BMS625 Foundations in Biostatistics and Computational Biology
Fall 2016, 2 credits
Dates 10/20/16 - 12/8/16
Course Coordinators:
Greg Carter, Ph.D. (Greg.Carter@jax.org)
Christine Duarte, Ph.D. (duartc@mmc.org)

Meeting time: Tuesdays and Thursdays 3:30-5 pm 10/20/16 - 12/8/16
Location: videoconference
Videoconference: University of Maine - 147 Hitchner, Maine Medical Center Research Institute conference room 5, UNE- No room, students participate via Jabber/Movi, MDIBL - Training Lab Conference Room (209.222.195.100), Jackson Laboratories - B1 Unit 4 Room 2160
Videoconference Details: Course BMS625; IP on UMS network 86658666486; *IP off UMS network* 169.244.81.115 then 8666486* ; SIP address 8666486@networkmaine.net; Telephone 2078666486

Pre-requisite: none

Description: Introduction to biostatistics with application to the biomedical sciences and genetics, and introduction to computational biology.

Format: The course will be a combination of lectures and data analysis exercises. Data analysis will be conducted in Excel and R or web-based software programs.

Grading: The following will be evaluated for your final course grade: discussion (20%) and grading of two final data analysis reports (40% each). Note: we understand the video-conferencing format has its limitations, especially for those of you at remote sites. We will elicit responses as much as we can, but the responsibility for participation is yours.

Faculty: This course will be taught by Christine Duarte at MMCRI and Greg Carter at Jackson Laboratory and broadcast through the videoconferencing facilities to participating University of Maine sites.

Readings: The primary textbook will be ***Using R for Introductory Statistics, Second Edition*** **by John Verzani**. Power point slides will be provided for this course. Students will be required to have access to Microsoft Excel and to the R programming language (available for free download at http://www.r-project.org/) for the completion of assigned homework. It is also recommended that they download R Studio https://www.rstudio.com/products/rstudio/download3/ which is a friendly environment for writing and running R code.

Student Learning Outcomes

Course Goals:
The goal of the first part of the course is for students to understand and apply the principles of statistics to appropriately design research studies, analyze the collected data, draw appropriate conclusions, and test assumptions. Students will learn basic modeling techniques such as linear regression, analysis of variance (ANOVA), and basic categorical data analysis. Students will be introduced to a sampling of advanced techniques and topics such as multivariate data analysis, missing data, multiple testing, survival analysis, and nonparametric methods. The overall goal is for students to have a broad understanding of good research design as well as understand some of the basic statistical methods available for the analysis of different types of data sets.

Instructional Objectives:
- Students will learn how to import, export, and manipulate data in R and Excel.
- Students will be able to perform standard statistical analysis methods on a sample data set (including linear regression, analysis of variance (ANOVA), and basic categorical data analysis).
- Students will be able to evaluate the results of statistical methods and draw appropriate conclusions after testing assumptions.
- Students will know how to choose which statistical method to apply to which type of research design and variable types.
- Students will learn how to formulate and evaluate hypothesis tests and how to interpret confidence intervals for

estimated parameters.

- Students will have a basic understanding of advanced topics such as missing data techniques, multiple testing, survival analysis, and nonparametric methods.
- Students will learn methods for complex trait analysis, multivariate analysis, and the analysis of large data sets, providing a basis for understanding current literature.
- Students will learn to present in written form the design and analysis of a research study with appropriate use of summary and descriptive statistics, tables and figures with appropriate legends, and a discussion of statistical methods used, assumptions tested, and a concise interpretation of results.

Student Learning Outcomes:

- Students will be able to design and analyze the results of a research study by formulating a hypothesis, selecting a design that tests that hypothesis, selecting and measuring all necessary explanatory and outcome variables, selecting a sample size that gives sufficient power to reject the null hypothesis, selecting a statistical method which is appropriate for the study design and variable types, analyzing the data and drawing appropriate conclusions, testing assumptions, and presenting tables, figures, and graphs that efficiently convey the results.

Course Syllabus, Fall 2016

Date	Topic	Recommended Reading	Exercises	Classes
Thursday 10/20/2016	Introduction to Statistics: sampling theory, distributions, summary statistics, types of variables, hypothesis testing, confidence intervals	Chapters 2, 6, 8	Install R & R Studio, UsingR package from CRAN	1
Tuesday 10/25/2016	Introduction to the R language, using Excel and R for basic statistical analysis, summary of graphing capabilities	Chapters 1, 4, 5	2.8 (p.46), 2.16 (p. 47), 2.31 (p. 81)	2
Thursday 10/27/2016	Regression and Correlation	Chapter 11	2.68 (p. 87), 11.2 (p. 368)	3
Tuesday 11/1/2016	**Review of Regression and R**			4
Thursday 11/3/2016	T-tests and ANOVA	Chapters 9, 12	9.14 (p. 310)	5
Tuesday 11/8/2016	Categorical data analysis, Multivariate Analysis	Chapters 3, 8, 10	9.21 (p. 319)	6
Thursday 11/10/2016	Study design and power analysis; Advanced topics: survival analysis, nonparametric methods, mixed models, missing data, multiple testing	Chapter 9, Chapters 7, 13	9.16 (p. 310)	7
Tuesday 11/15/2016	Introduction to genome-scale data: types of experiments, current technologies, experimental output.		**Assignment #1 due**	8
Thursday 11/17/2016	Analysis of genome-scale data: quality assessment and correction methods.			9
Tuesday 11/22/2016	Analysis of genome-scale data: dimensional reduction by data clustering.			10
~~Thursday 11/24/2016~~	~~Thanksgiving Break~~			
Tuesday 11/29/2016	Analysis of genome-scale data: functional assessment by integrating annotation data.			11

Thursday 12/1/2016	Basics of genetic association mapping and quantitative trait loci: data types and regression models.			12
Tuesday 12/6/2016	Advanced topics in genetic association studies: linkage disequilibrium, population stratification, and model systems.			13
Thursday 12/8/2016	Advanced topics in genetic association studies: epistasis and pleiotropy.			14

Assignment # 1: Using the sample data set given and the reference, "Comprehensive molecular characterization of clear cell renal cell carcinoma" by the TCGA research network (2013), perform analyses to support or refute whether FOXD1 mRNA expression is a potential biomarker for RCC disease severity. Here are some analyses to perform:

1. Construct a linear regression model for survival time as a function of FOXD1 expression. Hint: log or square root transformations may be needed. Summarize the results of the model and test assumptions. ~~Are SPRY1and SPRY4 significantly correlated with FOXD1 expression? Which statistical test should be used?~~
2. Does FOXD1 expression differ significantly between those that have survived or not? ~~Does it differ by stage or grade? Describe which tests or models you used for each question, and summarize results and state assumptions. You can collapse categories if you like, and you can choose which stage variable you prefer.~~
3. Stratify FOXD1 at the median, and report whether the proportion of high versus low FOXD1 individuals differs significantly by stage (you can pick which one), ~~grade, or survival (yes/no).~~
4. You would like to validate whether or not FOXD1 (log-transformed) can be used as a biomarker for distinguishing early (stage I or II) versus late (stage III or IV) stage RCC. You have a collaborator with a sample of 100 early stage and 100 late stage RCC patients that can measure FOXD1 for you. Is this sample size sufficient to validate the effect with 80% probability, assuming the same effect size and variance found in this data set? If it is not enough, what sample size do you need?
5. Bonus: perform a survival analysis of this data using FOXD1 stratified at the median. Are the Kaplan-Meir curves for high versus low expression significantly different? If you additionally adjust for stage and grade, is FOXD1 significant?