STAT3001 Notes

Mathematical Statistics

Michaela Cheong

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

School of Mathematics and Physics
The University of Queensland, Australia

As of Feb 2022

Contents

Preface

README

These notes are a complement to the Lecture Notes for STAT3001 Semester 1 2022 and are largely quoted or paraphrased from the references notes written by Dirk Kroese and adapted by Geoff McLachlan for the Semester 2 2020 running of STAT2004, plus other sources reference throughout. You can access these notes on blackboard by searching for the 2020 running of this course. The lecture notes

assume knowledge of the basics of probability, however, statistical inference hinges upon many basic probability notions. So, if you have trouble with any foundation skills in probability, probably have a reference for that whilst reading this. Dirk/Geoff's original reference notes are quite comprehensive anyway, if you can get access to those, or took STAT1201/1301/2003. The only issue with referring to the 2020 Reference notes for STAT2004 is that Geoff uses vectors for every formula in 2022, so that's something to keep in mind when referring to the 2020 notes. These notes also provide pointers to where concepts have been used in tutorial or

assignment questions in STAT2004 as of now, but these references will be updated for STAT3001. It's not rigorous right now bc I haven't been very diligent with citing that stuff but I aim for it to be rigorous. I'd like to note that these notes are

wonderful but they took a lot of time. If you know me, shout me a coffee sometime because I don't have a ko-fi hahaha.

How to read my notes

Here are some of the different text environments you will encounter:

Theorem 0.1 (THEOREM NAME). *The statement of a theorem.*

Proof. Here is where we might prove a property or theorem. ☐

Definition 0.1 (TERM BEING DEFINED). Here is a definition of a term.

Many terms will be introduced in the context of discussion, in which the key term will be **bolded** like so. This is done especially for simpler ideas which may have fewer mathematical properties to be investigated or used.

> **Example 0.1 (Name of example).**   Examples will either be presented as individual examples or as a list of examples.

Similarly, non-rigorous examples may simply be enumerated rather than given their own example environment.

1. Here is a simple example.

2. And here is another.

3. Since these are simple, it would not be of much benefit to give them a reference of their own.

> **Exercise 0.1 (Name of exercise).**   These are exercises left to the reader. I may add solutions in the appendix but ceebs tbh. These are also environments I may used to flag where a concept was used in a tutorial or assignment.

These are self-explanatory for the most part. Also, most things in this document are hyperlinked :)

Part I

Geoff's half of the course

1 | Background: Point estimation

1.1 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 pmf or pdf $f(\boldsymbol{x}; \boldsymbol{\theta})$, i.e.

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

1.2 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]

1.2.1 Error

Given

$$\mathbb{E}\left[|T(\boldsymbol{X}) - \theta|^a\right], \tag{1.1}$$

we call equation 1.1 the **average mean absolute error** if $a = 1$, or the **mean squared error** if $a = 2$.

1.2.2 Unbiased estimators

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

$$\mathbb{E}[\boldsymbol{T}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)] = \boldsymbol{\theta}, \forall \boldsymbol{\theta} \in \Omega \tag{1.2}$$

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,

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 also we need this condition to apply most of the theorems we encounter later in the course haha.

> **Definition 1.1.** The **bias** of an estimator $\boldsymbol{T}$ is defined as
>
> $$\text{bias}(\boldsymbol{T}) = \mathbb{E}[\boldsymbol{T}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) - \boldsymbol{\theta}] \tag{1.3}$$
>
> 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.

1.3 Method of Moments

With assistance from the folks at PennState, chapter 1.4.[3]
In order to use the method of moments, we take k sample moments and equate these to k population moments, also referred to as theoretical moments. We determine k by the number of unknown parameters we wish to estimate. The kth **sample moment** is defined as $\frac{1}{n}\sum_i^n X_i^k$. The kth **theoretical moment** is defined as $\mathbb{E}(X^k)$, and is a function of $\boldsymbol{\theta}$.

1.4 Likelihood function

From page 8, with assistance from the folks at PennState.[4]
Note: The following is almost directly quoted from Pennstate, with only formatting changes to suit Geoff's teaching.

1.4.1 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.[5]

---

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

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 1.8).

1.4.2   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 pmf or pdf $f_{\boldsymbol{X}_1,...,\boldsymbol{X}_n}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta})$. Our **likelihood function** is basically the joint pmf/pdf with a different name and notation, defined as follows:

> **Definition 1.2 (LIKELIHOOD FUNCTION).**   The **likelihood function** for a set of i.i.d. random variables is given by
>
> $$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}) \tag{1.4}$$

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 pmf or pdf 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 1.3 (SCORE STATISTIC).**   The **score statistic** is defined as
>
> $$S(\boldsymbol{x}_1, ..., \boldsymbol{x}_n); \theta) = \frac{\partial \log L(\theta)}{\partial \theta}, \tag{1.5}$$
>
> 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 1.4 (FISHER'S EXPECTED INFORMATION - SINGLE VARIABLE). When θ is a scalar, **Fisher's expected information**, often referred to as just the **expected information** is defined as

$$\mathscr{I}(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{\partial \log L(\theta)}{\partial \theta}\right)^2\right], \tag{1.6}$$

or,

$$\mathscr{I}(\boldsymbol{\theta}) = \text{Var}\left(\frac{\partial \log L(\theta)}{\partial \theta}\right). \tag{1.7}$$

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 1.5 (FISHER'S EXPECTED INFORMATION - MULTIVARIABLE). When $\boldsymbol{\theta}$ is a vector,

$$\mathscr{I}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot \left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right], \tag{1.8}$$

also given by

$$\mathscr{I}(\boldsymbol{\theta}) = \text{Cov}\left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right). \tag{1.9}$$

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

$$I(\boldsymbol{\theta}) = -\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \tag{1.10}$$

The definition of I feels a bit like abuse of notation... So, if it's too uncomfortable, think about it based on the Hessian definition. It's worth noting that we don't care about the positive Hessian: the negative of the Hessian is what we are interested in for statistics. Also, it's funny, because we don't have a name for this, but we do have a name for the mathematical object $I(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$.

Definition 1.6 (OBSERVED INFORMATION MATRIX - MULTIVARIABLE). When $\boldsymbol{\theta}$ is a vector, $I(\hat{\boldsymbol{\theta}})$ is called the **observed information matrix**. More simply, it is the negative of the Hessian of the log likelihood function with the ML estimate substituted into it.

Anyway, given this and under regularity conditions (see section section below), we have the following result which is technically a theorem and holds for distributions belonging to the regular exponential family (see section on regular exponential family):

$$\mathscr{I}(\boldsymbol{\theta}) = \mathbb{E}[I(\boldsymbol{\theta})] \tag{1.11}$$

1.5 Regularity conditions (I)

Page 9/10.

Here is a list of regularity conditions:
(literally in the notes, will add when I can be bothered, otherwise there is a nice explanation for them here).
Some of the results from the previous section only hold under the regularity conditions Geoff lists in the notes, namely the following results:

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

for any unbiased estimator T of the function $g(\theta)$. I think these extend to multivariable cases too ? sort of... well at least the first one might, the second one I'm pretty sure does and the third one would need covariance? There is a proof of the multivariate case in Lecture Note 1 for STAT3001 Sem 1 2022, starting at the end of page 1.

1.5.1 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 1.7 (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 $g(\theta)$,
>
> $$\frac{(g'(\theta))^2}{\mathscr{I}(\theta)}, \tag{1.12}$$
>
> satisfying the inequality
>
> $$\mathrm{Var}(T) \geq \frac{(g'(\theta))^2}{\mathscr{I}}., \tag{1.13}$$

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 1.1 (FACTORING THE SCORE STATISTIC).** *If we can factor the score statistic as*
> $$\frac{\partial \log L(\theta)}{\partial \theta} = k(\theta)(T - g(\theta)), \tag{1.14}$$
> *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
$$\mathscr{I}(\theta) = |k(\theta)g'(\theta)|, \tag{1.15}$$
or in the case where $g(\theta) = \theta$,
$$\mathscr{I}(\theta) = |k(\theta)|. \tag{1.16}$$

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 1.1.** 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.

1.6 Regular exponential family

From page 15, with assistance from the folks at MIT,[6] in Lecture 7.
Consider the case where we have p-dimensional observations, $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, a d-dimensional parameter 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 pmf or pdf $f(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta})$ belongs to the d-**parameter exponential family** if it has the form

$$L(\boldsymbol{\theta}) = f(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta}) = b(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \exp \boldsymbol{c}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta})/a(\boldsymbol{\theta}), \tag{1.17}$$

where $\boldsymbol{c}(\boldsymbol{\theta})$ 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. 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.

1.6.1   Parameter space

1.6.2   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 1.9 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 different theorems, such as Rao-Blackwell theorem 1.3, since completeness guarantees uniqueness in certain statistical procedures based on $\boldsymbol{T}$. I won't write a definition here because it's not even all that practical. Actually, just kidding, I will.

> **Definition 1.8 (COMPLETE).**   A sufficient statistic $\boldsymbol{T}$ is said to be **complete** if
> $$\mathbb{E}_{\boldsymbol{\theta}}[w(\boldsymbol{T})] = \boldsymbol{0}, \forall \boldsymbol{\theta} \in \boldsymbol{\Omega} \Rightarrow w\boldsymbol{T} \equiv \boldsymbol{0} \tag{1.18}$$
> except perhaps on a set of probability zero.

1.7   UMVU estimators

From page 17, with assistance from the folks at Stanford,[7] 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 1.1.**   The example here is basically question 1 of assignment 2.

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 1.10. Also, I think that if an estimator hits the Cramér-Rao lower bound for all values of θ, then it is also UMVU, but UMVU estimators don't necessarily need to have this as a property. So, you can factor the score statistic and then be like "the estimator must be UMVU".

---

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

1.8 Maximum Likelihood Method (ML)

Page 20, with help from PennState and that chonky textbook that costs $200 at the school locker (yikes).[89]

Let us think back to section 1.4 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 1.4,

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

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

> **Definition 1.9 (Maximum likelihood estimate).** If
>
> $$\hat{\boldsymbol{\theta}} = U(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \tag{1.19}$$
>
> 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.[10] From this, the definition of the **maximum likelihood estimator** follows.

> **Definition 1.10 (Maximum likelihood estimator).** Given an MLE such as above (eq. 1.19), the corresponding random variable
>
> $$U(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) \tag{1.20}$$
>
> is called the **maximum likelihood estimator**, also referred to as the **ML**

---

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

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

10. Rice.

> **estimator**.

How do we find these maximum likelihood things? Well, we can equate the score statistic to 0 and substitute θ for $\hat{\theta}$ to find the MLE. This actually makes a lot of sense since the score statistic is a derivative of the likelihood, so we are literally just using classic optimisation/maximisation techniques when using the ML method. When asked to find the estimator as opposed to the estimate, simply replace any observed values x_i with their corresponding random variables X_i.

> **Exercise 1.3.** Here are a bunch of places we do this: Tutorial 2 question 1 part (i) and (ii), tutorial 2 question 5, assignment 1 question (iii) and (v), assignment 2 question (b) part (ii).

1.9 Sufficiency

From page 24, with assistance from the folks at PennState.[11]
A **statistic** is a function of the data alone which does not rely on any other parameter.

> **Definition 1.11 (Sufficient).** Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be a random sample from a probability distribution with unknown parameter $\boldsymbol{\theta}$. Then, the statistic
>
> $$\boldsymbol{T} = U(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$$
>
> 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 1.2 (Fisher-Neyman factorisation theorem).** *Given the random variables $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ with joint pmf or pdf $f(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta})$. A statistic $\boldsymbol{T}$ is* **sufficient** *for $\boldsymbol{\theta}$ if and only if the joint pmf or pdf 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 pmf or pdf, the $\boldsymbol{T}$ we have found is a sufficient statistic.

---

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

1.10 I am not sure what to call this section

From page 30.

> **Theorem 1.3 (THEOREM 1 (RAO-BLACKWELL)).** *Let $\boldsymbol{X}$ be a random variable with a pmf or pdf $f(x; \boldsymbol{\theta})$ 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).*

The following theorem is referred to as the **Lehmann-Scheffé Theorem**.

> **Theorem 1.4 (THEOREM 2).** *Let $\boldsymbol{X}$ be a random variable with a pmf or pdf $f(x; \boldsymbol{\theta})$ 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).*

> **Theorem 1.5 (BASU'S THEOREM).** *Let $\boldsymbol{T}$ be a complete, sufficient statistic for $\boldsymbol{\theta}$. If the distribution of some statistic $V(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$ does not depend on $\boldsymbol{\theta}$, then $\boldsymbol{V}$ is said to be an **ancillary** statistic, i.e. a statistic that is distributed independently of $\boldsymbol{T}$.*

1.11 Jensen's Inequality

1.12 Kullback-Leibler Distance

1.13 Large Sample Theory

1.14 Consistency

> **Definition 1.12 (CONSISTENT).** A sequence of estimators T_n of $g(\boldsymbol{\theta})$ is said to be **consistent** if for every $\boldsymbol{\theta} \in \Omega$,
>
> $$T_n \overset{P_\theta}{\to} g(\boldsymbol{\theta}) \text{ as } n \to \infty;$$
>
> that is, given any $\varepsilon > 0$, then
>
> $$\mathbb{P}\left(|T_n(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) - g(\boldsymbol{\theta})| \geq \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

Theorem 1.6 (CONSISTENT). *If $Var(T_n) \to 0$ and $bias(T_n) \to 0$ as $n \to \infty$, then the sequence of estimates T_n is consistent for estimating $g(\boldsymbol{\theta})$.*

Proof. Given in the notes. $\square$

Exercise 1.4. Tutorial 2 question 4 asks you to present this proof, essentially.

1.15 Large-sample comparisons of estimators

1.16 Asymptotic efficiency

1.17 Maximum likelihood theorems

Page 48, a.k.a. "SOME THEOREMS THAT PROVIDE A BASIS FOR MAXIMUM LIKELIHOOD".

1.18 Asymptotic distribution theorems

Page 53, a.k.a. "Some further theorems concerning asymptotic distributions".

2 | Likelihood Equation: Regular Exponential Family

It just gets repeated here by Geoff.

Part II

Sharon's half of the course

Part III

Stuff from STAT2003, STAT2004 and such

3 | Formulas

From STAT2003 (thanks Nick if ur reading this because this is just all rewritten from your compilation). Under construction.

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

2. Inclusion-exclusion (2 sets): $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)\mathbb{P}(A \cap B)$

3. De Morgan's laws: $\mathbb{P}(A^c \cup B^c) = \mathbb{P}(A \cap B)^c, \mathbb{P}(A^c \cap B^c) = \mathbb{P}(A \cup B)^c$

4. Conditional probability: $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)\mathbb{P}(B)$

5. Law of total probability: $\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$, given $B_1, ..., B_n$ is a partition of Ω

6. Bayes' Rule: $\mathbb{P}(B_j|A) = \mathbb{P}(A|B_j)\mathbb{P}(B_j)Pni = 1P(A|B_i)\mathbb{P}(B_i)$

7. Sum rule: $\mathbb{P}(A) = \sum_{i=1}^{n} P(A \cap B_i)$

8. Product rule: $\mathbb{P}(A_1...A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)...P(A_n|A_1...A_{n-1})$

9. A_i independent $\iff$ $\forall k$ and any choice of $i_1, ..., i_k$, $\mathbb{P}(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2})...\mathbb{P}(A_{i_k})$

10. CDF of X: $F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)du, x \in \mathbb{R}$

11. CDF properties: $\lim_{x \to \infty} F(x) = 1$, $\lim_{x \to -\infty} F(x) = 0$, $\mathbb{P}(a < X \leq b) = F(b) - F(a) = \int_{b}^{a} f(u)du$

12. Pdf of X: $f(x) = F'(x), \int_{-\infty}^{\infty} f(x)dx = 1$

13. Pmf of X: $f(x) = \mathbb{P}(X = x)$

14. For discrete X, $\mathbb{P}(X \in B) = \sum_{X \in B} \mathbb{P}(X = x)$, i.e., $\mathbb{P}(X = 1, 2) = \sum_{x=1}^{2} \mathbb{P}(X = x)$

15. Expectation (discrete): $\mathbb{E}[g(X)] = \sum_x g(x)\mathbb{P}(X = x) = \sum_x g(x)g(x)f(x)$

16. Expectation (continuous): $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$

Important discrete distributions					
Distribution	pmf	$x \in$	$\mathbb{E}(X)$	$\text{Var}(X)$	PGF
$\text{Ber}(p)$	$p^x(1-p)^{1-x}$	$\{0,1\}$	p	$p(1-p)$	1 - p + zp
$\text{Bin}(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$\{0,1,...,n\}$	np	$np(1-p)$	$(1-p+zp)^n$
$\text{Geom}(p)$	$p(1-p)^{x-1}$	$\{1,2,...\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{zp}{1-z(1-p)}$
$\text{Hyp}(n,r,N)$	$\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$	$\{0,1,...\}$	$n\frac{r}{N}$	$n\frac{r}{N}\frac{(N-r)}{N}\frac{N-n}{N-1}$	

Important continuous distributions						
Distribution	pdf	cdf	$x \in$	$\mathbb{E}(X)$	$\text{Var}(X)$	PGF/MGF
$\text{U}[a,b]$	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$[a,b]$	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$	$\frac{e^{bs}-e^{as}}{s(b-a)}$
$\text{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$1 - e^{\lambda x}$	$\mathbb{R}^+$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-s}, s < \lambda$
$\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$		$\mathbb{R}$	μ	σ^2	$e^{s\mu+s^2\sigma^2/2}, s \in \mathbb{R}$