# STA410H1F/2102H1F Statistical Computation

**Instructor:** K. Knight; e-mail: keith@utstat.utoronto.ca

**Office hours:** To be determined. The easiest way to reach me is via e-mail and most questions can be addressed by this means.

**Textbook:** The required text is *Computational Statistics* by James Gentle (Springer). This book is available on-line via the UofT Library system; a link to the book is provided on Quercus.

This book gives a very comprehensive overview of computational methods used in probability and statistics; most of the material for this course is covered in chapters 2–7 and 11. I will also supplement the text with a number of handouts, particularly on topics that the text does not cover. (All of the handouts as well as additional material will be available on Quercus.)

**About the course.** The goal of this course is to give an overview of some of the computational methods that are useful in statistics. The first part of the course will focus on basic algorithms, such as the Fast Fourier Transform (and related methods) and methods for generating random variables. The second part of the course will focus on numerical methods for linear algebra and optimization (for example, computing least squares estimates and maximum likelihood estimates). Along the way, you will learn some basic theory of numerical analysis (computational complexity, convergence rates of algorithms) and you will encounter some statistical methodology that you may not have seen in other courses.

**Prerequisites:** The nominal prerequisites for this course are MAT223H/240H, STA302H and CSC108H/120H/121H/148H – these should give you the sufficient background in both statistics and computer programming to handle the course material. A solid foundation in linear algebra is very useful for this course.

**Computing:** All of the homework assignments and exams will involve some computational work. The R statistical language will be used - this is free software and can be downloaded (for Windows, Mac, and Linux operating systems) from cran.utstat.utoronto.ca. Documentation for R can also be found at www.r-project.org and this site also lists some books related to R. Of interest to many of you will be RStudio, which provides a very nice environment for using R; information on RStudio (including downloads) can be found at www.rstudio.com.

A useful book that gives a good introduction to R programming is

> *A First Course in Statistical Programming with R* by Braun and Murdoch (Cambridge University Press)

This book also covers many of the topics in this course although not at the same depth that we will require. (The text by Gentle also uses R to illustrate the implementation of various algorithms.) The strength of R lies in its flexibility and in its graphical facilities; however, the initial learning curve of R can be quite steep as R is very much a programming language (which makes it very useful for this course), albeit one with a lot of built-in functionality. No previous knowledge of this software is required and all relevant documentation will be made available to students as needed. (Students may also choose to use an alternative language such as Matlab or Python; however, no documentation will be provided.)

**Grading scheme:** Assignments (3 in total) 15%; midterm exam 35%; final exam 50%.

- Assignments will involve both mathematical exercises as well as applying numerical methods learned in class. Two assignments will be handed in before the midterm and one after.

- Students are encouraged to work together on assignments. However, note that any work submitted for grading must be your own original work. According to the University's policies on academic integrity, "working too closely with another student on an individual assignment so that the end result is too similar" is considered an academic offense. **Students should familiarize themselves with these policies, which can be found at** `www.artsci.utoronto.ca/osai/students`**.**

- The midterm exam will be a 24 hour "take-home" exam; it will be made available on October 28 (at 10am EDT) and will be due 24 hours later (10am on October 29). There will be no lecture on October 28. **If you are unable to complete the exam due to illness or other circumstances beyond your control (with appropriate medical or other documentation), the weight from the midterm will be carried over to the final exam.**

- The final exam will have a similar format to the midterm exam; it will be a 48 hour take-home exam and will take place at the end of term at a date and time to be determined later.

**Course outline:** The following is a list of topics to be covered in this course. The order of coverage may be slightly different.

1. Basics
   - floating point representation
   - round-off error and catastrophic cancellation
   - sparse matrices
   - Fast Hadamard and Fourier transforms

2. Generation of random variates
   - generating uniform random variates
   - methods for continuous distributions - inverse, rejection sampling
   - methods for discrete distributions
   - Markov Chain Monte Carlo
   - Monte Carlo integration and simulation

3. Numerical linear algebra
   - solving linear equations - Gaussian elimination
   - matrix factorizations - Cholesky, QR, singular value decomposition
   - iterative matrix methods
   - solving least squares problems
   - randomized numerical linear algebra

4. Optimization
   - fixed point, Newton-Raphson, gradient and coordinate descent algorithms, stochastic gradient descent
   - applications to parameter estimation
   - iterative reweighted least squares algorithm
   - EM and MM algorithms
   - interior and exterior point algorithms for constrained optimization
   - simulated annealing for optimization