BIOSTATISTICS 235
Statistical Computing - Basic Principles and Applications

Description:

The objective of this course is to prepare the student for effective use of the ever increasing computing power and introduce basic concepts and methods of computing as tools for statistical work and research.

The three main themes are software design, numerical methods and simulation methods. Topics include software development, algorithms and data structures, enumeration problems, some computational biology algorithms, function approximation, matrix computations, linear and nonlinear systems, the EM algorithm, simulation experiments, Gibbs sampling and Markov chain Monte-Carlo, symbolic computation and document preparation (LaTeX). Most topics will be implemented in actual working programs and documents.

Prerequisites: BIOS 160 and 161, basic statistical methods (example: BIOS 162 and 163), and familiarity with at least one computer system and programming language (R, SAS/IML, FORTRAN, PASCAL, etc.).

Instructor: Bahjat Qaqish, 966-7271, Room 3105-B, bahjat_qaqish@unc.edu
Required textbook: Monahan, J.F. (2001). Numerical Methods of Statistics.
Optional textbook: Robert, C.P. & Casella, G. (2000). Monte Carlo Statistical Methods.
Programming languages used in the course: C++, MATLAB, R.
Computer system: Sun workstations (running UNIX).
Class directory: /distrib/bios235/
Website: ftp://ftp.bios.unc.edu/distrib/bios235/

Outline:

Introduction to UNIX, introduction to C++, compiling and running programs.

Computer organization: bits, bytes, words, number systems, floating point representations (IEEE 754), floating point errors.

Elements of programming: data structures and algorithms, software design and development basics, program performance, worst case, best case, average case, profiling, examples (finding an array minimum, sorting algorithms).

Project maintainence with “make”, basic ideas of data abstraction, top-down design, bottom-up design, layering, modularization, testing, debugging, documentation.

Enumeration problems: Spearman’s correlation, the Wilcoxon rank sum test, 2×2 tables, $R \times C$ tables.

Dynamic programming: basic ideas, application to string matching of DNA sequences. Computations for hidden Markov chains.

Solving non-linear equations in one variable: bisection, Newton-Raphson, secant, error analysis, convergence rates, linear and quadratic convergence.

Function approximation: general principles, Taylor series, polynomial approximation of functions, polynomial interpolation - Lagrange's method, Hermite interpolation, cubic splines (natural, clamped), polynomial interpolation - error analysis.

Numerical integration and differentiation: open and closed rules (trapezoidal, Simpson's, ...), globally-adaptive quadratue, Gaussian quadratue, Gauss-Legendre quadratue, Gauss-Hermite quadratue, adaptive Gauss-Hermite quadratue, applications to generalized mixed models.

Matrix computations: linear spaces, vector and matrix norms, projectors, orthogonal projectors, QR decomposition, QR via the Gram-Schmidt algorithm, Householder reflectors, QR via Householder reflectors, SVD, eigenvalues and eigenvectors, Cholesky decompostion, back substitution, forward substitution, the least-squares problem, error analysis, condition number, accuracy and stability (general), accuracy and stability of least-squares algorithms, other LU factorizations (Doolittle, Crout), Gaussian elimination, applications to linear regression, iterative methods for linear systems (Jacobi, Gauss-Seidel), iteration and convergence for linear systems.

Solving non-linear systems: Newton-Raphson in several dimensions, applications to maximum-likelihood estimation, Fisher scoring, conjugate gradient.

The EM algorithm: definition, properties, advantages and disadvantages, computing the observed information, acceleration, examples, applications to mixture models, misclassified data, generalized mixed models.

Simulation: overview, random number generation and testing, design of simulation experiments, generating uni- and multi-variate, continuous and discrete, accept-reject methods, Markov-Chain Monte Carlo, coalescence, coupling from the past, the perfect sampler, Gibbs sampling, Metropolis algorithm, Metropolis-Hastings algorithm, importance sampling, Monte-Carlo integration, variance reduction, Rao-Blackwellization, sampling with and without replacement from a finite population, random permutations, parametric and non-parametric bootstrap sampling.

Example simulation experiment: The robustness of the one-sample t confidence interval.

Example simulation experiment: Simulating a stochastic process.

Introduction to symbolic computation (Mathematica).

Introduction to document preparation in LaTeX.

Linking C code to R.

Linking C code to MATLAB.

REFERENCES

Algorithms:

Ammeraal, L. (1992). Programs and Data Structures in C, 2nd ed. John Wiley & Sons, New York.

Cormen, T.H., Leiserson, C.E., Rivest, R.L. & Stein, C. (2001). Introduction to Algorithms, 2nd ed. MIT Press.

Knuth, D.E. (1998). The Art of Computer Programming, volumes 1-3. Addison-Wesley, New York. Other volumes are scheduled for 2005.

Sedgewick, R. (1990). Algorithms in C. Addison-Wesley, New York.

Sedgewick, R. (2001). Algorithms in C, 3rd ed. (5 parts). Addison-Wesley, New York. There are C++ and JAVA versions of this book.

Numerical methods:

Evans, M. & Swartz, T. (2000). Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford University Press.

Kincaid, D., & Cheney, W. (2001). Numerical Analysis: Mathematics of Scientific Computing. Brooks/Cole Publishing Company.

Ralston, A. & Rabinowitz, P. (2001). A First Course in Numerical Analysis, 2nd ed. Dover Publications. Original 2nd Edition 1978.

Matrix analysis:

Harville, D. A. (1997). Matrix Algebra from a Statistician's Perspective. Springer Verlag.

Horn, R. A. & Johnson, C. R. (1990). Matrix Analysis. Cambridge University Press.

Horn, R. A. & Johnson, C. R. (1994). Topics in Matrix Analysis. Cambridge University Press.

Meyer, C.D. (2001). Matrix Analysis and Applied Linear Algebra. Society for Industrial & Applied Mathematics.

Laub, A. J. (2004). Matrix Analysis for Scientists and Engineers. Society for Industrial & Applied Mathematics.

Searle, S. R. (1982). Matrix Algebra Useful for Statistics. Wiley-Interscience.

Matrix computations:

Bjorck, Ake (1996). Numerical Methods for Least Squares Problems. Society for Industrial & Applied Mathematics.

Golub, G.H. & Van Loan, C.F. (1996). Matrix Computations, 3rd ed. The Johns Hopkins University Press, Baltimore, MD.

Lawson, C.L. & Hanson, R.J. (1995). Solving Least Squares Problems (Classics in Applied Mathematics, No 15). Society for Industrial & Applied Mathematics. Originally published in 1974.

Stewart, G.W. (1998). Matrix Algorithms, Volume I: Basic Decompositions. Society for Industrial & Applied Mathematics.

Stewart, G.W. (2001). Matrix Algorithms, Volume II: Eigensystems. Society for Industrial & Applied

Mathematics.

Watkins, David S. (2002). Fundamentals of Matrix Computations, 2nd ed. John Wiley & Sons, New York.

C:

Ammeraal, L. (1992). C for Programmers, 2nd ed. John Wiley & Sons, New York.

Harbison, S. & Steele, G. (2002). C: A Reference Manual, 5th ed. Prentice-Hall, Englewood Cliffs, NJ.

King, K.N. (1996). C Programming: A Modern Approach. W.W. Norton & Company.

Reek, K.A. (1997). Pointers on C. Addison-Wesley, New York.

C++ (Introductory):

Koenig, A., & Moo, B.E. (2000). Accelerated C++: Practical Programming by Example. Addison-Wesley, New York.

Lippman, S.B., Lajoie, J. & Moo, B. (2005). C++ Primer, 4th ed. Addison-Wesley, New York.

C++ (Intermediate and Advanced):

Barton, J. J. & Nackman L. R. (1994). Scientific and Engineering C++: An Introduction with Advanced Techniques and Examples. Addison-Wesley, New York.

Josuttis, N. M. (1999). The C++ Standard Library: A Tutorial and Reference. Addison-Wesley Professional.

Josuttis, N. M. (2002). Object Oriented Programming in C++. John Wiley & Sons, New York.

Meyers, S. (2001). Effective STL: 50 Specific Ways to Improve Your Use of the Standard Template Library. Addison-Wesley Professional.

Meyers, S. (2005). Effective C++: 55 Specific Ways to Improve Your Programs and Designs (3rd Edition). Addison-Wesley Professional.

Stroustrup, B. (2000). The C++ Programming Language (Special 3rd Edition) Addison-Wesley Professional.

Yang, D. (2000). C++ and Object-Oriented Numeric Computing for Scientists and Engineers. Springer Verlag.

Software development:

Kernighan, B. & Pike, R. (1999). The Practice of Programming. Addison-Wesley, New York.

General statistical computing:

Lange, K. (1999). Numerical Analysis for Statisticians. Springer Verlag.

McCullogh, B.D. (1998). Assessing the reliability of statistical software: part I. *The American Statistician*, **52**, 358-366.

McCullogh, B.D. (1999). Assessing the reliability of statistical software: part II. *The American Statistician*, **53**, 149-159. Correction in **55**, 373.

Monahan, J.F. (2001). Numerical Methods of Statistics. Cambridge University Press.

Thisted, R.A. (1988). Elements of Statistical Computing: Numerical Computation. CRC Press.

Simulation:

Barndorff-Nielsen, O.E., Cox, D.R., & Kluppelberg, C. (2000). Complex Stochastic Systems. CRC Press.

Carlin, B.P. & Louis, T.A. (2000). Bayes and Empirical Bayes Methods for Data Analysis, 2nd ed. CRC Press.

Chen, M.H., Shao, Q.M. & Ibrahim, J.G. (2000). Monte Carlo Methods in Bayesian Computation (Springer Series in Statistics). Springer Verlag.

Fishman, G.S. (1996). Monte Carlo: Concepts, Algorithms, and Applications. Springer Verlag.

Robert, C.P. & Casella, G. (2000). Monte Carlo Statistical Methods. Springer Verlag.

Tanner, M.A. (1996). Tools for Statistical Inference : Methods for the Exploration of Posterior Distributions and Likelihood Functions. Springer Verlag.

Other topics:

Andrews, D.F. & Stafford, J.E. (2000). Symbolic Computation for Statistical Inference (Oxford Statistical Science Series, 21). Oxford University Press.

Dowd, K., & Severance, C.R. (1998). High-Performance Computing, 2nd ed. O'Reilly & Associates.

Griffiths, D.F. & Higham, D.J. (1997). Learning LaTeX. Society for Industrial & Applied Mathematics.

Higham, D.J. & Higham, N.J. (2000). MATLAB Guide. 2nd ed. Society for Industrial & Applied Mathematics.

Kopka H. & Daly, P.W. (2003). A guide to LaTeX. 4th ed. Addison-Wesley, New York.

Peek, J.D., Todino, G. & Strang, J. (2002). Learning the Unix Operating System, 5th ed. O'Reilly & Associates.

Stephen Wolfram (2003). The Mathematica Book, 5th ed. Wolfram Media, Inc.

Internet resources:

The Association of C & C++ Users: http://www.accu.org/

Blitz++ matrix library: http://www.oonumerics.org/blitz/

GSL, GNU Scientific Library: http://www.gnu.org/software/gsl/

The Matrix Template Library: http://www.osl.iu.edu/research/mtl/

National Institute of Standards and Technology: http://math.nist.gov/

Netlib collection of mathematical software, papers, and databases: http://www.netlib.org/

R Project for Statistical Computing: http://www.r-project.org/