \left(\frac{1}{b-a}\right)^n [\min\{X_1,...,X_n\} \geq a][\max\{X_1,...,X_n\} \leq b]$$

where $[\cdot]$ denotes the **Iverson Bracket** *which functions as an indicator variable (1 if condition in the bracket is true, 0 otherwise). From this, we can see that*

$$f(x_1,...,x_n|a,b) = v(r_1(x_1,...,x_n), r_2(x_1,...,x_n), a, b)$$

and $u(x_1,...,x_n) = 1$ which implies that the sufficient statistics are

$$T_1 = \min\{X_1,...,X_n\}$$
$$T_2 = \max\{X_1,...,X_n\}$$

are jointly sufficient for $\theta = (a,b)$

♥

3.1.4 Minimal Sufficient Statistics

There are often several sufficient statistics for a parameter θ. These are usually functions of the original sufficient statistics we find using the factorization theorem. We know that $X_1,...,X_n$ is a sufficient statistic for any parameter if

$$X_1,...,X_n \overset{iid}{\sim} P_\theta$$

However, these jointly sufficient statistics are not useful by themselves. We would like a summary form that is sufficient for θ, meaning they closely approximate the parameters in θ. In other words, for the normal example given above, we know that $X_1,...,X_n \overset{iid}{\sim} \mathcal{N}(\mu,\sigma^2)$ are jointly sufficient for μ, σ^2. However, we also found out that

$$T_1 = \sum X_i, \; T_2 = \sum X_i^2$$

are also jointly sufficient for μ, σ^2. **Which statistics do you think are better preferred, the data or the sums?** We note that $X_1, ..., X_n$ requires us to know n values but T_1, T_2 only requires us to know 2. We can show another example were we can have n sufficient statistics where there might be less that are needed. One such example is with *order statistics* which we state as a theorem.

Theorem 3.1.3 (Order Statistics Sufficiency)**.** *Order Statistics $Y_1, ..., Y_n$ from sample $X_1, ..., X_n$ are jointly sufficient for θ.*

Proof. Since $Y_1, ..., Y_n$ are order statistics, we know $y_1 \leq y_2 \leq \cdots \leq y_n$. The likelihood for the data is

$$f(x_1, ..., x_n|\theta) = \prod_{x=1}^{n} f(x_i|\theta)$$

and since the order of the data is ignored for this likelihood, we could also write

$$f(x_1, ..., x_n|\theta) = \prod_{x=1}^{n} f(y_i|\theta)$$

since all possible permutations of the sample are equally likely. Hence, $f(x_1, ..., x_n|\theta)$ only depends on the order statistics $y_1, ..., y_n$ and by the factorization criterion, we know $Y_1, ..., Y_n$ are joint sufficient statistics for θ. □

Remark 3.1.1. *Intuitively, we can always determine y_j from the data $x_1, ..., x_n$, but we can never recover x_j from the ordered data $y_1, ..., y_n$ for any j.*

♦

Using these ideas, we can define what a minimal sufficient statistic is.

Definition 3.1.2 (Minimal Sufficient Statistic)**.** *If T is sufficient and a function of every other sufficient statistic, then T is* **minimal sufficient**. *Further, the vector $(T_1, ..., T_k)$ is minimal sufficient when its elements are functions of every other jointly sufficient set of statistics.*

♣

Remark 3.1.2. *Statistics are data summaries and minimal sufficient statistics are summaries of every other sufficient statistic.*

♦

3.1.5 Bayes' Estimation, MLE, and Sufficiency

We now relate Bayes' Estimation and the MLE to the concept of Sufficiency.

Theorem 3.1.4 (MLE & Sufficient Statistics)**.** *Suppose $T = r(X_1, ..., X_n)$ is sufficient for θ, then the MLE $\hat{\theta}_{MLE}$ depends on $X_1, ..., X_n$ only through $T = r(X_1, ..., X_n)$.*

Proof. We know by the likelihood factorization theorem that

$$f(x_1, ..., x_n|\theta) = u(x_1, ..., x_n) \times v(r(x_1, ..., x_n), \theta)$$

Further, $\hat{\theta}_{MLE}$ maximizes this likelihood and if we examine the log-likelihood, we get

$$\ell(\theta|x_1, ..., x_n) = \log u + \log v(T, \theta)$$

since u does not depend on θ, we can ignore it when we find the maximum. This leaves the form of the maximum solely dependent on T. Therefore, $\hat{\theta}_{MLE}$ must be a function of $T = r(X_1, ..., X_n)$. □

This fact extends to multidimensional θ as well.

Theorem 3.1.5 (Bayes' Est. & Sufficient Statistics)**.** *Suppose $T = r(X_1, ..., X_n)$ is sufficient for θ. Then, the Bayes' Estimator $\delta^*(X_1, ..., X_n)$ is a function of $r(X_1, ..., X_n)$.*

Proof. We note the form of the posterior is

$$\begin{aligned} \xi(\theta|x_1, ..., x_n) &\propto f(x_1, ..., x_n|\theta)\xi(\theta) = u(x_1, ..., x_n)v(r(x_1, ..., x_n), \theta)\xi(\theta) \\ &\propto v(r(x_1, ..., x_n), \theta)\xi(\theta) \end{aligned}$$

since $u(x_1, ..., x_n)$ does not depend on θ. Notice, that the form of the Bayes' Estimator depends on the form of the posterior distribution. We know the form of the posterior distribution depends (is conditional) on $r(x_1, ..., x_n)$. Hence, the Bayes' Estimator also depends on $r(X_1, ..., X_n)$ as well. □

Fact: Let $\theta = (\theta_1, ..., \theta_k)$ be a real valued vector of parameters. Then also let

$$\begin{aligned} \hat{\theta}_{MLE} &= (\hat{\theta}_1, ..., \hat{\theta}_k) \qquad \text{Maximum Likelihood Estimates} \\ \delta^* &= (\delta^*_1, ..., \delta^*_k) \qquad \text{Bayes' Estimates} \end{aligned}$$

Then both $\hat{\theta}_{MLE}$ and δ^* depend on $X_1, ..., X_n$ only through $T_1, ..., T_k$. Also, $\hat{\theta}_1, ..., \hat{\theta}_k$ and $\delta^*_1, ..., \delta^*_k$ are all minimal sufficient statistics since they are functions of every set of sufficient statistics. ∞

3.2 Improving an Estimator

Now that we have developed the idea of sufficiency, we can use it to gauge how well an estimator $\delta(X_1, ..., X_n)$ performs relative to the true value θ or $h(\theta)$ we are interested in. To do this, we need some measure of this performance. One such measure is known as the **MSE** or **Mean Squared Error**. Before we do, we also define what a generalized expectation (for a function of the data) is.

Definition 3.2.1 (Generalized Expectation)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} P_\theta$ and let $Z = g(X_1, ..., X_n)$ some function of the data $X_1, ..., X_n$. Then, we define the generalized expectation $E(Z)$ as*

$$E(Z) = \int \cdots \int g(x_1, ..., x_n) f(x_1, ..., x_n|\theta) dx_1 \dots dx_n$$

♣

Now we can define the MSE.

Definition 3.2.2 (MSE)**.** *Suppose we have $X_1, ..., X_n \overset{iid}{\sim} P_\theta$ and we are interested in estimating $h(\theta)$ with $\delta(X_1, ..., X_n)$. Then, we define the* **MSE** *about $h(\theta)$ $R(h(\theta), \delta)$ as*

$$\begin{aligned} MSE = R(h(\theta), \delta) &= E([(\text{estimate}) - (\text{true value})]^2) \\ &= E([\delta(X_1, ..., X_n) - h(\theta)]^2) \end{aligned}$$

♣

Does observing a sufficient statistic change the efficacy of an estimate? We will investigate this and to note the concept, we will call this quantity the **Rao-Blackwell Estimate**.

Definition 3.2.3 (Rao-Blackwell Estimator (RBE))**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} P_\theta$ and $\delta(X_1, ..., X_n)$ and estimator of $h(\theta)$. Then, the* **Rao-Blackwell Estimator (RBE)** *is given by*

$$\delta_0(T) = E(\delta(X_1, ..., X_n)|T)$$

notice that it's random with respect to T.

♣

We wish to compare δ_0 to the original estimate $\delta(X_1, ..., X_n)$. A theorem by Rao and Blackwell let's us do this.

Theorem 3.2.1 (Rao-Blackwell)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} P_\theta$, $\delta(X_1, ..., X_n)$ is an estimator for*

$h(\theta)$, and $T = r(X_1, ..., X_n)$ is a sufficient statistic for θ. Then,

$$R(h(\theta), \delta_0) \leq R(h(\theta), \delta)$$

and if $R(h(\theta), \delta) < \infty$, then there is strict inequality unless

$$\delta(X_1, ..., X_n) = \text{function of } T$$

Proof. We will only prove the case where $R(h(\theta), \delta) < \infty$ since $R(h(\theta), \delta) = \infty$ implies that there is nothing to prove ($R(h(\theta), \delta_0) \leq \infty$ anyway). We note that since $V(Y) = E(Y^2) - [E(Y)]^2 \geq 0$ we know $[E(Y)]^2 \leq E(Y^2)$ for any random variable Y. Because of this, we can see that

$$E[(\delta(X_1, ..., X_n) - h(\theta))]^2 \leq E[(\delta(X_1, ..., X_n) - h(\theta))^2]$$

If we condition on T, then we get we also see

$$E[(\delta(X_1, ..., X_n) - h(\theta)|T)]^2 \leq E[(\delta(X_1, ..., X_n) - h(\theta))^2|T]$$

and since $E[\delta(X_1, ..., X_n) - h(\theta)|T] = E(\delta(X_1, ..., X_n)|T) - h(\theta) = \delta_0 - h(\theta)$ we can write the equivalent form

$$[\delta_0 - h(\theta)]^2 \leq E[(\delta(X_1, ..., X_n) - h(\theta))^2|T]$$

From here, we take the expected value with respect to T to marginalize it out, yielding

$$[E(\delta_0 - h(\theta))^2] \leq E\{E[(\delta(X_1, ..., X_n) - h(\theta))^2|T]\} = E[(\delta(X_1, ..., X_n) - h(\theta))^2]$$

where the equality is given by the law of iterated expectation. We can rewrite the above as

$$R(h(\theta), \delta_0) \leq R(h(\theta), \delta)$$

by definition of MSE. This proves the Rao-Blackwell Theorem and concludes the proof. □

We can also define another type of error to see how well our estimator performs known as the **Mean Absolute Error** or **MAE**. We define it below

Definition 3.2.4 (Mean Abs. Error (MAE))**.** *The* **Mean Absolute Error** *or* **MAE** *is defined as a mean deviation $R(\theta, \delta)$ such that*

$$R(\theta, \delta) = E(|\delta(X_1, ..., X_n) - \theta|)$$

♣

To see if an estimator is working above all others in terms of error we use the concepts of inadmissibility and admissibility.

Definition 3.2.5 (Inadmissible Estimator)**.** *For their MSE or MAE, we have an estimator δ is* **inadmissible** *if there is another estimator δ_0 such that*

$$R(\theta, \delta_0) \leq R(\theta, \delta) \qquad \forall \theta \in \Omega$$

with a strict inequality for at least one $\theta \in \Omega$. In other words, δ_0 has a lower MSE or MAE for at least one θ and never has higher MSE or MAE.

We also say in this case that δ_0 **dominates** *δ.*

♣

Definition 3.2.6 (Admissible Estimator)**.** *We say an estimator δ is* **admissible** *when there is no other estimator δ_0 that dominates δ.*

♣

Remark 3.2.1. *Notice if $\delta_0 = E[\delta(X_1, ..., X_n)|T]$, then δ_0 will dominate δ by the Rao-Blackwell Theorem. Therefore, if we can conclude that an estimator δ is a function of the data not only through a sufficient statistic it is dominated by δ_0. Hence, such an estimate is inadmissible.*

♦

We now give an example illustrating this process.

Example 3.2.1 (Customer Arrivals)**.** *Suppose the number of customer arrivals within some amount of time is modeled by the Poisson distribution. If we set X_i to be the ith count of arrivals for the ith interval check, then we say $X_1, ..., X_n \overset{iid}{\sim} Poisson(\theta)$ where θ is the unknown parameter we are interested in estimating. θ measures the rate of occurrence of customers arriving. We wish to estimate the chance of exactly one customer arriving in some time interval. To build an estimator for this chance, we first define*

$$Y_i = \begin{cases} 1 & \text{if } X_i = 1 \\ 0 & o.w. \end{cases}$$

It is then known that $Y_i \overset{iid}{\sim} Bernoulli(p)$ where $p = P(X_i = 1) = \theta e^{-\theta} = h(\theta)$. We estimate p with $\hat{p} = \sum Y_i/n$. Hence, $\delta(X_1, ..., X_n) = \sum Y_i/n$. Now, for the Poisson Process $X_1, ..., X_n$

we know $T = \sum X_i$ is sufficient. So, for one observation, we have

$$\begin{aligned} E(Y_i|T=t) &= P(X_i = 1|T=t) && (Y_i = [X_i = 1]) \\ &= \frac{P(X_i = 1, T = t)}{P(T=t)} \\ &= \frac{P(X_i = 1)P\left(\sum_{j\neq i} X_j = t-1\right)}{P(T=t)} \end{aligned}$$

Since the sum of independent Poisson distributions are Poisson themselves, we have

$$P(T=t) = \frac{e^{n\theta}(n\theta)^t}{t!}$$

and also

$$P(X_1 = 1)P\left(\sum_{j\neq i} X_j = t-1\right) = \theta e^{-\theta} \times \frac{e^{-[n-1]\theta}\left([n-1]\theta\right)^{t-1}}{(t-1)!}$$

taking the ratio of these two quantities gives

$$E(Y_i|T=t) = \frac{t}{n}\left(1-\frac{1}{n}\right)^{t-1}$$

notice that if $t = 0$, we know $E(Y_i|T=0) = 0$. The Rao-Blackwell estimator δ_0 is then

$$\begin{aligned} \delta_0(t) &= E(\delta(X_1, ..., X_n)|T=t) \\ &= E\left[\frac{1}{n}\sum Y_i \middle| T = t\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[Y_i|T=t\right] \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}\frac{t}{n}\left(1-\frac{1}{n}\right)^{t-1}\right] \\ &= \frac{t}{n}\left(1-\frac{1}{n}\right)^{t-1} \end{aligned}$$

In practice, this estimator for p performs better than $\delta(X_1, ..., X_n)$ since it has a lower MSE. Notice that $\delta(X_1, ..., X_n)$ is inadmissible.

♥

Chapter 4 — Sampling Distributions of Estimators

In this chapter we discuss how estimators (such as the sample mean) can have distributions and their various properties. Recall that if we have $X_1, ..., X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$, we can calculate some statistic $T = r(X_1, ..., X_n)$ that estimates θ or $T \in \Omega$ to be more precise. All estimators are random in nature, so they too have a probability distribution and we define them as follows.

Definition 4.0.1 (Sampling Distribution)**.** *Under any valid statistical model P_θ with statistic T, we call $f_T(t|\theta)$ the* **sampling distribution** *of T given θ.*

♣

Note: We could also calculate a quantity called $T(X_1, ..., X_n, \theta)$ and this would not be a statistic for an unknown θ, but as a function of $X_1, ..., X_n$ it has a probability distribution induced by the distribution of $X_1, ..., X_n$. ∞

One popular statistic for the probability distribution of T is $E_\theta(T)$ which is the mean of T.

Example 4.0.1 (Bernoulli Distribution)**.** *Suppose $X_1, ..., X_{40} \overset{iid}{\sim}$ Bernoulli(θ). Then,*

$$\hat{\theta} = \frac{\sum_{i=1}^{40} X_i}{40} = T$$

then, $40T \sim Bin(40, \theta)$ where $t \in \{0, 1/40, 2/40, ..., 1\}$.

♥

Example 4.0.2 (MLE of μ)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ and we estimate the mean with the MLE or $\hat{\mu} = \bar{x}$. We can show that*

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Note: *this is not by the CLT, this is an exact distribution.*

♥

4.1 Need for Sampling Distributions

In our setup so far, we have $\theta =$ the unknown parameter, $X_1, ..., X_n \overset{iid}{\sim} P_\theta$ and $T = r(X_1, ..., X_n)$ as a statistic and estimator of θ. We are now interested in an ideal $\hat{\theta} = T = r(X_1, ..., X_n)$, that is, one 'close' to θ. **How might we measure this?**

There are many ways. One possibility is to look at the chance

$$P(\left|\hat{\theta} - \theta\right| < \epsilon)$$

for some small value ϵ, e.g $\epsilon = 0.1$. Ideally, for small ϵ, we would like the probability above to be high. This makes $\hat{\theta}$ a 'close' estimator.

Example 4.1.1. *Suppose* $X_1, ..., X_3 \overset{iid}{\sim} \exp(\theta)$*. We know*

$$f(x) = \theta e^{-\theta x} \qquad x > 0; \theta > 0$$

and pick the statistic $T = \sum_i X_i \sim Gamma(3, \theta)$*. Further we have,*

$$\begin{aligned} \hat{\theta}_{MLE} &= \frac{1}{\bar{X}} = \frac{3}{\sum_i X_i} = \frac{3}{T} \\ \hat{\theta}_{Bayes'} &= \frac{4}{3+T} && (\theta_{prior} \sim Gamma(\alpha = 1, \beta = 2)) \end{aligned}$$

This makes

$$\begin{aligned} G(\theta) &= P(\left|\hat{\theta} - \theta\right| < \epsilon|\theta) \\ &= P(-\epsilon < \hat{\theta} - \theta < \epsilon|\theta) \\ &= F_{\hat{\theta}}(\theta + \epsilon) - F_{\hat{\theta}}(\theta - \epsilon) \approx dF_{\hat{\theta}} \end{aligned}$$

Now, we know

$$\begin{aligned} F_{\hat{\theta}}(t) = P\left(\hat{\theta} \le t|\theta\right) &= P\left(T \ge \frac{3}{t}\middle|\theta\right) \text{ for MLE} \\ &= P\left(T \ge \frac{4}{t} - 2\middle|\theta\right) \text{ for Bayes'} \end{aligned}$$

In practice, we will manually plug in values for θ for a chosen ϵ and generate a probability function G that describes the chance of observing each probability for a fixed θ. This works for the MLE. To find the value $P(\left|\hat{\theta} - \theta\right| < \epsilon)$, we take the mean of $E[\theta]$ using G as the *distribution function and use it in $P(|\hat{\theta}-\theta| < \epsilon)$. For the Bayes' estimate, we simply compute $E[G(\theta)]$ using the prior distribution $\xi(\theta)$ giving*

$$E[G(\theta)] = \int_{\Omega} G(\theta)\xi(\theta)d\theta = P(|\hat{\theta}-\theta| < \epsilon)$$

♥

Note: We can also calculate

$$P\left(\left|\frac{\hat{\theta}}{\theta}-1\right| \leq \epsilon \middle| \theta\right)$$

for a fixed ϵ, say $\epsilon = 0.1$. Now,

$$\frac{\hat{\theta}}{\theta} = \frac{3}{\theta T}$$

where $\theta T \sim \text{Gamma}(\alpha = 3, \beta = 1)$, whence

$$P\left(\left|\frac{\hat{\theta}}{\theta}-1\right| \leq \epsilon \middle| \theta\right) = 0.134$$

does not depend on θ. ∞

We will need several several sampling distributions that arise frequently in inferences.

1. χ^2_m distribution
2. t_m distribution

4.1.1 χ^2-distribution

The χ^2 distribution has its roots in the Gamma distribution and is a a Gamma distribution with special parameters. Suppose $X \sim \text{Gamma}(\alpha, \beta)$, then X can be described by the p.d.f

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad x > 0; \alpha, \beta > 0$$

Notice when $\beta = 1$, we the gamma integral $\Gamma(\alpha)$ becomes

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx = (\alpha - 1)!$$

typically we assume α is an integer for our calculations. If we let $\alpha = \frac{m}{2}$ and $\beta = \frac{1}{2}$, then we can rewrite the p.d.f as

$$\begin{aligned} f(x) &= \frac{\left(\frac{1}{2}\right)^{m/2} e^{-x/2} x^{m/2-1}}{\Gamma(\frac{m}{2})} \\ &= \frac{x^{m/2-1}e^{-x/2}}{2^{m/2}\Gamma(\frac{m}{2})} \end{aligned}$$

Further, if $m = 1$, then we have

$$f(x) = \frac{x^{-1/2}e^{-x/2}}{\sqrt{2}\Gamma(1/2)} \qquad (X \sim \chi_1^2 = \text{Gamma}(1/2, 1/2))$$

and we can show that

$$\Gamma(1/2) = \sqrt{\pi}$$

For the general distribution where $\alpha = \frac{m}{2}$ and $\beta = \frac{1}{2}$, we know the moment generating function (MGF) ψ is

$$\psi_X(t) = \left(\frac{1}{1-2t}\right)^{\frac{m}{2}}$$

and the mean and variance of $X \sim \text{Gamma}(m/2, 1/2)$ is

$$E(X) = m \qquad\qquad V(X) = 2m$$

The χ^2 distribution can also be related to the normal distribution since if $Y \sim \mathcal{N}(0,1)$ then $Y^2 \sim \chi_1^2$. We now give a proof of this fact.

Theorem 4.1.1 (Square of Normal)**.** *If* $Y \sim \mathcal{N}(0,1)$*, then* $Y^2 \sim \chi_1^2 = \textit{Gamma}(1/2, 1/2)$.

Proof. We begin with Y's c.d.f and use it to construct $W = Y^2$'s c.d.f too. We reason as follows

$$\begin{aligned} F(w) = P(W \le w) &= P(Y^2 \le w) \\ &= P(-\sqrt{w} \le Y \le \sqrt{w}) \\ &= \Phi(\sqrt{w}) - \Phi(-\sqrt{w}) \qquad (\Phi = \text{cum. norm. c.d.f}) \end{aligned}$$

then,

$$\begin{aligned}
f(w) &= F'(w) \\
&= \varphi(\sqrt{w})\left(\frac{1}{2}y^{-1/2}\right) + \varphi(-\sqrt{w})\left(\frac{1}{2}y^{-1/2}\right) && (\varphi = \text{norm. p.d.f}, \textbf{chain rule}) \\
&= \varphi(\sqrt{w})\left(y^{-1/2}\right) && (\varphi(-\sqrt{w}) = \varphi(\sqrt{w})) \\
&= (\sqrt{2\pi})^{-1} w^{-1/2} e^{-w/2} \\
&= \frac{w^{-1/2}}{\sqrt{2\pi}} e^{-w/2} \\
&= \frac{w^{\alpha-1}e^{-\beta w}}{2^{1/2}\pi^{1/2}} && (\beta = 1/2, \alpha = 1/2) \\
&= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \times w^{\alpha-1}e^{-\beta w}
\end{aligned}$$

The last form implies $W = Y^2 \sim \text{Gamma}(\alpha = 1/2, \beta = 1/2) = \chi_1^2$ as we sought to show. □

Why do we need the above theorem? It allows us to prove some facts about sums of squares of normal distributions. Among some, we can prove that if $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

Facts About Sums of Sq. of Normal

$$\sum_{i=1}^{n} X_i^2 \sim \chi_n^2 \qquad (\text{if } \mu = 0, \sigma^2 = 1)$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi_{n-1}^2$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi_n^2 \qquad (\mu\ \textbf{unknown})$$

4.1.2 *t*-distribution

Using the χ^2 distribution, we can create another type of distribution known as the **t-distribution**. Suppose $Y \sim \chi_m^2$, $Z \sim \mathcal{N}(0,1)$, and $Y \perp\!\!\!\perp Z$, then we say that $X \sim t_{n-1}$ when

$$X = \frac{Z}{\sqrt{Y/m}} \qquad (m = \text{degree of freedom})$$

The probability density function of a t-distribution with m degrees of freedom (df) can be written as

$$f(x) = \frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\Gamma(\frac{m}{2})}\left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}$$

The t-distribution has finite absolute moments up to its degree of freedom m. That is,

$$E\left(\left|X^k\right|\right) < \infty \qquad \text{for } k < m$$

Note, however, the MGF for a t-distribution does not exist. The t-distributions we will discuss are all *central* and have a mean of 0 or $E(X) = 0$. Further, the variance can be given as

$$V(X) = \frac{m}{m-2} \qquad \text{for } m > 2$$

To derive the p.d.f of t, we set the two random variables that make it up as independent or $Y \perp\!\!\!\perp Z$. As a reminder, $Y \sim \chi^2_m$ and $Z \sim \mathcal{N}(0,1)$. We then create two new random variables X and W such that

$$\begin{aligned} X &= \frac{Z}{\sqrt{Y/m}} \\ W &= Y \end{aligned}$$

To find the t-distribution density, we first find the joint distribution of X and W. Notice that both Y and Z can be written as functions of X and W or

$$Z = s_1(X, W) = X \times \left(\frac{W}{m}\right)^{1/2} \qquad Y = s_2(X, W) = W$$

So, when we compute the joint distribution of X, W we are using a transformation of variables. Hence, we can write

$$f_{X,W}(x, w) = f_{Y,Z}(s_1(x, w), s_2(x, w)) \times \det(\boldsymbol{J})$$

where

$$\boldsymbol{J} = \begin{pmatrix} \partial Z/\partial X & \partial Y/\partial X \\ \partial Z/\partial W & \partial Y/\partial W \end{pmatrix}$$

Now, $f_{X,W}(x, w) = f_W(w) \times f_{X|W=w}(x)$ by conditional probability and to obtain the marginal distribution of X integrate out W. That is, compute

$$\int_0^\infty f_W(w) \times f_{X|W=w}(x)dw = f_X(x)$$

Note: The t-distribution is symmetric like normal. Also, as $m \to \infty$ the t-distribution tends to a standard Normal distribution. Notice, that when $m = 1$, the t-distribution is a Cauchy distribution:

$$\begin{aligned} f(t|m=1) &= \frac{\Gamma(1)}{\sqrt{\pi}\Gamma(1/2)}\left(1+\frac{x^2}{1}\right)^{-1} \\ &= \frac{1}{\pi(1+x^2)} \end{aligned}$$

∞

4.1.3 Using χ^2 and t in Inferences

So, how to we use these distribution's in practice? We will show the following facts which are useful for statistical inference:

Facts for Statistical Inference

1. $(n-1)\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1} \cdot \sigma^2$
2. If $\bar{X} \sim \mathcal{N}(\mu, \sigma^2)$ and $\hat{\sigma}^2, \bar{X}$ are independent, then

$$\frac{\bar{X}-\mu}{\sqrt{\hat{\sigma}^2/n}} \sim t_{n-1}$$

☕

For these facts, we need the concept of an *orthogonal transformation* from linear algebra.

Orthogonal Transformation Facts

Definition 4.1.1 (Orthogonal Matrix)**.** *A matrix* $\boldsymbol{A}_{n\times n}$ *is* **orthogonal** *if*

$$\boldsymbol{A}' = \boldsymbol{A}^{-1}$$

In other words, an orthogonal matrix is a **norm-preserving operation** *for all vectors in the subspace it acts on. Each vector the makes up the column space of* $\boldsymbol{A}$ *forms an orthonormal basis. We can also say*

$$\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{A}'\boldsymbol{A} = I_{n\times n}$$

♣

Theorem 4.1.2 (Determinate Value)**.** *For an orthogonal matrix* $\boldsymbol{A}$*, we know* $|\det(\boldsymbol{A})| = 1$.

Proof. Since $\boldsymbol{A} = \boldsymbol{A}'$ by definition, we know $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$. Further, it known as property of determinants that for any matricies $\boldsymbol{A}$ and $\boldsymbol{B}$ that

$$\det(\boldsymbol{AB}) = \det(\boldsymbol{A}) \cdot \det(\boldsymbol{B})$$

It then follows that

$$\begin{aligned} \det(\boldsymbol{AA}') &= \det(\boldsymbol{A}) \det(\boldsymbol{A}') \\ &= (\det(\boldsymbol{A}))^2 = 1 && (\boldsymbol{A} \text{ is orthogonal}) \\ &\implies |\det(\boldsymbol{A})| = 1 \end{aligned}$$

as we sought to show. □

Theorem 4.1.3 (Norm Preservation)**.** *If the matrix* $\boldsymbol{A}$ *is orthogonal and* $\boldsymbol{A}x = y$ *for any vectors* x, y*, then* $||y|| = ||x||$*. Note that* $|| \cdot ||$ *signifies the norm or length of the vector.*

Proof. We compute the norm of y as follows

$$\begin{aligned} ||y||^2 = \sum_i y_i^2 = y'y &= (x'\boldsymbol{A}')(\boldsymbol{A}x) && (y = \boldsymbol{A}x) \\ &= x'\underbrace{\boldsymbol{A}'\boldsymbol{A}}_{I}x \\ &= x'x = \sum_i x_i^2 \\ &= ||x||^2 \end{aligned}$$

Which is what we sought to show. □

With these facts in mind, we can start with some statistical facts about normal distributions and their transformations.

Theorem 4.1.4 (P.D.F of Orthogonal Transformation)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(0,1)$ *and* $\boldsymbol{A}_{n\times n}$ *is orthogonal with* $y = \boldsymbol{A}x$*, then* $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(0,1)$*.*

Proof. We note the joint distribution of the $X_1, ..., X_n$'s is

$$f(x_1, ..., x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_i x_i^2\right)$$

In this setting, we perform the transformation $x = \boldsymbol{A}^{-1}y$ where $\boldsymbol{A}$ is orthogonal. Because of

Norm Preservation, we know $\sum_i x_i^2 = \sum_i y_i^2$, we can write the joint p.d.f of $Y_1, ..., Y_n$ as

$$\begin{aligned} f(y_1, ..., y_n) &= |\boldsymbol{J}| \cdot f(x_1, ..., x_n) \\ &= 1 \cdot \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_i y_i^2\right) \qquad \left(\boldsymbol{J} = \boldsymbol{A} \text{ and } \sum_i x_i^2 = \sum_i y_i^2\right) \end{aligned}$$

which is the joint distribution of $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(0, 1)$. This concludes the proof. □

Lemma 4.1.1 (Independece of $\bar{X}$ and S^2)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ *and we define* $S^2 = \sum_i (X_i - \bar{X})^2$ *and* $\bar{X} = \frac{1}{n}\sum_i X_i$*. Then,* $\bar{X} \perp\!\!\!\perp S^2$*.*

Proof. Suppose we have $Z_1, ..., Z_n \overset{iid}{\sim} N(0, 1)$. Then, let

$$u' = \left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$$

be a row vector. Then define an orthogonal transformation $\boldsymbol{A}$ such that

$$\boldsymbol{A} = \begin{pmatrix} u' \\ \text{row } 2 \\ \vdots \\ \text{row } n \end{pmatrix}$$

where it is constructed using the Gram-Schmidt method. Let $Y = \boldsymbol{A}Z$ and we have $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(0, 1)$ as well as

$$Y_1 = u'Z = \sum_{i=1}^{n} Z_i = \sqrt{n}\bar{Z}$$

which implies

$$\begin{aligned} \sum_{i=2}^{n} Y_i^2 &= \sum_{i=1}^{n} Y_i^2 - Y_1^2 \\ &= \sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2 \qquad \left(\sum Z_i^2 = \sum Y_i^2, Y_1 = \sqrt{n}\bar{Z}\right) \\ &= \sum_{i=1}^{n} (Z_i - \bar{Z})^2 \end{aligned}$$

Since each Y_i are mutually independent, we know $\sum_{i=2}^{n} Y_i^2 \perp\!\!\!\perp Y_1$ and by the alternative forms given above, we have

$$\sqrt{n}\bar{Z} \perp\!\!\!\perp \sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

which implies

$$\bar{Z} \perp\!\!\!\perp \sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

Now, when $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ we can write

$$Z_i = \frac{X_i - \mu}{\sigma} \implies \bar{Z} = \frac{1}{\sigma}\left(\bar{X} - \mu\right)$$

and

$$\sum_{i=1}^{n}(Z_i - \bar{Z})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

but since X_i's are the only random quantities in the equations above, we have

$$\bar{X} \perp\!\!\!\perp \sum_{i=1}^{n}(X_i - \bar{X})^2 = S^2$$

This concludes the proof. □

Theorem 4.1.5 (Distn. of S^2). *If $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $S^2 = \sum_i (X_i - \bar{X})^2 \sim \chi^2_{n-1} \cdot \sigma^2$.*

Proof. With the addition and subtraction of $\bar{X}$ in $(X_i - \mu)^2$ we can rewrite it as

$$(X_i - \mu)^2 = (X_i - \bar{X})^2 + (\bar{X} - \mu)^2 - 2(X_i - \bar{X})(\bar{X} - \mu)$$

the sum of the above indexing by i becomes

$$\sum_{x=1i}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

since $\sum_i (X_i - \bar{X}) = 0$. If we divide by the variance of the X_i's σ^2, then we get

$$\underbrace{\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2}_{C} = \underbrace{\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2}_{A} + \underbrace{\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2}_{B}$$

Now, the C is the sum of squared standard normals, so it is distributed χ^2_n and B is a squared standard normal, so it is distributed χ^2_1. We know $A \perp\!\!\!\perp B$ by Indepenence Lemma. Since only sums of squares of independent standard normal variates make up a χ^2 variate, it is then apparent that $A \sim \chi^2_{n-1}$ which implies that

$$S^2 = \sum_i (X_i - \bar{X})^2 \sim \chi^2_{n-1} \cdot \sigma^2$$

as we sought to show. □

With these facts in mind, we can derive one important fact about the sampling distributions of the standardized sample mean using the sample variance $\hat{\sigma}^2$ as an estimator of the population variance σ^2.

Theorem 4.1.6 (Distribution of t-statistic)**.** *The random quantity $(\bar{X} - \mu)/\sqrt{\hat{\sigma}^2/n}$ follows a t-distribution with $n-1$ degrees of freedom.*

Proof. We will find an alternative from of the expression

$$\frac{\bar{X} - \mu}{\sqrt{\hat{\sigma}^2/n}}$$

We reason as follows

$$\begin{aligned}
\frac{\bar{X} - \mu}{\sqrt{\hat{\sigma}^2/n}} &= \frac{\bar{X} - \mu}{\sqrt{\hat{\sigma}^2/n}} \cdot \frac{1/\sqrt{\sigma^2/n}}{1/\sqrt{\sigma^2/n}} \\
\frac{\bar{X} - \mu}{\sqrt{\hat{\sigma}^2/n}} &= \frac{(\bar{X} - \mu)/(\sqrt{\sigma^2/n})}{\sqrt{\hat{\sigma}^2/\sigma^2}} \\
&\sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-1}/n-1}} = t_{n-1}
\end{aligned}$$

where the last line came from the facts:

$$\begin{aligned}
E(\bar{X}) = \mu \qquad & V(\bar{X}) = \sigma^2/n \\
(n-1)\hat{\sigma}^2 = \frac{n-1}{n-1}S^2 &= S^2 \sim \chi^2_{n-1} \cdot \sigma^2 \\
\implies \frac{\hat{\sigma}^2}{\sigma^2} &\sim \chi^2_{n-1}/n-1
\end{aligned}$$

This concludes the proof. □

Chapter 5 — Confidence Intervals (CIs)

The basic idea behind a confidence interval is to find an interval $[A, B]$ such that

$$P\left(\theta \in [A, B]\right) \geq 1 - \alpha = \gamma$$

where we determine/choose the significance level α. Here, we have A and B as functions of the data we have soon to be collected $X_1, \ldots, X_n$. In other words,

$$\underbrace{A = A(X_1, \ldots, X_n) \qquad B = B(X_1, \ldots, X_n)}_{\text{functions of } X_1, \ldots, X_n}$$

The probability statement is about A, B as functions of random variables $X_1, \ldots, X_n$ not about θ, which is considered fixed.

Example 5.0.1 (CI for Log. Rainfall). *Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where*

$$\begin{aligned} X_i &= \textit{log rainfall} \\ \bar{X} &= \textit{sample avg/mean log rainfall} \\ \sigma' &= \textit{estimator of } \sigma \end{aligned}$$

Then, by our previous theorems about the t-distribution, we know

$$U = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma'} \sim t_{n-1}$$

where U does not depend on an unknown parameter, i.e. the distribution is t_{n-1} for all possible μ we may have. We can find values $[a, b]$ such that

$$P(a \leq U \leq b) \geq 1 - \alpha = \gamma$$

and by symmetry of the t-distribution we can choose $a = -b$; call it c. This then gives us

$$P(-c \leq U \leq c) \geq 1 - \alpha = \gamma$$

To get the equality in the above statement, we will set $c = T_{n-1}^{-1}(1 - \alpha/2)$ which is the

$(1-\alpha/2)$*th quantile of the t-distribution. With this, we can derive the interval as follows:*

$$P\left(-c \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma'} \leq c\right) = P\left(-c\sigma' \leq \sqrt{n}(\bar{X}-\mu) \leq c\sigma\right)$$
$$= P\left(\underbrace{\bar{X} - c\frac{\sigma'}{\sqrt{n}}}_{\text{random}} \leq \underbrace{\mu}_{\text{fixed}} \leq \underbrace{\bar{X} + c\frac{\sigma'}{\sqrt{n}}}_{\text{random}}\right) = 1-\alpha$$

This makes the $(1-\alpha/2)100\%$ *CI as*

$$[A, B] = \left[\bar{X} - c\frac{\sigma'}{\sqrt{n}}, \bar{X} + c\frac{\sigma'}{\sqrt{n}}\right]$$

If we let $1-\alpha = \gamma = 0.95$*, then we find* $c = T_{25}^{-1}(0.975) = 2.060$ *and* $c/\sqrt{n} = 2.060/\sqrt{26} \approx 0.404$*. Hence, the bounds for the interval are*

$$A = \bar{X} - 0.404\sigma'$$
$$B = \bar{X} + 0.404\sigma'$$

in this case.

♥

5.1 One/Two-Sided CIs

With this example, we can now give a formal definition of a confidence interval.

Definition 5.1.1 (Two-Sided CIs)**.** *Suppose* $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$ *and* $g(\theta)$ *is only a function of* θ*, then if*

$$P(A \leq g(\theta) \leq B) \geq 1-\alpha = \gamma$$

we can say the interval $[A, B]$ *is a coefficient* γ *confidence interval for* $g(\theta)$*. Further, if we have*

$$P(A \leq g(\theta) \leq B) = \gamma$$

the CI $[A, B]$ *is called* **exact**.

♣

Note: The γ confidence coefficient is a statement about the chance the sample statistics will lead to an interval containing $g(\theta)$. In addition, the statement $P(A \leq g(\theta) \leq B) \geq \gamma$ does not uniquely define an interval. For instance, any two values $A = T_{n-1}^{-1}(\gamma_1)$ and $B = T_{n-1}^{-1}(\gamma_2)$

such that $\gamma_2 - \gamma_1 = \gamma$ where $\gamma_2 > \gamma_1$ will create a $(\gamma)100\%$ CI. We can thus write,

$$P\left(T_{n-1}^{-1}(\gamma_1) \leq g(\theta) \leq T_{n-1}^{-1}(\gamma_2)\right) = \gamma$$

for any CI with CI coefficient γ. When the distribution is Normal, we can have an exact CI for the population mean μ. In practice, we choose $\gamma_1 = 1 - \gamma_2$ (equal areas on both ends) because it provides the **shortest interval**. ∞

If we are only interested in a one-sided interval (one with an upper or lower limit that is infinite) as we could be when we are estimating λ from an exponential distribution, then we have a definition of it as such:

Definition 5.1.2 (One-Sided CIs)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$ and $g(\theta)$ is only a function of θ, then if*

$$P(A \leq g(\theta)) \geq 1 - \alpha = \gamma$$

or

$$P(g(\theta) \leq B) \geq 1 - \alpha = \gamma$$

then, we have a one-sided coefficient γ CI where A is the **lower confidence limit** *and B is the* **upper confidence limit**.

♣

For the mean of a normal distribution μ, we have the upper and lower bounds for one-sided CIs as

$$A = \bar{X} + T_{n-1}^{-1}(\gamma)\frac{\sigma'}{\sqrt{n}}$$
$$B = \bar{X} - T_{n-1}^{-1}(\gamma)\frac{\sigma'}{\sqrt{n}}$$

∞

5.2 Pivotal Quantities

In order to construct CIs, we need to make sure we have what are known as **pivotal quantities**. These quantities allow us to isolate the desired quantity (parameter) and form a probability statement about it. We define a pivotal quantity as follows:

Definition 5.2.1 (Pivotal Quantity)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$, then if we can make a random variable $V(X_1, ..., X_n, \theta)$ whose distribution is the same for all θ, then V is called the*

pivotal quantity *or* **pivotal**. *Such a pivotal is useful for constructing a CI for $g(\theta)$ if there is a function r that can 'invert' the pivotal to isolate $g(\theta)$. In other words,*

$$r(V(X_1, ..., X_n, \theta), X_1, ..., X_n) = g(\theta)$$

♣

Example 5.2.1 (Identifying a Pivotal). *To give an idea for it we note here that $\frac{\bar{X}-\mu}{\sigma'/\sqrt{n}}$ is a pivotal quantity when the data is Normally distributed. For example, $U \sim T_{n-1}$ for all values of μ and the function's image that inverts this pivotal is given by*

$$r(V, X_1, ..., X_n) = g(\theta) = \mu = \bar{X} - u\frac{\sigma'}{\sqrt{n}}$$

♥

Using the idea of a pivotal, we can formally define the endpoints for a CI.

Theorem 5.2.1 (Endpoints for CI). *Suppose $X_1, ..., X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$ and a pivotal $V(X_1, ..., X_n, \theta)$ exists. Let $G =$ c.d.f of V that is,*

$$G(v) = P(V \leq v) \qquad (continuous)$$

If we then also assume r exists and is strictly increasing in V for each $X_1, ..., X_n$ and define the confidence limits as

$$0 < \gamma < 1 \qquad \gamma_2 > \gamma_1 \qquad \gamma_2 - \gamma_1 = \gamma$$

then the following statistics are endpoints of an exact γ coefficient CI for $g(\theta)$:

$$A = r\left(G^{-1}(\gamma_1), X_1, ..., X_n\right)$$
$$B = r\left(G^{-1}(\gamma_2), X_1, ..., X_n\right)$$

Proof. Since r is strictly increasing, we know that for any constants c_1, c_2 such that

$$c_1 < V < c_2$$

we equivalently have

$$r(c_1, X_1, ..., X_n) < g(\theta) < r(c_2, X_1, ..., X_n)$$

If we then set $c_i = G^{-1}(\gamma_i)$ for $i \in \{1, 2\}$, then we have

$$\begin{aligned} P\left(G^{-1}(\gamma_1) < V < G^{-1}(\gamma_2)\right) &= P\left(r(G^{-1}(\gamma_1), X_1, ..., X_n) < g(\theta) < r(G^{-1}(\gamma_2), X_1, ..., X_n)\right) \\ &= P\left(A < g(\theta) < B\right) = \gamma \end{aligned}$$

which implies

$$\begin{aligned} A &= r\left(G^{-1}(\gamma_1), X_1, ..., X_n\right) \\ B &= r\left(G^{-1}(\gamma_2), X_1, ..., X_n\right) \end{aligned}$$

as we sought to show. We can make a similar argument for a one-sided interval. □

Example 5.2.2 (CI for μ)**.** *For the CI for μ when we sample from Normal, we have $G = T_{n-1}$ and*

$$\begin{aligned} \gamma_1 &= \frac{\gamma}{2} \\ \gamma_2 &= \frac{\gamma}{2} \end{aligned}$$

since we want a symmetric interval. The inverting function r is given by

$$r = \bar{X} - T_{n-1}^{-1}(\cdot)\frac{\sigma'}{\sqrt{n}}$$

Notice it is only a function of $X_1, ..., X_n$ and $G^{-1} = T_{n-1}^{-1}$ which lets us write it as $r = r(G^{-1}(\cdot), X_1, ..., X_n)$. In previous examples we found out that

$$\begin{aligned} A &= \bar{X} - G^{-1}(\gamma_1)\frac{\sigma'}{\sqrt{n}} = r(G^{-1}(\gamma_1), X_1, ..., X_n) \\ B &= \bar{X} - G^{-1}(\gamma_2)\frac{\sigma'}{\sqrt{n}} = r(G^{-1}(\gamma_2), X_1, ..., X_n) \end{aligned}$$

which is the same form as we derived in Endpoints for CI

♥

Example 5.2.3 (Approx. CI for Poisson)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim}$ Poisson(θ) and we sample n large so $\bar{X} = \hat{\theta} \sim \mathcal{N}(\theta, \theta/n)$. We know by the Delta Method that $2\bar{X}^{1/2} \overset{appx}{\sim} \mathcal{N}(2\theta^{1/2}, 1/n)$ and this allows us to state*

$$P\left(\left|2\bar{X}^{1/2} - 2\theta^{1/2}\right| < c\right) \approx 2\Phi(cn^{1/2}) - 1$$

which implies that an approximate CI for $2\theta^{1/2}$ is

$$2\theta^{1/2} \in \left[-c + 2\bar{X}^{1/2}, c + 2\bar{X}^{1/2}\right]$$

For an interval for θ, we then perform inverse (monotonic) operations to isolate it:

$$\begin{aligned} 2\theta^{1/2} &\in \left[-c + 2\bar{X}^{1/2}, c + 2\bar{X}^{1/2}\right] \\ \implies \theta^{1/2} &\in \left[-c/2 + \bar{X}^{1/2}, c/2 + \bar{X}^{1/2}\right] \\ \implies \theta &\in \left[(-c/2 + \bar{X}^{1/2})^2, (c/2 + \bar{X}^{1/2})^2\right] \end{aligned}$$

If we want a 95% CI for θ assuming $n = 100$ we simply set $c = 0.196$ in the above expression.

♥

Chapter 6 — Credible Intervals (CDIs)

Credible intervals (CDIs) are a Bayesian analogue of Frequentist confidence intervals. However, since the population parameters are random variables, we interpret them differently. There are still endpoints A, B such that $\theta \in [A, B]$ only A, B are usually dependent on the data as well as parameters for θ's prior (or posterior) distribution. The probability statement we make is

$$P\left(\underbrace{A}_{\text{fixed}} \leq \underbrace{\theta}_{random} \leq \underbrace{B}_{\text{fixed}}\right) \geq 1 - \alpha = \gamma$$

and is interpreted as 'there is chance $\geq 1-\alpha$ that the interval $[A, B]$ contains θ'. Notice how this is different from the frequentist interpretation where the endpoints A, B are random and the one we observed is from a set that contains θ $(1-\alpha)100\%$ of the time. More formally, we define a credible interval as follows:

Definition 6.0.1 (Credible Interval)**.** *Suppose* $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$ *and* $\theta \sim \mathcal{D}$*. Then we say we have a* $1 - \alpha = \gamma$ *credible interval when*

$$P\left(A \leq \theta \leq B\right) = \gamma$$

where A, B *are quantiles of distribution* $\mathcal{D}$*.*

♣

Our focus with credible intervals are on samples from a Normal Distribution. So we will give a view of this distribution from the Bayesian perspective.

6.1 Bayesian Analysis of samples from a Normal Distribution

In Bayesian statistics, the variance gives information about how well our minds know what values a parameter can take. When the variance is high, we can say we have low precision about the parameter and when it is low we can say we high precision about the parameter. We formalize this concept below

Definition 6.1.1 (Precision). *The reciprocal of the variance σ^2 is known as the* **precision** *of the random quantity θ we are interested in. We denote the precision with τ and have*

$$\tau = \frac{1}{\sigma^2}$$

♣

If we have $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau)$, then in Bayesian Analysis μ, τ are random variables with prior distributions. We have seen estimates of μ based on the posterior distribution $\xi(\mu|x_1, ..., x_n)$ under the squared error loss and that the Bayes' Estimator was the mean of the posterior distribution.

When both μ, τ are unknown, however, we need to specify priors for both parameters. In Bayesian statistics, we specify the density of a normal variate X as

$$f(x|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}\tau(x-\mu)^2\right\}$$

and the joint density of $X_1, ..., X_n$ is thus

$$f(x_1, ..., x_n|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{n/2} \times \exp\left\{-\frac{1}{2}\tau\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

We desire a prior $\xi(\mu, \tau)$ that is a conjugate prior jointly for (μ, τ) so that a Bayesian update using the data yields a distribution that is of the same family as $\xi(\mu, \tau)$. By probability laws, we know $\xi(\mu, \tau) = \xi_1(\mu|\tau)\xi_2(\tau)$. To find $\xi_1(\mu|\tau)$ recall that when $\sigma^2 = \tau^{-1}$ was assumed known we showed that the family of Normal priors was a conjugate family of priors, i.e. it resulted in the posterior also being a Normal distribution. It is then natural to have $\xi_1(\mu|\tau)$ as a normal distribution with precision $\lambda_0\tau$.

More specifically, we write $\xi_1(\mu|\tau)$ as

$$\xi_1(\mu|\tau) = \left(\frac{\lambda_0\tau}{2\pi}\right) \exp\left\{-\frac{1}{2}\lambda_0\tau(\mu-\mu_0)^2\right\}$$

and set $\tau \sim \text{Gamma}(\alpha_0, \beta_0)$, making the p.d.f of it as

$$\xi_2(\tau) = \frac{\beta_0^{\alpha_0}\tau^{\alpha_0-1}e^{-\beta_0\tau}}{\Gamma(\alpha_0)}$$

and in terms of proportions this gives

$$\xi_2(\tau) \propto \tau^{\alpha_0-1}e^{-\beta_0\tau}$$

Note: $\beta_0^{\alpha_0}/\Gamma(\alpha_0)$ is a constant that allows the integral with respect to τ to integrate to 1—it does not affect the shape of the distribution. ∞

When we multiply $\xi_1(\mu|\tau)$ and $\xi_2(\tau)$ to make $\xi(\mu, \tau)$, we have what is known as a **Normal-Gamma Prior** with hyper-parameters $\mu_0, \lambda_0, \alpha_0, \beta_0$.

Under these specific priors, we will show that the family of joint priors (for μ, τ) is a conjugate family of joint distributions. This, in turn, with then imply that the joint posterior of μ, τ is also of the same family of distributions, i.e. we have a **Normal-Gamma Posterior** with hyper-parameters $\mu_1, \lambda_1, \alpha_1, \beta_1$. Before we do this, we give some notation used in the proof:

Notation: We use the following notation for averages (sample) and sums of square deviations

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \underbrace{\sum_{i=1}^{n}(X_i - \bar{X})^2}_{\text{not dividing by } n \text{ here}}$$

☕

Theorem 6.1.1 (Normal-Gamma Posterior)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau)$ *where* μ, τ *are both unknown and the supports (domains) for both are*

$$-\infty < \mu < \infty \qquad\qquad \tau > 0$$

Then, if $\mu|\tau \sim \mathcal{N}(\mu_0, \lambda_0\tau)$ *where* $\mu_0, \lambda_0 > 0$ *and* $\tau \sim \Gamma(\alpha_0, \beta_0)$ *where* $\alpha_0, \beta_0 > 0$ *we have*

$$(\mu|\tau, x_1, ..., x_n) \sim \mathcal{N}(\mu_1, \lambda_1\tau)$$
$$(\tau|x_1, ..., x_n) \sim \Gamma(\alpha_1, \beta_1)$$

where

$$\mu_1 = \frac{\lambda_0\mu_0 + n\bar{x}}{\lambda_0 + n} \qquad\qquad \lambda_1 = \lambda_0 + n$$

$$\alpha_1 = \alpha_0 + \frac{n}{2} \qquad\qquad \beta_1 = \beta_0 + \frac{1}{2}s^2 + \frac{n\lambda_0(\bar{x} - \mu_0)^2}{2(\lambda_0 + n)}$$

Proof. We will begin with the proportional form of the posterior distribution $\xi(\mu, \tau|x_1, ..., x_n)$:

$$\xi(\mu, \tau|x_1, ..., x_n) \propto f(x_1, ..., x_n|\mu, \tau)\xi_1(\mu|\tau)\xi_2(\tau)$$

$$\propto \tau^{n/2} \exp\left\{-\frac{1}{2}\tau\sum_{i=1}^{n}(x_i-\mu)^2\right\} \times \tau^{1/2}\exp\left\{-\frac{1}{2}\lambda_0\tau(\mu-\mu_0)^2\right\}$$
$$\times\, \tau^{\alpha_0-1}\exp\left\{-\beta_0\tau\right\}$$

$$= \tau^{\alpha_0+\frac{n+1}{2}-1}\exp\left\{-\frac{1}{2}\tau\left[\sum_{i=1}^{n}(x_i-\mu)^2+\lambda_0(\mu-\mu_0)^2\right]-\beta_0\tau\right\}$$
$$= \tau^{\alpha_0+\frac{n+1}{2}-1}\exp\left\{-\frac{1}{2}\tau\left[s^2+n(\bar{x}-\mu)^2+\lambda_0(\mu-\mu_0)^2\right]-\beta_0\tau\right\}$$
$$= \tau^{\alpha_0+\frac{n+1}{2}-1}\exp\left\{-\frac{1}{2}\tau\underbrace{\left[n(\bar{x}-\mu)^2+\lambda_0(\mu-\mu_0)^2\right]}_{A}\right\}\cdot\exp\left\{-\tau(s^2/2+\beta_0)\right\}$$

From here we note that A can be simplified as follows

$$\begin{aligned}
n(\bar{x}-\mu)^2+\lambda_0(\mu-\mu_0)^2 &= n(\bar{x}^2-2\bar{x}\mu+\mu^2)+\lambda_0(\mu^2-2\mu\mu_0+\mu_0^2)\\
&= \mu^2(n+\lambda_0)-2\mu(n\bar{x}+\lambda_0\mu_0)+(\lambda_0\mu_0^2+n\bar{x}^2)\\
&= (n+\lambda_0)\left[\mu^2-2\mu\left(\frac{\lambda_0\mu_0+n\bar{x}}{n+\lambda_0}\right)+\left(\frac{\lambda_0\mu_0^2+n\bar{x}^2}{n+\lambda_0}\right)\right]\\
&= (n+\lambda_0)\Bigg[\left(\mu-\frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}\right)^2+\frac{(\lambda_0\mu_0^2+n\bar{x}^2)(n+\lambda_0)}{(n+\lambda_0)^2}\\
&\qquad -\frac{(n\bar{x}+\lambda_0\mu_0)^2}{(n+\lambda_0)^2}\Bigg]\\
&= (n+\lambda_0)\left[\left(\mu-\frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}\right)^2+\frac{\mu_0^2(n\lambda_0)-2n\lambda_0\bar{x}\mu_0+\lambda_0 n\bar{x}^2}{(n+\lambda_0)^2}\right]\\
&= (n+\lambda_0)\left(\mu-\frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}\right)^2+\frac{n\lambda_0(\mu_0-\bar{x})^2}{(n+\lambda_0)}
\end{aligned}$$

this results in the simplification as

$$= \tau^{\alpha_0+\frac{n+1}{2}-1}\exp\left\{-\frac{1}{2}\tau\left[(n+\lambda_0)\left(\mu-\frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}\right)^2+\frac{n\lambda_0(\mu_0-\bar{x})^2}{(n+\lambda_0)}\right]\right\}\times$$
$$\exp\left\{-\tau(s^2/2+\beta_0)\right\}$$

and grouping terms for μ, τ's distributions we arrive at

$$\begin{aligned}
&= \tau^{1/2} \exp\left\{-\frac{1}{2}(n+\lambda_0)\tau\left[\left(\mu - \frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}\right)^2\right]\right\} \times \\
&\quad \tau^{\alpha_0+\frac{n}{2}-1} \exp\left\{-\tau\left(s^2/2+\beta_0+\frac{n\lambda_0(\mu_0-\bar{x})^2}{2(n+\lambda_0)}\right)\right\} \\
&\propto \xi_1(\mu|\tau, x_1, \dots, x_n) \cdot \xi_2(\tau|x_1, \dots, x_n) \qquad \text{(Normal − Gamma)}
\end{aligned}$$

The results above imply that

$$\mu_1 = \frac{\lambda_0\mu_0 + n\bar{x}}{\lambda_0+n} \qquad \lambda_1 = \lambda_0 + n$$

$$\alpha_1 = \alpha_0 + \frac{n}{2} \qquad \beta_1 = \beta_0 + \frac{1}{2}s^2 + \frac{n\lambda_0(\bar{x}-\mu_0)^2}{2(\lambda_0+n)}$$

as we sought to show. □

Now that we know that the Normal-Gamma is a conjugate prior, it is also useful (for credible intervals) to know the distribution of μ itself.

Theorem 6.1.2 (Distn. of μ)**.** *If* $\mu|\tau \sim \mathcal{N}(\mu_0, \sigma^2 = [\lambda_0\tau]^{-1})$ *and* $\tau \sim$ *Gamma*(α_0, β_0)*, then we have*

$$\left(\frac{\lambda_0\alpha_0}{\beta_0}\right)^{1/2}(\mu - \mu_0) \sim t_{2\alpha_0}$$

Proof. We begin with the random variable Z such that

$$Z = \frac{\mu-\mu_0}{\sigma} = \sqrt{\lambda_0\tau}(\mu-\mu_0)$$

Notice that $Z \sim \mathcal{N}(0,1)$ but not matter what the value of τ is, we will always have $Z \sim \mathcal{N}(0,1)$. This makes $Z \perp\!\!\!\perp \tau$, that is, if we hold μ as random and τ known, i.e. $f_1(z|\tau) = f(z)$. Now we create another random variable Y such that $Y = 2\beta_0\tau$ which implies $Y \sim \chi^2_{2\alpha_0}$. Since $Z \perp\!\!\!\perp \tau$, we have $Z \perp\!\!\!\perp Y$ too. This, by the definition of the t-distribution, gives the random variable U the form

$$U = \frac{Z}{\sqrt{Y/2\alpha_0}} = \frac{(\lambda_0\tau)^{1/2}(\mu-\mu_0)}{(2\beta_0\tau/2\alpha_0)^{1/2}} = \left(\frac{\lambda_0\alpha_0}{\beta_0}\right)^{1/2}(\mu-\mu_0) \sim t_{2\alpha_0}$$

which is what we sought to show.

Note: Accordingly, the marginal posterior of μ, i.e. $\mu|x_1, \dots, x_n$ is also $t_{2\alpha_1}$ due to Normal-Gamma distributions being a conjugate family. □

We can now give some facts about the marginalized distribution of μ.

Theorem 6.1.3 (Summary Stats of Marginal μ)**.** *Suppose that $\mu, \tau \sim$ Normal-Gamma with hyper-parameters $\mu_0, \lambda_0, \alpha_0, \beta_0$, then if $\alpha_0 > 1/2$*

$$E(\mu) = \mu_0$$

and if $\alpha_0 > 1$

$$V(\mu) = \frac{\beta_0}{\lambda_0(\alpha_0 - 1)}$$

Proof. Since $\left(\frac{\lambda_0\alpha_0}{\beta_0}\right)^{1/2}(\mu - \mu_0) = U$, we can also say

$$\mu = \left(\frac{\beta_0}{\lambda_0\alpha_0}\right)^{1/2} U + \mu_0$$

where we know $U \sim t_{2\alpha_0}$. Because of this we know from t-distribution facts that

$$E(U) = 0$$
$$V(U) = \frac{\alpha_0}{\alpha_0 - 1}$$

This then makes

$$E(\mu) = E\left\{\left(\frac{\beta_0}{\lambda_0\alpha_0}\right)^{1/2} U + \mu_0\right\} = \mu_0$$
$$V(\mu) = \frac{\beta_0}{\lambda_0\alpha_0}V(U) = \frac{\beta_0}{\lambda_0\alpha_0} \cdot \frac{\alpha_0}{\alpha_0 - 1}$$
$$= \frac{\beta_0}{\lambda_0(\alpha_0 - 1)}$$

which are the forms we sought to show.

Note: *for the posterior we have the same forms only we set $\mu_0, \lambda_0, \alpha_0, \beta_0$ to $\mu_1, \lambda_1, \alpha_1, \beta_1$.* □

6.2 Credible Intervals Based on Posterior Distribution

We can use the posterior distribution to find intervals (called **credible intervals**) by determining quantiles for the prior and/or posterior distributions.

Example 6.2.1 (Credible Interval for μ Prior & Post. (pg. 500 in book))**.** *Suppose $\mu, \tau \sim$*

Normal-Gamma where

$$\begin{aligned} \mu_0 &= 200 & \nu_0^2 &= 3150 \\ \alpha_0 &= 2 & \beta_0 &= 6300 \\ \lambda_0 &= 2 \end{aligned}$$

Then the random variable U follows a t_4 distribution since $2\alpha_0 = 4$. The form of U is then given by

$$U = \left(\frac{2 \cdot 2}{6300}\right)^{1/2} (\mu - 200) \approx 0.025(\mu - 200)$$

The 95% prior credible interval is then

$$P(-2.776 \leq 0.025(\mu - 200) \leq 2.776) = P(89 \leq \mu \leq 311) = 0.95$$

For a posterior interval, we receive the data

$$\bar{x} = 182.17 \qquad s^2 = 88678.5$$

which implies

$$\begin{aligned} \mu_1 &= 183.95 & \lambda_1 &= 20 \\ \alpha_1 &= 11 & \beta_1 &= 50925.37 \end{aligned}$$

Now the random variable $U|x_1, ..., x_n$ follows a t_{22} distribution since $2\alpha_1 = 22$. The form of $U|x_1, ..., x_n$ is then given by

$$U|x_1, ..., x_n = \left(\frac{20 \cdot 11}{50925.37}\right)^{1/2} (\mu - 183.95)$$

The posterior credible interval is then

$$P(-2.074 \leq U \leq 2.074|x_1, ..., x_n) = P(152.38 \leq \mu \leq 215.52|x_1, ..., x_n) = 0.95$$

. . . .

For comparison, the t condfidence interval bounds are

$$\begin{aligned} B &= \bar{x} + t_{17;0.975} \frac{\sigma'}{\sqrt{n}} \\ &= \bar{x} + 2.111 \sqrt{\frac{88678.5}{18}} \\ &= 218.09 \end{aligned}$$

$$\begin{aligned} A &= \bar{x} - t_{17;0.975}\frac{\sigma'}{\sqrt{n}} \\ &= \bar{x} - 2.111\sqrt{\frac{88678.5}{18}} \\ &= 146.25 \end{aligned}$$

♥

6.3 Improper Prior Distributions

We sometimes choose an "improper prior" in the sense that the prior is not a probability distribution. For example, if we are interested in the mean of a normal distribution μ, we could choose the prior as $\xi(\mu) = 1$ for $-\infty \leq \mu \leq \infty$, but this is improper since

$$\int_{-\infty}^{\infty} \xi(\mu) d\mu$$

is not defined. So, $\xi(\mu)$ is not a "real" prior p.d.f.

Similarly, we can choose an improper prior for σ. Typically this is $g(\sigma) = 1/\sigma$ for $0 \leq \sigma \leq \infty$ and again the integral

$$\int_{-\infty}^{\infty} g(\sigma) d\sigma$$

is not defined. If we wish to do Bayesian analysis with σ, then we know that since $\sigma = \tau^{-1/2}$, the improper p.d.f is

$$\xi(\tau) = \frac{1}{2}\tau^{-1} \qquad \text{for } \tau > 0$$

by the transformation of variables theorem[1]. If we, since the priors are improper, say $\mu \perp\!\!\!\perp \tau$, the joint improper p.d.f is then

$$\xi(\mu, \tau) = \frac{1}{2}\tau^{-1}$$

We can now compute the (proper[2]) posterior distribution from this improper prior as follows

$$\begin{aligned} \xi(\mu, \tau | x_1, \ldots, x_n) &\propto \xi(\mu, \tau) f(x_1, \ldots, x_n | \mu, \tau) \\ &\propto \tau^{-1}\tau^{n/2} \exp\left\{-\frac{\tau}{2}s^2 - \frac{n\tau}{2}(\mu - \bar{x})^2\right\} \\ &= \underbrace{\left[\tau^{1/2} \exp\left\{-\frac{n\tau}{2}(\mu - \bar{x})^2\right\}\right]}_{A} \times \underbrace{\left[\tau^{\frac{n-1}{2}-1} \exp\left\{-\tau\frac{s^2}{2}\right\}\right]}_{B} \end{aligned}$$

[1] since we want an improper prior for τ it is fair to write $\xi(\tau) = a\tau^{-1}$ for any real a too

[2] by Bayes' Theorem

Notice that A is proportional to the distribution $\mathcal{N}(\bar{x}, \underbrace{n\tau}_{\text{precision}})$ and B is proportional to the distribution Gamma$(\frac{n-1}{2}, \frac{s^2}{2})$. This implies that

$$\mu|\tau, x_1, ..., x_n \sim \mathcal{N}(\mu_1 = \bar{x}, \tau_1 = n\tau) \qquad (\lambda_1 = n)$$
$$\tau|x_1, ..., x_n \sim \text{Gamma}\left(\alpha_1 = \frac{n-1}{2}, \beta_1 = \frac{s^2}{2}\right)$$

Notice, for a proper Normal-Gamma prior we have

$$\mu_1 = \frac{\lambda_0\mu_0 + n\bar{x}}{\lambda_0 + n} \qquad \lambda_1 = \lambda_0 + n$$
$$\alpha_1 = \alpha_0 + \frac{n}{2} \qquad \beta_1 = \beta_0 + \frac{1}{2}s^2 + \frac{n\lambda_0(\bar{x} - \mu_0)^2}{2(\lambda_0 + n)}$$

and if we set $\lambda_0 = 0, \alpha_0 = -1/2$, and $\beta_0 = 0$, then we arrive at the posterior hyper-parameters using the improper prior. Also, if we set $\mu_0 = 0$ then we get a prior for μ centered at 0.

Example 6.3.1 (Improper Prior Credible Interval)**.** *We begin with the improper joint prior distribution we just created for our state of mind before collecting the data:*

$$\xi(\mu, \tau) = \frac{1}{2}\tau^{-1} \propto \tau^{-1}$$

and this is equivalent to choosing the 'improper prior hyper-parameters' $\lambda_0 = 0, \alpha_0 = -1/2$ *and* $\beta_0 = 0$. *Note, however, that this is different than individually choosing the improper priors for* μ *and* τ *as we discussed above.*[3]

We know that the posterior $\mu, \tau|x_1, ..., x_n$ *will be a Normal Gamma as we have already shown. So, when we collect data and observe*

$$n = 26 \qquad \bar{x} = 5.134 \qquad s^2 = 63.96$$

we observe the posterior hyper-parameters as

$$\mu_1 = \bar{x} = 5.134 \qquad \lambda_1 = n = 26$$
$$\alpha_1 = \frac{n-1}{2} = 12.5 \qquad \beta_1 = \frac{1}{2}s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = 31.98$$

which in turn allows us to calculate U *as*

$$\begin{aligned} U &= \left(\frac{\lambda_1\alpha_1}{\beta_1}\right)^{1/2}(\mu - \mu_1) \\ &= \left(\frac{26 \cdot 12.5}{31.98}\right)^{1/2}(\mu - 5.134) \\ &= 3.188(\mu - 5.134) \end{aligned}$$

Thus, the 95% credible interval for μ is obtained as follows

$$\begin{aligned} P(-2.060 \leq U \leq 2.060) &= P(-2.060 \leq 3.188(\mu - 5.134) \leq 2.060) \\ &= P\left(\frac{-2.060}{3.188} + 5.134 \leq \mu \leq \frac{2.060}{3.188} + 5.134\right) \\ &= P(4.488 \leq \mu \leq 5.78) = 0.95 \end{aligned}$$

. . . .

Note: *The abstract form of U using only the forms for the posterior hyper-parameters is given by*

$$\begin{aligned} U &= \left(\frac{n(n-1)/2}{\sum_{i=1}^{n}(x_i - \bar{x})^2/2}\right)^{1/2} (\mu - \bar{x}) \\ &= \left(\frac{n(n-1)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^{1/2} (\mu - \bar{x}) \\ &= \left(\frac{n}{\hat{\sigma}^2}\right)^{1/2} (\mu - \bar{x}) \sim t_{n-1} \end{aligned}$$

and the credible interval then becomes

$$\bar{x} - t^* \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + t^* \frac{\hat{\sigma}}{\sqrt{n}}$$

This has the same form as a t confidence interval for the population mean. Note, however, that the two are different in interpretation: credible intervals reflect our certainty about a parameter belonging to an interval and confidence intervals reflect our interval's probabilistic ability to capture the quantity in question.

. . . .

♥

[3] Both methods yield the same results in the end, however. So we are free to pick either.

Chapter 7 — Unbiased Estimators

In this chapter we discuss and give various examples of **unbiased statistics**, giving some advantages and disadvantages of these estimators. Before we talk about unbiasedness, we define what statistical bias is.

Definition 7.0.1 (Statistical Bias)**.** *The* **statistical bias** *or* **bias** *is the difference between the expected value of the estimator* $\delta(X_1, ..., X_n)$ *and the parameter to be estimated* $g(\theta)$*. In other words,*

$$bias = E\left[\delta(X_1, ..., X_n)\right] - g(\theta) = E(\delta(X_1, ..., X_n) - g(\theta))$$

♣

In a sense, the bias of an estimator tells us how 'off' or 'non-centered' that estimator is with respect to the parameter in question. The estimator is a random variable and will rarely be $g(\theta)$, but when it is centered about $g(\theta)$ in distribution we can have 'on average' a better estimate (most of the time). When the statistical bias in non-existent, we have what is known as **unbiasedness**.

Definition 7.0.2 (Statistical Unbiasedness)**.** *An estimator* $\delta(X_1, ..., X_n)$ *is* **unbiased** *when* $bias = 0$*. That is,*

$$\begin{aligned} bias = E\left[\delta(X_1, ..., X_n)\right] - g(\theta) = 0 \\ \implies E\left[\delta(X_1, ..., X_n)\right] = g(\theta) \end{aligned}$$

♣

Now that we know about bias, we can find an alternative form for the MSE that we defined in 3.

Corollary 7.0.1 (MSE Bias-Variance Form)**.** *If* $\delta(X_1, ..., X_n)$ *is an estimator of* $g(\theta)$ *and has* *finite variance, then*

$$MSE = [bias(\delta(X_1, ..., X_n))]^2 + V[\delta(X_1, ..., X_n)]$$

Proof. We begin with the definition of MSE and work from there:

$$\begin{aligned}
\text{MSE} &= E\left[(\delta(X_1, ..., X_n) - g(\theta))^2\right] \\
&= E\left[\delta(X_1, ..., X_n)^2 - 2\delta(X_1, ..., X_n)g(\theta) + g(\theta)^2\right] \\
&= E\left[\delta(X_1, ..., X_n)^2\right] - E\left[\delta(X_1, ..., X_n)\right]^2 + E\left[\delta(X_1, ..., X_n)\right]^2 \\
&\qquad\qquad - 2g(\theta)E\left[\delta(X_1, ..., X_n)\right] + g(\theta)^2 \\
&= V[\delta(X_1, ..., X_n)] + [E\left[\delta(X_1, ..., X_n)\right] - g(\theta)]^2 \\
&= V[\delta(X_1, ..., X_n)] + [\text{bias}(\delta(X_1, ..., X_n))]^2 \\
&= [\text{bias}(\delta(X_1, ..., X_n))]^2 + V[\delta(X_1, ..., X_n)]
\end{aligned}$$

as we sought to show. □

We now give some examples of unbiased estimators.

Example 7.0.1 (Sample Mean)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} P_\theta$ *and* $g(\theta) = E(X_i)$ *for any* $1 \leq i \leq n$, *then we find that* $\bar{X}$ *is an unbiased estimator for* $g(\theta)$. *This is because*

$$\begin{aligned}
E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} g(\theta) \\
&= g(\theta)
\end{aligned}$$

The MSE then becomes $MSE[\bar{X}] = V(\bar{X}) = V(X_i)/n$ *since the bias is 0.*

♥

Example 7.0.2 (Comparing Estimators)**.** *Suppose* $X_1, X_2, X_3 \overset{iid}{\sim} \exp(\theta)$ *where* $X_i =$

lifetimes of electronic components. We know the following from calculation

$$\hat{\theta}_{MLE} = \frac{3}{T} = \frac{3}{\sum_{i=1}^{3} X_i} = \frac{3}{X_1 + X_2 + X_3}$$
$$\hat{\theta}_{unbiased} = \frac{2}{X_1 + X_2 + X_3}$$
$$\hat{\theta}_{Bayes} = \frac{4}{2 + \sum_{i=1}^{3} X_i}$$

and

$$V\left(\hat{\theta}_{unbiased}\right) = \frac{2^2}{4}\theta^2 = \theta^2$$
$$V\left(\hat{\theta}_{MLE}\right) = \frac{9}{4}\theta^2$$

which can yield

$$MSE\left(\hat{\theta}_{unbiased}\right) = \theta^2$$
$$MSE\left(\hat{\theta}_{MLE}\right) = \frac{9}{4}\theta^2 + \frac{\theta^2}{4} = 2.5\theta^2$$

Notice that $MSE\left(\hat{\theta}_{unbiased}\right) < MSE\left(\hat{\theta}_{MLE}\right)$, suggesting that the unbiased statistic has a lower chance of error than the MLE. In practice $MSE\left(\hat{\theta}_{Bayes}\right)$ cannot be computed directly so we use simulation of the sampling distribution of $\hat{\theta}_{Bayes}$ to obtain it. We find, however, that $MSE\left(\hat{\theta}_{Bayes}\right) < MSE\left(\hat{\theta}_{unbiased}\right)$. So, in terms of MSE alone a Bayes' Estimate has lower error chance and an unbiased estimator is chosen above the MLE.

♥

7.1 Unbiased Estimation of Variance

Using the concept on unbiasedness, we can construct an unbiased estimate for any variance σ^2.

Theorem 7.1.1 (Unbiased Estimator for σ^2)**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$ and $g(\theta) = V(X_i) = \sigma^2$ for any $1 \leq i \leq n$, then*

$$\hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is unbiased for $g(\theta)$.

Proof. Consider $\sigma_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. We know that

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

and this allows us to write the form of $E\left(\sigma_0^2\right)$ as

$$\begin{aligned} E\left(\sigma_0^2\right) &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] - E\left[(\bar{X} - \mu)^2\right] \end{aligned}$$

Since $E(X_i) = \mu$ and $V(X_i) = E\left([X_i - E\left(X_i\right)]^2\right) = \sigma^2$ we have

$$\begin{aligned} E\left([X_i - \mu]^2\right) &= \sigma^2 \\ \implies E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] &= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \sigma^2 \end{aligned}$$

Similarly, $E(\bar{X}) = \mu$ and $V(\bar{X}) = \sigma^2/n$ since $X_1, ..., X_n$ are i.i.d and this implies

$$\begin{aligned} E\left[(\bar{X} - \mu)^2\right] &= \sigma^2/n \\ \implies E\left[\sigma_0^2\right] &= \sigma^2 - \sigma^2/n = \frac{n-1}{n}\sigma^2 \end{aligned}$$

To make an unbiased estimator, then, we multiply σ_0^2 by $\left(\frac{n-1}{n}\right)^{-1} = \left(\frac{n}{n-1}\right)$ to yield

$$\sigma_1^2 = \left(\frac{n}{n-1}\right)\sigma_0^2$$

as unbiased for σ^2. □

Example 7.1.1 (Sampling from Normal Distribution)**.** *If $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $\hat{\sigma}^2_{MLE}$ is a biased estimator for σ^2. We know this because*

$$\begin{aligned} \hat{\sigma}^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \hat{\sigma}_0^2 = \textit{biased from theorem above} \end{aligned}$$

♥

Sometimes, we can have multiple unbiased estimators as the following example illustrates.

Example 7.1.2 (Sampling from Poisson). *Suppose $X_1, ..., X_n \overset{iid}{\sim}$ Poisson(θ), then we know*

$$E(X_i) = \theta \qquad\qquad V(X_i) = \theta$$

If we wish to unbiasedly estimate θ we have

$$\begin{aligned} \hat{\theta}_1 &= \bar{X} && \text{(unbiased)} \\ \hat{\theta}_2 = \hat{\sigma}_1^2 &= \frac{1}{n-1}(X_i - \bar{X})^2 && \text{(unbiased)} \end{aligned}$$

and further for any α such that $-\infty \leq \alpha \leq \infty$ the estimator

$$\hat{\theta}_\alpha = \alpha\bar{X} + (1-\alpha)\hat{\sigma}_1^2$$

is unbiased as

$$\begin{aligned} E\left(\hat{\theta}_\alpha\right) &= E\left[\alpha\bar{X} + (1-\alpha)\hat{\sigma}_1^2\right] \\ &= \alpha\theta + (1-\alpha)\theta = \theta \end{aligned}$$

Hence, there can be many unbiased estimates; they are not unique.

♥

Remark 7.1.1. *Given that we have several choices/options for estimators that are unbiased, which one do we choose?*

$\longrightarrow$ *Typically, we choose the estimator with the smallest variance.*

♦

Example 7.1.3 (Minimum MSE $\hat{\sigma}^2$). *Suppose $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ and consider estimating σ^2. We know*

$$\begin{aligned} \hat{\sigma}_0^2 = \hat{\sigma}_{MLE}^2 &= \frac{1}{n}(X_i - \bar{X})^2 && \text{(biased)} \\ \hat{\sigma}_1^2 &= \frac{1}{n-1}(X_i - \bar{X})^2 && \text{(unbiased)} \end{aligned}$$

Does $\hat{\sigma}_1^2$ have a smaller MSE among all estimators? We can begin to get an answer if we look at the general case

$$T_c = c\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad (c > 0)$$

Recall that $\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \sigma^2\chi_{n-1}^2$ which implies that $c\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim c\sigma^2\chi_{n-1}^2$. This *makes its summary statistics as*

$$\begin{aligned} E\left(T_c\right) &= c\sigma^2(n-1) \\ V\left(T_c\right) &= c^2\sigma^4 2(n-1) \end{aligned}$$

by properties of the χ^2 *distribution. The MSE for* T_c *then becomes*

$$\begin{aligned} MSE(T_c) &= bias(T_c)^2 + V(T_c) \\ &= \left(E\left[T_c\right] - \sigma^2\right)^2 + V(T_c) \\ &= \left(c\sigma^2(n-1) - \sigma^2\right)^2 + c^2\sigma^4 2(n-1) \\ &= \sigma^4\left[(c(n-1)-1)^2 + 2c^2(n-1)\right] \end{aligned}$$

and is minimized by choosing $n = 1/n + 1$.

Therefore, the estimator of the variance with the lowest MSE is

$$T_{1/n+1} = \frac{1}{n+1}(X_i - \bar{X})^2$$

and this is true for all σ^2. *Our results show that* $T_{1/n}$ *and* $T_{1/n-1}$ *are both inadmissible because they are dominated, with respect to squared error loss, by* $T_{1/n+1}$. *In addition, C. Stein (1964) further showed that* $T_{1/n+1}$ *is also inadmissible.*

♥

7.2 Issues with Unbiasedness

While unbiasedness can help as a deciding factor for choosing estimates, it has limitations. For instance, unbiased estimators do not always exist for every statistical situation and they can sometimes lead to 'silly' estimators that do not make intuitive sense.

Example 7.2.1 (Geometric Distribution)**.** *Suppose* p = *chance of success,* X = *#of failures before a success and all trials are independent of each other, then* $X \sim Geo(p)$ *and the mass function is*

$$P(X = x) = p(1-p)^x \qquad (x = 0, 1, 2, ...)$$

An unbiased estimator $\hat{p}$ *of* p *is one such that* $E(\hat{p}) = p$. *One quick way to obtain* $\hat{p}$ *is to simply set* $\hat{p}$ *to 1 when* $x = 0$ *and 0 otherwise giving*

$$\hat{p} = \begin{cases} 1 & x = 0 \\ 0 & x > 0 \end{cases}$$

As verification,

$$E(\hat{p}) = \sum_{x=0}^{\infty} \hat{p}(x) \cdot p(1-p)^x = p + 0 = p$$

It can be shown that this is the only unbiased estimator. But there is something off about it. If we observe a success without any failures, then we think that a success will always happen, i.e. $p = 1$. Also, just observing a single failure before one success leads us to believe there is no chance of a success, i.e. $p = 0$ when we just had one success. This is counterintuitive.

♥

Example 7.2.2 (Poisson Distribution)**.** *Suppose $X \sim$ Poisson(λ) and we want to estimate $e^{-2\lambda}$ where $\lambda > 0$. We know*

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

and by examining $e^{-2\lambda}$'s Taylor series we can come up with the unbiased estimator

$$\widehat{e^{-2\lambda}} = \begin{cases} 1 & x \text{ even} \\ -1 & x \text{ odd} \end{cases} = (-1)^x$$

As verification, we note the Taylor series for $e^{-2\lambda}$ is

$$\begin{aligned} e^{-2\lambda} = e^{-\lambda}e^{-\lambda} &= \left[\sum_{x=0}^{\infty} \frac{(-\lambda)^x}{x!}\right] \cdot e^{-\lambda} \\ &= \sum_{x=0}^{\infty} \frac{(-1)^x(\lambda)^x e^{-\lambda}}{x!} \end{aligned}$$

and the expectation of $\widehat{e^{-2\lambda}}$ is

$$\begin{aligned} E\left(\widehat{e^{-\lambda}}\right) &= \sum_{x=0}^{\infty}(-1)^x \cdot e^{-\lambda}\frac{\lambda^x}{x!} \\ &= \sum_{x=0}^{\infty} \frac{(-1)^x\lambda^x e^{-\lambda}}{x!} \end{aligned}$$

Since both forms match, we have $E\left(\widehat{e^{-\lambda}}\right) = e^{-2\lambda}$ which fits the definition of an unbiased estimator.

♥

Chapter 8 — Fisher Information

The Fisher Information $\mathcal{I}(\theta)$ of a parameter θ measures how helpful an observation or set of observations are in determining (estimating) the true value of θ. The higher $\mathcal{I}(\theta)$ is, the more the data gives us an idea of what θ could be. The lower $\mathcal{I}(\theta)$ is, the less the data gives us an idea of what θ could be and we would need many more samples to correctly (within probability) estimate θ. To define the Fisher Information of a data point or a sample, we first examine the likelihood curve $f(x|\theta)$. To make the math easier, we take the logarithm of this curve and set it as $\lambda(x|\theta) = \log f(x|\theta)$. When this curve has high curvature (or high peaks), the data is helpful in estimating θ—low curvature means the opposite.

Intuitively, then, this would mean that the Fisher Information is related to $\frac{d\lambda(x|\theta)}{d\theta}$ in that higher values of it mean more information. Since $-\infty < \lambda'(x|\theta) < \infty$, we can square it and marginalize out the data (by expectation) to obtain some form of information the data point has on θ. Hence, we can set $\mathcal{I}(\theta) = E\left[(\lambda'(x|\theta))^2 \,|\theta\right]$. This is an intuitive explanation of the concept of 'Fisher Information'. We now formalize this idea below.

8.1 Fisher Information for a Single Random Variable

We can define Fisher Information as follows.

Definition 8.1.1 (Fisher Information for one r.v.). *Suppose $X \sim P_\theta : \theta \in \Omega$ and its p.d.f (or p.m.f.) is $f(x|\theta) > 0$ for all $x \in S, \theta \in \Omega$ for θ unknown. We further assume $\Omega \subset (a,b) \subset \mathbb{R}$ where (a,b) is an open interval on the real line $\mathbb{R}$. We then define*

$$\lambda(x|\theta) = \log[f(x|\theta)]$$

and assume f is twice differentiable in θ. This means that

$$\lambda'(x|\theta) = \frac{\partial\lambda(x|\theta)}{\partial\theta} \qquad (\textit{exists})$$

$$\lambda''(x|\theta) = \frac{\partial^2\lambda(x|\theta)}{\partial\theta^2} \qquad (\textit{exists})$$

We then define the **Fisher Information** *as*

$$\mathcal{I}(\theta) = E\left\{[\lambda'(x|\theta)]^2 \middle| \theta\right\}$$

in other words,

$$\mathcal{I}(\theta) = \int_{x\in S} [\lambda'(x|\theta)]^2 f(x|\theta)dx$$

♣

In practice, the above integral can be complex to calculate, so we have the following theorem.

Theorem 8.1.1 (Fisher Information Variance Form)**.** *If we assume we can calculate*

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\int_{x\in S}\lambda(x|\theta)dx &= \int_{x\in S}\frac{\partial^2\lambda(x|\theta)}{\partial\theta^2}dx\\ &= \int_{x\in S}\lambda''(x|\theta)dx\end{aligned}$$

and the integrals exist, then

$$\mathcal{I}(\theta) = -E\left[\lambda''(x|\theta)|\theta\right] = V[\lambda'(x|\theta)|\theta]$$

Proof. We will first show $E\left[\lambda'(x|\theta)|\theta\right] = 0$. We work from definition and note here that

$$\lambda'(x|\theta) = \frac{f'(x|\theta)}{f(x|\theta)}$$

and differentiation under the integral sign is allowed. We can then have

$$\begin{aligned}E\left[\lambda'(x|\theta)|\theta\right] &= \int_{x\in S}\frac{f'(x|\theta)}{f(x|\theta)}f(x|\theta)dx\\ &= \int_{x\in S} f'(x|\theta)dx\\ &= \frac{\partial}{\partial\theta}\int_{x\in S} f(x|\theta)dx\\ &= \frac{\partial}{\partial\theta}1 = 0\end{aligned}$$

We can now state $\lambda'(x|\theta) = \lambda'(x|\theta) - E\left[\lambda'(x|\theta)|\theta\right]$ and the definition of Fisher Information can be seen as

$$\begin{aligned}\mathcal{I}(\theta) &= E\left\{[\lambda'(x|\theta)]^2 \middle| \theta\right\} \\ &= E\left\{(\lambda'(x|\theta) - E[\lambda'(x|\theta)|\theta])^2 \middle| \theta\right\} \\ &= V[\lambda'(x|\theta)|\theta] && \text{(def. of variance)}\end{aligned}$$

For the next equality, we compute the form of $\lambda''(x|\theta)$ as follows

$$\begin{aligned}\lambda''(x|\theta) &= \frac{f(x|\theta)f''(x|\theta) - [f'(x|\theta)]^2}{[f(x|\theta)]^2} && \text{(quotient rule)} \\ &= \frac{f''(x|\theta)}{f(x|\theta)} - \lambda'(x|\theta)^2\end{aligned}$$

By this fact, we can see if we take expectations assuming some value of θ

$$\begin{aligned}E[\lambda''(x|\theta)|\theta] &= E\left[\frac{f''(x|\theta)}{f(x|\theta)}\middle|\theta\right] - E\left[\lambda'(x|\theta)^2\middle|\theta\right] \\ &= \int_{x\in S} \frac{f''(x|\theta)}{f(x|\theta)} f(x|\theta)dx - \mathcal{I}(\theta) \\ &= \int_{x\in S} f''(x|\theta)dx - \mathcal{I}(\theta) \\ &= \frac{\partial^2}{\partial\theta^2}1 - \mathcal{I}(\theta) \\ &= -\mathcal{I}(\theta)\end{aligned}$$

which implies

$$\mathcal{I}(\theta) = -E[\lambda''(x|\theta)|\theta]$$

too. This leaves $\mathcal{I}(\theta) = -E[\lambda''(x|\theta)|\theta] = V[\lambda'(x|\theta)|\theta]$ as we sought to show. □

Example 8.1.1 (Bernoulli $\mathcal{I}(\theta)$)**.** *Suppose $X \sim$ Bernoulli(p), then $f(x|\theta) = p^x(1-p)^{1-x}$ which implies that*

$$\begin{aligned}\lambda(x|p) &= x\log(p) + (1-x)\log(1-p) \\ \implies \lambda'(x|p) &= \frac{x}{p} - \frac{1-x}{1-p} \\ \implies \lambda''(x|p) &= -\frac{x}{p^2} - \frac{1-x}{(1-p)^2} = -\left(\frac{x}{p^2} + \frac{1-x}{(1-p)^2}\right)\end{aligned}$$

This then makes the Fisher Information as

$$\begin{aligned}\mathcal{I}(p) = -E\left\{\lambda''(x|p)\middle| p\right\} &= E\left[\frac{x}{p^2} + \frac{1-x}{(1-p)^2}\right] \\ &= \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}\end{aligned}$$

♥

8.1.1 Multiple Parameters

If we have multiple parameters $\theta = (\theta_1, ..., \theta_n)$ then we develop what is known as a **Fisher Information Matrix** $\boldsymbol{\mathcal{I}}_\theta$. For the two parameter case $\theta = (\theta_1, \theta_2)$, this matrix looks like

$$\boldsymbol{\mathcal{I}}_\theta = -E\left[\begin{pmatrix} \dfrac{\partial^2}{\partial\theta_1^2}\lambda(x|\theta) & \dfrac{\partial^2}{\partial\theta_1\partial\theta_2}\lambda(x|\theta) \\ \dfrac{\partial^2}{\partial\theta_1\partial\theta_2}\lambda(x|\theta) & \dfrac{\partial^2}{\partial\theta_2^2}\lambda(x|\theta)\end{pmatrix}\right]$$

and is a measure of how well the observation X can help us estimate θ_1, θ_2, and θ_1, θ_2 jointly. More generally, the Fisher Information Matrix for $\theta = (\theta_1, ..., \theta_n)$ can be given by

$$\begin{aligned}\boldsymbol{\mathcal{I}}_\theta &= E\left[\nabla_\theta\lambda(x|\theta)\nabla_\theta\lambda(x|\theta)'\right] \\ &= E\left[\begin{pmatrix}\dfrac{\partial}{\partial\theta_1}\lambda(x|\theta) \\ \vdots \\ \dfrac{\partial}{\partial\theta_n}\lambda(x|\theta)\end{pmatrix}\begin{pmatrix}\dfrac{\partial}{\partial\theta_1}\lambda(x|\theta) & \dots & \dfrac{\partial}{\partial\theta_n}\lambda(x|\theta)\end{pmatrix}\right] \\ &\overset{(*)}{=} \begin{pmatrix}\mathcal{I}(\theta_1) & \dots & \mathcal{I}(\theta_1, \theta_n) \\ \vdots & \ddots & \vdots \\ \mathcal{I}(\theta_n, \theta_1) & \dots & \mathcal{I}(\theta_n)\end{pmatrix}\end{aligned}$$

where the prime denotes vector transpose. The equality $(*)$ can be derived via the method [2] gave:

Theorem 8.1.2 (Fisher Information Hessian Form)**.** *The Fisher-Information Matrix can be written as*

$$\boldsymbol{\mathcal{I}}_\theta = \begin{pmatrix}\mathcal{I}(\theta_1) & \dots & \mathcal{I}(\theta_1, \theta_n) \\ \vdots & \ddots & \vdots \\ \mathcal{I}(\theta_n, \theta_1) & \dots & \mathcal{I}(\theta_n)\end{pmatrix}$$

where

$$\mathcal{I}(\theta_i,\theta_j) = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\lambda(x|\theta)\right] \qquad i,j\in\{1,\dots,n\}$$

in this case. Note that $\mathcal{I}(\theta_i,\theta_i)=\mathcal{I}(\theta_i)$ by definition.

Proof. We define $D_i = \frac{\partial}{\partial\theta_i}$ and $D_{i,j} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}$. With this we find out the form of any entry in $\boldsymbol{\mathcal{I}}_\theta$ or $D_{i,j}\lambda(x|\theta)$:

$$\begin{aligned} D_{i,j}\lambda(x|\theta) = D_i\left[D_j\lambda(x|\theta)\right] &= D_i\left[\frac{D_jf(x|\theta)}{f(x|\theta)}\right] \\ &= \frac{D_{i,j}f(x|\theta)}{f(x|\theta)} - \frac{D_if(x|\theta)}{f(x|\theta)}\frac{D_jf(x|\theta)}{f(x|\theta)} \qquad \text{(quotient rule)} \\ &\overset{(*)}{=} \frac{D_{i,j}f(x|\theta)}{f(x|\theta)} - \frac{\partial}{\partial\theta_i}\lambda(x|\theta)\frac{\partial}{\partial\theta_j}\lambda(x|\theta) \end{aligned}$$

If we allow differentiation under the integral sign, then we can compute the expectation of $\dfrac{D_{i,j}f(x|\theta)}{f(x|\theta)}$ as

$$\begin{aligned} E\left[\frac{D_{i,j}f(x|\theta)}{f(x|\theta)}\right] &= \int_{x\in S}\frac{D_{i,j}f(x|\theta)}{f(x|\theta)}f(x|\theta)dx \\ &= \int_{x\in S} D_{i,j}f(x|\theta)dx \\ &= D_{i,j}\int_{x\in S} f(x|\theta)dx \\ &= D_{i,j}1 = 0 \end{aligned}$$

With this, we are ready to compute the expectation of the equation $(*)$:

$$\begin{aligned} E\left[D_{i,j}\lambda(x|\theta)\right] &= E\left[\frac{D_{i,j}f(x|\theta)}{f(x|\theta)}\right] - E\left[\frac{\partial}{\partial\theta_i}\lambda(x|\theta)\frac{\partial}{\partial\theta_j}\lambda(x|\theta)\right] \\ &= -E\left[\frac{\partial}{\partial\theta_i}\lambda(x|\theta)\frac{\partial}{\partial\theta_j}\lambda(x|\theta)\right] = -\mathcal{I}(\theta_i,\theta_j) \\ &\implies \mathcal{I}(\theta_i,\theta_j) = -E\left[D_{i,j}\lambda(x|\theta)\right] = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\lambda(x|\theta)\right] \end{aligned}$$

as we sought to show. □

With our understanding of Fisher Information for a single random variable X, we can write the steps to finding such information:

Steps to Finding $\mathcal{I}(\theta)$

Step 1 Find the p.d.f or p.m.f. of X, $f(x|\theta)$

Step 2 Take its logarithm to get $\lambda(x|\theta) = \log\left[f(x|\theta)\right]$

Step 3 Take the second derivative of $\lambda(x|\theta)$ to get $\lambda''(x|\theta)$. For the two variable case, compute the mixed derivative $\dfrac{\partial^2}{\partial\theta_1\partial\theta_2}\lambda(x|\theta)$.

Step 4 Find the expectation of this quantity with respect to X: $E\left[\lambda''(x|\theta)\middle|\theta\right]$

Step 5 Flip the sign of this expectation, this is the Fisher Information $\mathcal{I}(\theta) = -E\left[\lambda''(x|\theta)\middle|\theta\right]$

Example 8.1.2 (Normal Distribution $\mathcal{I}_\theta$). *Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. We split our analysis into 3 cases.*

Case 1: μ unknown

- *In this case we know that*

$$f(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

which implies

$$\begin{aligned}\lambda(x|\mu) &= -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{(x-\mu)^2}{2\sigma^2}\\ \implies \lambda'(x|\mu) &= \frac{x-\mu}{\sigma^2}\\ \implies \lambda''(x|\mu) &= -\frac{1}{\sigma^2} && (a\ constant)\\ \implies \mathcal{I}(\mu) &= \frac{1}{\sigma^2}\end{aligned}$$

Case 2: σ^2 unknown

- *In this case we compute $\lambda''(x|\sigma)$ with respect to σ:*

$$\begin{aligned}\lambda'(x|\sigma^2) &= \frac{\partial}{\partial\sigma}\left\{-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{(x-\mu)^2}{2\sigma^2}\right\}\\ &= -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}\end{aligned}$$

Which then leads to

$$\lambda''(x|\sigma^2) = \frac{1}{\sigma^2} - \frac{3}{\sigma^4}(x-\mu)^2$$

This allows us to compute the Fisher Information as

$$\begin{aligned}\mathcal{I}(\sigma^2) = -E\left[\lambda''(x|\sigma^2)|\sigma^2\right] &= -\frac{1}{\sigma^2} + \frac{3}{\sigma^2}E\left[\left(\frac{x-\mu}{\sigma}\right)^2\right] \\ &= -\frac{1}{\sigma^2} + \frac{3}{\sigma^2} = \frac{2}{\sigma^2}\end{aligned}$$

Case 3: both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ unknown

- *In this case we create the Fisher Information Matrix. We evaluate* $\frac{\partial^2}{\partial\sigma^2}\lambda(x|\mu,\sigma^2)$ *and* $\frac{\partial^2}{\partial\mu^2}\lambda(x|\mu,\sigma^2)$ *which have the same forms as in the cases 1 and 2. In addition, we have to evaluate the mixed derivative* $\frac{\partial^2}{\partial\mu\partial\sigma}\lambda(x|\mu,\sigma)$. *We proceed as follows*

$$\begin{aligned}\frac{\partial^2}{\partial\mu\partial\sigma}\lambda(x|\mu,\sigma) &= \frac{\partial}{\partial\mu}\left[\frac{\partial}{\partial\sigma}\lambda(x|\mu,\sigma)\right] \\ &= \frac{\partial}{\partial\mu}\left[-\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}\right] \\ &= \frac{-2(x-\mu)}{\sigma^3}\end{aligned}$$

This gives us the joint Fisher Information as

$$\mathcal{I}(\mu,\sigma) = E\left[\frac{\partial^2}{\partial\mu\partial\sigma}\lambda(x|\mu,\sigma)\bigg|\mu,\sigma\right] = E\left[\frac{-2(x-\mu)}{\sigma^3}\bigg|\mu,\sigma\right] = 0$$

The Fisher Information Matrix becomes

$$\boldsymbol{\mathcal{I}}_{\mu,\sigma} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$$

♥

8.2 Fisher Information for a Random Sample

Now that we know how to compute the Fisher Information for a single sample, we can discuss how we would (in most practical applications) use it for a random sample $X_1, ..., X_n$. Since each sample provides $\mathcal{I}(\theta)$ amount of information, it would make sense that n of those samples contains the sum of all the information given or $n\mathcal{I}(\theta)$. We formalize the concept below.

Definition 8.2.1 (Fisher Information for a$X_1, ..., X_n$)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$, $f(x_i|\theta)$ *is the p.d.f or p.m.f. for each variate, and* $\theta \in \underbrace{(a, b)}_{\text{open interval}} \subset \Omega$. *We define* $\lambda(x_1, ..., x_n|\theta)$ *as*

$$\lambda(x_1, ..., x_n|\theta) = \log\left[f(x_1, ..., x_n)\right]$$

The Fisher Information for $X_1, ..., X_n$ *is then defined as*

$$\begin{aligned}\mathcal{I}_n(\theta) &= E\left\{\left[\lambda_n'(x_1, ..., x_n|\theta)\right]^2 \middle| \theta\right\} \\ &= \int_{X \in \mathcal{S}} \left[\lambda_n'(x_1, ..., x_n|\theta)\right]^2 f(x_1, ..., x_n|\theta)dX\end{aligned}$$

where $X = (X_1, ..., X_n)$ *and* $\mathcal{S} = S \times \cdots \times S$. *As before we can show*

$$\begin{aligned}\mathcal{I}_n(\theta) &= V\left[\lambda_n'(x_1, ..., x_n|\theta)\middle|\theta\right] \\ &= -E\left[\lambda_n''(x_1, ..., x_n|\theta)\middle|\theta\right]\end{aligned}$$

♣

Theorem 8.2.1 (Fisher Information for $X_1, ..., X_n$ Simplification)**.** *If* $\mathcal{I}(\theta) = -E\left[\lambda_n''(x|\theta)\middle|\theta\right]$ *is the Fisher Information for a single random variable* X *and set of parameters* $\theta = (\theta_1, ..., \theta_n)$, *then for a random sample* $X_1, ..., X_n$ *from this population we have*

$$\mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$$

Proof. We begin with the likelihood function $f(x_1, ..., x_n|\theta)$ and create a new form for it

$$f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

This implies

$$\begin{aligned}
\lambda_n(x_1, ..., x_n|\theta) &= \log\left[\prod_{i=1}^{n} f(x_i|\theta)\right] \\
&= \sum_{i=1}^{n} \log\left[f(x_i|\theta)\right] \\
&= \sum_{i=1}^{n} \lambda(x_i|\theta)
\end{aligned}$$

This can be used to give new forms for λ_n' and λ_n''

$$\begin{aligned}
\lambda_n'(x_1, ..., x_n|\theta) &= \sum_{i=1}^{n} \lambda'(x_i|\theta) \\
\lambda_n''(x_1, ..., x_n|\theta) &= \sum_{i=1}^{n} \lambda''(x_i|\theta)
\end{aligned}$$

We can now compute the Fisher Information

$$\begin{aligned}
\mathcal{I}_n(\theta) &= -E\left[\lambda_n''(x_1, ..., x_n|\theta)\middle|\theta\right] \\
&= -E\left[\sum_{i=1}^{n} \lambda''(x_i|\theta)\right] \\
&= \sum_{i=1}^{n} -E\left[\lambda''(x_i|\theta)\right] \\
&= n\mathcal{I}(\theta)
\end{aligned}$$

as we sought to show. Note that this works for mixed derivatives as well. □

Corollary 8.2.1 (Fisher Information Matrix for $X_1, ..., X_n$)**.** *If we take a random sample $X_1, ..., X_n$ with multiple parameters $\theta = (\theta_1, ..., \theta_n)$, then the Fisher Information Matrix $\boldsymbol{\mathcal{I}}_{\theta,n}$ is given by $\boldsymbol{\mathcal{I}}_{\theta,n} = n\boldsymbol{\mathcal{I}}_{\theta}$.*

Proof. Since $\mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$ and $\mathcal{I}_n(\theta_i, \theta_j) = n\mathcal{I}(\theta_i, \theta_j)$, we have

$$\begin{aligned}
\boldsymbol{\mathcal{I}}_{\theta,n} &= \begin{pmatrix} n\mathcal{I}(\theta_1) & \dots & n\mathcal{I}(\theta_1, \theta_n) \\ \vdots & \ddots & \vdots \\ n\mathcal{I}(\theta_n, \theta_1) & \dots & n\mathcal{I}(\theta_n) \end{pmatrix} \\
&= n\boldsymbol{\mathcal{I}}_{\theta}
\end{aligned}$$

as we sought to show. More samples, give us a directly proportional amount of information.

□

Example 8.2.1 (Customer Arrivals)**.** *Suppose we own a store and wish to study the rate at which people come in. We have two sampling plans*

- **Sampling Plan 1:** *fix $n = \#$ of customers and note the time it took to observe that many of them or $X =$ time until first n customers. If we assume a rate of θ customers arriving in a fixed unit of time and $W_i \sim \exp(\theta)$ inter-arrival times, we have*

$$X = \sum_{i=1}^{n} W_i \sim Gamma(n, \theta)$$

- **Sampling Plan 2:** *fix a time t and observe $Y = \#$ of customers arriving at time t. Then we have*

$$Y \sim Poisson(\theta t)$$

  where θ is the arrival rate (same one as with plan 1).

We can, with some calculation, prove that the Fisher Information for X and Y are

$$\mathcal{I}_X(\theta) = \frac{n}{\theta^2} \qquad\qquad \mathcal{I}_Y(\theta) = \frac{t}{\theta}$$

We can see from the above information that $\mathcal{I}_X(\theta) = \mathcal{I}_Y(\theta)$ only when $n = t \cdot \theta$. This means we have the same information only when the number of customers we wait for in plan 1 is equivalent to the amount of time we fixed in plan 2 multiplied by the rate θ that is the same for both plans. That is, the number of customers we wait for is the average amount that would come if we fix some time t.

♥

8.3 Cramér-Rao Lower Bound (Information Inequality)

Using the idea of Fisher Information, we can prove the minimum value a variance can have from some estimator $T = r(X_1, ..., X_n)$. This is known as the **Cramér-Rao Lower Bound** or **Information Inequality.**

Theorem 8.3.1 (Cramér-Rao Lower Bound (one parameter))**.** *Suppose $X_1, ..., X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$ where $\theta \in (a, b) \subset \Omega \subset \mathbb{R}$. We then define a statistic $T = r(X_1, ..., X_n)$ with finite variance and set*

$$m(\theta) = E(T) \qquad (m(\theta) \text{ differentiable in } \theta)$$

then,

$$V(T) \geq \frac{[m'(\theta)]^2}{n\mathcal{I}(\theta)}$$

Proof. We know

$$\lambda_n'(x_1, ..., x_n|\theta) = \frac{f'(x_1, ..., x_n|\theta)}{f(x_1, ..., x_n|\theta)}$$

This makes its average as

$$\begin{aligned} E\left[\lambda_n'(x_1, ..., x_n|\theta)\right] &= \int_{\mathcal{S}} f'(x_1, ..., x_n|\theta) dX \\ &= 0 \end{aligned}$$

Since $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$, we can state $\text{Cov}\left[\lambda_n'(x_1, ..., x_n|\theta), T\right]$ as

$$\begin{aligned} \text{Cov}\left[\lambda_n'(x_1, ..., x_n|\theta), T\right] &= E(T\lambda_n'(x_1, ..., x_n|\theta)) - E(T)\underbrace{E(\lambda_n'(x_1, ..., x_n|\theta))}_{0} \\ &= E(T\lambda_n'(x_1, ..., x_n|\theta)) \end{aligned}$$

From here we expand $E(T\lambda_n'(x_1, ..., x_n|\theta))$ to find another form for it

$$\begin{aligned} E(T\lambda_n'(x_1, ..., x_n|\theta)) &= \int_{\mathcal{S}} t\lambda_n'(x_1, ..., x_n|\theta) f(x_1, ..., x_n|\theta) dX \\ &= \int_{\mathcal{S}} tf'(x_1, ..., x_n|\theta) dX \\ &= \int_{\mathcal{S}} r(x_1, ..., x_n) f'(x_1, ..., x_n|\theta) dX \end{aligned}$$

If we allow differentiation under the integral sign, then we have

$$\begin{aligned} E(T\lambda_n'(x_1, ..., x_n|\theta)) &= \frac{\partial}{\partial\theta} \int_{\mathcal{S}} r(x_1, ..., x_n) f(x_1, ..., x_n|\theta) dX \\ &= \frac{\partial}{\partial\theta} E(T) = m'(\theta) \end{aligned}$$

Hence, $m'(\theta) = \text{Cov}\left[\lambda_n'(x_1, ..., x_n|\theta), T\right]$. If we note the Cauchy-Schwartz inequality,

$$|\text{Cov}(X, Y)|^2 \leq V(X)V(Y)$$

we can see that

$$\text{Cov}\left[\lambda_n'(x_1, ..., x_n|\theta), T\right]^2 \leq V(\lambda_n'(x_1, ..., x_n|\theta)) \cdot V(T)$$

but, $m'(\theta) = \text{Cov}\left[\lambda_n'(x_1, ..., x_n|\theta), T\right]$ so this is equivalent to

$$[m'(\theta)]^2 \leq \underbrace{V(\lambda_n'(x_1, ..., x_n|\theta))}_{\mathcal{I}_n(\theta)} \cdot V(T)$$

$$\implies V(T) \geq \frac{[m'(\theta)]^2}{\mathcal{I}_n(\theta)} = \frac{[m'(\theta)]^2}{n\mathcal{I}(\theta)}$$

as we sought to show. □

We arrive at a nice fact for the variance of unbiased estimators

Corollary 8.3.1 (Unbiased Estimators C-R Lower Bound)**.** *If* $T = r(X_1, ..., X_n)$ *is unbiased for* θ*, then*

$$V(T) \geq \frac{1}{n\mathcal{I}(\theta)}$$

Proof. Since T is unbiased, we have $E(T) = m(\theta) = \theta$ which implies that $m'(\theta) = 1$. Using the C-R bound we just proved we see

$$V(T) \geq \frac{1}{n\mathcal{I}(\theta)}$$

as we sought to show. □

Example 8.3.1 (Exponential Distribution)**.** *Suppose* $X_1, ..., X_n \overset{iid}{\sim} \text{exp}(\beta)$*, then we know* $f(x_i) = \beta e^{-\beta x_i}$ *for* $x_i, \beta > 0$*. We wish to estimate* β *with* $T = (n-1)/\sum X_i$ *and see how much it varies. For the denominator in* T *we know* $\sum X_i \sim$ *Gamma*(n, β) *and this lets us see the average and variance for* T *as*

$$\begin{aligned} E(T) &= (n-1)E\left[\frac{1}{\sum X_i}\right] = \cancel{n-1}\frac{\beta}{\cancel{n-1}} = \beta \\ V(T) &= (n-1)^2 V\left[\frac{1}{\sum X_i}\right] = \cancel{(n-1)^2}\frac{\beta^2}{\cancel{(n-1)^2}(n-2)} = \frac{\beta^2}{n-2} \end{aligned}$$

This shows T *is unbiased for* β *as well as its variance as* $\beta^2/(n-2)$*. Let's see if it meets the*

Cramér-Rao lower bound. First, we compute the Fisher Information for this sample

$$\begin{aligned}
&\lambda(x|\beta) = \log\beta - \beta x \\
&\implies \lambda'(x|\beta) = \frac{1}{\beta} - x \\
&\implies \lambda''(x|\beta) = -\frac{1}{\beta^2} \\
&\implies \mathcal{I}(\beta) = -E\left[\lambda''(x|\beta)|\beta\right] = \frac{1}{\beta^2} \\
&\implies \mathcal{I}_n(\beta) = \frac{n}{\beta^2}
\end{aligned}$$

Since, T is unbiased, we have $m'(\beta) = 1$ and the C-R inequality is

$$V(T) \geq \frac{1}{n\mathcal{I}(\theta)} = \frac{\beta^2}{n}$$

and indeed $V(T) = \beta^2/(n-2) \geq \beta^2/n$, but notice $V(T) \neq \beta^2/n$ and is higher than the lowest possible variance we can have for an estimator (the lower bound).

If we estimate β with $T^* = \bar{X} \sim Gamma(n,\beta)/n$, then we find

$$\begin{aligned}
E(T^*) &= m(\beta) = \frac{1}{\beta} \\
V(T^*) &= \frac{1}{n\beta^2}
\end{aligned}$$

Let's see what the C-R lower bound is. We know $m'(\beta) = -1/\beta^2$ and this makes the inequality

$$\begin{aligned}
V(T^*) \geq \frac{[m'(\beta)]^2}{\mathcal{I}_n(\beta)} &= \frac{[-1/\beta^2]^2}{n/\beta^2} \\
&= \frac{1}{n\beta^2}
\end{aligned}$$

But, $V(T^*) = 1/(n\beta^2)$, so $T^* = \bar{X}$ achieves the C-R lower bound and is the best estimator (in terms of variance) amongst all that have $m(\beta) = 1/\beta$.

♥

8.4 Efficient Estimators

When an estimator achieves its C-R lower bound, it is said to be *efficient*. More precisely, we define it as follows.

Definition 8.4.1 (Efficient Estimator)**.** *T is an* **efficient estimator** *of $m(\theta)$ when*

$$V(T) = \frac{[m'(\theta)]^2}{\mathcal{I}_n(\theta)} \qquad \forall \theta \in \Omega$$

♣

In the following discussion, we assume $X_1, ..., X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$ and $f(x|\theta)$ satisfies the conditions for the Information Inequality.

Theorem 8.4.1 (Efficient Estimator Distribution)**.** *Under the conditions given above and assuming $|m'(\theta)| > 0$, we have*

$$\frac{\sqrt{\mathcal{I}_n(\theta)}}{m'(\theta)}(T - m(\theta)) \sim \mathcal{N}(0, 1)$$

as $n \to \infty$.

Proof. We note that $\lambda'_n(x_1, ..., x_n|\theta) = \sum_{i=1}^n \lambda'(x_1, ..., x_n|\theta)$ and by the Lindeberg and Lévy central limit theorem, we can conclude

$$\frac{\lambda'_n(x_1, ..., x_n|\theta)}{\sqrt{\mathcal{I}_n(\theta)}} \sim \mathcal{N}(0, 1) \qquad n \to \infty$$

We also know

$$E(T) = m(\theta)$$
$$V(T) = \frac{[m'(\theta)]^2}{\mathcal{I}_n(\theta)} \qquad\qquad (T \text{ is efficient})$$

and since $V(\lambda'_n(x_1, ..., x_n|\theta)) = \mathcal{I}_n(\theta)$ as well as $V(T) = \rho(\mathcal{I}_n(\theta))$ for some function ρ, we must have $T = \epsilon(\lambda'_n(x_1, ..., x_n|\theta))$ for ϵ as

$$\epsilon(u(\theta), v(\theta)) = u(\theta)\lambda'_n(x_1, ..., x_n|\theta) + v(\theta)$$

For our purposes, we set

$$v(\theta) = E(T)$$
$$u(\theta) = \frac{m'(\theta)}{\mathcal{I}_n(\theta)}$$

This makes T as

$$T = \frac{m'(\theta)}{\mathcal{I}_n(\theta)}\lambda'_n(x_1, ..., x_n|\theta) + E(T)$$

which implies

$$\begin{aligned}\frac{T - E(T)}{m'(\theta)} &= \frac{\lambda_n(x_1, ..., x_n|\theta)}{\mathcal{I}_n(\theta)} \\ \implies \frac{\sqrt{\mathcal{I}_n(\theta)}}{m'(\theta)}(T - m(\theta)) &= \frac{\lambda_n(x_1, ..., x_n|\theta)}{\sqrt{\mathcal{I}_n(\theta)}} \qquad (E(T) = m(\theta))\end{aligned}$$

Which leads to

$$\frac{\sqrt{\mathcal{I}_n(\theta)}}{m'(\theta)}(T - m(\theta)) \sim \mathcal{N}(0, 1)$$

as we sought to show. □

Corollary 8.4.1 (Efficient MLE's). *For efficient MLE's $\hat{\theta}_n = \hat{\theta}_{MLE}$, we have as the immediate consequence*

$$\sqrt{\mathcal{I}_n(\theta)}(\hat{\theta}_{MLE} - \theta) \sim \mathcal{N}(0, 1)$$

Proof. Since we are talking about MLE's, $m(\theta) = \theta$ and $m'(\theta) = 1$. We then substitute these values for the ones in the above theorem. □

Side-Note: Even if the MLE is not efficient, we can prove that if it is determined by solving

$$\lambda_n'(x_1, ..., x_n|\theta) = 0$$

$\lambda''(x_1, ..., x_n|\theta)$, $\lambda'''(x_1, ..., x_n|\theta)$ exist and certain regulatory conditions are satisfied, then

$$\sqrt{\mathcal{I}_n(\theta)}(\hat{\theta}_{MLE} - \theta) \sim \mathcal{N}(0, 1) \qquad n \to \infty$$

such an MLE's are known as **asymptotically efficient**.

From a Bayesian Perspective, we can show that if $\hat{\theta}_n$ is the MLE and θ has a prior defined on it, that under large samples, the posterior has the following distribution

$$(\theta|X_1, ..., X_n) \sim \mathcal{N}\left(\hat{\theta}_n, 1/\mathcal{I}_n(\hat{\theta}_n)\right)$$

such a result is sometimes known as the '**Bayesian Central Limit Theorem**'.

8.5 Cramér-Rao for Multiple Parameters

We now expand on the information inequality for $\theta = (\theta_1, ..., \theta_n)$.

Theorem 8.5.1 (Cramér-Rao (Multi-Parameter))**.** *If $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta : \theta \in \Omega$ where $\theta = (\theta_1, \ldots, \theta_n) \in (a, b)^n \subset \Omega \subset \mathbb{R}^n$ and we then define a statistic $T = r(X_1, \ldots, X_n)$ with finite variance and set $E(T) = m(\theta)$ that is differentiable, then we have*

$$V(T) \geq [\nabla_\theta m(\theta)]' \boldsymbol{\mathcal{I}}_{n,\theta}^{-1} [\nabla_\theta m(\theta)] =$$

$$= \begin{pmatrix} \frac{\partial}{\partial \theta_1} \lambda(x|\theta) & \ldots & \frac{\partial}{\partial \theta_n} \lambda(x|\theta) \end{pmatrix} \boldsymbol{\mathcal{I}}_{n,\theta}^{-1} \begin{pmatrix} \frac{\partial}{\partial \theta_1} \lambda(x|\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_n} \lambda(x|\theta) \end{pmatrix}$$

Proof. Beyond scope of course. However, it can be completed if one knows some linear algebra. If you are interested, please see [1] for a more in-depth analysis. □

References

[1] John Duchi. *Chapter 8: Fisher Information*. 2020, p. 73. url: https://web.stanford.edu/class/stats311/Lectures/lec-09.pdf.

[2] Mark Reid. *Fisher Information and the Hessian of Log Likelihood*. Apr. 2012. url: http://mark.reid.name/blog/fisher-information-and-log-likelihood.html.