Health Economics by Distance Learning

Module 5

Statistics for Health Economics

Module Workbook

2018 - 2019 academic year
Module 5 (ECO0052M)

Statistics for Health Economics

Module Workbook

2018 - 2019 academic year

Authors: Paul Contoyannis, Nigel Rice


Date of printing: December 7, 2018

Copyright © 2018. The University of York, Heslington, York YO10 5DD. All rights reserved. For use only with the Postgraduate Certificate and Diploma programmes in Health Economics for Health Care Professionals. No part of this workbook may be reproduced, stored in a retrieval system or transmitted in any form or by any means, without written permission from the University of York.
Module 1
Basic Economic Concepts

Module 2
Health Economics: Concepts and Analysis

Module 3
Introduction to Health Care Evaluation

Module 4
The Economics of Health Care Systems

Module 5
Statistics for Health Economics

Module 6
Further Topics in Economic Evaluation

Module 7
Assessing the Impact of Medical Technologies on Health

Module 8
Outcome Measurement and Valuation

Module 9
Decision Analysis for Health Technology Assessment

PG Certificate

PG Diploma

MSc
Structure of Module 5: Statistics for Health Economics

Unit 5.1
Random Variables and Probability Distributions

Unit 5.2
Descriptive Statistics

Unit 5.3
Sampling, Sampling Error and Confidence Intervals

Unit 5.4
Hypothesis Testing

Unit 5.5
Correlation and Regression
Academic members of the Programme Team:

Martin Forster, Programme Leader, PG Certificate and PG Diploma in Health Economics for Health Care Professionals, Department of Economics and Related Studies
Helen Weatherly, Programme Leader, MSc in Economic Evaluation for HTA, Centre for Health Economics
Stuart Bell, Dean, Faculty of Social Sciences
Andrew Jones, Professor of Economics, Department of Economics and Related Studies
Subir Chattopadhyay, Chair, Graduate School Board, Department of Economics and Related Studies
Mark Sculpher, Professor of Health Economics, Centre for Health Economics
Matthew Taylor, Director, York Health Economics Consortium

Thanks also to the following peer reviewers for this module:
(Affiliations correct at time of review)

Christene Abrams, MSc Health Economics student, University of York
Paul Burrows, Reader in Economics, Department of Economics and Related Studies, University of York
Karen Croucher, Research Fellow, York Health Economics Consortium
Tony Culver, Head of Department of Economics and Related Studies, University of York
Martin Grossman, PGCert Health Economics for Health Care Professionals, 2009-10 academic year
Andrew Jones, Director of the MSc in Health Economics, University of York
Martin Land, Consultant, KPMG Consulting
Michael Place, Research Fellow, York Health Economics Consortium
Nigel Rice, Senior Research Fellow, Centre for Health Economics, University of York
James Ryan, Research Fellow, York Health Economics Consortium
Diana Sanderson, Senior Research Fellow, York Health Economics Consortium
Trevor Sheldon, Head of Department of Health Studies, University of York
Andrew Street, Research Fellow, Centre for Health Economics, University of York
Ian Watt, Professor of Primary and Community Care, Department of Health Studies, University of York
Peter West, Director, York Health Economics Consortium
Jean Wood, Distance Education Developer, Centre for Teaching, Learning and Assessment, University of Edinburgh

This workbook is produced using LaTeX at the Department of Economics and Related Studies, University of York. Sincere thanks to all of those who make this software and the associated packages freely available.
## Contents

5 [Guide to Module 5](#) ................................................................. 1

Module overview ................................................................. 9

**Unit 5.1 Random variables and probability distributions** .................. 13

- 5.1.1 *Introduction* ............................................................. 16
- 5.1.2 *Sampling from the population* ....................................... 17
- 5.1.3 *Probabilistic models* ................................................... 20
- 5.1.4 *Conditional probability and independence* ....................... 23
- 5.1.5 *Discrete random variables* .......................................... 32
- 5.1.6 *Continuous random variables* ...................................... 42
- 5.1.7 *Other continuous random variables* ............................... 52

**Unit 5.2 Descriptive statistics** ............................................. 57

- 5.2.1 *The scope of descriptive statistics* ................................ 60
- 5.2.2 *Types of variable* ..................................................... 62
- 5.2.3 *Graphical and tabular approaches* ................................ 64
- 5.2.4 *Numerical approaches* ............................................... 73

**Unit 5.3 Sampling, sampling error and confidence intervals** .......... 81

- 5.3.1 *Introduction* ............................................................. 84
- 5.3.2 *Sampling from a population* ....................................... 85
- 5.3.3 *Sampling variation* ................................................... 88
- 5.3.4 *Confidence interval for the population mean* .................... 95
- 5.3.5 *Distribution of the difference between two means* ............... 103
- 5.3.6 *Sampling distribution of a variance* ................................ 104
- 5.3.7 *Sampling distribution of a proportion* ............................ 106
Unit 5.4  Hypothesis testing  ........................................... 109
  5.4.1 Introduction .................................................. 112
  5.4.2 The null hypothesis and hypothesis testing .......... 113
  5.4.3 Hypothesis test of one mean ......................... 121
  5.4.4 Comparison of two means ............................ 128
  5.4.5 More than two means ................................ 136
  5.4.6 Comparison of proportions .......................... 144
  5.4.7 The chi-squared test for association ................ 151
  5.4.8 The design of clinical trials ....................... 156
  5.4.9 Bayesians and frequentists ....................... 156

Unit 5.5  Correlation and regression ............................ 161
  5.5.1 Introduction ............................................. 162
  5.5.2 Correlation ............................................. 163
  5.5.3 Simple regression .................................. 170
  5.5.4 Association and causation .......................... 185
  5.5.5 Multiple regression ................................. 187

Appendix: The principle of least squares ...................... 193
  Information on one random variable Y .................. 193
  Information on two variables Y and X .................. 194
## List of Figures

5.1 Sampling $n = 1$ and $n = 2$ items from a population of five. .......................... 19
5.2 The event $A = \{2, 4, 6\}$ when tossing a fair die once. ................................. 21
5.3 The event $A \cap B$. ......................................................................................... 22
5.4 Probability Mass Function for a simple coin-tossing experiment. ......................... 37
5.5 Sequential description of the coin tossing problem. ............................................. 39
5.6 PDF for a normal random variable with $\mu_X = 130$ and $\sigma_X = 20$. ............. 44
5.7 Standard normal PDF ......................................................................................... 45
5.8 Changing the standard deviation of a normal random variable. ......................... 46
5.9 Standard normal cumulative distribution function. .......................................... 48
5.10 Four PDFs for a $\chi^2_n$ random variable. ...................................................... 53
5.11 Left skewed probability density function. ......................................................... 54
5.12 Bar chart for self-assessed health (data in Table 5.5). ....................................... 66
5.13 Pie chart for self-assessed health (data in Table 5.5). ....................................... 66
5.14 Relative frequency histogram (ten bins, data in Table 5.5). ............................. 69
5.15 Relative frequency histogram (five bins, data in Table 5.5). ............................ 69
5.16 Scatterplot of age and exercise (data in Table 5.5). ......................................... 71
5.17 Scatterplot of age and blood pressure (data in Table 5.5). ............................... 71
5.18 Sampling distributions of the mean for our example. ........................................ 91
5.19 Uniform distribution for weights example. ........................................................ 96
5.20 95% confidence interval for the sample mean $\bar{x}$. ........................................ 97
5.21 Four 95% confidence intervals for the population mean. ............................... 98
5.22 Distributions of $\bar{x}$ under null and alternative hypotheses. ......................... 126
5.23 Hypothetical distributions for three groups. ..................................................... 137
5.24 Hypothetical distributions with greater variation. ............................................ 138
5.25 Positive association between $X$ and $Y$. ....................................................... 164
5.26 Joint probability density function ‘from above’. ............................................. 165
5.27 Scatterplots and correlation coefficients. ......................... 168
5.28 Regression line showing \( u \) conditional upon \( X = x \). ............... 172
5.29 Scatterplot of age and blood pressure. ............................... 175
5.30 Decomposition of the variation about the mean. ..................... 179
5.31 Output of a regression package. ................................. 183
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Elements in the sample space for a population of size 5</td>
<td>19</td>
</tr>
<tr>
<td>5.2</td>
<td>Blank table with cells to be filled in</td>
<td>25</td>
</tr>
<tr>
<td>5.3</td>
<td>Numbers testing positive and negative for osteoporosis</td>
<td>27</td>
</tr>
<tr>
<td>5.4</td>
<td>Probabilities derived from Table 5.3</td>
<td>27</td>
</tr>
<tr>
<td>5.5</td>
<td>Data for Exercise 5.6</td>
<td>61</td>
</tr>
<tr>
<td>5.6</td>
<td>Frequency table for gender (data in Table 5.5)</td>
<td>65</td>
</tr>
<tr>
<td>5.7</td>
<td>Contingency table for self-assessed health and gender</td>
<td>65</td>
</tr>
<tr>
<td>5.8</td>
<td>Frequency table for age (data in Table 5.5)</td>
<td>67</td>
</tr>
<tr>
<td>5.9</td>
<td>Sample means and their expected values</td>
<td>92</td>
</tr>
<tr>
<td>5.10</td>
<td>Respiratory health by smoking status</td>
<td>131</td>
</tr>
<tr>
<td>5.11</td>
<td>Respiratory health by smoking status</td>
<td>141</td>
</tr>
<tr>
<td>5.12</td>
<td>Decomposition of sums of squares for analysis of variance</td>
<td>143</td>
</tr>
<tr>
<td>5.13</td>
<td>Outcome for two treatments for asthma</td>
<td>149</td>
</tr>
<tr>
<td>5.14</td>
<td>Outcome for two treatments for asthma, paired observations</td>
<td>149</td>
</tr>
<tr>
<td>5.15</td>
<td>Outcomes after treatment for asthma</td>
<td>152</td>
</tr>
<tr>
<td>5.16</td>
<td>Expected outcomes (data in final column of Table 5.15)</td>
<td>152</td>
</tr>
<tr>
<td>5.17</td>
<td>ANOVA table for the regression presented in this section</td>
<td>181</td>
</tr>
</tbody>
</table>
Guide to Module 5

Learning aims and objectives

This is a Level 7(M) module and is compensatable in certain circumstances. Please refer to the Programme Handbook, section 4, for full details.

Learning aims

To introduce you to the foundations of statistical methods for health economics.

Learning objectives

Upon successful completion of the module you should be able to:

- Explain the basic ideas underlying the theory of probability and classical statistical analysis, including random variables and their probability distributions, descriptive statistics, sampling and sampling error.

- Explain the difference between association and causation and the role played by randomisation in identifying causal effects.

- Take a practical problem and a sample of data, define the problem in a way that is amenable to statistical analysis and explain why the approach adopted is reasonable.
• Perform the relevant computations for the statistical methods covered and be able to provide intuitive explanations of the methods and results, showing how the results are derived.

• Use a software package such as Excel to carry out data management, descriptive statistics and inferential statistics through to, and including, multiple regression.

You should note that, in the formal piece of assessed work, you will normally be asked to supply the full derivations of, and explanations for, your results. It will not be sufficient for you to use an in-built spreadsheet routine, or other piece of statistical software, to analyse a set of data.

In addition to the above learning objectives, you should ensure you can meet the learning objectives outlined at the start of each unit.

Credits, timing and assessed work

The module is worth 20 credits and should take approximately 200 hours to complete over 12 weeks of study, including the module assessment (see below). The following is a guide to the amount of time you might spend on each unit, although you are of course free to fit your study time to your own circumstances.


• week commencing 21/1/2019. Week 4. Unit 5.2.


Module assessment

The two pieces of assessed work to be submitted on, or before, the dates detailed above are as follows:

- Piece 1. 2017–2018 examination paper question 1 (including appropriate commentary).
- Piece 2. 2017–2018 examination paper, all remaining questions (including appropriate commentary).

Your submission should follow the instructions on the exam paper. In particular, ensure that you separate the report from the commentary and respect the word count. The past paper and data file for the assessed work are available on Yorkshare’s Content system.

Students from past cohorts have requested additional clarification about what should be contained within the report and commentary. As well as reading the instructions in the rubric of the exam, please note the following: The general rule is that the REPORT contains your main descriptive and inferential results, explained in a manner which can be understood by people who are not particularly statistically-minded, but who nevertheless must understand the key results of your analysis (e.g. managers in your organisation). The technical derivations, statistical explanations and so on, should go in the commentary. Markers have quite a lot of latitude in looking at the overall quality of the submitted piece of work, so if there is some material that you have included in the report and which the examiners think should be in the commentary (or vice versa), this won’t count against you. However, if you do not split your submission into the two sections (i.e. REPORT and COMMENTARY) as instructed and make some effort to place material in the appropriate sections, this will count against you; most likely you will be marked out of half the marks available for the questions, meaning you will not pass Module 5. Hence, the bottom line is that you should please make sure that your submission is clearly separated into a section entitled ‘Report’ and a section entitled ‘Commentary’.

Please note that programme staff are under no obligation to mark assessed work that is submitted late. This applies to all modules.

The assessment for this module is by a two day piece of work completed in the final week.
Note that each student will receive a different data set for the assessment and that the data will be supplied to you in a file in ASCII or Excel format.

Please read the Programme Handbook carefully for the regulations regarding assessment and reward at Postgraduate level.

Many of the statistical problems discussed here can be solved using software such as spreadsheets or dedicated statistical packages. You are encouraged to explore these as you progress through the module. As mentioned above, it is important to note, however, that questions in the formal piece of assessed work taken at the end of the module will normally be asking for full derivations of your results. This means that it will not be sufficient for you to apply a spreadsheet routine, or use another piece of statistical software, to analyse the data. You will also have to be able to derive test statistics, plots, regression results etc. from first principles, applying the formulae that we cite in the workbook and in the readings.

Please note that a word limit of 2,500 words (excluding tables, figures and references) applies to the final examination and you should bear this in mind when completing your assessed work. Clear rules exist for penalising work which is over the word limit and these are noted in the exam paper itself. Note also the distinction between the report and commentary that is asked for in the past exam paper. If you require clarification regarding the difference between each, please contact your tutor.

---

**Online learning: Yorkshire**

The module is supported by the Yorkshire virtual learning environment located at:

[vle.york.ac.uk](http://vle.york.ac.uk)

You can access this using your username and password which are supplied separately. Exercises which should be discussed on Yorkshire are flagged in the text with the computer icon.
The main teaching materials are contained within this workbook. References to books and papers are noted in this workbook by the book icon. These may be ordered from the University of York’s library. If you work in the NHS, your local professional librarian may prove helpful.

The following textbook is supplied to you to accompany your study:


This book has been due for a revision for a number of years, which has never been forthcoming. We have, therefore, updated some of the exercises and added additional web links to this workbook, to keep the material up-to-date.

The Altman textbook is a good complement to the workbook, with a useful mathematical appendix, to which you should refer if you encounter unfamiliar mathematical expressions. It also has an appendix containing statistical tables and the end of each chapter contains useful exercises (with selected answers given at the end of the text) to supplement those contained within this workbook.

For material on making causal inferences contained within Unit 5.4, we also make reference to:


This is available as a free electronic resource from york.ac.uk/library. Log-in, search for the book, and you should be able to access it.

Although the workbook, accompanying text and web-based readings should be sufficient for most of the material in this module, you might wish to purchase or borrow one or more of the following books:

Module 5: Statistics for Health Economics


We also recommend that you visit the British Medical Journal’s (BMJ) website and browse the series of short papers entitled *Statistical Notes*, authored by Martin Bland and Douglas Altman, which appear in the *Education and Debate* section of the journal. Notes which complement the content of this module are referred to in the text.

Finally, for those of you who like the material in Unit 5.1, the following is an excellent reference which your module leader frequently consults (and which was referenced when writing that unit):


The first chapter of this book, which is useful for Unit 5.1, is downloadable for free from: http://athenasc.com/probbook.html

Reading that is strongly recommended is marked with a *.

Web-based resources

There are many web-based resources that you might find helpful:

- Onlinestatbook. This site contains a statistics book, simulations using java applets, and case studies. We make reference to this a number of times during the module. www.onlinestatbook.com

- The Royal Statistical Society has ‘Statslife’ pages which give current news and views as well as a link to their publication *Significance*, which has very readable articles on the application of statistics in real life (and which is available from the University of York as an e-journal): www.statslife.org.uk

- Wolfram Demonstrations Project and Mathworld. More technical material, for those of you who like it, may be found by linking to a number of demonstrations from the
Wolfram website:

deemonstrations.wolfram.com/index.html You need the Wolfram player:
http://demonstrations.wolfram.com/download-cdf-player.html to run the demonstrations.

We thank the authors of all of the web-based resources for making their materials freely available for educational purposes.

Software

Working with a suitable software package of your choosing will help you during the module. If you work with MS Excel, you are advised to load the Analysis Toolpak to help you with material in Unit 5.2 onwards. Instructions may be found here:

https://support.office.com/en-gb/article/Load-the-Analysis-ToolPak-6a63e598-cd6d-42e3-9317-6b40ba1a66b4?ui=en-US&rs=en-GB&ad=GB