Mathematics and Statistics Undergraduate Handbook

Supplement to the Handbook

Honour School of Mathematics and Statistics Syllabus and Synopses for Part C 2012–2013 for examination in 2013

Contents

Note:

(a) **The MS2b course will be in Michaelmas Term.** (In previous years it has usually been in Hilary Term.)

(b) **MS1b will be assessed by written examination.** (In the last few years it has been assessed by mini-project.)

(c) **There are two new courses:**

- **MS5a Probability and Statistics for Network Analysis**
- **MS6b Advanced Simulation Methods**

(d) The course MS1a Graphical Models and Inference will NOT run in 2012-13.

Every effort is made to ensure that the list of courses offered is accurate at the time of going online. However, students are advised to check the up-to-date version of this document on the Department of Statistics website.
Notice of misprints or errors of any kind, and suggestions for improvements in this booklet should be addressed to the Academic Administrator in the Department of Statistics.

Updated September 2012

1 Honour School of Mathematics and Statistics

See the current edition of the Examination Regulations at http://www.admin.ox.ac.uk/examregs/ for the full regulations governing these examinations.

The examination conventions can be found at http://www.stats.ox.ac.uk/current_students/bammath/examinations

Students staying on to take the four-year course will take the equivalent of two units from Part C in their fourth year, and will also offer a dissertation on a statistics project. Of the two units from Part C, at least half a unit will be from the schedule of 'Statistics' units for Part C.

This booklet describes the half-units available in Part C. Information about dissertations/ statistics projects is available on the Department of Statistics website at http://www.stats.ox.ac.uk/current_students/bammath/projects

All of the units described in this booklet are "M-level".

We ask that you register by the end of week 10 Trinity Term 2012 for classes for the Mathematics/ Statistics courses that you wish to take. A registration form is attached to these synopses. Some combinations of subjects are not advised and lectures in these subjects may clash. However, when timetabling lectures we will aim to keep clashes to a minimum.

Language Classes: If spaces are available, Mathematics and Statistics students are also invited to apply to take classes in a foreign language. In 2012-2013, French and German language classes will be offered. Students' performance in these classes will not contribute to the degree classification in Mathematics and Statistics. However, upon successful completion of the course, students will be issued with a certificate of achievement which will be sent to their college. See http://www.maths.ox.ac.uk/current-students/undergraduates/handbooks-synopses/maths for further information.

2 Statistics units and half units

2.1 **MS1b: Statistical Data Mining** - 12HT plus 4 1-hour computer practical classes

Level: M-level
Method of Assessment: Written examination.
Weight: Half- unit.

Recommended Prerequisites
Part A Probability and Statistics. BS1 *Applied Statistics* would be an advantage.

Aims & Objectives
'Data mining' is now widely used to find interesting patterns in large databases, for example in insurance, in marketing and in many scientific fields. With large amounts of data we can search for quite subtle patterns. This course concentrates on the statistical tools used to identify patterns, and then to identify those which are interesting not just the result of chance associations.

Synopsis
Fundamentals of pattern recognition, machine learning and data mining.

Exploratory methods: principal components analysis, biplots, independent component analysis, multidimensional scaling.

Cluster Analysis: K-means, hierarchical methods, vector quantisation, self-organising maps.

Linear discriminant analysis, logistic discrimination, linear separation.

Feed-forward neural networks, Classification trees, ensemble methods, V-fold cross-validation. trees, tree ensembles

Reading
C. Bishop, *Neural Networks for Pattern Recognition*, Oxford UP (1995).
D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*, MIT Press (2001).
T Hastie, R Tibshirani, J Friedman, *Elements of Statistical Learning,* Springer (2009)

Further Reading
B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge UP (1996).

2.2 **MS2a: Bioinformatics and Computational Biology**- 16MT

Level: M-level
Method of Assessment: This course is assessed by mini-project.
Weight: Half- unit

Recommended Prerequisites
None. In particular, no previous knowledge of Genetics will be necessary.

Aims & Objectives
Modern molecular biology generates large amounts of data, such as sequences, structures and expression data, that needs different forms of statistical analysis and modelling to be properly interpreted. The field of Computational Biology is viewed as the study of the models, statistical methodology and algorithms needed to do bioinformatics analysis. This course aims to present core topics of these fields with an emphasis on modelling and computation.

Synopsis
Stochastic Models of Sequence and Genome Evolution including models of single nucleotide/amino acid/codon evolution.
Phylogenies: enumerating phylogenies, the probability of sequences related by a specified phylogeny, the minimal number of events needed to explain a data set (Parsimony).

Alignment Algorithms. Comparing 2 strings, an arbitrary number of strings, find segments of high similarity in 2 strings. The analogous algorithms for probabilistic models of insertion-deletions [statistical alignment].

Genome annotation – how are proteins and RNA genes predicted for a DNA sequence? Common patterns in a set of sequences. How are identical patterns found in independent strings? How are conserved patterns found by evolutionary analysis (footprinting)?

Network Inference and Network Evolution. Networks can be inferred since they govern observable phenomena like expression levels, concentrations of metabolites. Network evolve over time in analogy with sequences and their evolution must be modelled.

Comparative Biology of Structures, Patterns, Shapes and Dynamical Systems. Any biological object that can be called homologous has an evolutionary model that needs stochastic modelling and this is not confined to sequences and networks.

Integrative Genomics – models analysing multiple types of high-throughput data simultaneously.

Reading
Z Yang, *Computational Molecular Evolution*, Oxford University Press (2004).
C. Semple and M. Steel, *Phylogenetics*, Oxford University Press (2003).
Durbin et al., *Biological Sequence Analysis*, Cambridge University Press (1998).
T. Jiang et al., (editors) *Current Topics in Computational Biology*, MIT Press, (2003).
M.S. Waterman et al., *Computational Genome Analysis: An Introduction*, Springer (2004).

2.3 **MS2b: Stochastic Models in Mathematical Genetics –** 16MT

Level: M-level
Method of Assessment: written examination
Weight: Half- unit

Aims & Objectives
The aim of the lectures is to introduce modern Stochastic models in Mathematical Population Genetics and give examples of real world applications of these models. Stochastic and Graph theoretic properties of coalescent and gene trees are studied in the first eight lectures. Diffusion processes and extensions to model additional key biological phenomena, are studied in the second eight lectures.

Synopsis
Evolutionary models in Mathematical Genetics:
The Wright-Fisher model. The Genealogical Markov chain describing the number ancestors back in time of a collection of genes.

The Coalescent process describing the stochastic behaviour of the ancestral tree of a collection of genes. Mutations on ancestral lineages in a coalescent tree. Inferring the time to the most recent common ancestor in a sample of genes from the number of mutations occurring to the genes. Models with a variable population size.

The frequency spectrum and age of a mutation. Ewens' sampling formula for the probability distribution of the allele configuration of genes in a sample in the infinitely-many-alleles model. Hoppe's urn model for the infinitely-many-alleles model.

The infinitely-many-sites model of mutations on DNA sequences. Gene trees as perfect phylogenies describing the mutation history of a sample of DNA sequences. Graph theoretic constructions and characterizations of gene trees from DNA sequence variation. Gusfield's construction algorithm of a tree from DNA sequences. Examples of gene trees from data. The probability distribution of a gene tree.

Modelling biological forces in Population Genetics:

Recombination. The effect of recombination on genealogies. Detecting recombination events under the infinitely-many-sites model. Hudson's algorithm. Haplotype bounds on recombination events.

Modelling recombination in the Wright-Fisher model. The coalescent process with recombination: the ancestral recombination graph. Properties of the ancestral recombination graph. Introduction to diffusion theory. Tracking mutations forward in time in the Wright-Fisher model. Modelling the frequency of a neutral mutation in the population via a diffusion process limit. The generator of a diffusion process with two allelic types. The probability of fixation of a mutation. The connection between the coalescent process and diffusion process as a duel process.

Genic selection. Extension of results from neutral to selection case. Behaviour of selected mutations.

Reading
R. Durrett, *Probability Models for DNA Sequence Evolution*, Springer (2008).
A.Etheridge, Some Mathematical Models from Population Genetics. Ecole d'Eté de Probabilities de Saint-Flour XXXIX-2009, Lecture Notes in Mathematics 2012.
W. J. Ewens, *Mathematical Population Genetics,* 2nd ed, Springer (2004).
J. R. Norris, *Markov Chains*, Cambridge University Press (1999).
M. Slatkin and M. Veuille, *Modern Developments in Theoretical Population Genetics*, Oxford Biology (2002).
S. Tavaré and O. Zeitouni, *Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilities de Saint-Flour XXXI - 2001*, Lecture Notes in Mathematics 1837. Springer (2004).

2.4 MS4b/C11.1b: **Probabilistic Combinatorics** - 16HT

[In the synopses booklet for Mathematics Part C, this course appears in the Mathematics Department Units section; in this booklet it is in the Statistics section. For any Mathematics and Statistics student taking this half-unit, it will count as a Statistics half-unit. Note that a recommended prerequisite is C11.1a Graph Theory, which is available to Mathematics and Statistics students in Section 3 – C11.1a counts as a Mathematics half-unit for Mathematics and Statistics students.]

Level: M-level
Method of Assessment: written examination
Weight: Half- unit

Recommended Prerequisites
C11.1a *Graph Theory*. Part A Probability.

Learning outcomes
To develop an appreciation of probabilistic methods in discrete mathematics.

Aims and objectives
Probabilistic combinatorics is a very active field of mathematics, with connections to other areas such as computer science and statistical physics. Probabilistic methods are essential for the study of random discrete structures and for the analysis of algorithms, but they can also provide a powerful and beautiful approach for answering deterministic questions. The aim of this course is to introduce some fundamental probabilistic tools and present a few applications.

Synopsis
First-moment method, with applications to Ramsey numbers, and to graphs of high girth and high chromatic number.
Second-moment method, threshold functions for random graphs.
Lovasz Local Lemma, with applications to two-colourings of hypergraphs (property B), and to Ramsey numbers.
Cherno_ bounds, concentration of measure, Janson's inequality.
Branching processes and the phase transition in random graphs.
Clique and chromatic numbers of random graphs.

Reading
N. Alon and J.H. Spencer. *The Probabilistic Method*, Second edition, Wiley, 2000.

Further reading:
B. Bollobas, *Random Graphs*, second edition, CUP, 2001.
M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, B. Reed, ed., P*robabilistic Methods for Algorithmic Discrete Mathematics*, Springer, 1998.
S.Janson, T. Luczak and A.Rucinski, *Random Graphs*, John Wiley and Sons, 2000.
M. Mitzenmacher and E. Upfal. *Probability and Computing : Randomized Algorithms and Probabilistic Analysis,* Cambridge University Press, New York (NY), 2005.
M. Molloy and B. Reed, *Graph Colouring and the Probabilistic Method* (Springer, 2002).
R. Motwani and P. Raghavan, *Randomized Algorithms* (CUP, 1995).

2.5 **MS5a Probability and Statistics for Network Analysis** – 16 MT

Level: M-level
Method of Assessment: written examination
Weight: Half- unit

Number of classes: 4 – weeks 2, 4, 6, 8
Number of computer practicals: 3- weeks 3, 5, 7

Recommended prerequisites: Part A *Probability* and *Statistics*

Aims and Objectives
Many data come in the form of networks, for example friendship data and protein-protein interaction data. As the data usually cannot be modelled using simple independence assumptions, their statistical analysis provides many challenges. The course will give an introduction to the main problems and the main statistical techniques used in this field. The techniques are applicable to a wide range of complex problems. The statistical analysis benefits from insights which stem from probabilistic modelling, and the course will combine both aspects.

Synopsis
Exploratory analysis of networks. The need for network summaries. Degree distribution, clustering coefficient, shortest path length. Motifs.
Probabilistic models: Bernoulli random graphs, geometric random graphs, preferential attachment models, small world networks, inhomogeneous random graphs, exponential random graphs. Small subgraphs: Stein's method for normal and Poisson approximation. Dense graphs: normal approximations, limiting behaviour.Sparse graphs: Poisson approximations, limiting behaviour. Branching process approximations: moment-generating functions, threshold behaviour. Application of branching process approximations: giant component, shortest path. Weighted graphs, directed graphs.
Statistical analysis of networks:Parameter estimation for models: maximum-likelihood estimation, method of moments, pseudo-likelihood, computer-intensive approaches. Inference from networks: vertex characteristics and missing edges. Nonparametric graph comparison: subgraph counts, subsampling schemes, MCMC methods. Examples: protein interaction networks, social ego-networks.

Reading:
R. Durrett: *Random Graph Dynamics*. Cambridge University Press 2007.

Further reading:
S.N. Dorogovtsev and J.F.F. Mendes: *Evolution of Networks*. Oxford University Press 2003.
M. Newman: *Networks: An Introduction*. Oxford University Press 2010.
M. Newman, A.-L. Barabasi, D.J. Watts (eds.). *The Structure and Dynamics of Networks*. Princeton University Press 2006.
S. Wasserman and K. Faust: *Social Network Analysis*. Cambridge University Press 1994.

2.6 **MS6b Advanced Simulation Methods** - 16 lectures HT

Level: M-level
Methods of Assessment: This course is assessed by written examination.

Recommended Prerequisites
Part A Probability and Statistics. Part A *Simulation* and BS3a *Applied Probability* would be an advantage but are not necessary.

Aims & Objectives
The aim of the lectures is to introduce modern simulation methods.
This course concentrates on Markov chain Monte Carlo (MCMC) methods and Sequential Monte Carlo (SMC) methods. Examples of applications of these methods to complex inference problems will be given.

Synopsis
Classical methods: inversion, rejection, composition.
MCMC methods: Metropolis-Hastings algorithm, Gibbs sampling, elements of discrete-time general state-space Markov chains theory.

Advanced MCMC methods: slice sampling, tempering/annealing, path sampling, perfect simulation, adaptive MCMC.
Importance sampling, sequential importance sampling.

SMC methods: nonlinear filtering, SMC samplers, SMC/MCMC, elements of theory. Likelihood-free methods.

Reading
C.P. Robert and G. Casella, Monte Carlo Statistical Methods, Springer-Verlag.
Further reading
J.S. Liu, Monte Carlo Strategies in Scientific Computing, Springer-Verlag.

3 Mathematics units and half units

The Mathematics units and half units that students may take are drawn from Part C of the Honour School of Mathematics. For full details of these units and half-units, see the Syllabus and Synopses for Part C of the Honour School of Mathematics, which are available on the web at
http://www.maths.ox.ac.uk/current-students/undergraduates/handbooks-synopses/maths

The Mathematics units and half-units that are available are as follows:
C1.1a Model Theory
C1.1b Gödel's Incompleteness Theorems
C1.2a Analytic Topology
C1.2b Axiomatic Set Theory
C2.1a Lie Algebras
C2.1b Representation Theory of Symmetric Groups
C2.2a Commutative Algebra
C2.2b Homological Algebra
C2.3b Infinite Groups
C3.1a Algebraic Topology
C3.2b Geometric Group Theory
C3.3b Differentiable Manifolds
C3.4a Algebraic Geometry
C3.4b Lie Groups
C4.1a Functional Analysis
C4.1b Banach and C* -Algebras
C5.1a Methods of Functional Analysis for PDEs
C5.1b Fixed Point Methods for Nonlinear PDEs
C5.2b Calculus of Variations
C6.1a Solid Mechanics
C6.1b Elasticity and Plasticity
C6.2a Statistical Mechanics
C6.3a Perturbation Methods
C6.3b Applied Complex Variables
C6.4a Special Topics in Fluid Mechanics
C6.4b Stochastic Modelling of Biological Processes
C7.1b Quantum Theory and Quantum Computers
C7.2a General Relativity I
C7.2b Relativity II
C8.1a Mathematics of Geoscience
C8.1b Mathematical Physiology
C9.1a Modular Forms
C9.1b Elliptic Curves
C10.1a Stochastic Differential Equations
C10.1b Brownian Motion in Complex Analysis
C11.1a Graph Theory
C12.1a Numerical Linear Algebra
C12.1b Continuous Optimization
C12.2b Finite Element Methods for Partial Differential Equations.
C12.3b Approximation of Functions

4 Registration

We ask that students register in advance for the classes they wish to take, by the end of week 10 Trinity Term 2012, using the form overleaf.

Because of the large number of options which are available in Part C, some lectures will clash. See the Syllabus and Synopses for Part C of the Honour School of Mathematics for information on which lectures may clash.

FHS MATHEMATICS AND STATISTICS
REGISTRATION FORM: PART C CLASSES 2012-2013

SURNAME . . . . . .FIRST NAME . . . . . .

EMAIL ADDRESS . . . . . .

COLLEGE . . . . . .

Note: As described in Section 1, you need to do a total of 2 units in Part C (in addition to doing a dissertation on a statistics project). At least half a unit will be from the schedule of 'Statistics' units for Part C

Please give details of the subjects in which you wish to take classes.
I wish to take classes in the following subjects: [Please Tick]

☐ MS1b Statistical Data Mining (HT)

☐ MS2a Bioinformatics and Computational Biology (MT)

☐ MS2b Stochastic Models in Mathematical Genetics (MT)

☐ MS4b Probabilistic Combinatorics (HT)

☐ MS5a Probability and statistics for network analysis (MT)

☐ MS6b Advanced Simulation Methods (HT – this course will run if teaching resources allow)

For Mathematics units or half-units, please list the unit or half-unit code and name:
Unit code Unit name

. . . . . .

. . ..

. . . . . .

Please return this form to the Academic Administrator, Department of Statistics, 1 South Parks Road, by the end of week 10 Trinity Term 2012.