BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Probability and Distribution Theory (PDT)

Semester 1, 2017

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

Tel: (03) 9903 0580

E-mail: Andrew.Forbes@monash.edu Rory.Wolfe@monash.edu

MONASH University

Unit summary and study guide

Objectives of Probability and Distribution Theory

At the completion of this unit the student will be able to:

- Demonstrate an understanding of the meaning and laws of probability
- Recognise common probability distributions and their properties
- Apply calculus-based tools to derive key features of a probability distribution, such as mean and variance
- Manipulate multivariate probability distributions to obtain marginal and conditional distributions
- Obtain mean, variance and the probability distribution of transformations of random variables
- Understand properties of parameter estimators and the usefulness of large sample approximations in statistics
- Appreciate the role of simulation in demonstrating and explaining statistical concepts.

Background and method of learning

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.

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 is required and we cover it in our first module; 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.

PDT is organized in five modules

Module 1: Probability
Module 2: Discrete random variables
Module 3: Continuous random variables
Module 4: Multiple random variables
Module 5: Estimation

Most of 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 software and the Wolfram Alpha web-based algebra program (details below) and we do not assume that you have previously used either of these.

Past students have told us that PDT is a challenging subject, 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 subjects, after successfully completing PDT you should feel confident about taking on the technical material in other units in the BCA program.

Instructors and method of teaching

Andrew Forbes has primary responsibility for this unit and will deal with administrative aspects of the course. He will contribute to class discussions on e-learning and respond there to content-related questions. Rory Wolfe may contribute occasionally too. Further, Emily Karahalios, a biostatistics postdoctoral research fellow, will assist with monitoring of discussions and the marking of submitted material. All of us are based in the Biostatistics Unit of the Department of Epidemiology and Preventive Medicine at Monash University.

All content-related questions should be posted to one of the Discussions topics in the PDT Blackboard on the BCA's eLearning site that is hosted by Sydney University. You should have received instructions on accessing eLearning from the BCA administrative coordinator Erica Jobling. It is important that questions, responses from instructors, and any ensuing discussion is available via the PDT Blackboard for all students to access.

We will use the PDT Blackboard site for posting course materials. In addition we will send out this study guide, and modules 1-5 material in paper form at the start of semester.

Instructors will contribute to Blackboard during work hours on most week days (Monday to Friday). We encourage you to respond to each others' postings, if possible to provide help, but also just to echo difficulties (it is important that students are aware whether they are “alone” in struggling with a particular topic or whether most of the class share their struggle).

Any questions about administrative matters relating to PDT can be emailed directly to Andrew.Forbes@monash.edu and Andrew can be contacted for any urgent PDT matters during business hours: +61 (03) 9903 0580.

Assessment

Assessment will include two written assignments worth 35% each, to be completed within 2-3 weeks of being made available. These assignments will be posted on the Blackboard site together with an Announcement announcing their availability. In addition, students will be required to submit solutions to selected practical exercises from each module worth 5% per module (10% for Module 3) for a total of 30%. See the timetable below for more details.

We adhere to standard BCA policy for late penalties for submitted work, i.e. a 5% deduction from the earned mark for each day the assessment is late, up to a maximum of 50%. Of course, extensions are possible, but these need to be applied for (by email) as early as possible.

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 Blackboard. See the BCA Assessment Guide document (link provided below) 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 PDT notes among themselves, via Blackboard. 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.

You should submit all your assessment material via Blackboard unless otherwise advised. We have instigated the use of Turnitin within Blackboard for the submission process. For more detail please see the BCA Student Assessment Guide (which includes links to Universities' Plagiarism Policies) at:
`http://www.bca.edu.au/currentstudents.html#assessmentguide`
This guide will also be included in hardcopy in your package of notes.

Timetable

Module	Title	Duration	Teaching dates	Exercises due
1	Probability	2 weeks	6 Mar – 19 Mar	Sunday 26 March
2	Discrete RV	2 weeks	20 Mar – 2 April	Sunday 2 April
3	Continuous RV	3 weeks	3 April - 23 Apr	Sunday 23 April
4	Multiple RV	2 weeks	1 May – 14 May	Sunday 14 May
5	Estimation	3 weeks	15 May - 4 Jun	Sunday 4 June

Timetable notes:

- Note that Modules 3 and 5 are each of 3 weeks duration. There is the Easter break in the midst of Module 3, with Good Friday being April 14.
- Assignment 1: There is a week with no new material on 24-30 April to facilitate completion of Assignment 1 which is due for submission on Sunday 30 April. This assignment will cover material from Modules 1-3 only.

- Assignment 2: There is no new material in the week 5-11 June with Assignment 2 being due for submission on Wednesday 14 June. This assignment will primarily cover Modules 4 and 5 although the entire semester's material may be relevant.
- Assignments and exercises from modules may be submitted at any time up to midnight on the due date.

Books

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 must be purchased. We will refer to this textbook as WMS throughout the PDT material.

Please note that 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., ISBN-13 978-0-495-11081-1.

Other books from which we provide readings 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

Use of the Stata statistical package is required for PDT. Our notes assume the use of release 14 of Stata which was released in April 2015. The Stata commands used in PDT will work in in older Stata versions, specifically versions 11-13.

If you have not used Stata before (we assume this applies to most of you), there is introductory instructional material available to assist you. First of all, there is a "Getting Started" instructional manual that is installed with your copy of Stata. To access this, open Stata, click on Help, then PDF Documentation, and a pdf file will open. On the left hand side click on "[GS] Getting Started" and the Getting Started manual will open for you to read online or print. There are also resources available from a website at UCLA: `http://www.ats.ucla.edu/stat/stata/`

There are three flavours of Stata: Small, Stata/IC and Special Edition (SE). We recommend use of Stata/IC, and the typical installation will be the "32 bit edition" or "Stata IC/32". This usually needs to be manually selected when you install Stata. In terms of the other flavours, Stata/SE is for enormous datasets and our commands will work fine if you happen to have that flavour. We can't guarantee that our commands will work in Small Stata.

We will make use of the Wolfram Alpha (WA) web-based program to perform certain algebraic calculations. Some of you may have used it in MBB but we are not assuming you have done so. It is important to realise that much of the work in PDT can be completed without it although some of the more complex integration in modules 3 and 4 will certainly require its use. We anticipate that it might also be useful for checking your answer to certain questions involving calculus.

Changes to PDT since last delivery

PDT was last delivered in Semester 2 2016. There have been only minor changes since that delivery in the form of typos and minor edits for greater clarification of the text. PDT did undergo more substantial changes between semesters 1 and 2 in 2015; in Modules 3 and 4 the need to use integration by substitution or integration by parts was removed by requiring the use of Wolfram Alpha to solve such integrals, in Module 4 the complexity of double integral regions of integration was reduced such that only rectangular regions are considered, and in Module 5 the material was expanded to include additional detail on properties of estimators and introductions to the Delta method and confidence intervals.

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 $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$
- 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.

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 and Sarah Arnup for development of some of the video mini-lectures.