NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

MSDS 401-DL Applied Statistics with R

Contact Information

Instructor: **Syamala Srinivasan, Ph.D.**

Teaching Assistant: Todd Peterson MS

Email: Syamala.srinivasan@northwestern.edu; todd.peterson@northwestern.edu

Office Hours: By appointment

Response Time: I will respond to emails within 24-48 hours.

Course Description

This course teaches fundamentals of statistical analysis. This includes evaluating statistical information, performing data analyses, and interpreting and communicating analytical results. Students will learn to use the R language for statistical analysis, data visualization, and report generation. Topics covered include descriptive statistics, central tendency, exploratory data analysis, probability theory, discrete and continuous distributions, statistical inference, correlation, multiple linear regression, contingency tables, and chi-square tests. Selected contemporary statistical concepts, such as bootstrapping, are introduced to supplement traditional statistical methods. Recommended prior course: MSDS 400-DL Math for Data Scientists.

Course Objectives

At the conclusion of the course, students will be able to:

- Perform statistical analyses.
- Interpret and evaluate statistical information.
- Prepare technical reports.
- Use the language R for data analysis.

Prerequisites

MSDS 400 is a prerequisite for MSDS students who began studies fall term 2014 and thereafter.

Required Resources

Required Readings

- Black, K. (2013). *Business Statistics.* 9th Hoboken, NJ: John Wiley & Sons, Inc. [ISBN-13: 978-1119334965]
- Kabacoff, R. I. (2015) *R in Action, 2nd* Shelter Island, NY: Manning Co. [ISBN-13: 978-1617291388]
- Course Reserves from Wilcox and others – Details are in Canvas

The biweekly tests and final exam will be based on the material in the texts by Black and Wilcox. The text by Kabacoff deals with R programming and statistical applications. It is intended as a reference aid, for the two data analysis assignments, the programming in R tests plus the *Lessons in R.*

Purchasing Options

Business Statistics 9th ed. may be purchased on the MSDS 401 Course Site along with an access code for WileyPLUS. Wiley offers different purchasing options including an e-book and loose-leaf versions of the text. All required materials are also available from the Abbott Hall Bookstore. Be aware, the ISBN number shown above for *Business Statistics* is used solely by the Abbott Hall Bookstore. A third party vendor will use a different ISBN number. If you purchase from a third party vendor, be sure to purchase an access code bundled with *Business Statistics 9th ed.* If you purchase an access code separately from a third party vendor, there is a risk it will not work with WileyPLUS on our course site.

- *Basic Statistics* is available in hardcopy and electronic versions from various vendors.
- *R in Action, 2nd ed.* can be obtained at https://www.manning.com/books/r-in-action-second-edition.
- http://www.greenteapress.com/thinkbayes/thinkbayes.pdf

MSDS 401 will use the high-level language R (http://cran.r-project.org/). Instructions are given on the course site. RStudio is not required but is highly recommended for new users of R. Everything students need to do in this course can be accomplished using the standard R console with a plain text editor.

RStudio http://www.rstudio.com/provides an integrated environment, and installer packages are easily available online. The installation process is straightforward. Tutorials dealing with R are available on the course site. Students are encouraged to review the *R Tutorial Materials* module on the course site, and to start studying R early in the quarter. There are tests during the quarter dealing with R programming. Check the syllabus and course site for details on date.

Assignment Overview and Grading Breakdown

The students' final grade will be based on the following:

Discussion Board	100 possible points	20%
4 self-administered Tests	100 possible points	20%
2 R Programming Assignments	125 (50+75) possible points	25%
2 Data Analysis Assignments	125(50+75) possible points	25%
Final Exam (proctored)	50 possible points	10%

There will be two programming assignments. For the second assignment students will be given a choice between two different versions of this assignment. Both versions carry the same point total.

Assignment Preparations:

- The self-administered tests are due at the end of the weeks of 2, 4, 6, and 8.
- The 4 assignments (2 on R, 2 on Data Analysis) are due at the end of the weeks of 3, 5, 7, and 9.
- Make sure that you start working on the 4 assignments 2 weeks prior to the due dates.

Sync Sessions

There will be five scheduled sync sessions. Dates and times are listed in the Course Schedule and announced. Please note that attendance at any scheduled synchronous or "live" meetings is considered optional. Your attendance is highly encouraged. Recordings are made available along with handouts.

Final Exam Proctoring

Examity, an independent organization provides proctoring services for the final exam. Students are responsible for scheduling their exam with Examity and paying all fees. The exam is taken within Canvas using the course site. The proctor must be able to monitor the student using a webcam and also view the screen images seen by the student. **Only one computer screen is allowed for the final exam.**

Final exam consists of conceptual questions only and hence no need for any computations. Final exam is closed book, closed notes. **Access to the internet (outside of the approved resources) during the exam is not permitted. Separate portable devices such as iPads and Kindles are not permitted.** Since only one computer screen may be used for the exam, any student who uses a device such as an iPad or Kindle or multiple screens must plan ahead and be prepared to migrate the necessary files to the computer used for the final exam.

Students with disabilities working through AccessibleNU must discuss reasonable accommodations, including use of non-approved technology, with the instructor and the proctors well before their exam. Please see the Canvas course site for more information.

Grading Scale

A	**93-100%**	**465-500 points**
A-	**90-92%**	**450-464 points**
B+	**87-89%**	**435-449 points**
B	**83-86%**	**415-434 points**
B-	**80-82%**	**400-414 points**
C+	**77-79%**	**385-399 points**
C	**73-76%**	**365-384 points**
C-	**70-72%**	**350-364 points**
F	**0-69%**	**000-349 points**

Late Work Policy

When stating due dates and times for work, the abbreviation "CST" is used. "CST" means Chicago, IL clock time. This defines "course time". Canvas will adjust what students see as deadlines according to the time zone specified by the student in personal settings for Canvas. (Thus an 12 pm CST deadline is a 10 pm deadline on the West Coast, and so forth depending on time zone.) Deadlines for all work are stated in this syllabus and posted on the Course Site. This includes exams, reports and participation in the discussions. Assignments are to be submitted prior to the deadlines. **Without prior arrangement, any late assignment will receive a 1% point deduction for each hour late, totaling to a maximum 50% deduction.** For example, ten hours late will result in a 10% point deduction from the total possible assignment points. Prior communication with the instructor is essential.

Do not fall behind. We cover a great deal of material in this course, and falling behind is the primary reason why students have difficulty, particularly toward the end of the course. To that end, the syllabus and course site give due dates for the entire course.

Online Communication and Interaction Expectations

Discussion Forums

The purpose of the discussion boards is to allow students to freely exchange ideas. It is imperative to remain respectful of all viewpoints and positions and, when necessary, agree to respectfully disagree. While active and frequent participation is encouraged, cluttering a discussion board with inappropriate, irrelevant, or insignificant material will not earn additional points and may result in receiving less than full credit. Frequency matters, but contributing content that adds value is

paramount. Please remember to cite all sources—when relevant—in order to avoid plagiarism. Please post your viewpoints first and then discuss others' viewpoints.

The quality of your posts and how others view and respond to them are the most valued. A single statement mostly implying "I agree" or "I do not agree" is not considered to be a post. Explain, clarify, politely ask for details, provide details, persuade, and enrich communications for a great discussion experience. Please note, there is a requirement to respond to at least one post from a class member. I'm looking for insightful analysis, probing questions, and *constructive* suggestions to each other. Keep thinking from the perspective—how can I *add something useful?* It may be an experience you have had professionally or a quote from an article/website you come across. If it is the latter, cite it properly.

You are expected to participate actively with other students in one discussion each week. This discussion will be focused on a topic is related to the course assignments. You are also expected to post a comment to a second review and reflections question. This latter question provides for comments about the course, and how your studies are progressing. You are expected to participate in both forums with polished, well-structured and APA- compliant posts each week, adding references as needed. Be sure to check spelling and grammar.

It is highly desirable that your initial comments be posted Thursday so that follow-up comments can be made. The discussion forum is intended for exchange of ideas between students. **Discussion Board responses are DUE by 11:59 pm on every Sunday. Five points are available for the first discussion topic, two points for class participation and three points for the reflections topic making ten points total per week. Only one grade is entered each week for the Discussion Board.**

Participation and Attendance

This course will not meet at a particular time each week. All course goals, session learning objectives, and assessments are supported through classroom elements that can be accessed at any time. To measure class participation (or attendance), your participation in threaded discussion boards is required, graded, and paramount to your success in this course. Please note that any scheduled synchronous meetings are optional. While your attendance is highly encouraged, it is not required and you will not be graded on your attendance or participation.

Academic Integrity at Northwestern

Students are required to comply with University regulations regarding academic integrity. If you are in doubt about what constitutes academic dishonesty, speak with your instructor or graduate coordinator before the assignment is due and/or examine the University Web site. Academic dishonesty includes, but is not limited to, cheating on an exam, obtaining an unfair advantage, and plagiarism (e.g., using material from readings without citing or copying another student's paper). Failure to maintain academic integrity will result in a grade sanction, possibly as severe as failing and being required to retake the course, and could lead to a suspension or expulsion from the program. Further penalties may apply. For more information, visit The Office of the Provost's Academic Integrity page.

Course Schedule

Week 1 – Complete by Sunday, Jun 28, 2020

Learning Objectives

After this week, the student will be able to:

- List examples of statistical applications in business.
- Explain the difference between variables, measurement and data.
- Define and compare four different levels of data.
- Construct a frequency distribution and different data displays.
- Construct and interpret two-variable tables and scatter plots.
- Write simple programs using the language R.

Reading

- Black, K. *Business Statistics* Chapter 1 Sections 1.1-1.2 & Chapter 2 Sections 2.1-2.4.

R Installation:

- Installation of R and completion of *The Quick Start Guide to R* is expected this week.

Videos:

- Levels of Data Measurement
- Stem-and-Leaf Plot
- R Lesson Video

Assignments

- Install R
- Study *The Quick Start Guide to R* and complete the exercises.

Sync Session

There will be a sync session the first week of class Monday, Jun 22, 2020 from 7 PM to 9 PM CST. Attendance is optional. A recording of the session will be posted in class the next day.

Week 2 – Complete by Sunday, Jul 5, 2020

Learning Objectives

After this week, the student will be able to:

- Calculate and apply measures of central tendency and variability.
- Describe a data distribution using a box-and-whisker plot.
- Interpret graphical data displays.
- Detect outliers using box plots.
- Perform calculations to trim data.

Reading

- Black, K. *Business Statistics* Chapter 3 Sections 3.1-3.5.

  Course Reserves:

- Wilcox R. R. *Basic Statistics* Chapter 2: Pages 22-29, Chapter3: Pages 34-45
- Chihara & Hetersberg: *Mathematical Statistics with Resampling and R* Chapter 2 (EDA)

Videos

- Computing Variance and Standard Deviation
- Understanding and Using the Empirical Rule
- R Lesson Video

Assignments

- Test 1

Week 3 -- Complete by Sunday, Jul 12, 2020

Learning Objectives

After this week, the student will be able to:

- Describe probability.
- Articulate the different methods of assigning probabilities.
- Understand and apply axioms and properties of probability.
- Compute probabilities under different conditions.
- Understand conditional probability and Bayes' theorem.
- Determine the mean, variance and standard deviation for a discrete variable.
- Solve problems using binomial and Poisson probability distributions.

Reading

- Black, K. *Business Statistics* Chapter 4 Sections 4.1-4.7 & Chapter 5 Sections 5.1-5.5.
- Downey, A. B. *Think Bayes* Chapter 1 pages 1-10 (Course Reserves on the course site.)

Videos

- Constructing and Solving Probability Matrices
- Solving Probability Word Problems
- Solving Binomial Distribution Problems, Part I
- Solving Binomial Distribution Problems, Part II
- R Lesson Video

Assignments

- R Programming Assignment

Sync Session

There will be a sync session Monday, Jul 6, 2020 from 7 PM to 9 PM CST. Attendance is optional. A recording of the session will be posted in class the following day.

Week 4 -- Complete by Sunday, Jul 19, 2020

Learning Objectives

After this week, the student will be able to:

- Explain what is a probability density function for a continuous variable.
- Compute the expected mean value and variance.
- Describe a standard normal distribution and its properties
- Use the standard normal distribution to find z-scores
- Convert distributions to standard normal.
- Use the normal distribution as an approximation to the binomial distribution.
- Explain different types of sampling plans.
- Explain the central limit theorem.

Reading

- Black, K. *Business Statistics* Chapter 6 Sections 6.1-6.4 & Chapter 7 Sections 7.1-7.3.

Course Reserves:

- Wilcox R. R. *Basic Statistics* Chapter 4 Section 7 (Pages 70-76)
- Wilcox R.R. *Fundamental of Modern Statistical Methods* – Chapter 3

Videos

- Solving Problems Using the Normal Curve
- Solving for Probabilities of Sample Means using the z Statistic
- R Lesson Video

Assignments

- Test 2

Week 5 -- Complete by Sunday, Jul 26, 2020

Learning Objectives

After this session, the student will be able to:

- Estimate a population mean and a proportion.
- Define the t-distribution and determine probabilities given degrees of freedom.
- Use the chi-square distribution to estimate a population variance.
- Determine the sample size needed to estimate a population mean and a proportion.
- State what is a confidence interval and how it is used for statistical inference.
- Compute confidence intervals for a mean and a proportion.

Reading

- Black, K. *Business Statistics* Chapter 8 Sections 8.1-8.5.

Course Reserves:

- Wilcox R. R. *Basic Statistics* Chapter 6 Section 5 (Pages 121-129)
- Wilcox R.R. *Fundamental of Modern Statistical Methods* – Chapter 6

Videos

- Confidence Intervals
- R Lesson Video

Assignments

- Data Analysis Assignment #1

Sync Session

There will be a sync session Monday, Jul 20, 2020 from 7 PM to 9 PM CST. Attendance is optional. A recording of the session will be posted in class the following day.

Week 6 -- Complete by Sunday, Aug 2, 2020

Learning Objectives

After this session, the student will be able to:

- Develop one- and two-tailed hypotheses that can be tested.
- Develop test critical regions.
- Reach conclusions based on hypothesis tests
- Explain Type I and Type II errors.
- Perform hypothesis tests on means and proportions.
- Use p-values for hypothesis testing.
- Discuss statistical significance versus practical significance.

Reading

- Black, K. *Business Statistics* Chapter 9 Sections 9.1-9.6 and Chapter 16 Sections 16.1-16.2.

Course Reserves:

- Wilcox R.R. *Fundamental of Modern Statistical Methods* – Chapter 5

Videos

- Establishing Hypothesis
- Two-Tailed Tests
- Type I and Type II Errors
- Hypothesis Testing Using the z Statistic
- Understanding p-values
- R Lesson Video

Assignments

- Test 3

Week 7 -- Complete by Sunday, Aug 9, 2020

Learning Objectives

After this session, the student will be able to:

- Develop hypotheses for testing the difference in means or proportions of two populations.
- Use the z-statistic to develop confidence intervals for the difference in two means.
- Perform two-sample t-tests, paired t-tests and construct confidence intervals.
- Develop confidence intervals for the difference in two population proportions.
- Test hypotheses about the difference in variance between two populations.

Reading

- Black, K. *Business Statistics* Chapter 10 Sections 10.1-10.5.

  Course Reserves:

- Wilcox R. R. *Basic Statistics* pages Chapter 9 Section 9.1 pages 184-193.

Videos

- Determining Which Inferential Technique to Use, Part I: Confidence Intervals
- Determining Which Inferential Technique to Use, Part II: Hypothesis Tests
- t Test for Two Samples
- R Lesson Video

Assignments

- R Programming Assignment #2

Sync Session

There will be a sync session Monday, Aug 3, 2020 from 7 PM to 9 PM CST. Attendance is optional. A recording of the session will be posted in class the following day.

Week 8 – Complete by Sunday, Aug 16, 2020

Learning Objectives

After this session, the student will be able to:

- Describe what is a designed experiment.
- Use a single factor AOV model for analysis.
- Recognize a randomized block design.
- Explain the advantages of a two-way AOV.
- Compute sums of squares and mean squares
- Use multiple comparison tests.
- Explain what is an interaction.
- Calculate correlations.
- Fit a simple linear regression equation.

Reading

- Black, K. *Business Statistics* Chapter 11 Sections 11.1-11.5 & Chapter 12 Sections 12.1-12.3.

Course Reserves:

- Wilcox R. R. *Basic Statistics* Chapter 10 Section 10.1 pages 210-217.

Videos

- Computing and Interpreting a One-Way ANOVA
- Testing the Regression Model I: Predicted Values, Residuals, and Sum of Squares of Error
- R Lesson Video

Assignments

- Test 4

Week 9 – Complete by Sunday, Aug 23, 2020

Learning Objectives

After this session, the student will be able to:

- Explain a simple linear regression model.
- Determine the equation of a simple linear regression line.
- Specify the two parameters of a straight line.
- Discuss the risks of extrapolation.
- Perform inference about regression coefficients.
- Calculate the Pearson product-moment correlation coefficient.
- Calculate standard errors and confidence intervals for regression coefficients.
- Test the overall model.
- Assess Model Adequacy.

Reading

- Black, K. *Business Statistics* Chapter 12 Sections 12.4-12.7 & Chapter 13 Sections 13.1-13.3.

Course Reserves:

- Wilcox R. R. *Basic Statistics* Chapter 8 Section 8.3 pages 172-176.

Videos

- Testing the Regression Model II—Standard Error of the Estimate and r squared
- R Lesson Video

Assignments

- Data Analysis Assignment #2

Students should be aware that the proctored final exam opens at 12:01 am CST Monday, Aug 24, 2020. The Final Examination is due by 11:59 pm CST Sunday, Aug 30, 2020. **You are responsible for scheduling this proctored exam with Examity. Please be aware access to the internet during the exam is not permitted. Separate portable devices such as iPads and Kindles are not permitted. Only one computer screen may be used for the exam. See the course site for instructions.**

Sync Session

There will be a sync session Monday, Aug 17, 2020 from 7 PM to 9 PM CST. Attendance is optional. A recording of the session will be posted in class the following day. This will be a Q&A session with practice problems in preparation for the final.

Week 10 – Complete Saturday, Aug 30, 2020

Learning Objectives

- No new learning objectives.

Assignments

- Final Exam

The Final Examination opens at 12:01 am CST Monday, August 24, 2020. The Final Examination is due by 11:59 pm CST Saturday, August 30, 2020. **You are responsible for scheduling this proctored exam with Examity. Please be aware access to the internet during the exam is not permitted. Separate portable devices such as iPads and Kindles are not permitted. Only one computer screen may be used for the exam. See the course site for instructions.**

There is a practice test with solutions posted in Canvas under Week 10. ***Practice Problems*** are posted in the weekly modules. These are practice problems and carry no point value. They may be attempted at any time during the course.