**MATH7340: Statistics for Bioinformatics**

**Credit Hours**: 4

**Format**: Online. The materials are delivered online.

**Term:** Spring 2017

**Instructor Information:** **Prasanth George**

**Email:** [p.george@neu.edu](mailto:p.george@neu.edu)

Office: 532 Nightingale Hall

**Course Prerequisites:**  Bioinformatics majors only

**Course Description:** Introduces the concepts of probability and statistics used in Bioinformatics applications, particularly the analysis of microarray data. Statistical computation using the open-source R program will be used throughout the course. The topics include: maximum likelihood; Monte Carlo simulations; false discovery rate adjustment; nonparametric methods including bootstrap and permutation tests; correlation, regression, ANOVA and generalized linear models; preprocessing of microarray data and gene filtering; visualization of multivariate data; machine learning techniques such as clustering, principal components, support vector machine, neural networks and regression tree.

**Course Outcomes:** By the end of this course you will be able to:

* Apply the concepts of data analysis and statistical and probability vocabulary to bioinformatics applications
* Formulate a problem using a statistical model
* Analyze the process used in the derivation of a statistical model for its use in bioinformatics
* Solve a given bioinformatics problem by applying statistical methods
* Produce relevant outputs in R for the statistical procedures
* Interpret the results of a given output to a bioinformatics application

**Course Summary:** This course provides the fundamentals of probability and statistics, as the use of statistical thinking and problem solving is necessary for bioinformatics. We will examine basic concepts, terminology, computations, and procedures that are often used in the field. We will be using the open-source program R to conduct standard statistical data analyses; however, the interpretation of the problem and drawing conclusions in the context of bioinformatics is most substantial. In this course you will have the opportunity to practice the statistical and programming applications that are essential for the degree program, which is designed around quantitative and computational skills used in the field.

**Course Format & Methodology:** This course runs for a total of 15 weeks beginning January 9, 2017 and is delivered online via the NU Online Blackboard (Bb) system accessible at: [nuonline.neu.edu](http://nuonline.neu.edu/). Each week (or module) contains one or more lessons that you begin on Monday and complete by Sunday of the same week. Lessons are organized around specific course topics, and contain readings and multimedia presentations, practice activities that you complete individually or in small groups (online), and assessments. P**lease note that all due dates and times are specified according to the Eastern Standard Time zone (EST);** plan to complete and submit all assignments accordingly.

**Required Textbook & Materials:**

[Statistics Using R with Biological Examples](http://cran.r-project.org/doc/contrib/Seefeld_StatsRBio.pdf)

Kim Seefeld and Ernst Linder

[Applied Statistics for Bioinformatics using R](http://cran.r-project.org/doc/contrib/Krijnen-IntroBioInfStatistics.pdf)

Wim P. Krijnen (WPK)

**Required Software:** [**R**](http://www.r-project.org/)(Instructions for downloading available within Blackboard)

**Participation and Engagement:** Because this course is taken online, your presence in peer-to-peer activities, and your performance on assignments, will serve as indicators of your level of engagement and effort throughout the course. The instructor will provide frequent and varied opportunities for students to receive constructive feedback, help, and clarification on course material throughout the term. Students that complete the self-check assignments, and utilize feedback provided by instructor, can be successful in this course.

**Communication/Submission of Work:** Guidelines for completing and submitting each assignment are posted along with each assignment in Blackboard. Please note that if you are unable to complete an activity within the period it is assigned, a documented compelling excuse (such as hospitalization) is required. In general, assignments will not be accepted late. If there are extenuating circumstances, please contact the instructor beforehand, and late submission or resubmissions can discussed.

**Course Activities and Assignments:** This course includes the following required activities and assignments:

* **Weekly lessons:** Weekly readings and lessons provide the background knowledge, terminology, and practical examples you need in order to understand and correctly apply fundamental course concepts. You are responsible for completing the assigned textbook and other readings and for viewing the presentations and demonstrations included in the lessons. All materials should be completed in the order in which they are presented, and by the due dates specified, within the weekly module.
* **Self-checks:** Each week, you must complete required self-checks that live within the lessons, aimed at enhancing your current understanding, and/or ability to correctly apply concepts covered in the weekly lessons. These self-checks will provide feedback with suggested areas for review.
* **Discussions: Discussions on the blackboard forum are optional and will not be graded.**
* **Homework:** Each module will include homework based on the lessons and readings. Instructions for downloading and submitting the homework are located in Blackboard.
* **Exams:** There will be two exams in this course, a midterm and a final**.** The concepts in this course are interrelated and often build on each other. At these two points during the term, you will be assessed on your cumulative understanding of course topics. Exams will be administered both electronically, via Blackboard, and there will also be a take home portion. You will be provided with detailed guidelines that describe the examination protocol.

**Course Grading Criteria:**

Homework: 60%

Midterm Exam covering Modules 1-7: 15%

Final Exam (cumulative): 25 %

Note that homework grade will be the average of your weekly grades, after dropping the lowest one grade.

**Class Schedule / Topical Outline:***Please note:* for more information about specific assignments and due dates, see instructions within your course site.

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Topics | Readings | Assignments |
| 1  **Getting Started with R** | Installing R  R Basics  R-scripts and Functions | *Statistics Using R with Biological Examples:* Chapter 2  *Applied Statistics for Bioinformatics using R*: Chapter 1 | Introductory Discussion  Homework |
| 2  **Working with Data Sets in R** | Graphical Displays  Data Manipulation  Data Frames | *Statistics Using R with Biological Examples:* Pages 17-24, 45-59  *Applied Statistics for Bioinformatics using R*: Pages 10-12, 17-26 | Homework |
| 3  **Probability Distributions** | Probability Distribution Functions (PDFs)  Calculating Probabilities Using R  Common Probability Distributions  Mean, Variance, and Linear Transformations  Generate Data from Distributions | *Statistics Using R with Biological Examples:* Pages 71-107  *Applied Statistics for Bioinformatics using R*: Chapter 3 (pages 31-43) | Homework |
| 4  **Monte Carlo Simulations** | Generate Data from Multivariate Distributions  Linear Combination of Random Variables  Simulations for Verifying or Deriving Properties  Central Limit Theorem | *Statistics Using R with Biological Examples:* Pages 105-107, 121-130, 217-221  *Applied Statistics for Bioinformatics using R*: Chapter 3 | Discussion: Simulations to Replace Theorems  Homework |
| 5  **Statistical Inferences** | Parameter Estimation  Sample Distributions  Confidence Intervals | *Statistics Using R with Biological Examples:* Chapters 13 and 14.  *Applied Statistics for Bioinformatics using R*: Chapter 4: section 4.1 | Discussion: Bootstrap Method  Homework |
| 6  **Hypothesis Testing** | Hypothesis Testing  T-test for the Mean  Two Sample Tests  Test for a Proportion  Multiple Testing | *Statistics Using R with Biological Examples:* Chapter 14 (pages 240-256)  *Applied Statistics for Bioinformatics using R*: Chapter 4: pages 47-58  Supplemental Reading | Discussion: Interval Estimates  Homework |
| 7  **Nonparametric Hypothesis Testing** | Wilcoxon Tests  Chi-square Test  Fisher’s Exact Test  Tests for Normality and Outlier | *Statistics Using R with Biological Examples:* Chapter 14 (pages 256-262)  *Applied Statistics for Bioinformatics using R*: Chapter 4 (pages 59-69) | Homework  Midterm Exam: Begin |
| 8  **Analysis of Variance ANOVA** | One-way Analysis of Variance  Linear Model and Testing for Pairwise Difference  Two-way ANOVA  Diagnostic Tests and Robust Tests | *Statistics Using R with Biological Examples:* Chapter 15 (pages 263-274)  *Applied Statistics for Bioinformatics using R*: Chapters 5 (pages 73-90). | Homework  Midterm Exam: Due |
| 9  **Regression Analysis** | Correlation Measures  Simple Linear Regression  Diagnostic Tests and Transformations  Multiple Linear Regression | *Statistics Using R with Biological Examples:* Chapter 15 (pages 275-286)  *Applied Statistics for Bioinformatics using R*: Chapter 7: Section 7.3 (pages 130-133) | Homework |
| 10  **Microarray Data** | Microarray Data and Preprocessing  Gene Filtering and the Application of Linear Models  Annotation and the GO Numbers | *Applied Statistics for Bioinformatics using R*:  Chapter 6 (pages 91-110) | Homework |
| 11  **Clustering Methods** | Clustering Analysis and Distance Measures  Hierarchical Clustering  K-means and K-Medoids Clustering | *Statistics Using R with Biological Examples:* Chapter 16 (pages 310-318)  *Applied Statistics for Bioinformatics using R*: Chapter 7: Sections 7.1-7.2 (pages 117-130) | Homework |
| 12  **Principal Components Analysis** | Multivariate Distribution and Multivariate Data Display  Principle Components Analysis | *Statistics Using R with Biological Examples:* Chapter 16 (pages 288-301)  *Applied Statistics for Bioinformatics using R*: Chapter 7: Section 7.4 (pages 133-141) | Homework |
| 13  **Classification Methods** | Classification and Logistic Regression  Evaluation of Classification Methods: ROC Curves and Validation  Support Vector Machines (SVM)  Classification Tree and Comparison of Classifiers | *Statistics Using R with Biological Examples:* Chapter 16 (pages 301-310)  *Applied Statistics for Bioinformatics using R*: Chapter 8 (pages 145-172) | Homework |
| 14  **Final Review** | Review of Topics  Use of Statistical Methods for Microarray Data |  | Final Exam: Begin |
| 15  **Final Exam** |  |  | Final Exam: Due |

**Special Accommodations**: If you have specific physical, psychiatric or learning disabilities that may require accommodations for this course, please contact Northeastern's Disabilities Resource Center (DRC) at (617) 373-2675. The DRC can provide you with information and assistance to help manage any challenges that could affect your performance in the course. The University requires that you provide documentation of your disabilities to the DRC so that they may identify what accommodations are required, arrange with the instructor to provide those on your behalf, as needed.

**Honor Code**: All students must adhere to the Northeastern University honor code available on the Northeastern web site   
(see http://www.northeastern.edu/osccr/academicintegrity/index.html) and the graduate student handbook.

**Northeastern University Copyright Statement**

This course material is copyrighted and all rights are reserved by Northeastern University. No part of this course material may be reproduced, transmitted, transcribed, stored in a retrieval system, or translated into any language or computer language, in any form or by any means, electronic, mechanical, magnetic, optical, chemical, manual, or otherwise, without the express prior written permission of the University.