

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Online Statistics Teaching and Learning

Citation for published version:
Cetinkaya-Rundel, M, Ju, M & Albert, J 2020, Online Statistics Teaching and Learning. in JP Howard II & JF Beyers (eds), *Teaching and Learning Mathematics Online.* Chapman and Hall, New York. https://doi.org/10.1201/9781351245586

Digital Object Identifier (DOI):
10.1201/9781351245586

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Teaching and Learning Mathematics Online

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 06. Nov. 2026

CHAPTER 6

Online Statistics Teaching and Learning

Jim Albert, Mine Cetinkaya-Rundel, and Jingchen Hu*

CONTENTS

6.1 Introduction 100
6.2 Description of the Online Courses 101
6.2.1 An Online Course in Exploratory Data Analysis 101
6.2.2 A Bayesian Statistics Course for Cross-Campus Share 102
A Five-Course MOOC Specialization: Statistics with R 104
6.3 Online Course Design 106
6.3.1 Presentation of Content 106
6.3.2 Course Material in Video Form 107
6.3.3 Learning Objectives 107
6.3.4 Suggested Readings and Practice 108
6.4. Use of Technology 108
6.4.1 R Package 108
6.4.2 EDA Blog, YouTube Videos 108
6.4.3 Shiny Activities 109
6.4.4 Use of Technology in Lectures 109
6.4.5 Computing Lab 109
6.5 Assessments and Engagement 110
6.5.1 Weekly Data Analysis Assignments 110
6.5.2 Final Project 110
6.5.3 Homework 111
6.5.4 Case Studies 111
6.5.5 Quizzes 112
6.6 Interaction 112
6.7 Challenges 113

* July 13, 2019Jingchen Hu was supported by The Andrew W. Mellon Foundation Anne McNiff Tatlock '61 Endowment for Strategic Faculty Support at Vassar College and the Liberal Arts Collaborative for Digital Innovation (LACOL).

TNF_06_K338312_C006_docbook_new_indd.indd 99 06-03-2020 14:57:43

6.8 Concluding Comments 114
Presentation of Content 115
Interaction 115
Assessment 115
Using Software 115
References 116

6.1 INTRODUCTION

Statistics is probably one of the most active fields to embrace and engage online teaching and learning. Numerous massive open online courses (MOOC) have been developed and have attracted a great number of online learners. At the time of writing, there are 346 courses and specializations with the keyword "statistics" on Coursera,* one of the most popular online learning platforms. DataCamp,† a growing online learning platform specializing in data science education, has 159 courses, 22 tracks, and 122 instructors. A broad survey of online statistics education is described in [1], and a discussion of building an online statistics curriculum is contained in [2].

For statistics courses at all levels, teaching and learning online pose challenges in different aspects. Particular online challenges include how to effectively and interactively conduct exploratory data analyses, how to incorporate statistical programming, how to include individual or team projects, and how to present mathematical derivations efficiently and effectively.

This chapter draws from the authors' experience with seven different online statistics courses to address some of the aforementioned challenges. Section 2.1 is an online exploratory data analysis course taught at Bowling Green State University. Section 2.2 is an upper-level Bayesian statistics course taught at Vassar College and shared among ten liberal arts colleges through a hybrid model. Section 2.3 describes a five-course MOOC specialization on Coursera, offered by Duke University.

All of these courses are designed for undergraduate or graduate students with calculus backgrounds. The general aim in this chapter is to provide overviews of these online courses, discuss challenges and approaches, and provide general guidelines for statistics educators interested in online teaching and learning of statistics.

Yang [3] provides an overview of the components of an online statistics course and uses student feedback to gain insight into the particular components that appear effective for learning the statistics material. Everson and Garfield [4] discuss the use of student discussions in an online statistics course and focus on the types of discussion that appear to facilitate understanding of the statistics concepts. Dunwill [5] provides general comments about the challenges of teaching in an online format, and Everson [6] discusses her experiences in teaching online after teaching face-to-face courses in statistics.

* For more information, visit www.coursera.org/.
† For more information, visit datacamp.com.

Section 6.2 gives an overview of the seven online statistics courses focusing on the intended audience and the course content. Sections 6.3 through 6.6 focus on particular components of the online course, and discuss how each course addresses the particular component. Section 6.3 discusses the problem of the online course design. How is the instructional content presented and organized, keeping in mind the learning objectives of the course? Section 6.4 focuses on the use of technology in each course. A statistics course will typically include the use of a software package. What types of software are used in each class, and how is the software integrated with the learning of the conceptual material? Section 6.5 describes the different forms of assessment for each course. What types of assessments such as homework or projects are used, and do students have the opportunity to work together on assignments? Section 6.6 describes how the students interact with the instructor in the course and how students interact with other students. How does a student get help in the course? Not all of the presented online courses run smoothly, and Section 6.7 describes the challenges that the instructors faced when implementing each class. Section 6.8 summarizes the general features and challenges of the courses, directed towards the instructor who is developing their first online statistics class.

6.2 DESCRIPTION OF THE ONLINE COURSES

6.2.1 An Online Course in Exploratory Data Analysis

This online course, Exploratory Data Analysis at Bowling Green State University, focuses on the principles of exploring data following ideas from John Tukey's EDA book [7]. The audience consists of graduate and undergraduate students majoring in statistics, and there is a probability prerequisite.

Generally, the main intent of the course is to describe an exploratory philosophy in the analysis of data. One does not wish to impose any assumptions such as normal sampling distributions or equality of variances between groups. Instead, one wishes to explore the data, looking for patterns in distributions and relationships. There are four "R"s in EDA that summarize the general philosophy in data exploration. "Revelation" means that EDA often uses graphical displays in data discovery. "Resistance" means that it is desirable to use statistical methods that are resistant or non-sensitive to outlying values. "Reexpression" means that it is sometimes useful to reexpress variables by a nonlinear transformation such as a log or square root. Last, "residual" means that one usually wishes to look at the deviations from a statistical fit.

Table 6.1 shows the main units for the EDA course. The course begins with a discussion of graphical displays and resistant summaries for a single batch of measurement data. The next general topic is the comparison of batches of measurement data and the use of reexpressions to equalize spreads across batches. Properties of the Box-Cox power family of transformations [8] are explored in Unit 4, and this family is used to perform an appropriate reexpression to make a data distribution symmetric. Scatterplots of two measurement variables are introduced in Unit 5, and Tukey's resistant line is applied as a general method of fitting a straight line to data. In cases where the scatterplot pattern is nonlinear, a running-median smoother is described as a simple way of smoothing a scatterplot to assess the general pattern. Unit 6 explores a two-way table where one summarizes a measurement variable over two categorical variables. Median

TABLE 6.1 Unit and Lectures for the Online EDA Course

Unit	Lecture
1. Introduction	1.1 Introduction to EDA I 1.2 Introduction to EDA II
2. Single Batch	2.1 Displays 2.2 Summaries
3. Comparing Batches	3.1 Boxplots 3.2 Spread Level Plots 3.3 Comparing Batches III
4. Transformations	4.1 Transformations 4.2 Reexpressing for Symmetry 4.3 Reexpressing for Symmetry II 4.4 Transformations Summary
5. Plotting	5.1 Introduction to Plotting 5.2 Resistant Line 5.3 Plotting II 5.4 Straightening 5.5 Smoothing
6. Two-Way Analyses	6.1 Median Polish 6.2 Plotting Additive Fit 6.3 Multiplicative Fit 6.4 Extended Fit
7. Counts and Fractions	7.1 Binning Data 7.2Binning Data II 7.3 Fraction Data
EDA Project	

polish is a resistant method of applying an additive fit by the use of a logarithmic transformation; this method can also be used to apply a multiplicative fit to these two-way data structures. The course concludes in Unit 7 by describing methods for binning measurement data, assessing if the histogram has a Gaussian shape, and exploring batches of fraction data.

6.2.2 A Bayesian Statistics Course for Cross-Campus Share

The past decades have seen great methodological, computational, and inferential advancement of Bayesian statistics. While Bayesian statistics continues to gain attention and becomes ever more popular among data analysts and researchers, the topic itself is rarely available to students, especially at the undergraduate level. In most liberal arts colleges, with the staffing constraints, offering a topic course on Bayesian statistics can at most be an occasional luxury. More commonly, such a course is not offered at all.

Vassar College had the chance to offer an undergraduate-level Bayesian statistics course in Fall 2016. The extremely positive experience with a small group of motivated students has encouraged the instructor to think beyond the boundary of a physical college location. Vassar is a member of the Liberal Arts Consortium for Online Learning (LACOL).* Under

* LACOL is a partnership of ten liberal arts colleges in the United States, founded in 2014. By leveraging the power of consortial relationships, LACOL focuses on utilizing and adapting emerging technologies to promote excellent and innovative teaching, learning, and research in the liberal arts. For more information about LACOL, visit http://lacol.net/about-the-consortium/.

TNF_06_K338312_C006_docbook_new_indd.indd 102 06-03-2020 14:57:43

TABLE 6.2 Section and Main Topics for the Cross-Campus Shared Bayesian Statistics Course

Section	Topics
Inference	Bayes theorem, conjugate prior, posterior distribution, HPD interval, predictive distribution
Computation	Monte Carlo approximation, Markov chain Monte Carlo (MCMC), Gibbs sampler, Metropolis-Hastings alogrithm, MCMC diagnostics, JAGS
Applications	Bayesian hierarchical modeling, Bayesian linear regression, latent class modeling, Bayesian cognitive modeling

the Upper Level Math & Stats Project,* starting from Fall 2017, the Bayesian Statistics course at Vassar College is taught locally at Vassar while shared among the LACOL colleges through a hybrid model.

The Upper Level Math & Stats Project focuses on sharing upper-level mathematics and statistics courses among participating campuses, to supplement existing and probably limited offerings while maintaining the liberal arts flavor. Vassar's Bayesian Statistics course in Fall 2017 is one of the three courses in the pilot study (the other two are upper-level mathematics courses offered by two other member colleges).

The student audience consists of junior and senior students. The prerequisite includes multivariate calculus, linear algebra, and probability. The textbook is *A First Course in Bayesian Statistics Methods* by Peter D. Hoff [9], a book mainly used at the graduate level. The instructor intentionally borrows more applied material from *Bayesian Cognitive Modeling: A Practical Course*, written by Michael D. Lee and Eric-Jan Wagenmakers [10], making the course more accessible to undergraduate students. Occasionally, advanced material from *Bayesian Data Analysis* by Andrew Gelman and others [11] is used to supplement.

There are three general sections of the course: inference, computation, and applications, with main topics in each section listed in Table 6.2. Through the introduction of one-parameter models such as beta-binomial and normal-normal, the inference section covers the inferential basics. Moving to multi-parameter models such as normal with two unknown parameters, computation techniques are covered. Students are equipped with the skills of writing up the Markov chain Monte Carlo (MCMC) sampler when possible, as well as the use of Just Another Gibbs Sampler (JAGS) through the `rjags` R package. Ultimately, through various applications, students are exposed to more advanced models. They are motivated to understand and construct Bayesian models in each application, perform simulation by MCMC with appropriate computation techniques, and answer inferential questions in context.

Participating in LACOL's Upper Level Math & Stats Project is the first time that a Bayesian statistics course is ever shared among colleges from different geographic locations. In addition to motivating and cultivating students' learning of such advanced statistics topics, the instructor needs to redesign an existing face-to-face course to adapt to a hybrid model. Challenges include using software to provide both synchronous and asynchronous access to the lectures, identifying what material is suitable for being moved online, holding online office hours for remote students, coordinating with local faculty liaisons from each remote

* For more information, visit http://lacol.net/category/collaborations/projects/upper-level-math/.

campus, and creating a learning community involving all students, among other things. This course uses R extensively for simulations and data analysis, and how to effectively incorporate R programming through a hybrid instruction model is also challenging.

A Five-Course MOOC Specialization: Statistics with R

Statistics with R is a specialization offered on Coursera (`www.coursera.org/specializations/ statistics`) comprised of five massive open online courses (MOOCs) designed and sequenced to help learners master foundations of data analysis and statistical inference and modeling. The specialization also has a significant hands-on computing component. The target audience is learners with no background in statistics or computing.

The first four courses in the specialization are Introduction to Probability and Data, Inferential Statistics, Linear Regression and Modeling, and Bayesian Statistics. These courses cover exploratory data analysis, study design, light probability, frequentist and Bayesian statistical inference, and modeling. A major focus of all of these courses is hands-on data analysis in R; each course features computing labs in R where learners create reproducible data analysis reports as well as fully reproducible data analysis projects demonstrating mastery of the learning goals of each of the courses. The fifth course is a capstone, where learners complete a data analysis project that answers a specific scientific/business question using a large and complex dataset. This course is an opportunity for learners to practice what they learned in the first four courses in the specialization.

Table 6.3 shows the modules and associated topics for each of the first four courses in this specialization. Each subsequent course assumes learners have either completed the previous course(s) or have background knowledge equivalent to what is covered in them. Each module is designed to be completed in one week, though learners have the flexibility to extend this if they need to.

Course 1, Introduction to Data, introduces sampling and exploring data, as well as basic probability theory and Bayes' rule. In this course learners examine various sampling methods, and discuss how such methods can impact the scope of inference. In addition, a variety of exploratory data analysis techniques are covered, including using data visualization and summary statistics to explore relationships between two or more variables. Another key learning goal for this course is the use of statistical computing, with R, for hands-on data analysis. The concepts and techniques introduced in this course serve as building blocks for the inference and modeling courses in the specialization.

Course 2, Inferential Statistics, introduces commonly used statistical inference methods for numerical and categorical data. Learners learn how to set up and perform hypothesis tests and construct confidence intervals, interpret p-values and confidence bounds, and communicate these results correctly, effectively, and in context without relying on statistical jargon. Building on computing skills they acquired in the previous course, learners conduct these analyses in R.

In Course 3, Linear Regression and Modeling, introduces simple and multiple linear regression. Learners learn the fundamental theory behind linear regression and, through data examples, learn to fit, examine, and utilize regression models to examine relationships

TABLE 6.3 Modules and Topics for the First Four Courses in the Statistics with R Specialization

Course 1: Introduction to Probability and Data	
1.1 *Introduction to data*	Data basics, observational studies and experiments, sampling and sources of bias, experimental design
1.2 *Exploratory data analysis and introduction to inference*	Visualizing data, measures of center and spread, robust statistics, transformations, exploring bi-/multivariate relationships, introduction to inference via simulation
1.3 *Introduction to probability*	Independent and disjoint events, conditional probability, Bayes' rule, introduction to Bayesian inference
1.4 *Probability distributions*	Normal and binomial distributions, assessing normality
1.5 *Data analysis project*	Exploratory data analysis of data from the Behavioral Risk Factor Surveillance System
Course 2: Inferential Statistics	
2.1 *Confidence intervals*	Sampling variability and Central Limit Theorem, confidence intervals for a mean, accuracy vs. precision
2.2 *Inference and significance*	Hypothesis testing for a mean, decision errors, statistical vs. practical significance
2.3 *Inference for means*	t-distribution, inference for a mean and for comparing two or more means, multiple comparisons, bootstrapping
2.4 *Inference for proportions*	Sampling variability and CLT for proportions, confidence intervals and hypothesis tests for two or more proportions, randomization tests for small samples
2.5 *Data analysis project*	Inference on data from the Behavioral Risk Factor Surveillance System
Course 3: Linear Regression and Modeling	
3.1 *Linear regression*	Correlation, residuals, least squares line, prediction and extrapolation
3.2 *More on linear regression*	Outliers. Inference for regression, variability partitioning
3.3 *Multiple linear regression*	Multiple predictors, adjusted R^2, collinearity and parsimony, inference for MLR, model selection and diagnostics
3.4 *Data analysis project*	EDA and single and multiple regression for movies data
Course 4: Bayesian Statistics	
4.1 *Basics of Bayesian statistics*	Conditional probabilities and Bayes' rule, diagnostic testing, Bayes updating, Bayesian vs. frequentist definitions and inference, effect size and significance
4.2 *Bayesian inference*	From discrete to continuous, elicitation, conjugacy, Gamma-Poisson and normal-normal conjugate families, non-conjugate priors, credible intervals, predictive inference
4.3 *Decision making*	Loss functions, minimizing expected loss, Monte-Carlo sampling, prior choice and reference priors, MCMC
4.4 *Bayesian regression*	Bayesian simple and multiple regression, model uncertainty and averaging, decisions under model uncertainty
4.5 *Perspectives*	Interviews with statisticians on how they use Bayesian statistics in their work
4.6 *Data analysis project*	Bayesian inference and regression for movies data
Course 5: Statistics with R Capstone	
5.1 *Exploratory data analysis*	
5.2 *Basic model selection*	
5.3 *Model Selection and Diagnostics*	
5.4 *Out of Sample Prediction*	

between multiple variables. Model fitting and assessment is done in R, and a substantial amount of examples are focused on interpretation and diagnostics for model checking.

These first three courses were originally offered as a single, much longer, MOOC, for two years, before being split into shorter courses to be bundled up in a specialization. Course 4, Bayesian Statistics, was added to the sequence at this point, in order to make

this introductory specialization more complete by adding a different point of view for approaching statistical analysis.

Course 4, Bayesian Statistics, introduces learners to the underlying theory and perspective of the Bayesian paradigm and shows end-to-end Bayesian analyses that move from framing the question to building models to eliciting prior probabilities to implementing in R. The course also introduces credible regions, Bayesian comparisons of means and proportions, Bayesian regression and inference using multiple models, and discussion of Bayesian prediction.

The last course in the specialization is a capstone course. The materials provided for this course are designed to serve as a reminder of learning goals of earlier courses or expand on them ever so slightly. A large and complex dataset is provided to the learners, and the analysis requires the application of a variety of methods and techniques introduced in the previous courses, including exploratory data analysis through data visualization and numerical summaries, statistical inference, and modeling as well as interpretations of these results in the context of the data and the research question. Learners are encouraged to implement both frequentist and Bayesian techniques and discuss in the context of the data how these two approaches are similar and different, and what these differences mean for conclusions that can be drawn from the data.

6.3 ONLINE COURSE DESIGN

6.3.1 Presentation of Content

The EDA course was originally taught face-to-face in the classroom where the instructor would introduce and demonstrate the EDA methods in class and the students would work on weekly data analysis assignments. The online course was designed to follow the same format as the face-to-face version.

1. The lecture material for the class is posted online as PDF documents. Students have had difficulties understanding the material by reading directly from [7], and so the lecture material seems to be a reasonable substitute for the book material. The core EDA material is contained in a series of 23 PDF documents. A particular "lecture" motivates and describes the particular EDA method with an illustration using the R programming language.
2. There are weekly data analysis assignments and no written exams in the course. It is well-known that exams can be challenging to administer in an online format.
3. The work on the assignments is a blend of statistical work such as tables and graphs and interpretation of the results. These assignments are turned in online by the use of R Markdown documents saved in html format.
4. All of the R code in the lecture notes is made available to the students by a collection of R scripts.

In the Bayesian statistics course, regular face-to-face lectures are delivered in the classroom with students present at Vassar College. Every lecture is broadcast and recorded by

Zoom, a video conferencing software.* Remote students can join the lecture in real time with a Zoom meeting ID. Otherwise, they can watch the recorded videos after the videos are posted on the same day of the lecture.

6.3.2 Course Material in Video Form

This hybrid model in the Bayesian statistics course made lectures available in video form. With the flexibility of making videos and the added familiarity of learning through videos on the students' end, other course material has also been turned into video form. This material has conventionally been available as a Word- or PDF document.

For example, when an example is not fully developed during the lecture due to time constraints, a short video on this example is created and made available to students to review if necessary. As another example, when many students are having problems with the same homework question (based on observation from office hour visits), a short video providing hint on this homework question is created and made available.

R programming demonstrations are very suited to video form. By watching a video with a step-by-step demonstration of programming, students are able to pause when needed, see things in action, and practice along the way. Several R programming videos are created for students in this course.

In the Five-Course MOOC, each module includes seven to ten videos roughly four to seven minutes in length. Most of these videos introduce new concepts and the remaining provide additional examples and worked-out problems. The slides that serve as the background in the videos are created in Keynote (Apple's presentation software application) and feature a substantial amount of animations such that text, visualizations, and calculations showed on the slides follow the pace of speech in the videos. Many learners have expressed in their course feedback that these features make the videos more engaging and easier to follow compared to videos in many other MOOCs. Sample videos from the Inferential Statistics course are hosted on YouTube (`bit.ly/2LrO6KZ`).

6.3.3 Learning Objectives

Each module of the Five-Course MOOC featured a set of learning objectives. A sampling of learning objectives from the Inferential Statistics course is shown below:

- Explain how the hypothesis testing framework resembles a court trial.
- Recognize that in hypothesis testing we evaluate two competing claims: the null hypothesis, which represents a skeptical perspective or the status quo, and the alternative hypothesis, which represents an alternative under consideration and is often represented by a range of possible parameter values.
- Define a p-value as the conditional probability of obtaining a sample statistic at least as extreme as the one observed given that the null hypothesis is true: p-value = P(observed or more extreme sample statistic | H_0 true).

* For more information about Zoom, visit https://zoom.us/

These learning objectives are constructed using verbs from the revised Bloom's Taxonomy [12] and aim to keep learners organized and focused while watching the videos. The learners are recommended to have the learning objectives handy while watching the videos and revisit sections of the videos and/or suggested readings for any learning objectives that they feel like they have not mastered at the end of the module.

The learning objectives are provided as separate stand-alone documents, and after each batch of related learning objectives are a few simple conceptual questions for learners to check their understanding before moving on.

6.3.4 Suggested Readings and Practice

Suggested readings for the first three courses of the MOOC course came from *OpenIntro Statistics* [13]. This book is free and open-source, meaning that learners enrolled in the MOOC do not need to additionally purchase a textbook. The readings are optional as the videos explicitly introduce and cover all required topics for the course; however many learners have reported in their feedback that they really like having a reference book that closely follows the course material. Practice problems are also suggested from the end of chapter exercises in this book.

For the fourth course on Bayesian statistics, readings are suggested from *An Introduction to Bayesian Thinking* [14]. This textbook has been written by the Bayesian Statistics course development team (faculty and PhD students) specifically as a companion to this course and is also freely available on the web.

6.4. USE OF TECHNOLOGY

6.4.1 R Package

A special R package `LearnEDAfunctions` was written for the EDA class. This package contains all of the datasets used in the lecture notes and the assignments. In addition, the package contains special functions for implementing the EDA computations. For example, the function `rline` computes Tukey's resistant line and the function `fit.gaussian` fits a Gaussian comparison curve to histogram data and outputs the rootogram residuals from the Gaussian fit.

6.4.2 EDA Blog, YouTube Videos

Different methods in the EDA course were used to provide weekly communications with the students. For several iterations of the course, the instructor posted weekly articles on the blog "Exploratory Data Analysis" (`https://exploredata.wordpress.com/`).

In a typical post, the instructor would give an example of the weekly EDA concept and give advice on common problems in applying the interpretation of the EDA method. Blog postings from previous years are made available for the student who wishes to see additional illustrations of the statistical methods. Since some students expressed preference for learning by watching videos instead of reading notes, the instructor added videos at the YouTube channel `www.youtube.com/user/bayesball2/videos`. A particular video would show the implementation of a particular EDA method using R and the `LearnEDAfunctions` package.

This Bayesian course extensively used videos to deliver content. One of the most creative uses of making videos for this hybrid course is to create guest lectures. Conventionally, guest lectures are delivered in the physical classroom. Now, to include remote students, guest lectures can be created in video form and made available to students online. This practice also saves class meeting time when possible. For example, a guest lecture video by a cognitive science professor at Vassar College is created and used as an introduction to Bayesian hierarchical modeling. Students watch the guest lecture before the class meeting, become familiar with the topic by themselves outside of the class, then when meeting in class, the lecture and discussion can follow from the common ground of the material from the guest lecture directly. This practice also exposes students to applications of Bayesian statistics at various stages of their learning.

6.4.3 Shiny Activities

In addition to the weekly data analysis assignments, the EDA class also contains several activities where the student uses sliders and other interactive tools in exploring data. For example, in choosing the "correct" power of a reexpression, the student can choose a value of the power on a slider and see the immediate impact of that particular reexpression in a graph of the reexpressed data.

6.4.4 Use of Technology in Lectures

Instead of writing on the chalkboard or presenting lecture slides on the projector in front of the classroom using a computer, the instructor in the Bayesian course uses an iPad as a participant of the Zoom meeting, brings up lecture slides inside the Zoom software, shares the screen to present the slides to the class on the projector, and uses an Apple pencil to write on the slide or a whiteboard. When doing an R programming demonstration, the instructor joins the Zoom meeting with a laptop, then shares the screen of R/RStudio on the laptop to the projector. Zoom records the video from the projector, and all audio from the lecture. Sometimes, the instructor could use a directional mic on the iPad to improve sound capturing.

6.4.5 Computing Lab

Each module in the MOOC featured a computing lab in R. The objective of the labs is to give learners hands-on experience with data analysis using modern statistical software, R, as well as providing them with tools that they will need to complete the data analysis projects successfully.

The statistical content of the labs matches the learning objectives of the respective modules they appear in, and the application examples (i.e. datasets and research questions) are primarily from social and life sciences. The labs also make heavy use of an R package, `statsr`, which was designed specifically as a companion for the specialization [15]. Two other important aspects of the labs are that (1) they use the `tidyverse` syntax and (2) they are completed as reproducible R Markdown reports. The tidyverse is an opinionated collection of R packages designed for data science, meaning that the grammar used in the packages is optimized for working with data specifically for wrangling,

cleaning, visualizing, and modeling data [16]. The choice of the tidyverse syntax for beginners is rooted in wanting to get learners exploring real and interesting data and building informative and appealing visualizations and drawing useful conclusions as much as possible [17].

R Markdown provides an easy-to-use authoring framework for combining statistical computing and written analysis in one document [18]. It builds on the idea of *literate programming*, which emphasizes the use of detailed comments embedded in code to explain exactly what the code is doing [19]. The primary benefit of R Markdown is that it restores the logical connection between statistical computing and statistical analysis by synchronizing these two parts in a single reproducible report. From an instructional perspective this approach has many advantages: reports produced using R Markdown present the code and the output in one place, making it easier for learners to learn R and locate the cause of an error and learners keep their code organized and workspace clean, which is difficult for new learners to achieve if primarily using their R console to run code [20]. Each lab is provided to the learners in an R Markdown template that they can use as a starting point for their lab report. Earlier labs in the specialization include lots of scaffolding, and almost have a fill-in-the-blanks feel to them. As the course progresses the scaffolding in the templates is removed, and by the end of the first course learners are able to produce a fully reproducible data analysis project that is much more extensive than any of their labs. All labs in the specialization are hosted in a publicly available GitHub repository at `https://github.com/StatsWithR/labs`.

6.5 ASSESSMENTS AND ENGAGEMENT

6.5.1 Weekly Data Analysis Assignments

In a typical data analysis assignment of the EDA course, the student works on a particular EDA method using a specific dataset or another suitable dataset chosen by the student. One challenge for the student is to find a suitable data structure to implement the EDA method. For example, if the EDA task is to symmetrize a dataset by the use of a power expression, the student needs to find a strongly skewed dataset that could benefit with a reexpression.

6.5.2 Final Project

In a final capstone project of the EDA course, the students select their own dataset, state questions of interest, and explore the dataset using several of the EDA methods discussed in class. The focus of this project is not on the implementation of the methods but rather on the interpretation of the results in light of the questions that were originally posed.

The Bayesian course has a final project component, and students can choose from one of the following.

- A Bayesian data analysis on a topic of your choosing
- A new Bayesian methodology or theoretical finding
- A Bayesian research paper or a book chapter (choose from a provided list)

Students submit a project proposal after the first midterm exam. They are encouraged to meet with the instructor to discuss their project ideas and their progress along the way.

The final project presentation has two parts: a two-min video on Moodle, and a poster at the poster session. The choice of a poster session is to accommodate a relatively large class (16 local students). However, it mimics real research settings, as it has become common for academic conferences to have a poster session for graduate students and junior researchers. Students in the class overall enjoy being able to talk to audience as a small group. The amount of interaction between presenters and the audience is much more than a regular presentation.

Each course in the MOOC ends with a data analysis project, the focus of which is summarized in Table 6.3, and the specialization wraps up with an extended capstone project. Each student who turns in a project evaluates three other students' work using a peer evaluation rubric. Learners are also strongly encouraged to seek informal feedback on their projects in the course discussion forums. All data analysis projects appearing in the courses in this specialization are hosted in a publicly available GitHub repository at `https://github.com/StatsWithR/projects`.

6.5.3 Homework

Homework for the Bayesian course is on a biweekly basis. The assignment usually consists of a set of derivation exercises to enhance the understanding of Bayesian methodology, and a set of application-based exercises, which require the use of R programming. There are two midterm exams and no final exam.

The teaching assistant for this course holds regular office hours at Vassar. While these office hours are held online too, remote students rarely make use of them. Instead, the instructor meets with the remote students together during a separately scheduled online office hour, also through Zoom.

Homework submission from the remote students is done through scans and emails. Exams for remote students are proctored by the local faculty liaison. Exam papers are sent to the instructor by scans and emails too. All grading is either done by the instructor or the teaching assistant. Graded homework and exams are returned to the remote students by scans and emails.

6.5.4 Case Studies

For the Bayesian course, towards the latter part of the semester when students have been exposed to and gained experience with Bayesian inference, students are grouped to do case studies with real data applications. These case studies are all open-ended. Students are given the chance to freely explore the datasets and come up with their methods and realize their inferences through MCMC computation techniques.

Prior to the case study class meeting, groups need to post their analyses onto the Moodle discussion forum to receive credit. Students in the same group take turns to be the leading writer of the analyses. Such a practice ensures that everyone is prepared to discuss the approaches and findings from the group, and the class meetings usually turn out to be great discussion sessions and ideas bounce back and forth.

6.5.5 Quizzes

Each module of the MOOC course features two sets of multiple-choice quizzes, one formative and one summative. Each question is encoded with feedback that points learners back to relevant learning objectives. Learners can attempt the summative quizzes multiple times with slightly modified versions of the questions.

6.6 INTERACTION

Currently, there is limited interaction in the EDA course. Students communicate with the instructor by means of personal meetings or email or messages sent through the learning management system. There is no regularly planned interaction between students such as an outline chat session, but students are asked to read and review the project presentations of two other students in the class.

In the Bayesian course, to create and foster an online learning community, there are extensive uses of the online discussion forum on Moodle.*

At the beginning of the semester, students make self-introduction posts about their basic information (name, year, and school), prior statistics exposure, prior R exposure, and potential final project interests.

During the semester, online discussion forums are created whenever sharing of information and making comments are needed, and they are for credit sometimes. For example, when covering the Gibbs sampler, the class reads a research paper "Explaining the Gibbs Sampler" by George Casella and Edward I. George [21]. A reading guide for this paper with six questions is provided to the students. Prior to the class meeting, students need to respond to any one question on the online discussion forum to receive participating credit. After the class discussion, students need to make another response to receive credit. Such a requirement not only helps students in reading a statistics research paper outside of class, but also helps facilitate in-class discussion both by more engagement prior to and post class meeting.

There is a final project for students in the course. In addition to presenting their projects in a poster session at the end of the semester, students need to make a two-min video post about their projects on the online discussion forum. Watching a two-min pitch talk prior to the poster session helps other students to arrange their poster visits. These videos also help students to succinctly present their projects in a manner as appealing as possible.

While the use of the online discussion forum on Moodle is motivated first mostly to make remote students feel included, it ultimately engages local students much better as well.

In-person student interaction, or the lack thereof, is often a major challenge in online courses. However in MOOCs the discussion forums are often a major strength of the course. Given that at any point thousands of students are enrolled in the course, even if a small percentage of these students choose to browse the discussion forums, and an even smaller percentage of them interact with other learners on the course discussion forums, this still results in a large number of learners interacting with each other.

* Moodle is the course management system used at Vassar College.

Additionally, over the years of the MOOC being offered, a growing number of very knowledgeable and helpful course mentors emerged from the discussion forums. These are learners who took the courses at an earlier time and now volunteer their time to answer student questions and provide direction for new learners. Their contribution is essential to the success of the course discussion forums. In addition to answering student questions directly, they also escalate any questions that would benefit from additional input to the course instructor. This workflow makes it feasible for the instructor to keep tabs on the discussion forums when most needed, but not get overwhelmed by the number of questions posted each day by the learners.

6.7 CHALLENGES

There are several current challenges in the current version of the EDA online course.

- Interaction with the student

  It is beneficial if the student can interact with the instructor and fellow students in an online course. Unfortunately, attempts for interaction such as online chat sessions or online message boards have not been effective in this particular class. Therefore most of the communication is done through email and personal meetings. There is an effort to try out new methods of interaction when they become available.

- Technology issues

  Students can get frustrated with technology issues such as installing software or getting their R markdown files to "knit" properly. It is best to address these issues early in the course so the course is more about the EDA concepts and less about the associated technology.

- Balance of computation and interpretation in assignment work

  In a typical assignment, the student will turn in a Markdown file that blends output from the R system and written text that interprets the R output in the context of the particular applied problem. Since the course is really focusing on the interpretation rather than the implementation of the EDA methods, the assignment should emphasize the interpretation component. Depending on the background, the student may emphasize instead the computation component, but hopefully the students will learn what is expected in future assignments.

Although the instructor of the Bayesian course has been faced with various challenges, advice, suggestions, and sharing from multiple parties have been tremendously helpful. To improve the teaching and learning mode of the hybrid model, lecture videos can be edited and shortened when resources permit. If done properly, the lecture videos can well be a set of learning material for anyone (not necessarily from the LACOL member colleges) who is interested in an undergraduate-level introduction to Bayesian statistics. More thought and consideration on turning course material into video form can further enhance the

teaching and learning. While students' interaction can be maintained using an online discussion forum on Moodle, other forms of interaction can be explored and developed to further enhance the overall interaction in the course.

For the five-course MOOC, there are three main challenges with offering this content on an online platform; two of these are associated with the labs and the other is associated with the data analysis projects.

- Autograding: Given that this is a course with thousands of learners enrolled at any given point, human-grading is simply not feasible. The lab assessments are set up as multiple-choice questions. Learners complete the lab exercises by generating R Markdown reports in which they analyze a dataset. Then, they answer a series of multiple-choice questions about the data analysis results. The challenge is that the multiple-choice questions do not assess the full spectrum of the skills we want learners to acquire via these labs—they assess whether they can obtain the correct results using R, however they do not assess mastery of R syntax, reproducibility of their analysis, etc.
- Computing infrastructure: Our preferred method for getting students with no computing background started with R is cloud-based access to RStudio in order to avoid challenges around local installation and to provide a uniform computing environment for all learners. However, it is not feasible to offer a centralized cloud-based solution to all learners enrolled in an MOOC, and hence students have to locally install R and RStudio and the correct versions of all packages they use in the labs.
- As a partial solution for this challenge, we offer students the option to complete the labs in the first course of the specialization on DataCamp (`www. datacamp.com`), an online learning platform that provides in-browser access to RStudio. This helps students struggling with software installation issues early on in the course to get started with data analysis and go back to tackling software challenges once they feel a little bit more confident with R.
- Peer evaluation: Autograding is not feasible for open-ended data analysis projects, and hence peer evaluation is the only solution for grading of these projects. Even with a very detailed rubric, consistency in grading is difficult to attain, and it is challenging for learners who are just learning the material themselves to evaluate others' work. Additionally, variability in the quality and depth of feedback provided can leave learners frustrated. The option to share their projects on the discussion forums and get feedback can be helpful for some learners, but others are not so keen on publicly sharing their projects.

6.8 CONCLUDING COMMENTS

Although these is current interest in teaching online introductory statistics courses, the online statistics courses described here are directed towards specific groups of students and all of the comments may not be directly applicable to the introductory class. For

example, the students in the EDA online course are primarily masters-level or advanced undergraduate students who are comfortable in working independently on assignments and projects, and this particular course design may not be suitable for an introductory statistics class with a minimal mathematics prerequisite. Similarly, the Five-Course MOOC has been designed for a global audience, assuming no common background and varying levels of access to computing resources. The delivery of the course can be simplified for a university audience where assumptions about common access and background might be more appropriate. Nonetheless, there are general common elements to these courses which would be helpful for the instructor who is designing the first online statistics course.

Presentation of Content

Although the EDA course features written instructional content, much of the content delivery in all of these online courses happens in videos. The use of videos makes it possible for several instructors to get involved in the presentation of content. In addition, it is possible for the student to learn from the video at their own pace, replaying parts of the video to help them understand the material.

Interaction

It is important to develop a cooperative learning environment among students in the online course. These courses suggest useful methods for facilitating this type of environment. Discussion forums, as described by the shared Bayesian course and the Five-Course MOOC, are one good way to foster communication between students. Another good opportunity for collaboration is through statistics projects where groups of students explore data on case studies.

Assessment

It should be noted that traditional forms of assessment such as multiple-choice exams play a limited role in assessment for these online statistics courses. Instead, these courses feature data analysis projects, interactive computer lab assignments, and other projects where the student carries out an exploration into a new method or finding that is not covered in the curriculum. However, the necessity for peer evaluation of the projects in the Five-Course MOOC presents its own challenges.

Using Software

All of these online courses use technology or software that may not be familiar to the student. Specifically, since these are statistics courses, one would typically use the R language together with specialized R packages. Introducing this technology creates challenges since the students can vary greatly with their experience with programming in general as well as with the R ecosystem. These courses have presented various ways to mitigate the technology challenges by creating special packages (e.g. the `LearnEDA` package in the EDA course, the `rjags` package in the shared Bayesian course, and the `statsr` package in the Five-Course MOOC) that include all of the datasets and special functions needed for the course. The Five-Course MOOC provides some suggestions to help some of these

technology issues such as providing cloud-based access to R via a third-party company, DataCamp. The instructor teaching an online statistics course should think carefully about the use of software, especially from the viewpoint of the student who is inexperienced with technology.

REFERENCES

1. Mills, J. D. and D. Raju, "Teaching statistics online: A decade's review of the literature about what works," *Journal of Statistics Education*, vol. 19, no. 2, 2011.
2. Young, D. S., G. F. Johnson, M. Chow and J. L. Rosenberger, "The challenges in developing an online applied statistics program: Lessons learned at penn state university," *The American Statistician*, vol. 69, no. 3, pp. 213–220, 2015.
3. Yang, D. "Instructional strategies and course design for teaching statistics on- line: Perspectives from online students," *International Journal of STEM Edu-cation*, vol. 4, no. 34, pp. 1–15, 2017.
4. Everson, M. and J. Garfield, "An innovative approach to teaching online statistics courses," *Technology Innovationss in Statistics Education*, vol. 2, no. 1, 2008.
5. Dunwill, E., "Teaching principles transferred to online courses: Strategies to use," *eLearning Best Practices*, 2016.
6. Everson, M, "10 things i learned about teaching online," *eLearn Magazine*, 2009.
7. Tukey, J. W., *Exploratory Data Analysis*. Pearson, 1977.
8. Sakia, R. M., "The box-cox transformation technique: A review," *The Statistician*, vol 41, no. 2 pp. 169–178, 1992.
9. Hoff, P. D., *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics, New York: Springer-Verlag, 2009.
10. Lee, M. D. and E. Wagenmakers, *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, 2014.
11. Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Ve-htari and D. B. Rubin, *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science, 3rd ed., 2013.
12. Anderson, L. W., D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths and M. C. Wittrock, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Blooms Taxonomy of Educational Objec-tives, Abridged Edition*. White Plains, NY: Longman, 2001.
13. Diez, D. M., C. D. Barr and M. Cetinkaya Rundel, *OpenIntro Statistics*. Cre- ateSpace, 3rd ed., 2014. https://www.openintro.org/.
14. Clyde, M., M. Cetinkaya Rundel, C. Rundel, D. Banks, C. Chai and L. Huang, *An Introduction to Bayesian Thinking*. GitHub, 1st ed., 2018. https:// statswithr.github.io/book/.
15. Rundel, C., M. Cetinkaya-Rundel, M. Clyde and D. Banks, Statsr*: Companion Package for Statistics with R*, 2018. R package version 0.1-0.
16. Wickham, V, *Tidyverse: Easily Install and Load the 'Tidyverse'*, 2017. R package version 1.2.1.
17. Robinson, D., "Teach the tidyverse to beginners," 2017. http://varianceexplained. org/r/ teach-tidyverse/.
18. Xie, Y., J. Allaire and G. Grolemund, *R Markdown: The Definitive Guide*. Boca Raton, FL: CRC Press, 2018.
19. Knuth, D. E., "Literate programming," *The Computer Journal*, vol. 27, no. 2, pp. 97–111, 1984.
20. Etinkaya-Rundel, M. C. and C. Rundel, "Infrastructure and tools for teaching computing throughout the statistical curriculum," *The American Statistician*, vol. 72, no. 1, pp. 58–65, 2018.
21. Casella, G. and E. I. George, "Explaining the gibbs sampler," *The American Statistician*, vol. 46, pp. 167–174, 1992.