**An Introduction to Practical Data Analysis in Medicine and Public Health**
**Syllabus**

## Course Information

| | |
|---|---|
| Term: | Spring 2015 |
| Course Time: | Tuesdays and Thursdays, 10:30 – 11:45 AM |
| Course Location: | Homewood Campus, Krieger 309 |
| Office Hours: | TBA |


| | |
|---|---|
| Instructor: | Therri Usher |
| Contact Information: | tusher1@jhu.edu |
| Office: | 615 North Wolfe Street, E3148 (JHSPH) |


| | |
|---|---|
| Course Credit: | 3 |
| Prerequisites: | Completion of the undergraduate statistics course for Homewood public health majors and familiarity with the statistical software R |

## Course Description

The course is designed to introduce undergraduate public health majors to the methodology of data analysis, such as how to apply previously learned statistical methods in the performance of data analysis in medical and public health research. This course is unique in that it focuses on all parts of the data analysis process, from formulating a research question to synthesizing the results and creating a data product. While the emphasis is placed on developing and implementing various methods of data analysis, the course will also address interpreting and evaluating the strengths and limitations of existing data analyses. Students' understanding will be solidified through in-class active learning activities that explore the process of data analysis and evaluations of analyses in the scientific literature, culminating in a final data analysis project relevant to their own areas of expertise in order to incorporate knowledge gained from the course into their own research.

## Course Learning Objectives

Upon successfully completing this course, students will be able to:

1. Understand the purpose, process, and methodology of quantitative data analysis with regards to the design, interpretation, and evaluation of medical and public health research;
2. Create a sequence of data analysis steps based on a developed research question and available data;
3. Implement the planned sequence in order to perform data analysis and adjust the plan accordingly;
4. Analyze and evaluate methods of data analysis performed in various projects by fellow medical and public health researchers as well as his or her own data analysis;
5. Internalize the complexity and inherent ambiguity of data analysis into his or her understanding of medical and public health research;
6. Incorporate knowledge of data analysis into his or her research agendas in order to identify innovative and possibly effective public health interventions.

## Course Format

The course is structured with the belief that students learn best when they are actively engaged in the acts of doing and creating. As a result, the course uses various formats to stimulate active learning:

- Interactive Lectures: Some of the course content is communicated through presentations in the form of lectures. However, the lectures will incorporate informal active learning activities, such as the "think-pair-share" technique, that will require students to discuss, analyze, and evaluate relevant topics, individually or in groups.
- Discussions: Class-wide discussions will be utilized to deepen students' understanding of the material as well as to provide feedback about the learning process to the teacher. Since the class contains students interested in different areas of medicine and public health, it is highly recommended that everyone bring their interest and excitement for their respective areas in order to enrich and energize the discussions.
- Debates: Like many fields, data science has ethical questions that need to be addressed, particularly with regards to reproducible research and data collection. Students will be separated into two groups with each group advocating either for or against an ethical issue relevant to data analysis.
- Paper Evaluations: In order to prepare for the graded evaluation of an academic paper, certain papers will be evaluated during class. It is the student's responsibility to review the paper beforehand and to come to class ready to participate in the evaluation.
- Data Analyses: In order to prepare for the final data analysis, students will practice data analysis skills by performing aspects of data analysis in class. As a show of good faith, the instructor will also participate in a data analysis and will update students on its progress during class time. This gives students an opportunity to observe and evaluate data analysis as it is in progress.

Each class will have at least one active learning activity in order to actively engage students and promote higher-level cognitive development.

## Course Atmosphere

While it is the responsibility of the instructor to provide an interactive, student-oriented learning environment and to guide students in the learning process, it is the responsibility of students to work outside of class hours and to actively participate in the discussion and active learning activities within class hours and to inform the instructor when he/she is having difficulty with the material. Participation and collaboration are key elements of the course. While much of the learning is done outside of the classroom, participation inside the classroom is imperative for developing a high-level knowledge of data analysis. Attendance is a vital part of success in the course and will be factored into the participation grade. With both parties fulfilling their responsibilities, the course will have an open, interactive, and encouraging atmosphere in which students can begin to craft their data analysis skills.

## Intended Audience

Upper-level undergraduate students that have completed an undergraduate biostatistics course, have experience with R, and are interested in the role and application of quantitative data analysis in medical and public health research. Maximum enrollment is 19 students.

## Textbooks

There is no one book that fully explains data analysis. As a result, there is no required textbook for this class. Students are, however, welcome to refer to data analysis books and discuss in class the material they gain from the books. The following books may be beneficial in applying statistical methods to data analysis:

1. Wickham, Hadley. ggplot2: elegant graphics for data analysis. Springer, 2009.
2. Rosner, Bernard. Fundamentals of Biostatistics, 7th edition. Cengage Learning, 2010.

The textbooks used in the undergraduate biostatistics course will be useful as well.

**Resources**

The following websites may be helpful in learning more about data analysis as well as obtaining material for the course:

1. Open Intro: http://www.openintro.org/
2. Introduction to statistical learning with R: http://www-bcf.usc.edu/~gareth/ISL/
3. SwirlStats: http://swirlstats.com/
4. DataCamp: https://www.datacamp.com/
5. Coursera: http://www.coursera.com/
6. JHU Blackboard: https://blackboard.jhu.edu/

NOTE: This list of resources is by no means exhaustive. If you find a resource that is not listed in the syllabus that you believe to be beneficial, please feel free to share with the instructor and other students.

**Graded Assignments**

R Review - 10%
Research Question and Processing Data Quiz - 10%
Exploratory Data and Statistical Analyses Quiz - 10%
Class Participation - 10%
Data Analysis Project - 35%

- Progress Report - 5%
- Paper - 15%
- Presentation - 15%

Academic Paper Evaluation: Letter to the Editor - 25%

*R Review*

An in-class review will be included in the beginning of the course to refresh students on the syntax and capabilities of R. The students are required to submit an assigned review that they will complete on their own in order to become re-familiar with R and to start becoming familiar with RStudio if they have not used it before. The students are allowed to use resources in the completion of the review.

*Quizzes*

There will be two quizzes during the course. Each quiz should be completed using Blackboard. Students can use their notes, textbooks, and resources. Students will have until the Friday of the week after the material is fully covered in class, or as indicated by the instructor, to complete the quiz. Students can access each quiz one (1) time and students can take the quiz one (1) time.

*Class Participation*

Students are expected and encouraged to participate in class discussion and active learning activities. Participation in both will help students develop their own deeper knowledge of quantitative data analysis.

*Data Analysis Project*
Students will complete their own data analysis based on a question they seek to answer. Students should complete all steps up to devising ways to synthesize their findings into a useful product. (Implementation of the devised method is optional for this project.) They will then write their project and findings in the form of a scientific paper.

Near the midpoint of the semester, each student must submit a progress report on their project. The report should include a summary of what they have done, using elements of reproducible research, as well as a breakdown of what needs to be done to complete the project, including any questions and/or potential obstacles they might have.

The final paper should contain introduction, methods, analysis, results, discussion, and conclusion sections. The paper should touch on the themes of the course, such as reproducibility, data cleaning and processing, exploratory data analysis, model evaluation, and interpretation of results. It should also address the potential medical or public health implications.

Each student will present his/her data analysis in an 8-minute presentation to the class while students complete a written critique of the presenter's analysis. The presentation should address the chosen process of data analysis but can also include an application of the findings in action, such as a statistical package. A short question and answer session will follow each presentation. Students are welcome to use one of the methods discussed in synthesizing research to present their project. A rubric for the paper and presentation will be provided.

*Academic Paper Evaluation*
Students will be given an academic paper to evaluate. Students will then have to write their evaluation in the form of a letter addressed to an editor of a scientific journal pertaining to a paper they published. The evaluation is expected to address the themes of the course and to summarize the quality of the data analysis. A rubric for the evaluation will be provided.

**Attendance and Late Work**
Attendance is factored in to the participation score for the course. For a 13-week course, there are two free absences allowed without affecting one's participation grade for the course. Any absence afterwards that is not attributable to illness or academic events such as conferences or lectures can lead to a decrease of the participation score. Full credit for participation in a class can be acquired by attending the class and actively participating, whether by participating in the discussions or in the active learning activity for that day.

If a student is absent due to illness or other academic engagements, it is strongly recommended that he or she informs the instructor at least 24 hours ahead of time and completes the active learning activity for the day by submitting it to the instructor within a week of the missed class. Failure to do so may negatively affect the student's participation score.

Assigned work, such as the quizzes, paper evaluation, and final project, will be due by 11:59 PM Eastern time of the given due date. These assignments will only be accepted late if there were extraordinary circumstances occurring, such as a family or medical emergency. If a student feels that this might happen, it is best to inform the instructor as soon as possible so that the two of them may discuss a plan of action.

**Electronics**

As this is a data analysis course, students are allowed and even encouraged to bring their laptops to class so that they can perform aspects of data analysis in class with the assistance of the instructor. However, laptop use should be restricted to course-related activities. Additionally, laptops, phones, and other electronic devices should be silenced so as not to disturb the instructor or other students.

**Course Schedule**

| Week | Topic Overview | Learning Activity | Due Dates |
|---|---|---|---|
| 1 (1/27, 1/29) | Course Overview | | |
| | Statistical Programming Using R & RStudio | Review R software | R Review due at 11:59 PM |
| 2 (2/3, 2/5) | Reproducible Research | Debate the usefulness of reproducible research | |
| | Forming a Research Question | Create & evaluate research questions | |
| 3 (2/10, 2/12) | Collecting and Processing Data | Extract data from a chosen website | |
| 4 (2/17, 2/19) | | Debate the ethics of data scraping | |
| 5 (2/24, 2/26) | Exploratory Data Analysis | Perform EDA using R's ggplot2 package | Quiz 1 due Friday 2/27 at 11:59 PM |
| 6 (3/3, 3/5) | | | |
| 7 (3/10, 3/12) | Review / In-Class Paper Evaluation | Evaluate a designated academic paper | Progress Report due Friday 3/13 at 11:59 PM |
| 8 (3/24, 3/26) | Statistical Analysis | Fit multiple statistical models on an existing data set in R | |
| 9 (3/31, 4/2) | | | |
| 10 (4/7, 4/9) | Model Checking and Evaluation | Perform sensitivity analysis on previously fit models | Quiz 2 due Friday 4/10 at 11:59 PM |
| 11 (4/14, 4/16) | Interpreting and Challenging Results | Assess analysis & results in Class 9's academic paper | |

| | | | |
|---|---|---|---|
| 12<br>(4/21, 4/23) | Synthesizing Results & Creating a Data Product | Devise a potential data product | |
| 13<br>(4/28, 4/30) | Student Presentations / Course Wrap-Up | Finish up student presentations | Paper evaluation and data analysis due Friday 5/8 at 11:59 PM |

**Course Plan**

**Week 1: Course Overview**
Overview of the Course
Logistics of the Course
What is Data Science?
What is Data Analysis?

*Resources*
Top 10 Tips for Data Analysis to Make Your Research Life Easier!: www.statsmakemecry.com/smmctheblog/top-ten-tips-for-data-analysis-to-make-your-research-life-ea.html
What is the Best Way to Analyze Data?: simplystatistics.org/2013/06/27/what-is-the-best-way-to-analyze-data/

After this week, students will be able to:
- Discuss the overview and logistics of the course
- Describe the flow of data analysis
- Understand the purposes of data analysis as compared to statistical analysis, especially with regards to research

**Week 1 (continued): Statistical Programming Using R and RStudio**
Why Use R?
Review of R
RStudio

*Resources*
R's website: cran.r-project.org
R's Wikibook: en.wikibooks.org/wiki/R_Programming
RStudio's website: www.rstudio.com
UCLA's "Resources to help you learn and use R": www.ats.ucla.edu/stat/r/
Quick-R: www.statmethods.net

After this week, students will be able to:
- Perform calculations and create basic functions in R
- Operate RStudio to perform tasks using R

**Week 2: Reproducible Research**

Replication
Reproducible Research
Methods of Reproducible Research

*Resources*
GitHub: https://github.com
Karl Broman's reproducible research tutorials: https://github.com/kbroman

After this week, students will be able to:
- Explain ways to create reproducible research
- Compare reproducible research to replication
- Apply the tenets of reproducible research to his/her research endeavors

**Week 2 (continued): Forming a Research Question**
Interesting Question vs. Research Question
Characteristics of a Research Question

*Resources*
Wikipedia's research question page: en.wikipedia.org/wiki/Research_question

After this week, students will be able to:
- Understand the process and importance of creating a research question
- Assemble a research question that can be investigated through quantitative data analysis
- Evaluate existing research questions and alter them if needed so that may be appropriately investigated

**Weeks 3 and 4: Collecting and Processing Data**
Primary and Secondary Data
Raw vs. Processed Data
Avenues of Obtaining Data
Data Scraping
Characteristics of Research Data
Tidy Data

*Resources*
XML package manual: cran.r-project.org/web/packages/XML/XML.pdf
Tidy Data: vita.had.co.nz/papers/tidy-data.pdf

After these weeks, students will be able to:
- Compare primary versus secondary data and raw versus processed data
- Understand the avenues through which data can be obtained
- Obtain data responsibly through "data scraping" internet sources
- Formulate an individual stance on the ethics of data scraping
- Process obtained data so as it can be used for data analysis

**Weeks 5 and 6: Exploratory Data Analysis**

Plotting
SVD/PCA
Clustering
Other Methods of EDA
Interpreting Graphs
Evaluating Graphs

*Resources*
Tukey, John W. "Exploratory data analysis." (1977).
Wickham, Hadley. ggplot2: elegant graphics for data analysis. Springer, 2009.

After these weeks, students will be able to:
- Understand the uses of exploratory data analysis
- Select appropriate tools of exploratory data analysis to use on a given data set
- Apply exploratory data analysis techniques to existing data
- Construct a preliminary plan of statistical analysis based on findings from the exploratory data analysis
- Interpret and evaluate existing output generated through exploratory data analysis

**Week 7: Review / In-Class Paper Evaluation**
Evaluate an Academic Paper
Address Questions About Material and Projects

*Assigned Paper:*
Li, Donghui, et al. "Body mass index and risk, age of onset, and survival in patients with pancreatic cancer." Jama 301.24 (2009): 2553-2562.

After this week, students will be able to:
- Evaluate the data analysis performed in an existing research project, particularly its steps up to and including exploratory data analysis

**Weeks 8 and 9: Statistical Analysis**
Review Statistical Methods
Parametric vs. Nonparametric
Regression
Bootstrapping
Tools of Inference
Simulation
Prediction

*Resources*
Kuhn, Max, and Kjell Johnson. Applied predictive modeling. New York: Springer, 2013.
www.appliedpredictivemodeling.com
Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

The Elements of Statistical Learning:
www.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

After these weeks, students will be able to:
- Differentiate between inference, simulation, and prediction
- Decide the proper goal of the statistical analysis based on the research question and the exploratory data analysis
- Compare the advantages and disadvantages of various statistical methods
- Construct a detailed plan of statistical analysis
- Implement the plan on existing data
- Evaluate methods of statistical analyses used in existing research

**Week 10: Model Checking and Evaluation**
Assess Model Fit
Sensitivity Analysis
Evaluate Model Fit
Limitations of Model Fit
Student Presentations

*Resources*
See suggested textbooks for the course and resources from Weeks 8 and 9.

After this week, students will be able to:
- Describe the process of assessing model checking and evaluation, including sensitivity analysis
- Evaluate the statistical analysis's fit on the data and its effectiveness in deriving the information desired
- Acknowledge and discuss the limitations of the data analysis, particularly its statistical methods
- Evaluate an existing project's statistical analysis

**Week 11: Interpreting and Challenging Results**
Interpreting Model Output
Improving Statistical Analysis
Student Presentations

*Resources*
See suggested textbooks for the course and resources from Weeks 8 and 9.

After this week, students will be able to:
- Interpret the output from statistical analyses into scientific-based results
- Assess results of existing projects and provide feedback on the data analysis process for improvement

**Week 12: Synthesizing Results and Creating a Data Product**
Statement of Results
Presentation of Results
Data Products

LaTeX
Markdown

*Resources*
googleVis package: cran.r-project.org/web/packages/googleVis/vignettes/googleVis.pdf
shiny: www.rstudio.com/shiny/
slidify: slidify.org

After this week, students will be able to:
- Create a statement of results in terms of the scenario of the data analysis
- Understand different avenues of presenting results
- Compare and contrast the presentation of results in different forms
- Discuss different ways to create data products based on findings

**Week 13: Student Presentations / Course Wrap Up**
Benefits of Data Analysis
Resources for Improving at Data Analysis
Student Presentations

After this week, students will be able to:
- Analyze whether data analysis can be beneficial for public health researchers
- List available resources that are useful in learning more about data analysis

**Academic Ethics Statement**

The strength of the University depends on academic and personal integrity. In this course, you must be honest and truthful. Ethical violations include cheating on exams, plagiarism, reuse of assignments, improper use of the Internet and electronic devices, unauthorized collaboration, alteration of graded assignments, forgery and falsification, lying, facilitating academic dishonesty, and unfair competition.

Report any violations you witness to the instructor. You may consult the Associate Dean of Student Affairs and/or the Chairperson of the Ethics Board beforehand. See the guide on "Academic Ethics for Undergraduates (http://e-catalog.jhu.edu/undergrad-students/student-life-policies/) for complete information.

**Statements on Accommodations**

If you are a student with a disability or believe you might have a disability that requires accommodations, please contact Dr. Brent Mosser, in Student Disability Services, 385 Garland, (410) 516-4720, studentdisabilityservices@jhu.edu. Students that require an accommodation must obtain an accommodation letter from Student Disability Services. The office website is http://web.jhu.edu/disabilities/index.html

If you believe you need other accommodations for assignments or examinations, please contact the course instructor ahead of time to discuss the matter privately.

Student athletes are responsible for submitting their semester schedule in writing during the first week of class. The only excused absences for athletic related purposes will be for competition related events.

Students who heed the advice of health professionals to stay home due to illness and thus miss class will be accommodated. Students who must miss a class or an examination because of a religious holiday must inform the instructor as early in the semester as possible in order to make up any work that is missed.

## Statement of Diversity and Inclusion

Johns Hopkins University is a community committed to sharing values of diversity and inclusion in order to achieve and sustain excellence. We believe excellence is best promoted by being a diverse group of students, faculty and staff who are committed to creating a climate of mutual respect that is supportive of one another's success. Through its curricula and clinical experiences, the University purposefully supports this goal of diversity, and in particular, works towards an outcome of best serving the needs of students. Faculty and candidates are expected to demonstrate an understanding of diversity as it relates to planning, instruction, management, and assessment.

## Cancelation of Class

Due to weather or unforeseen circumstances, lecture and/or discussion session may be canceled. If JHU has canceled classes on the Homewood campus during the course meeting time, then class will be canceled. Please refer to JHU resources for emergency notifications at http://esgwebproxy.johnshopkins.edu/notice/ and 410-516-7781 and 1-800-548-9004. If class is canceled by the instructor for any reason, we will notify you by email as soon as possible.

## In Case of Emergency

- Emergency information and weather alerts,http:/esgwebproxy.johnshopkins.edu/notice/, 410-516-7781, 1-800-548-9004
- In case of medical emergency, dial 911
- In case of fire, pull alarm, then dial 911
- Homewood Campus Safety & Security, Emergency 410-516-7777, Information and Walking
- Escorts 410-516-4600, Email Safety.and.Security@jhu.edu
- Hopkins Emergency Alert (JHEA) system sends a text message to the cell phones of those who have subscribed to the service. If you are not yet a JHEA subscriber, you can sign up on the https://my.johnshopkins.edu/ portal. After signing in with your JHED ID, go to the "My JHED" tab then update your emergency alert information. Remember to click "Save" when you are done.

This syllabus and lecture schedule is subject to change and revision at the instructor's discretion.