STAT2004 Notes

Statistical Modelling and Analysis

Michaela Cheong

Adapted to Geoff McLachlan's STAT2004 2021 Sem 2 delivery,
based on reference notes by Geoff McLachlan and Dirk Kroese.

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

School of Mathematics and Physics
The University of Queensland, Australia

As of August 2021

Contents

Preface

README

These notes are a complement to the Lecture Notes for STAT2004 Semester 2 2021 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 this course. You can access these notes on blackboard by searching for the 2020 running of this course. These notes also borrow terminology from the Lecture notes from 2020, such as the introduction of sigma algebras for the abstraction/generalisation of events. The lecture notes assume knowledge of the basics of probability, however,

statistical inference hinges upon many basic probability notions. The second half of these notes contain the foundations of probability and go through some more rigorous formulas and theorems that will be needed. Well, I did some of that stuff since I think sigma algebra notation is useful, however, I did end up abandoning making that section of notes since Dirk/Geoff's original reference notes are quite comprehensive anyway. The only issue with referring to the 2020 Reference notes is that Geoff uses vectors for every formula in 2021, 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. 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 Lec 1: Introduction to statistical inference

STAT2004 is a course focussed on **statistical inference**. A framework for understanding statistical inference is given in the lecture notes.

2 Lec 2: Some examples in calculating probabilities with rolling of dice

2.1 Is probability easier now than in 1560? (Carden's Counting)

In this worked example, we investigate the key problem:

"How many ways can we roll three dice such that we obtain a sum of 3?"

This means that the set of three dice rolled either contains a three on one of its faces (e.g. the outcome $(3, ..., ...)$), or that any combination of the faces sum to three (e.g. $(2, 1, x)$).

In order to solve this question, we must employ combinatoric strategies.

2.1.1 Combinatorics

Some key issues that come into play when trying to solve this question include overcounting or missing cases.

2.2 Efron's Dice

3 Lec 3: Examples using Baye's theorem

Definition 3.1 (Conditional Probability). $\mathbb{P}(A|B)$ denotes the **conditional probability** of A if B has occured, for $A, B \in \mathcal{F}$, and is defined by

$$\mathbb{P}(A|B) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{3.1}$$

A helpful interpretation of this definition is that in order for A to occur, if B is given then the event must lie within the intersection of A and B. Then, we incorporate the information from B already having occured. Or, we consider B our new "sample space" and are looking for the likelihood that the event lands in the event A.

Next, we introduce the **Law of Total Probability**.

Theorem 3.1 (Law of Total Probability). *If $\{B_1, ..., B_n\}$ be a partition of Ω with $\mathbb{P}(B_i) > 0$ for all i, then*

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i). \tag{3.2}$$

When we combine the Law of Total Probability (equation 3.2) together with the definition of conditional probability (equation 3.1), we obtain the following:

Theorem 3.2 (Baye's Theorem). *Let $\{B_1, ..., B_n\}$ be a partition of Ω with $\mathbb{P}(B_i) > 0$ for all i and assume $\mathbb{P}(A) > 0$. Then,*

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(B_j)\mathbb{P}(A|B_j)}{\sum_{i=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A|B_i)}. \tag{3.3}$$

Now, we consider Baye's Theorem in terms of applications to cancer screenings. Here, $B = \{\text{diagnosis is positive}\}$, $A_i = \{\text{the } i^{\text{th}} \text{ person has cancer}\}$.

4 Lec 4: Some important probability density functions

Some probability distributions						
Distribution	Notation	pmf/pdf	$x \in$	$\mathbb{E}(X)$	$\mathrm{Var}(X)$	PGF/MGF
Uniform	$\mathrm{U}[a,b]$	$\frac{1}{b-a}$	$[a,b]$	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$	$\frac{e^{bs}-e^{as}}{s(b-a)}$
Exponential	$\mathrm{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$\mathbb{R}^+$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-s}, s<\lambda$
Normal, Gaussian	$\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$	$\mathbb{R}$	μ	σ^2	$e^{s\mu+s^2\sigma^2/2}, s\in\mathbb{R}$

note: needs fixing, I know.

5 Lec 5: Examples of inference from Efron

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

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

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 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 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 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 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 pmf or pdf $f(\boldsymbol{x}_1, ..., \boldsymbol{x}_n; \boldsymbol{\theta})$ 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 $\boldsymbol{c}(\boldsymbol{\theta})$ 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.

6.6.1 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.

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

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

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

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

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

How do we find these maximum likelihood things? Well, we can equate the score statistic to 0 and substitute $\boldsymbol{\theta}$ for $\hat{\boldsymbol{\theta}}$ 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 6.2. 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).

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

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

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 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*

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

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.

6.10 I am not sure what to call this section

From page 30.

Theorem 6.3 (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 6.4 (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 6.5 (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 $\boldsymbol{T}$.*

6.11 Jensen's Inequality

6.12 Kullback-Leibler Distance

6.13 Large Sample Theory

6.14 Consistency

Definition 6.10 (Consistent). A sequence of estimators T_n of $g(\boldsymbol{\theta})$ is said to be **consistent** if for every $\boldsymbol{\theta} \in \Omega$,

$$T_n \xrightarrow{P_{\boldsymbol{\theta}}} g(\boldsymbol{\theta}) \text{ as } n \to \infty;$$

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 6.6 (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. □

Exercise 6.3. Tutorial 2 question 4 asks you to present this proof, essentially.

6.15 Large-sample comparisons of estimators

6.16 Asymptotic efficiency

6.17 Maximum likelihood theorems

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

6.18 Asymptotic distribution theorems

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

Part II

Alan's half of the course

Part III

Stuff from STAT2003 and such

7 Random experiments and probability models

7.1 Random experiments

The basic notion in probability is that of a **random experiment**: an experiment whose outcome cannot be determined in advance, but is nevertheless subject to analysis. Some examples of random experiments are:

1. tossing a die,
2. counting the number of ducks walking along a lake in a given hour, or
3. measuring the length of 5 footlong subs purchased from a Subway.

We wish to describe these experiments via a mathematical model. This model consists of three building blocks: a *sample space*, a set of *events* and a *probability*. We will now describe each of these building blocks.

7.2 Sample spaces

Although we cannot predict the outcome of a random experiment with certainty we usually can specify a set of possible outcomes. This gives the first ingredient in our model for a random experiment.

> **Definition 7.1 (Sample space).** The **sample space Ω** of a random experiment is the set of all possible outcomes of the experiment.

The following are examples of random experiments with their sample spaces.

1. Casting two die consecutively,
$$\Omega = \{(1,1),(1,2),...,(1,6),(2,1),...,(6,6)\}.$$
2. The number of ducks that might walk past on a given day,
$$\Omega = \{0,1,2,...\} = \mathbb{Z}^{\geq 0}.$$
3. The length of 5 footlong subs purchased from a Subway,
$$\Omega = \{(x_1,...,x_5) \mid x_i \geq 0, i = 1,...,5\}.$$

Note that for modelling purposes it is often easier to take the sample space larger than necessary. For example, it is not likely for a footlong sub to be 100m long even though the sample space would include this measurement, and we would not expect 1000 ducks to walk past a regular lake.

7.3 Events and sigma algebras

Often we are not interested in a single outcome but in whether one of a *group* of outcomes occurs. Such subsets of the sample space are called **events**, denoted by $A, B, C, ...$ We say that event A occurs if the outcome of the experiment is one of the elements in A. Examples of events are:

1. The event that the sum of two dice is 10 or more,
$$A = \{(5,5),(5,6),(6,5),(6,6)\}.$$
2. The event that we see at most a dozen ducks today,
$$A = \{0, 1, ..., 12\}.$$
3. The event that a Subway footlong is an acceptable length (in inches),
$$A = [11.5, 12.5].$$
4. The event that out of fifty selected people, 5 are left-handed,
$$A = \{5\}.$$

Whilst the notion of events is familiar, in order to further extend the notion of events, we introduce the abstraction of sets of events, i.e. **sigma algebras**.

Definition 7.2 (Sigma algebra). A **sigma algebra** $\mathcal{F}$ is a collection of subsets of the sample space Ω if:

1. $\mathcal{F}$ is non-empty
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. $A_1, A_2, ... \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

For an experiment we say that event A **occurs** if its outcome $\boldsymbol{\omega}$ is an element of A, i.e. $\omega \in A$.

Example 7.1 (Sigma algebras). Some examples of sigma algebras include:

1. If Ω is a sample space, the power set (the set of all subsets of a set where $|\mathcal{P}(\Omega)| = 2^{|\Omega|}$) is an event space.
2. $\{\varnothing, \Omega, \{1,2\}, \{3,4\}, ...\}$ when you throw a dice multiple times.

Exercise 7.1. Prove these.

In the context of sigma algebras, an event is an element of a sigma algebra and itself is a set. So, we can apply our usual set operations:

For $A, B \in \mathcal{F}$,

1. the set $A \cup B$ (A **union** B) is the event that either A *or* B occur,
2. the set $A \cap B$ (A **intersection** B) is the event that A *and* B both occur,
3. the event A^c (A **complement**) is the event that A does *not* occur, and
4. if $A \subset B$ (A is a **subset** of B) then event A is said to *imply* event B.

Two events A and B which have no outcomes in common, i.e. $A \cap B = \varnothing$, are called **disjoint** events.

Sigma algebras are useful as they allow us to investigate larger collections of events and their properties in a nicely defined context. With them, we can discuss events more generally and establish rigorous definitions and axioms for probability measures.

7.4 Probability

Definition 7.3 (Probability). A probability $\mathbb{P}$ is a rule (function) which assigns a positive number to each event and satisfies the following axioms:
Axiom 1: $\mathbb{P}(A) \geq 0 \forall A \in \mathcal{F}$
Axiom 2: $\mathbb{P}(\Omega) = 1$
Axiom 3: For any sequence $A_1, A_2, ...$ of *disjoint* events we have

$$\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i). \tag{1.3}$$

7.5 Conditional probability and independence

7.5.1 Chain rule

Theorem 7.1 (Baye's Theorem). *Wow it's Baye's!*

See 3.2

7.5.2 Law of total probability and Baye's theorem

7.5.3 Independence

8 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	$x \in$	$\mathbb{E}(X)$	Var(X)	PGF
Ber(p)	$p^x(1-p)^{1-x}$	$\{0, 1\}$	p	$p(1-p)$	1 - p + zp
Bin(n, p)	$\binom{n}{x}p^x(1-p)^{n-x}$	$\{0, 1, ..., n\}$	np	$np(1-p)$	$(1-p+zp)^n$
Geom(p)	$p(1-p)^{x-1}$	$\{1, 2, ...\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{zp}{1-z(1-p)}$
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	$x \in$	$\mathbb{E}(X)$	$\mathrm{Var}(X)$	PGF/MGF
$\mathrm{U}[a,b]$	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$[a,b]$	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$	$\frac{e^{bs}-e^{as}}{s(b-a)}$
$\mathrm{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$1-e^{\lambda x}$	$\mathbb{R}^+$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-s}, s<\lambda$
$\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$		$\mathbb{R}$	μ	σ^2	$e^{s\mu+s^2\sigma^2/2}, s\in\mathbb{R}$

Preface

README

These notes are a complement to the Lecture Notes for STAT2004 Semester 2 2021 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 this course. You can access these notes on blackboard by searching for the 2020 running of this course. These notes also borrow terminology from the Lecture notes from 2020, such as the introduction of sigma algebras for the abstraction/generalisation of events. The lecture notes assume knowledge of the basics of probability, however,

statistical inference hinges upon many basic probability notions. The second half of these notes contain the foundations of probability and go through some more rigorous formulas and theorems that will be needed. Well, I did some of that stuff since I think sigma algebra notation is useful, however, I did end up abandoning making that section of notes since Dirk/Geoff's original reference notes are quite comprehensive anyway. The only issue with referring to the 2020 Reference notes is that Geoff uses vectors for every formula in 2021, 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. 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 8.1 (Theorem name). *The statement of a theorem.*

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

Definition 8.1 (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 8.1 (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 8.1 (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 :)