BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Probability and Distribution Theory (PDT)

Semester 1, 2021

Prepared by:

Andrew Forbes, Rory Wolfe
Department of Epidemiology & Preventive Medicine
Monash University

Tel: (03) 9903 0580

E-mail: Andrew.Forbes@monash.edu

MONASH University

Contents

Probability and Distribution Theory (PDT)
Semester 1, 2021

Instructor contact details

Andrew Forbes	Jessica Kasza
Department of Epidemiology and Preventive Medicine, Monash University Alfred Hospital Tel: (03) 9903 0580	Department of Epidemiology and Preventive Medicine, Monash University Alfred Hospital Tel: (03) 9903 0555
Email: Andrew.Forbes@monash.edu	Email: Jessica.Kasza@monash.edu

Andrew Forbes will be the coordinator of the unit this semester, with Jessica Kasza assisting with online discussions and assessments.

Background

To obtain a sound understanding of the statistical methods used in the design and analysis of medical and health studies, it is essential to have a thorough knowledge of the theoretical basis for these techniques. This unit will focus on applying the calculus-based techniques learned in Mathematical Background for Biostatistics (MBB) to the study of probability and statistical distributions. These two units, together with Principles of Statistical Inference (PSI), will provide the core prerequisite mathematical statistics background required for the study of later subjects in the BCA program.

Unit summary

In PDT, we will harness your existing knowledge and understanding of mathematical methods and apply them to statistical distribution theory. One further area of mathematics beyond the MBB unit is required and we cover it from the first module onwards; namely probability theory. Wherever possible we demonstrate the real-world applicability of the theoretical results that we cover. The PDT material is interspersed with exercises for you to attempt and hence gain a deeper understanding of the theory and methods covered. There will be videos posted on Canvas to summarise content as you work through each module of the unit.

The PDT modules make extensive use of the prescribed textbook, “WMS” (see below for details of this book and other books relevant to PDT). You will be directed to readings from the WMS book and to complete selected exercises. We intend the PDT

material to be a comprehensive guide to reading WMS and anticipate that you will have our module notes and the WMS book side by side. We encourage you to use our notes to guide you through the WMS book rather than just plunging straight into WMS.

Worked solutions will be made available for the exercises during semester. Student Solution Manuals for WMS do exist and these provide worked solutions for all odd numbered exercises in the WMS textbook; you may consider purchasing one of these manuals to enable you to have solutions to extra exercises that we don't set in PDT. This would allow you to undertake extra practice and in the past some students have said that this was helpful to them. Note that the solutions we provide in PDT are more detailed than the solutions provided in Student Solution Manuals but we only provide solutions for a careful selection of WMS exercises, and some other exercises that we set ourselves.

In PDT we make use of Stata and R software and the Wolfram Alpha web-based algebra program (details below) and we do not assume that you have previously used any of these. You are free to chose whether to use Stata or R, or indeed a combination.

Past students have told us that the content of PDT makes it a challenging unit, so you too should expect to feel challenged by the material and anticipate plenty of hard work in the coming semester. The good news is that since PDT lays the foundation for most future units, after successfully completing the unit you should feel confident about taking on the technical material in other units in the BCA program.

Workload requirements

The expected workload for this unit is 10-12 hours per week on average, consisting of guided readings, discussion posts, independent study and completion of assessment tasks.

Prerequisites

Mathematical Background for Biostatistics (MBB)

Co-requisites

Nil

Learning Outcomes

At the completion of this unit students should be able to:

1. Demonstrate an understanding of the meaning and laws of probability
2. Recognise common probability distributions and their properties
3. Apply calculus-based tools to derive key features of a probability distribution, such as mean and variance
4. Obtain mean, variance and the probability distribution of transformations of random variables
5. Manipulate multivariate probability distributions to obtain marginal and

conditional distributions
6. Understand properties of parameter estimators and the usefulness of large sample approximations in statistics
7. Appreciate the role of simulation in demonstrating and explaining statistical concepts.

Unit content

The unit is divided into 5 modules, summarised in more detail below. Each module will involve 2 or 3 weeks of study and generally includes the following material:

1. Module notes describing concepts and methods, and including exercises of a more theoretical nature.
2. Selected readings from the textbook.
3. One or more extended examples illustrating the concepts/methods introduced in the notes and including some practically oriented exercises.
4. One or more online videos to summarise the module content
5. An online 'live' tutorial, with date and time chosen by a Doodle poll of student preferences. The recording of the tutorial will be made available online for students unable to make it to the live session, or for later viewing.

Study materials for all Modules are downloadable from the Canvas unit site. No hardcopy of notes will be posted. Assignments, and supplementary material such as datasets, will be made available on the unit site. Please note that we are not able to upload copies of copyright material (journal articles and book extracts)—for these you will have to rely on resources from your home university's library.

Recommended approaches to study

Students should work through each module systematically, following the module notes and any readings referred to, and working through the accompanying exercises. *You will learn a lot more efficiently if you tackle the exercises systematically as you work through the notes.* You are encouraged to post any content-related questions to the Canvas site, whether they relate directly to a given exercise, or are a request for clarification or further explanation of an area in the notes. You should also work through all of the computational examples in the notes for yourself on your own computer.

Solutions to the exercises in each module (except those to be submitted for assessment, as described below) are included in the study materials.

Method of communication with coordinator(s)

Questions about administrative aspects or course content can be emailed to coordinator(s), and when doing so please use "PDT:" in the Subject line of your email to assist in keeping track of our email messages. Coordinator(s) will be available to

answer questions related to the module notes and practical exercises, and to address any other issues that require clarification. However, please note that instructors are not necessarily available every day of the week and you should expect that it may take a day or so to respond to questions (possibly longer over weekends and during breaks).

We strongly recommend that you post content-related questions to the Discussions tool in the PDT Canvas site. You may be familiar with the Canvas system from previous BCA units, and will receive any specific instructions on using the Canvas site this semester from the BCA Coordinating Office. There is also a "Getting Started" document available on the Student Resources page of the BCA website.

Module descriptions

Below is an outline of the study modules, followed by a timetable and assessment description table.

Each module begins on a Monday and concludes on a Sunday. **The due date for submission of the required exercises from each module is 11:59pm on the day indicated in the assessment table below.**

Module 1: **Probability**

- Use of set notation (Venn diagrams) including null set, union, subset, intersection, complement, and mutually exclusive.
- Definitions of events, simple event, sample space, discrete sample space, and probability of an event.
- Calculation of the probability of an event by listing all simple events
- Calculation of the probability of an event by using combinatorial methods
- Application of conditional probability, independent events, the multiplicative law of probability, the additive law of probability, Bayes rule (particularly to diagnostic/screening tests), and the law of total probability.

Module 2: **Discrete Random variables**

- Definition of a random variable
- Definition and application of a probability distribution for a discrete
- random variable
- The Bernoulli, Binomial and Poisson distributions
- Additional discrete probability distributions
- The expectation (mean) and variance of a discrete probability distribution
- Moment generating functions

Module 3: **Continuous random variables**

- The cumulative distribution function and probability density function of a continuous random variable
- The mean and variance of a continuous random variable given its density function
- Probability calculations using the Uniform, Normal and Lognormal distributions
- The form and potential uses of other continuous distributions

- The definition and use of moment generating functions for continuous random variables
- Transformation of a random variable
- The expectation of a transformation of a random variable
- The distribution function of a transformation of a random variable using either moment generating functions or the method of transformations

Module 4: **Multiple random variables**

- Definitions of correlation, covariance and independence
- Joint, conditional and marginal distributions in the context of multiple random variables
- The Bivariate Normal and Multinomial distributions
- Application of the definitions of conditional expectation and variance to obtain marginal means and variances without the need for obtaining the marginal distribution
- The expectation and variance of transformations of multiple random variables
- The distribution of a transformation involving more than one random variable using the method of moment generating functions.

Module 5: **Estimation: Concepts and properties of estimators**

- The meaning of a parameters, estimators and estimates
- The concept of a sampling distribution and standard error of an estimator
- Unbiasedness, consistency and efficiency of an estimator
- The general purpose of large sample theory
- Application of the Central Limit Theorem
- Computation and interpretation of confidence intervals

Unit schedule

Semester 1, 2021 starts with Module 1 on Monday March 1st. Note that Modules 3 and 5 are each of 3 weeks duration. The mid-semester break is April 5-9 (plus Good Friday of the week before), and the week commencing April 26 has no new unit content to allow for Major Assignment 1 completion.

Week	Week commencing	Module	Topic	Assessment
1	March 1	Module 1	Probability	
2	March 8	Module 1		Mod 1 Exercises due March 21 [extra time]*
3	March 15	Module 2	Discrete random variables	
4	March 22	Module 2		Mod 2 Exercises due March 28
5	March 29	Module 3	Continuous random variables	
6	April 5		**Mid-semester break (1 week)**	
7	April 12	Module 3		
8	April 19	Module 3		Mod 3 Exercises due April 25
9	April 26		**Major Assignment 1**	**Due date: May 2**
10	May 3	Module 4	Multiple random variables	
11	May 10	Module 4		Mod 4 Exercises due May 16
12	May 17	Module 5	Estimation	
13	May 24	Module 5		
14	May 31	Module 5		Mod 5 Exercises due June 6
15	June 7		**Major Assignment 2**	**Due date: June 13**

*An extra week is allowed for Module 1 exercises submission.

Assessment

Assessment will include 2 major written assignments worth 35% each, to be made available in the middle and at the end of the semester, and to be completed within approximately 2 weeks. These assignments will be posted on the Canvas site together with an online Announcement broadcasting their availability. In addition, students will be required to submit solutions to selected practical exercises from each module,

worth a total of 30%, by deadlines specified throughout the semester (see table below).
Assignment 1 will cover material from Modules 1-3 only. Assignment 2 will cover the entire semester's material, but with emphasis on Modules 4 and 5. Assignments and exercises from modules may be submitted at any time up to midnight on the due date.

Assessment name	Assessment type	Coverage	Learning objectives	Weight
Module 1 exercises	Assignment	Module 1	1	5%
Module 2 exercises	Assignment	Module 2	1,2	5%
Module 3 exercises	Assignment	Module 3	1,2,3,7	10%
Major Assignment 1	Assignment	Modules 1-3	1,2,3,4,7	35%
Module 4 exercises	Assignment	Module 4	1,2,3,4,5	5%
Module 5 exercises	Assignment	Module 5	1,2,3,4,5,6	5%
Major Assignment 2	Assignment	Modules 1-6	1,2,3,4,5,6	35%

In general you are required to submit your work typed in Word or similar (e.g. using Microsoft's Equation Editor for algebraic work) and we strongly recommend that you become familiar with equation typesetting software such as this. If extensive algebraic work is involved you may submit neatly handwritten work, however please note that marks will potentially be lost if the solution cannot be understood by the markers due to unclear or illegible writing. This handwritten work should be scanned and collated into a *single pdf file* and submitted via the Canvas site. See the BCA Assessment Guide document for specific guidelines on acceptable standards for assessable work.

The instructors will generally avoid answering questions relating directly to the assessable material until after it has been submitted, but we encourage students to discuss the relevant parts of the notes among themselves, via Canvas. However **explicit solutions to assessable exercises should NOT be posted for others to use**, and each student's submitted work must be clearly their own, with anything derived from other students' discussion contributions clearly attributed to the source.

Submission of assessments and academic honesty policy

You should submit all your assessment material via the Canvas site unless otherwise advised. For more detail please see the BCA Assessment Guide.

The BCA pays great attention to academic honesty procedures. Please be sure to familiarise yourself with these procedures and policies at your university of enrolment. Links to these are available in the BCA Student Assessment Guide. When submitting assessments on Canvas (using Turnitin) you will need to indicate your compliance with the plagiarism guidelines and policy at your university of enrolment before making the submission.

Late submission of assessments and extension procedure

We adhere to standard BCA policy for late penalties for submitted work, i.e, unless otherwise stated, a student can submit an assessment up to 10 days after the due date. A late penalty of 5% per day will be applied (including weekends and public holidays). The maximum penalty which can be applied is a reduction to 50% of the total assessment mark. Extensions are possible, but these need to be applied for (by email) as early as possible. The Unit Coordinator is not able to approve extensions beyond three days; for extensions beyond three days you need to apply to your home university, using their standard procedures.

Learning resources

The prescribed textbook for PDT is

> Wackerley DD, Mendenhall W, Schaeffer RL. Mathematical Statistics with Applications. 7th edition. 2008 Thomson Learning, Inc. (Duxbury, Thomson Brooks/Cole) ISBN-13: 978-0-495-11081-1

This textbook is central to this subject and you must have unrestricted access to this book throughout semester. We refer to this textbook as "WMS" in PDT material.

- There are several international editions of WMS available, however they are not identical and we do not recommend purchase of any edition which has a different ISBN to that of the edition in the BCA textbook and software guide, i.e., please only purchase ISBN-13 978-0-495-11081-1.
- Please be very careful if ordering this textbook online to ensure that the correct ISBN appears at each step of your ordering process – some websites advertise the above book and ISBN but their purchasing process, after displaying the correct book initially, subsequently skips to a different ISBN before you have completed the purchase.

Other books which cover similar material and that we recommend are:

- Rosner B. Fundamentals of Biostatistics 4th edition.
    - *A textbook suitable for introductory courses in medical statistics that also touches on more advanced topics.*
- Larsen RJ & Marx ML. An Introduction to Mathematical Statistics and its Applications, Fourth Edition. 2006 Pearson International Edition.
    - *A direct competitor to WMS, this book is a useful source as an alternative to WMS; in general we prefer the WMS presentation and progression of topics but there are places where Larsen & Marx is better.*
- Casella G & Berger RL. Statistical Inference 2nd edition. 2002 Wadsworth Group (Duxbury / Thomson Learning, Inc.)
    - *This book covers similar ground to WMS but at a more advanced level.*
- Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics 3rd edition. 1963 International Student Edition, McGraw-Hill Kogakusha
    - *An old classic that is at a more advanced level than WMS.*

Software

For this subject you will need to have access to the Stata or R software packages. For Stata, we expect most of you would be using Stata 15 or 16, the latter of which was released in July 2019. We are not aware of any major differences between Stata versions that affect the material, but minor issues will be pointed out in Canvas postings. Whichever version you are using, we recommend that you perform the online update to the latest update of that version. (Use the command `update query`). For R, we assume you are using at least R version 3.6, and we expect many of you will be using version 4.0.0 or later. You can check the version by typing `R.version`. We do not expect any differences in results between versions of R.

Feedback

Our feedback to you:

The types of feedback you can expect to receive in this unit are:

- Formal individual feedback on submitted module exercises
- Responses to questions posted on Canvas

Your feedback to us:

One of the formal ways students have to provide feedback on teaching and their learning experience is through the BCA student evaluations at the end of each unit. The feedback is anonymous and provides the BCA with evidence of aspects that students are satisfied with and areas for improvement.

Required mathematical background

We list here the mathematical techniques that will be used during PDT. All of these techniques were covered to differing levels of detail in MBB. If you are unfamiliar or lack confidence with any of these techniques, now is the time to do some revision since most of them won't be used until later modules of PDT.

- Functions and their inverse; one parameter *f(x)* and two parameter *f(x,y)*.
- Absolute values $|.|$, exponential and logarithm.
- Increasing and decreasing functions; one-to-one transformations and concept of "onto".
- Summations, especially $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$, and the fact that $\left(\sum_{i=1}^{n} x_i \right)^2 \neq \sum_{i=1}^{n} x_i^2$
- Solving quadratic functions, $ax^2 + bx + c = 0$, by "completing the square" or by obtaining roots with use of $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.
- Differentiation: Chain Rule, Product Rule, Second derivatives.

- Maximizing (and minimizing) a given function *f(x)* using the derivative $\frac{df}{dx}$ and ensuring, for a maximum (minimum), that the second derivative $\frac{d^2f}{dx^2}$ is negative (positive).

- A familiarity with Integration by parts (e.g. Anton 7th ed; Section 8.2) and Integration by Substitution and changing limits (see "Method 2" in Anton 7th Ed; Section 6.8 Evaluating definite integrals by substitution) that is sufficient to comprehend output from Wolfram Alpha.

- Double integration involving rectangular regions, e.g. Anton 7th Ed Section 15.1 (although most technicalities of $\lim_{n\to\infty}()$ can be skipped).

- Taylor Series.

Changes to PDT since last delivery, including changes in response to student evaluation

PDT was last delivered in Semester 2 2020. R code for all aspects of the unit have been added since that delivery. Otherwise there have been minor changes for greater clarification of the text.

Acknowledgments

The material for PDT was developed by Rory Wolfe and Andrew Forbes. We would like to acknowledge some sources of help that are not otherwise acknowledged in the material.
We thank Professor Phil Prescott of Southampton University, UK for helpful discussions and access to material from MATH1024. We thank John Carlin for the use of existing BCA material for LCD and LMR that he developed with Andrew Forbes. We thank Ian Marschner and subsequent PSI unit coordinators for the use of existing BCA material for PSI. We thank Jessica Kasza, Emily Karahalios and Sarah Arnup for development of videos.