Part I

Building on Lecture Notes

6 Lec 6: Point estimation

6.1 Introduction

Suppose we have a set of random variables $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ and their observed values $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. We say that these are independently and identically distributed with a common p.f. or p.d.f. $f(\boldsymbol{x}; \boldsymbol{\theta})$, i.e.

$$\boldsymbol{X}_1, ..., \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} F_{\boldsymbol{\theta}}$$

6.2 Estimate and estimator

From Reference Notes
"We have a model of the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ in terms of the random variables $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ whose distribution is completely specified up to an unknown parameter vector $\boldsymbol{\theta}$. We wish to estimate $\boldsymbol{\theta}$ on the basis of $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ only. Specifically, we wish to find a function $\boldsymbol{T}$ of the data such that the vector $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ is close to the unknown $\boldsymbol{\theta}$. $\hat{\boldsymbol{\theta}}$ is called an **estimate** of $\boldsymbol{\theta}$. The corresponding *random variable* $\boldsymbol{T}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$ is called an **estimator** of $\boldsymbol{\theta}$. A function such as $\boldsymbol{T}$ above that only depends on the data but *not on any unknown parameter* is called a **statistic**."[1]

6.2.1 Unbiased estimators

From page 3 and PennState.[2]
An **unbiased estimator** satisfies

$$\mathbb{E}[\boldsymbol{T}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)] = \boldsymbol{\theta}, \forall \boldsymbol{\theta} \in \boldsymbol{\Omega} \tag{6.1}$$

That is, an unbiased estimator of $\boldsymbol{\theta}$ has an expectation of $\boldsymbol{\theta}$. We are concerned with finding unbiased estimators since we don't really want our estimators to be biased, but also we need this condition to apply most of the theorems we encounter later in the course haha.

Definition 6.1. The **bias** of an estimator $\boldsymbol{T}$ is defined as

$$\text{bias}(\boldsymbol{T}) = \mathbb{E}[\boldsymbol{T}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) - \boldsymbol{\theta}] \tag{6.2}$$

1. Geoff McLachlan, *Geoff's Lecture Notes for STAT2004 2021 Semester 2*, If ur not a student well good for you bro haha, 2021.
2. The Pennsylvania State University, *STAT 415 Introduction to Mathematical Statistics*, https://online.stat.psu.edu/stat415/lesson/introduction-stat-415, Of great assistance. 2021.

but since θ functions as a constant, this can also be thought of as

$$\text{bias}(\boldsymbol{T}) = \mathbb{E}[\boldsymbol{T}] - \boldsymbol{\theta}$$

Using this definition, we can construct unbiased estimators for new functions of θ using the following process. Since $\text{bias}(\boldsymbol{T}) = \mathbb{E}[\boldsymbol{T}] - \boldsymbol{\theta}$, we can rearrange this to $\mathbb{E}[\boldsymbol{T}] = \text{bias}(\boldsymbol{T}) - \boldsymbol{\theta}$. Defining $g(\theta) := \text{bias}(\boldsymbol{T}) - \boldsymbol{\theta}$, we now have that T is an unbiased estimator of $g(\theta)$. hax. This is used in tutorial 2 question 3.

6.3 Method of Moments

6.4 Likelihood function

From page 8, with assistance from the folks at PennState.[3]
Note: The following is almost directly quoted from Pennstate, with only formatting changes to suit Geoff's teaching.

6.4.1 Motivation and intuition

Suppose we have a random sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ whose assumed probability distribution depends on some unknown parameter $\boldsymbol{\theta}$. Our primary goal here will be to find a point estimator $U(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$, such that $U(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ is a "good" point estimate of $\boldsymbol{theta}$, where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are the observed values of the random sample. For example, if we plan to take a random sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ for which the $\boldsymbol{X}_i$ are assumed to be normally distributed with mean μ and variance σ^2, then our goal will be to find a good estimate of μ, say, using the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ that we obtained from our specific random sample.[4]

We could probably say that a good estimate of unknown parameter $\boldsymbol{\theta}$ would be the value of $\boldsymbol{\theta}$ that **maximises** the probability, i.e. the **likelihood**, of getting the data we observed. This is where we get the idea of **"maximum likehood"** (see section 6.8).

6.4.2 The meat

How do we even start to think about implementing this practically? Going back to our motivation, suppose we have a random sample of i.i.d. $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ with the joint p.f. or p.d.f. $f_{\boldsymbol{X}_1,...,\boldsymbol{X}_n}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta})$. Our **likelihood function** is basically the joint p.f./p.d.f. with a different name and notation, defined as follows:

Definition 6.2 (Likelihood function). The **likelihood function** for

3. The Pennsylvania State University, *STAT 415 Introduction to Mathematical Statistics*.
4. The Pennsylvania State University.

a set of i.i.d. random variables is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}) \tag{6.3}$$

Notice that we don't tend to write that the function considers $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ as variables. Whilst we consider the joint p.f. or p.d.f. as a function primarily of the observed values of our random sample, i.e. a function of $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, and also a function of $\boldsymbol{\theta}$, we consider $L(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ as the variable of interest, given that we already have $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, i.e. these values are not "varying". Alternatively, we consider $L(\boldsymbol{\theta})$ as a realisation of the random variable $L(\boldsymbol{\theta}; \boldsymbol{X}_1, ..., \boldsymbol{X}_n)$. We just consider the likelihood function as a function of $\boldsymbol{\theta}$ since we are only interested in maximising $L(\boldsymbol{\theta})$ by finding the best value of $\boldsymbol{\theta}$ for the job.

In the following definitions, we make use of the **log likelihood**, which is exactly what it sounds like: the log of the likelihood function, $\log L(\boldsymbol{\theta})$.

Definition 6.3 (Score statistic). The **score statistic** is defined as

$$S(\boldsymbol{x}_1, ..., \boldsymbol{x}_n); \theta) = \frac{\partial \log L(\theta)}{\partial \theta}, \tag{6.4}$$

where $\boldsymbol{\theta}$ is a scalar.

This is the derivative of the log likelihood. The score statistic is used to find many other things in statistics. We actually have a nice property of the single-variable score statistic: that the expectation is 0, i.e.

$$\mathbb{E}\left(\frac{\partial \log L(\theta)}{\partial \theta}\right) = 0.$$

For the multivariable case, you can find the score statistic as a vector with each entry as the first partial derivative in terms of each entry of $\boldsymbol{\theta}$.

Definition 6.4 (Fisher's expected information - single variable). When $\boldsymbol{\theta}$ is a scalar, **Fisher's expected information**, often referred to as just the **expected information** is defined as

$$\mathscr{I}(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{\partial \log L(\theta)}{\partial \theta}\right)^2\right], \tag{6.5}$$

or,

$$\mathscr{I}(\boldsymbol{\theta}) = \mathrm{Var}\left(\frac{\partial \log L(\theta)}{\partial \theta}\right). \tag{6.6}$$

I.e. for a single variable parameter, we have that the expected information is the expectation of the square of the score statistic. We obtain the second definition of the expected information from the fact that the expectation of the score is 0 (see above). We have an analogue definition for the multivariable case.

Definition 6.5 (Fisher's expected information - multivariable). When $\boldsymbol{\theta}$ is a vector,

$$\mathscr{I}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \log L(\theta)}{\partial \boldsymbol{\theta}} \cdot \frac{\partial \log L(\theta)}{\partial \boldsymbol{\theta}}^T\right], \tag{6.7}$$

also given by

$$\mathscr{I}(\boldsymbol{\theta}) = \mathrm{Cov}\left(\frac{\partial \log L(\theta)}{\partial \boldsymbol{\theta}}\right). \tag{6.8}$$

Here, we introduce $I(\boldsymbol{\theta})$, which becomes useful later in the maximum likelihood section.

$$I(\boldsymbol{\theta}) = -\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \tag{6.9}$$

That definition feels a bit like abuse of notation... idk that should be single variable stuff. For the multivariable case, this is also just the negative of the Hessian of the log likelihood function, which is what we are interested in for statistics. Anyway, given this, we have the following result which is technically a theorem?:

$$\mathscr{I}(\boldsymbol{\theta}) = \mathbb{E}[I(\boldsymbol{\theta})] \tag{6.10}$$

6.5 Regularity conditions (I)

Page 9/10.

Some of the results from the previous section only hold under the stuff Geoff lists in the notes, namely the following results:

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial \log L(\theta)}{\partial \theta}\right) &= 0 \\
\mathscr{I}(\theta) &= \mathbb{E}[I(\theta)] \\
\mathrm{Var}(T) &\geq \frac{(g'(\theta))^2}{\mathscr{I}(\theta)}
\end{aligned}$$

for any unbiased estimator T of the function $g(\theta)$. I could add references for these but like just scroll up my dude. I think these extend to multivariable cases too ? sort of... well at least the first one doesn't, the second one I'm pretty sure does and the third one would need covariance?

6.5.1 Cramér-Rao lower bound

Consider the unbiased estimator T of $g(\theta)$. We have that this estimator is unbiased, but wouldn't it be nice if it also had as little variance as possible? Luckily we have a handy dandy formula for the smallest possible value of the variance of any unbiased estimator.

Definition 6.6 (Cramér-Rao lower bound). The **Cramér-Rao lower bound**, or the **minimum variance bound (MVB)**, is the lower bound of the variance of some unbiased estimator T of $g(\theta)$,

$$\frac{(g'(\theta))^2}{\mathscr{I}(\theta)}, \tag{6.11}$$

satisfying the inequality

$$\text{Var}(T) \geq \frac{(g'(\theta))^2}{\mathscr{I}}., \tag{6.12}$$

So, if the variance of T attains this value, we know that it has the smallest variance possible. There are two ways we can determine whether T attains this lower bound:

The first method is introduced in a way that I don't really understand in the notes: "The unbiased estimator T of $g(\theta)$ **attains the MVB**, if and only if there is equality in the Cauchy-Schwarz inequality applied to the score statistic and T." What does this mean? Idk, but here is the conclusion we get from it

Theorem 6.1 (Factoring the score statistic). *If we can factor the score statistic as*

$$\frac{\partial \log L(\theta)}{\partial \theta} = k(\theta)(T - g(\theta)), \tag{6.13}$$

*where $k(\theta)$ is some function of only θ, i.e. it is not dependent on T, then we know that the variance of T **attains the MVB**. This is a necessary and sufficient condition for an estimator to attain the MVB.*

A fun thing from this is that we can find the expected information from this factored form, where

$$\mathscr{I}(\theta) = |k(\theta)g'(\theta)|, \tag{6.14}$$

or in the case where $g(\theta) = \theta$,

$$\mathscr{I}(\theta) = |k(\theta)|. \tag{6.15}$$

Another method we can use to determine whether an unbiased estimator T for $g(\theta)$ attains the MVB is to calculate its variance and see if it is equal to the expected information $\mathscr{I}(\theta)$ (see expected information section). I.e. calculate the variance of T, and the value of

Exercise 6.1. See tutorial 2 for useful exercises in this. Questions 1 (i) and (ii) cover factoring the score statistic. Question 5 is an exercise in determining whether an unbiased estimator attains the MVB using both of the above methods.

6.6 Regular exponential family

From page 15, with assistance from the folks at MIT,[5] *in Lecture 7.*
Consider the case where we have p-dimensional observations, $\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta}$, a d-dimensional parametor vector $\boldsymbol{\theta}$, and a q-dimensional sufficient statistic $\boldsymbol{T}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta})$ where $q \geq d$. The likelihood function $L(\boldsymbol{\theta})$ or equivalently the joint p.f. or p.d.f. $f(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta})$ belongs to the **d-parameter exponential family** if it has the form

$$f(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta}) = L(\boldsymbol{\theta}) = b(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \exp \boldsymbol{c}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta}) / a(\boldsymbol{\theta}), \tag{6.16}$$

where $bmc(\boldsymbol{\theta})$ is a $q \times 1$ vector function of $\boldsymbol{\theta}$ and $a(\boldsymbol{\theta})$ and $b(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ are non-negative scalar functions.

6.6.1 Complete???

I am not sure why it never says this anywhere in the notes explicitly, but being able to find that a likelihood function belongs to the regular exponential family shows that the corresponding $\boldsymbol{T}$ is a complete and sufficient statistic for $\boldsymbol{\theta}$ (see section 6.9 for more on sufficiency). I don't even understand what it means to be complete ngl (though there is a definition in the notes on p. 24), but we need for our statistics to be complete in order to apply them. I won't write a definition here because it's not even all that practical.

6.6.2 Parameter space

$\boldsymbol{X}_1, ..., \boldsymbol{X}_n$

6.7 UMVU estimators

From page 17, with assistance from the folks at Stanford,[6] *in Lecture 4 2016.*
An estimator is said to be **UMVU (uniform minimum variance unbiased)** estimator of $g(\theta)$ if its variance is a minimum for all values of θ in the class of all unbiased estimators of $g(\theta)$. The variance of a UMVU estimator does not necessarily have to attain the MVB, but it certainly can. If an unbiased estimator T attains the MVB, then it is a **MVB estimator** (it must attain the MVB for all values of θ).

Example 6.1. The example here is basically question 1 of assignment 2.

5. Peter Kempthorne, *Mathematical Statistics - Lecture Notes*, https://ocw.mit.edu/courses/mathematics/18-655-mathematical-statistics-spring-2016/lecture-notes/, Thank goodness for the internet. 2016.

6. Jiantao Jiao and Tsachy Weissman, *EE378A Statistical Signal Processing*, https://web.stanford.edu/class/ee378a/lecture-notes/, thx, 2016.

Note that if a function $g(\theta)$ of θ has an unbiased estimator that attains the MVB, then any other function of θ with an unbiased estimator that attains the MVB must be a linear function of $g(\theta)$.
To find a UMVU estimator or determine whether an estimator is UMVU, we have a couple of methods. A UMVU estimator can be found by implementing Theorem 2 from section 6.10.1.

6.8 Maximum Likelihood Method (ML)

Page 20, with help from PennState and that chonky textbook that costs $200 at the school locker (yikes).[7][8]
Let us think back to section 6.4 where we discussed the likelihood function. Why introduce the likelihood function separately to maximum likelihood estimation? I don't know. But, to reintroduce the idea of maximum likelihood estimation, I will quote myself:

"We could probably say that a good estimate of unknown parameter $\boldsymbol{\theta}$ would be the value of $\boldsymbol{\theta}$ that **maximises** the probability, i.e. the **likelihood**, of getting the data we observed. This is where we get the idea of **"maximum likehood"**."

So, recalling the definition of the likelihood function from equation 6.3,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}),$$

let's suppose we have some likelihood function $L(\boldsymbol{\theta})$ and proceed to define **maximum likelhood estimates and estimators**.

Definition 6.7 (Maximum likelihood estimate). If

$$\hat{\boldsymbol{\theta}} = \boldsymbol{U}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \tag{6.17}$$

maximises the corresponding likelihood $L(\boldsymbol{\theta})$, then $\hat{\boldsymbol{\theta}}$ is called the **maximum likelihood estimate**. This is also referred to as the **ML estimate** or simply the **MLE**.

In plainer terms, the maximum likelihood estimate is the value of $\boldsymbol{\theta}$ which makes the observed data most probable or "most likely" to occur.[9] From this, the definition of the **maximum likelihood estimator** follows.

Definition 6.8 (Maximum likelihood estimator). Given an MLE

7. The Pennsylvania State University, *STAT 415 Introduction to Mathematical Statistics*.
8. John Rice, *Mathematical Statistics and Data Analysis*, 3rd ed. (Cengage Learning, 2007).
9. Rice.

such as above (eq. 6.17), the corresponding random variable

$$\boldsymbol{U}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) \tag{6.18}$$

is called the **maximum likelihood estimator**, also referred to as the **ML estimator**.

How do we find these maximum likelihood things?

6.9 Sufficiency

From page 24, with assistance from the folks at PennState.[10]
A **statistic** is a function of the data alone which does not rely on any other parameter.

Definition 6.9 (Sufficient). Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be a random sample from a probability distribution with unknown parameter $\boldsymbol{\theta}$. Then, the statistic

$$\boldsymbol{T} = \boldsymbol{U}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$$

is said to be **sufficient** for $\boldsymbol{\theta}$ if the conditional distribution of $X_1, ..., X_n$, given the statistic $\boldsymbol{T}$, does not depend on the parameter $\boldsymbol{\theta}$.

This is kinda weird, yeah. In a sense, this means that $\boldsymbol{T}$ contains all the information about $\boldsymbol{\theta}$ contained in the observed sample $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$.

The definition hopefully makes more sense after the previous example. Though, in practice, finding the conditional distribution of $X_1, ..., X_n$ given $\boldsymbol{T}$ is neither convenient nor practical. As such, we do not often use the formal definition of sufficiency to identify or verify sufficient statistics. Instead, we employ the following theorem.

Theorem 6.2 (Fisher-Neyman factorisation theorem). *Given the random variables $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ with joint p.f. or p.d.f. $f(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta})$. A statistic $\boldsymbol{T}$ is **sufficient** for $\boldsymbol{\theta}$ if and only if the joint p.f. or p.d.f. can be written as*

$$f(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta}) = h_1(\boldsymbol{T}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta}); \boldsymbol{\theta}) h_2(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta}), \forall \boldsymbol{\theta} \in \Omega$$

So, if we are able to find such a factorisation for a joint p.f. or p.d.f., the $\boldsymbol{T}$ we have found is a sufficient statistic.

6.10 Rao-Blackwell Theorem

From page 30.

10. The Pennsylvania State University, *STAT 415 Introduction to Mathematical Statistics*.

Theorem 6.3 (Theorem 1 (Rao-Blackwell)). *Let $\boldsymbol{X}$ be a random variable with a p.f. or p.d.f. $f(x;\boldsymbol{\theta})$ and suppose that $\boldsymbol{T}$ is a complete, sufficient statistic for θ. If $U(\boldsymbol{T})$ is an unbiased estimator of θ with finite variance, then, $U(\boldsymbol{T})$ is a UMVU estimator of θ and it is unique (see Geoff's notes for particularities).*

6.10.1 Theorem 2.

Theorem 6.4 (Theorem 2). *Let $\boldsymbol{X}$ be a random variable with a p.f. or p.d.f. $f(x;\boldsymbol{\theta})$ and suppose that $\boldsymbol{T}$ is a complete, sufficient statistic for θ. If $U(\boldsymbol{T})$ is an unbiased estimator of θ with finite variance, then, $U(\boldsymbol{T})$ is a UMVU estimator of θ and it is unique (see Geoff's notes for particularities).*

Bibliography

Jiao, Jiantao, and Tsachy Weissman. *EE378A Statistical Signal Processing*. https://web.stanford.edu/class/ee378a/lecture-notes/. Thx, 2016.

Kempthorne, Peter. *Mathematical Statistics - Lecture Notes*. https://ocw.mit.edu/courses/mathematics/18-655-mathematical-statistics-spring-2016/lecture-notes/. Thank goodness for the internet. 2016.

McLachlan, Geoff. *Geoff's Lecture Notes for STAT2004 2021 Semester 2*. If ur not a student well good for you bro haha, 2021.

Rice, John. *Mathematical Statistics and Data Analysis*. 3rd ed. Cengage Learning, 2007.

The Pennsylvania State University. *STAT 415 Introduction to Mathematical Statistics*. https://online.stat.psu.edu/stat415/lesson/introduction-stat-415. Of great assistance. 2021.