Mathematics and Statistics Undergraduate Handbook
Supplement to the Handbook

Honour School of Mathematics and Statistics
Syllabus and Synopses for Part C 2013–2014
for examination in 2014

Contents

Note:

(a) The MS2b course will run in Michaelmas Term.

(b) There is a new course: MS6a Modern Survival Analysis

(c) MS1b has been renamed Statistical Data Mining and Machine Learning

(c) The courses MS1a Graphical Models and Inference and MS2a Bioinformatics and Computational Biology will NOT run in 2013-14.

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 June 2013

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

In Part C,

- each candidate shall offer **five units** from the schedule of units for Part C
- and each candidate shall also offer **a dissertation** on a statistics project.

Of the five units from Part C, at least one unit should be from the schedule of 'Statistics' units.

Note: The dissertation is the equivalent of 3 units, so Part C is the equivalent of 8 units in total (5 from lecture courses, 3 from dissertation).

Units from the schedule of 'Mathematics Department units' for Part C of the Honour School of Mathematics are also available – see Section 3.

This booklet describes the 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 2013 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 2013-2014, French and Spanish 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

2.1 **MS1b: Statistical Data Mining and Machine Learning** – 16 HT

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

Recommended Prerequisites
Part A *Probability* and *Statistics*.
BS1 *Applied Statistics* would be an advantage, BS2a *Statistical Inference* would be helpful, but these are not essential.

Aims & Objectives
Data mining and machine learning are now widely used across many scientific and engineering disciplines. These encompass methods to find interesting patterns and to predict accurately in large datasets.

This course introduces the essential data mining and machine learning tools, how they are used in practice, what they are used for, and their underpinning statistical principles. The course will also cover computational considerations and how these tools can scale to large datasets.

Synopsis
Fundamentals of machine learning and data mining:
Statistical learning theory, bias/variance tradeoff, generalization and overfitting, regularization, decision theory. Evaluating learning methods with training/test sets and cross-validation.

Unsupervised learning:
Dimension reduction and visualization: principal components analysis, biplots.
Clustering: mixture models, K-means, hierarchical clustering.
Probabilistic latent variable models and the EM algorithm.

Supervised learning:
Generative methods: naive Bayes, linear discriminant analysis.
Discriminative methods: K-nearest neighbours, logistic regression, neural networks, support vector machines and the kernel trick, decision trees.
Ensemble methods: bagging, boosting, random forests.

Reading
C. Bishop, *Pattern Recognition and Machine Learning*, Springer (2007).
T. Hastie, R. Tibshirani, J Friedman, *Elements of Statistical Learning*, Springer (2009)
B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge UP (1996).

Further Reading
D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*, MIT Press (2001).
P. Hart, D. Stork, R. Duda, *Pattern Classification*, Wiley-Interscience (2000).

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

Level: M-level
Method of Assessment: written examination
Weight: Unit

Recommended Prerequisites
Part A *Probability*.
Part B *Applied Probability* would be helpful.

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 genealogical 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 DNA sequences.

The Coalescent process describing the stochastic behaviour of the ancestral tree of a collection of DNA sequences. Mutations on ancestral lineages in a coalescent tree. 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 DNA sequences 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.

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. 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).

2.3 **MS4b/C11.1b: Probabilistic Combinatorics** – 16 HT

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

Recommended Prerequisites:
Part B *Graph Theory* and Part A *Probabilit*y. C11.1a *Combinatorics* is not an essential prerequisite for this course, though it is a natural companion for it.

Overview
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.

Learning Outcomes
The student will have developed an appreciation of probabilistic methods in discrete mathematics.

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.
Lovász Local Lemma, with applications to two-colourings of hypergraphs, 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* (third edition, Wiley, 2008).

Further Reading
B. Bollobás, *Random Graphs* (second edition, Cambridge University Press, 2001).

M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, B. Reed, ed., *Probabilistic 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* (Cambridge University Press, 1995).

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

Level: M-level
Method of Assessment: Written examination
Weight: Unit

For this course, 2 lectures and 2 intercollegiate classes are replaced by 2 practical classes. (The total time for this course is the same as for other Part C courses.)

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: shortest path.

Statistical analysis of networks: Parameter estimation for models: maximum-likelihood estimation, method of moments, 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.
R.v.d. Hofstad, Random Graphs and Complex Networks, Manuscript available at
http://www.win.tue.nl/~rhofstad/RGCN.html
E.D Kolaczyk, *Statistical Analysis of Network Data*, Springer 2009
M. Newman: *Networks: An Introduction*. Oxford University Press 2010.

Further reading:
S.N. Dorogovtsev and J.F.F. Mendes: *Evolution of Networks*. Oxford University Press 2003.
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.

2.5 **MS6a Modern Survival Analysis** – 16 MT

Level: M-level
Method of Assessment: written examination
Weight: Unit

Recommended prerequisites: Part A *Probabilit*y and *Statistics*. BS3a *Applied Probability* and BS3b *Statistical Lifetime Models* would be helpful. Basic computer skills, including some familiarity with R – this is on a level that an interested student with no R experience could acquire in a few hours.

Learning outcomes
Students will learn how to use the basic mathematical tools that are used to evaluate survival models. They will learn both mathematical facts about standard survival models currently in use, and how to use standard R packages to fit these models to data. They will learn how to interpret models critically, and how to choose an appropriate model.

Aims and Objectives
Survival analysis is a branch of statistics in huge demand, particularly in the biomedical area, but also for applications to engineering, demography, and ecology. New experimental methods and new applications have transformed the subject in recent decades. This course aims to develop a solid ground of techniques for analysing a wide range of data types, including longitudinal covariates, which are increasingly common, and complicated observational settings. Starting from a review and development of the theory of counting processes, this course presents the theory of survival and event-history analysis in a general framework. Cox proportional hazards regression and additive hazards are the key models that will be developed, but the emphasis is on general principles of hypothesis testing, model selection and diagnostics principles that can be applied to a wide range of models and a wide range of data types.

Synopsis
Point processes and compensators. Introduction to martingales.
Non-parametric estimation. Semi-parametric estimation and the Cox model.
Additive hazards regression. Varieties of data, such as current-status and randomly truncated data.
Model selection: Hypothesis testing and information criteria. Testing shapebased

hypotheses (in particular, increasing hazards). Isotonic approaches (introduction to the concept).
Model Diagnostics: Graphical methods, Cox-Snell Residuals. Martingale Residuals. Deviance Residuals, Schoenfeld residuals.
Influence and robustness in survival models.
Measurement error and longitudinal data: Modified partial likelihoods and introduction to joint modeling.
Recurrent event models. Frailty models.

Reading
John P. Klein and Melvin L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, 2d edition
Terry M. Theneau and Patricia M. Grambsch, *Modeling Survival Data: Extending the Cox Model*
Odd O. Aalen, Ornulf Borgan, Hakon K. Gjessing, *Survival and Event History Analysis: A Process Point of View*

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

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

Recommended Prerequisites
Part A Probability and Statistics. Part A *Simulation* and BS3a/B12a *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, reversible jump MCMC, pseudo-marginal MCMC.

Importance sampling, sequential importance sampling.
SMC methods: nonlinear filtering.

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

The Mathematics units that students may take are drawn from Part C of the Honour School of Mathematics. For full details of these 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 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 Linear Operators
C5.1a Methods of Functional Analysis for PDEs
C5.1b Fixed Point Methods for Nonlinear PDEs
C5.2b Calculus of Variations
C5.3b Hyperbolic Equations
C6.1a Solid Mechanics
C6.1b Elasticity and Plasticity
C6.2a Statistical Mechanics
C6.2b Networks
C6.3a Perturbation Methods
C6.3b Applied Complex Variables
C6.4a Topics in Fluid Mechanics
C6.4b Stochastic Modelling of Biological Processes
C6.5b Mathematical Mechanical Biology
C7.1b Quantum Theory and Quantum Computers
C7.2a General Relativity I
C7.2b General Relativity II
C8.1a Mathematical Geoscience
C8.1b Mathematical Physiology
C9.1a Modular Forms
C9.1b Elliptic Curves
C9.2a Analytic Number Theory
C10.1a Stochastic Differential Equations
C10.1b Brownian Motion and Conformal Invariance
C11.1a Combinatorics
C12.1a Numerical Linear Algebra
C12.1b Continuous Optimization
C12.2a Approximation of Functions
C12.2b Finite Element Methods for Partial Differential Equations.

4 Registration

We ask that students register in advance for the classes they wish to take, by the end of week 10 Trinity Term 2013, 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 2013-2014

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

EMAIL ADDRESS . . . . . .

COLLEGE . . . . . .

Note: As described in Section 1, you need to do a total of 5 units in Part C (in addition to doing a dissertation on a statistics project). At least one 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 and Machine Learning – HT
	MS2b Stochastic Models in Mathematical Genetics – MT
	MS4b/C11.1b Probabilistic Combinatorics - HT
	MS5a Probability and Statistics for Network Analysis – MT
	MS6a Modern Survival Analysis – MT
	MS6b Advanced Simulation Methods – HT

For Mathematics units, please list the unit code and name:

. . . . . .

. . ..

. . . . . .
Please return this form to the Academic Administrator, Department of Statistics, 1 South Parks Road, by the end of week 10 Trinity Term 2013.