Applications of technology and large data in statistics education and statistical graphics

by

Karsten Tait Maurer

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Heike Hofmann, Major Professor

Robert Stephenson

Max Morris

Mark Kaiser

Alejandro Andreotti

Iowa State University

Ames, Iowa

2015

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEDICATION

I dedicate this thesis to my love, Madeline, for the support she has given me in the pursuit of my goals. I am so excited to marry her this summer and begin our life together. Watch out world, here we come!

ACKNOWLEDGEMENTS

First, I would like to thank Dr. Heike Hofmann for the guidance she has provide during my time at Iowa State University. Her patient tutelage has helped me immeasurably in my winding – often stubbornly veering – pursuit of a doctorate.

Great appreciation also extends to my Ph.D. committee members, who have graciously given their time and advice. A special thanks to Dr. Bob Stephenson, whom also served as a study advisor and mentor in statistics education.

I would also like to thank my undergraduate advisor, Dr. Jon Anderson, whom sparked my interest in statistics and fostered that early curiosity.

Lastly, I owe an enormous debt of gratitude to my friends and family – especially my mother and father, Joyce and Clayton Maurer – for their love and support throughout my life. Thank you all.

CHAPTER 1. LITERATURE REVIEW

1.1 Introduction

This dissertation is a composite of research preformed in the fields of statistics education and statistical graphics. The three body chapters stand as the pillars of this work; tied together by the common theme of overcoming challenges and grasping opportunities that are posed by emerging technologies and prodigious data sources. In this first chapter a review of the literature is conducted to lay the foundation upon which the work of the following chapters is built.

We begin by investigating literature on the technological history of statistical education and a review of current uses of technological tools in the undergraduate statistics classroom. Development and application of educational technology in similar STEM disciplines are also explored to identify general pedagogical and design principles for effective implementation of technology in statistics curricula. As a note, Section 1.2 is intended to be submitted for publication to the Journal of Statistics Education (JSE) as a stand alone review of literature on technology in statistics education.

The remaining sections of the literature review then investigate work from fields pertinent to the individual body chapters of this dissertation. Chapter 2 is an educational experiment comparing the learning outcomes from simulation-based and traditional statistical inference curricula. Literature to support this study come from the subjects of simulation-based inference curriculum development, learning assessment, comparative educational studies and experimentalism in education. Chapter 3 studies the development of a `shiny` (RStudio and Inc., 2014) application to connect students to large data; thus literature on developing and evaluating educational technology, as well as existing tools for statistics education and data science, are

reviewed. Chapter 4 is tangential to the work with large data found in the `shiny` application, however it contributes to research on binned scatterplots as a graphical tool for visualizing large data. Pertinent literature for this research include works on binning strategies, statistical graphics, and perceptual psychology. Each body chapter in this dissertation is intended be submitted for publication individually; therefore, the works discussed in this comprehensive literature review will also be found cited throughout the respective chapters.

1.2 Technology in Statistics Education

We will begin with a review of the evolution of technology in the statistics education reformation movement. We then discuss specific implementations of technologies, including data tools, software applications, and hardware, within undergraduate statistics education. Literature on educational technology from the fields of mathematics and physics is also examined to gain broader perspective on general principles for development and application of educational technology.

1.2.1 Role of Technology in the Statistics Education Reform Movement

The role of technology in the undergraduate statistics classroom expanded in the wake of the statistics education reform movement of the mid 1990s. The reform movement was a push to modernize curricula in terms of what statistics courses taught and how they approached the material. Moore (1997) chronicled the pedagogical shift toward constructivism and democratization within the statistics classroom in the 1990s. Introductory statistics was dramatically altered during this movement with data collection, graphical display and a focus on application and interpretation added to the previously theory oriented curriculum. Moore argued that this shift in pedagogy and content needed to be accompanied with a larger focus on using technology in order to maximize student learning.

The statistics education reform movement of the 1990s changed the landscape of statistics education. Teaching practices advocated for by reformers are summarized by a panel of experts in statistics education selected by the American Statistical Association in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE): College Report* (Aliaga et al.,

2005). It should be mentioned that another panel of experts also convened under the GAISE name to compile a list of recommendations for developing a Pre-K-12 curriculum framework for statistics education and hold similar recommendations as those found in the College Report (Franklin et al., 2005). The GAISE College Report provides the following six recommendations for teaching introductory statistics courses:

1. Emphasize statistical literacy and develop statistical thinking
2. Use real data
3. Stress conceptual understanding, rather than mere knowledge of procedures
4. Foster active learning in the classroom
5. Use technology for developing conceptual understanding and analyzing data
6. Use assessment to improve and evaluate student learning

The role of technology in these principles is explicit in item 5 and it can be argued that technology plays a supporting role in all other recommendations. Since the GAISE guidelines were proposed, technology has been increasingly implemented by statistics educators. Hassad surveyed 227 introductory statistics faculty members on their attitudes and use of technology in the classroom in 2005 then again in 2013. He found that technology use in statistics curricula has been increasingly embraced by the majority of statistics faculty. For instance, 76% of faculty report "involving students in using a statistical software program" in 2013 as compared to 50% in 2005 (Hassad, 2013). These works illuminate why the role of technology in the statistic classroom underwent massive changes since the early 1990s, but the question of how those changes unfolded is a complex story.

To evaluate how technology has evolved in recent years, we begin with existing reviews on the subject. Chance et al. (2007) reflect on how the statistics education reform movement impacted the role of technology in the classroom. They evaluated that, as of 2007, technology was being widely used to help with educational administration through course management software. Currently, the most popular learning platforms include Blackboard (Blackboard Inc.,

2014), Moodle (Moodle, 2014) and Desire2Learn (Desire2Learn, 2014) that account for 44%, 23% and 11% market share, respectively (Green, 2013). Chance et al. found that general purpose statistical analysis software, such as *R* (R Core Team, 2013), *SAS* (SAS Institute Inc., 2014), *SPSS* (IBM Corp., 2013a), *STATA* (StataCorp LP, 2013) and *Minitab* (Minitab, Inc., 2010), had gained a large footing within statistics curricula to support the increased focus on applied data analysis. They also found that software and hardware had developed specifically to facilitate the change in content and pedagogy (Chance et al., 2007). Rubin (2007) evaluated the changing trends of technology in statistics education from 1992 to 2007. She illustrated how several software programs were implemented to introduce new data and graphical displays to the statistics classroom. She concluded that although new data sources and user interfaces allow for a more advanced methods for statistics education, problems with student understanding persist, stating "a general caution for readers; as amazing and inspiring as these technologies may seem, none of them can have any educational effect without carefully constructing curriculum and talented teaching" (Rubin, 2007, p.2).

1.2.2 Statistics Education Technologies

We now inspect specific technologies used in statistics education. To reflect the current field of statistics education, this review on specific technological tools narrows in scope to literature and tools produced, or at least updated, within the past decade. Literature is organized into topics of data oriented technologies, educational software, statistical and mathematical software extensions, and hardware used for statistics education.

1.2.2.1 Data Technologies

The second GAISE principle argues that the use of real data for statistics education is superior to using fabricated data because it fosters student engagement and the context emphasizes how and why data is collect and analyzed (Aliaga et al., 2005). This call for using authentic data is backed by Neumann et al. (2013) who conducted open-ended student interviews in a course implementing only real-life data sources. Students were asked for their general thoughts on the use of real data and several themes were recurring in the answers from large percentages

of the students: real life relevance (63%), interest (58%), learning and memory (50%), motivation (37%), engagement (32%), understanding (24%). Finzer et al. (2007) argue that data needs to be engaging and interesting in addition to being real. They state "(w)hat seems to be missing (in introductory statistics courses) are data sets – especially large and highly multivariate data sets – that are ripe for exploration and conjecture driven by students' intrigue, puzzlement and desire for discovery" (Finzer et al., 2007, p.1).

While some educators may have access to real data from their own research projects, many need to seek data from other sources to use in their courses. Online data repositories exist to facilitate this search for data. Curators of data repositories gather data on many subjects then provide data descriptions and easy access. Hundreds of data repositories exist to organize data on various subjects.

There are data repositories curated for statistics education, where data sets are organized by applicable statistical topics. These include Consortium for the Advancement of Undergraduate Statistics Education Online Resources (CAUSE, 2014), Many Eyes (IBM Corp., 2013b), Journal of Statistics Education Data Archives (American Statistical Association, 2014), The Data and Story Library (DASL Project, 1996), and Australasian Data and Story Library (OzDASL) (Smyth, 2011). Massive data archives from other fields of study have been collected and made available through universities, such as University of Michigan's Inter-university Consortium for Political and Social Research (ICPSR, 2014), University of Connecticut's Roper Center Public Opinion Archives (Roper Center, 2014), and the General Social Survey by the National Opinion Research Center at the University of Chicago (NORC, 2014). The multitude of data repositories has led to web directories entirely devoted to organizing links to repositories, such as the Statistical Science Web (Smyth, 2014) and the University of Illinois – Urbana's Applied Technologies for Learning in the Arts & Sciences (ATLAS, 2014). With this plethora of data sources, statistics instructors hardly have an excuse to use boring or fake data examples within classes.

There are other major sources of data for the classroom that are free to access but require more technical skills to retrieve. The Freedom of Information Act was expanded in 1996 to ensures that non-classified data from the United States Government is accessible online (U.S.C.,

1996). The online government resources at www.data.gov/ (U.S. General Services Administration, 2014), www.census.gov/ (U.S. Census Bureau, 2014), www.nhtsa.gov/ (National Highway Transportation Safety Administration, 2015), and www.cdc.gov/ (Center for Disease Control, 2014) are all locations of massive data stores. Other large online databases have also been made available by non-governmental organizations such as public health data from the World Health Organization (WHO, 2014) and poverty data from the World Bank (World Bank, 2014).

Schutes presents a related idea on how to stretch a single data set into individualized data sets for students using a random sub-selection process. Sub-selecting using automation gives all students data with identical context and interpretability, but different subsets of data values. This is ideal for projects where students are asked to individually complete a task but have the freedom to discuss context and approach freely with classmates (Shutes, 2009). This topic could be extended to include the automated generation of solution sets from each individualized data set using the `knitr` package in *R* (Xie, 2013).

The enormity and accessibility of data is a wonderful opportunity for statistics education but Nicholson et al. (2013) argue that it does not improve anything in statistics education unless educators actually use it. Ridgeway et al. (2013) propose use of the "symantic web" as a framework to support the use of large data but seem to fall short of convincingly arguing how this technology would get better multivariate data into classrooms. Finzer et al. (2007) argue that large real-life data often fails to make it into the classroom because educators tend to lack the technical skills and experience for getting those types of data into the classroom. The combination of these opinions seems to indicate that statistics educators must be technologically savvy to seize the opportunities that large data can provide.

1.2.2.2 Educational Software

Many software applications have been developed specifically for statistics education. The following review is not meant to serve as an exhaustive list of all software applications, but is a representation of the tools being employed in statistics education in recent years. Some applications are learning objects focused on teaching statistical methodology. One such tool is *The*

Island, an online data collection simulator which allows students to collect medical records or conduct experiments with virtual residents of a fiction island (Bulmer and Haladyn, 2011). In order to emphasize a realistic portrayal of experimental design and data collection, *The Island* requires students to collect data in a laborious step-by-step manner, instead of providing a simple one-step data pull. Baglin et al. (2013) conducted a student survey about using the tool and found that students perceive *The Island* to be engaging, easy to use and beneficial to their learning. Another software tool, *GeoVista* was implemented to bring Geographic Information System (GIS) data into a statistics course (Forbes, 2012). Other software packages have been developed to help students understand mathematically complex probability concepts. *Tinkerplots* has been developed as a tool for teaching students about the model fitting process and how to connect data collection to the probability concepts used for analysis (Konold and Kazak, 2008). The *Probability Explorer* software allows for students to simulate discrete outcomes from theoretical distributions with the goal of teaching the interplay of ideas that knowing empirical observations inform on how to model the theoretical distributions and knowing the theoretical distribution informs on what to expect from empirical observations (Lee and Lee, 2009).

Online discussion boards and multimedia integration have been implemented widely within online and traditional classrooms. Although not specifically developed for statistics education, literature on the use of these tools within statistics education yields quality insight. Everson and Garfield (2008) implemented a completely online course in introductory statistics with heavy emphasis on discussion boards. They give many administrative pointers on how to ensure that online discussion boards are beneficial to student learning. Schmidt (2013) also discusses how to foster meaningful discussion about statistical concepts based on experience in an introductory biostatistics course. She finds that "(t)he main challenge to implementing these activities is the time commitment required to read and respond to posts... but the opportunities and benefits far outweigh this" (Schmidt, 2013, p.9).

Multimedia tools integrate audio, video, images and reading materials into learning modules to engage students in a comprehensive manner. McDaniel and Green (2012a) have developed `i^3`, an open-source Java applet for teaching distributional, inferential and probabilistic concepts. In another paper, McDaniel and Green (2012b) use pre/post testing to find that using

`i^3` to supplement traditional instruction increases student understanding of sampling distributions. Harraway (2012) researched the pedagogy of combined use of multimedia and the *GenStat for Teaching and Learning (GTL)* software. He argues that the implementation of this combined approach with New Zealand high schoolers was successful with the multimedia instruction and menu driven software allowing students to focus on learning statistical concepts instead of syntax details and programming.

1.2.2.3 Statistical and Mathematical Software Extensions

Research has also been conducted on employing existing statistical and mathematical software for the purposes of statistics education. *R* has been utilized in a number of ways within statistics education; through package development, data analysis and extensions. An example is the `LearnBayes` package (Albert, 2014) developed by Albert (2009) for teaching basic discrete Bayesian modeling. Another option through *R* is for instructors to make web based applications through the `shiny` package that are tailored to their specific course needs (RStudio and Inc., 2014). The *Visual Interactive Statistical Analysis (VISA)* program is built on top of *Excel* (Microsoft Corp., 2013) to give students menu driven functionality; however, the authors are not convincing in arguing that the *VISA* interface adds much pedagogical value over the underlying *Excel* programming (Shaltayev et al., 2010).

Hoff et al. (2012) employ *Mathematica* (Wolfram, 2015) to teach students about the Central Limit Theorem; specifically how the interplay between sample size and population distribution effect the normal approximation of the true sampling distribution for the sample mean. Unfortunately, the implementation within an assignment amounted more to a test of technical computing skill than a learning opportunity. A visual comparison of the normal approximation and the true sampling distribution would provide a more direct way to establish the conceptual understanding. Harlow et al. (2009) designed the *Visual Cognitive Tool (VCT)*, a graphical user interface built upon *MATLAB* (MATLAB, 2014), to teach mathematical statistics. The *VCT* program allows students to visualize complex probabilistic concepts, such as Galton's binomial Quincunx and the trinomial Septcunx. These applications can expose students to an

understanding of analysis software that holds value in a future professional setting, but it is of foremost importance that they hold pedagogical value for learning statistical concepts.

1.2.2.4 Hardware

There is little literature on technological hardware developed specifically for statistical education due to the fact the most technological tools used in the classroom are software. There are however a few notable handheld electronic devices being used within statistics classrooms; graphing calculators and clickers. The graphing calculator is falling out of favor at the university level but is still the only allowed tool for statistical computation for Advanced Placement Exam following Advanced Placement statistics courses in American high schools (The College Board, 2014). Graphing calculators and wireless response "clickers" have also been implemented with some success within large lecture sections of introductory statistics courses as an attempt to drive student engagement and learning through involvement in a data simulation process (Kaplan, 2011). A large scale designed experiment at University of Michigan exposed 48 sections of an introductory statistics course to combinations of treatment by McGowan and Gunderson (2010). They found little evidence that clickers drive engagement but some evidence that they improve learning. They emphasize that the technology alone will not improve student learning unless it is thoughtfully used for a pedagogical reason.

1.2.3 Educational Technologies in Similar STEM Disciplines

To gain perspective on the use of technology in statistics education, we may also look to the development and use of educational technology in other STEM fields. This exploration will consider software and hardware that are either designed for the classroom or professional tools applied within the classroom. For most direct comparison, we discuss two disciplines that strongly resemble statistics: mathematics and physics. Following the specific examples educational technologies within each field is a discussion of general principles of development and application of technology that can extend to statistics education.

1.2.3.1 Educational Technologies in Mathematics

The National Council of Mathematics Teachers periodically issue guidelines for effective mathematics education – akin to the GAISE guidelines in statistics education – that emphasize the importance of integrating technologies in mathematics curricula (NCTM, 2000). The widespread use of electronic technology in mathematics education started in the 1960's with the accessibility of four-function calculators, then followed by scientific calculators (Trouche and Drijvers, 2010). Early implementation of calculator use was limited to the role of automating computation. In the late 1980's and early 1990's handheld technology in the mathematics classroom advanced substantially with the advent of the programmable graphing calculator. Curricula started to be developed with the graphing calculator as part of pedagogy. Burrill et al. (2002) state that "Integrating, not simply adding, the use of handheld graphing technology within the context of the mathematics being studied can help students develop essential understandings about the nature, use, and limits of the tool and promote deeper understanding of the mathematical concepts involved." Kaput and Thompson (1994) use a wave metaphor to characterize the research on technology in mathematics educational from the 1970's to the 1990's; with *wave level* studies quantifying the improvement in computational precision, *swell level* studies exploring the cognitive advantages of learning with technology, and *tidal* changes in mathematical curricula.

The modern mathematics classroom has greatly expanded the type and power of digital technologies in the past decade. The graphing calculator still plays an important role but many other digital teaching tools have emerged, most notable being the plethora of software developed for use on general purpose computing hardware (Trouche and Drijvers, 2010). The many specialized software applications are developed for specific mathematical learning objectives. Examples abound in geometry (e.g. CaRMetal (Hakenholz, 2010), Cinderella (Richter-Gebert and Kortenkamp, 2013), The Geometer's Sketchpad (Jackiw, 2002), and 3-D Applets (Boon, 2009)) and algebra (e.g. Digital Mathematics Environment (Freudenthal Institute, 2014), Maple TA (Waterloo Maple Inc., 2015b), Webwork (Gage et al., 2002) and Wims (Gang, 2015)). Also, many general use mathematics software programs have been increasing incorporated into the

mathematics classroom; such as Mathematica (Wolfram, 2015), Maple (Waterloo Maple Inc., 2015a), Wolfram Alpha (Wolfram Alpha LLC, 2015), Sage (Sage Developement Team, 2015), and MATLAB (MATLAB, 2014). Most recently, widespread ownership of smart-phones has also been harnessed for mathematics education. For instance, the MobileMath game was developed for younger students to learn geometric concepts through a smart-phone GPS orienteering game (Wijers et al., 2010).

With the large scale development of digital technologies for mathematics educations there has been growing research into properly designing and evaluating these tools. Drijvers et al. (2012) argues that the design of technology for education must also include the design of accompanying activities. This philosophy is argued to lead to software that more effectively integrated into curricula (Freiman, 2014). Stohl Drier et al. (2000) provide five guidelines for technology-based activity development: (1) introduce technology in context, (2) address worthwhile mathematics with appropriate pedagogy, (3) take advantage of technology, (4) connect mathematical concepts, and (5) incorporate multiple representations. Research is also being done to formally evaluate existing digital technologies. For instance, Bokhove and Drijvers (2010) evaluate several of the algebra learning objects mentioned above using a set of criterion constructed for digital tools in mathematics education; finding the Digital Mathematics Environment (Freudenthal Institute, 2014) to be the strongest tool based on adherence to quality software and pedagogical design.

1.2.3.2 Educational Technologies in Physics

Flick and Bell (2000) provide general guidelines for using technology in science education – echoing the five principles from Stohl Drier et al. (2000) nearly verbatim – that advise for technology to be introduced within context and leverage appropriate pedagogy to help develop students' understanding of scientific concepts. Rios and Madhavan (2000), and more recently Bryan (2006), have provided overviews on the use of technology within physics education. They both highlight the three primary applications of technology: computer interfacing equipment used to collect and process data, modeling, and graphical simulation. Conducting physics experiments allows students to actively explore the mathematical concepts through col-

lected data. This is often done by digitally collecting measurements with physical sensors (e.g. Calculator-Based Laboratory (Texas Instruments, 2015), and PASPORT Probeware (PASCO, 2015)) or through the use of video analysis software (e.g. Logger Pro (Vernier Software and Technology, 2015), and Tracker (Brown, 2015)). Learning objects for modeling or graphical simulation are often developed as web-based applets for accessibility (e.g. Physics Education Technology (PhET), Interactive Simulation (PhET, 2015) and Physion (Xanthopoulos, 2010)). In his overview, Bryan (2006) concludes that these technologies can aid in student learning and are most successfully implemented when they are developed to adhere to the pedagogical principles and clearly incorporate curriculum concepts.

1.3 Comparison of Statistical Inference Curricula

The research done in Chapter 2 of this dissertation is a designed experiment to compare the learning outcomes of students receiving traditional and simulation-based curricula for teaching statistical inference in an introductory statistics course. The term *traditional* refers to a curriculum that introduces statistical inference through probability distributions and application of mathematical theory in empirical scenarios. This approach is characterized by the use of distributional approximations, cumulative probability tables and formulas for confidence intervals and test statistics. In contrast, the *simulation-based* approach to teaching statistical inference relies on methodology driven by resampling and permutation instead of theoretical probability distributions. This allows for the concepts of statistical inference to be learned through methods such as bootstrap confidence intervals and randomization testing of hypotheses. The following subsections will outline the existing literature on simulation-based inference curricula and their comparison to traditional inference curricula, then explore comparative studies within statistics education and lastly discuss assessment tools designed for measuring student learning outcomes.

1.3.1 Simulation-Based Inference

George Cobb is widely regarded as the progenitor of the recent call to use simulation-based methods for teaching introductory statistics. Cobb argued that calculation and graphics were

automated and streamlined through an expansion in the use of technology during the statistics education reform movement but nothing was done to change how inference concepts were being taught. Computation has become so cheap and widely available that simulation-based methods that rely heavily on repeated simulation are now becoming practically feasible. He also argues that the traditional theory-based inference methods continue being taught, not because they are better, but because historical momentum drives to their perpetuation. He uses the shift from a earth centered to a sun centered view of the solar system as a metaphor for the the shift from tradition to simulation-based methods for teaching inference; it is not only simpler but it is more accurate to the core concepts of inference (Cobb, 2007).

Since 2007 a growing number of educators have joined Cobb in advocating for teaching simulation-based methods in introductory statistics courses. There are a few established groups of authors producing course materials to support the wave of educators adopting the simulation-based approach to inference. *Statistics: Unlocking the Power of Data* is a textbook that follows a traditional progression through data collection, summarization and graphical topics, then uses randomization tests and bootstrap confidence intervals to teach the concepts of statistical inference (Lock et al., 2013). The *Introduction to Statistical Investigation* textbook approaches the restructure of the entire introductory course by using the "spiral approach" to introduce statistical inference from the very beginning of the course, then adds complexity and related ideas as students repeatedly revisit the inference process (Tintle et al., 2014). Although not in a textbook format, the *Change Agents for Teaching and Learning Statistics (CATALST)* group has a large online collection of freely available lesson plans and course materials that use the simulation-based approach to teaching inference (CATALST, 2012).

Simulation-based inference has been gaining popularity among statistics educators in recent years. At the 9^{th} International Conference on Teaching Statistics (ICOTS9) more than a dozen educators presented on their research or experience with simulation-based inference curricula (ICOTS9, 2014). There are only a small number of published studies available on the efficacy of the simulation-based approach. Carver (2011) offers a case study on the use of a curriculum structure that heavily integrated simulation-based teaching, but provided no evidence that the methods were superior to the traditional approach. (Budgett et al., 2013) also attempt to

argue that students' inferential learning is improved by using a simulation-based methods using pre/post test data on student learning outcomes; but again there is no data from students in a traditional curriculum on which comparisons of the curricula can be made.

Quality comparisons of students in traditional and simulation-based inference curricula have been made in two articles by authors associated with the *Introduction to Statistical Investigation* textbook. Both studies measured learning outcomes, through pre/post testing, from groups of students at Hope College who received either a traditional or simulation-based approach to inference during an introductory statistics course. One study found that students of the simulation-based inference had a larger improvement in understanding of hypothesis testing than the traditionally taught students (Tintle et al., 2011). The other study used a follow-up examination four months after the course ended to test memory retention and found that students of the traditional inference curriculum had significantly worse retention of inference concepts (Tintle et al., 2012).

1.3.2 Experiments in Statistics Education

The pair of studies by Tintle et al. present strong preliminary evidence in support of using simulation-based methods for teaching statistical inference. However, these studies suffer from a common problem in educational literature; educational conditions are administered to entire classes of students but comparisons are made – potentially improperly – using measurements from individual students (Ragasa 2008, Baglin and DaCosta 2013, Williams 2012, Carlson and Winquist 2011). Using experimental design terminology, we would consider the class of students in these studies the experimental units and the individual student the observational unit. The issue is that causal inference cannot be made about the treatment effects on the observational units, only on the experimental units. In other words, in the Tintle studies there is no guarantee that the curriculum treatment led to the improvement in understanding and retention scores for students, because the treatment effect is inextricable confounded with all other class related factors.

There are two approaches that can remedy the mismatch of the experimental and observational units; randomly assign multiple classes to each treatments then compare whole class

measurements, or randomly assign multiple students to each treatments then compare individual student measurements. The former requires a large investment of resources because it involves multiple classes of students. This approach was used in a study on the effectiveness of clicker use in an introductory statistics course at University of Michigan where 48 sections – each section containing 25 students – were randomly assigned in a 4X3 full factorial designed experiment and thus had four sections per comparison groups (McGowan and Gunderson, 2010). McGowan (2011) reminds the reader that if this design is used, then random assignment of multiple classes to each comparison group is necessary so that treatment effects are not confounded with group factors.

The alternative, randomly assigning individual students to treatment groups, has not been utilized in statistics education research. This is in large part due to logistical impracticality of randomly allocating students to different treatments in a comparison of educational techniques. Either multiple educators are needed to handle the comparison groups simultaneously, or students must be reassigned to different class times; both options are far from ideal in a university settings. The experimental design for our curriculum study was able to accomplish random assignment of students to curriculum without affecting scheduled class times through an innovative combination of co-teaching and room scheduling.

There are arguments to be made against the use of experiments in educational studies. Cook (2002) argues that experimental logic strongly supports causal conclusions but outlines several practical reasons why he believes they are not used in education. He argues that there are often insurmountable problems with properly implementing experiments in schools, results tend to trade external validity for causal isolation, and that often observational studies may establish strong enough associations to make the necessary policy decisions. There are also those who are more philosophically opposed to experiments in education. Howe (2004) argues that the quantitative methods used in experimentation are not capable of measuring student learning in any meaningful way and that qualitative methods are the only appropriate approach to assess learning. Aside from his astoundingly imprecise and inept criticism of the experimental methodologies, his assertion that we cannot quantify understanding or learning is fundamentally incorrect. Educators – perhaps imperfectly – regularly quantify student knowledge through the

grading of assessments and substantial research has been done on tools designed to measure student learning in statistics.

1.3.3 Assessments for Statistical Inference Learning Outcomes

In the experiment comparing simulation-based and traditional statistics inference curricula we need a valid metric for student learning outcomes. There exist a number of tools developed to measure students learning of introductory statistics concepts, some specific to statistical inference. The *Comprehensive Assessment of Outcomes in a first Statistics course (CAOS)* test is a general purpose test that evaluates student understanding in a wide range of introductory statistics topics (DelMas et al., 2007). The *Reasoning about P-values and Statistical Significance (RPASS)* scale was developed to assess student understanding of statistical concepts surrounding hypothesis testing (Lane-Getaz, 2013). The *Assessment Resource Tools for Improving Statistical Thinking (ARTIST)* project which produced the *CAOS* test also produced several banks of topic specific questions for testing understanding of particular concepts in introductory statistics. These ARTIST question sets were developed through an iterative process using the *Context, Input, Process, Product (CIPP)* evaluation model to improve the quality of learning assessment (Ooms and Garfield, 2008). The intent of our study is to compare the learning outcomes specific to hypothesis testing and confidence intervals, so we elected to use the ARTIST questions sets for these two topics as measures on which to make comparisons of learning outcomes from the statistical inference curricula.

1.4 Development of the Shiny Database Sampler

Chapter 3 of this dissertation details the construction and evaluation stages in the development of the Shiny Database Sampler; a point-and-click web-based software tool for selecting random subsamples from large databases. The Shiny Database Sampler was created to help fill the void in data technologies discussed in Subsection 1.2.2.1, connecting introductory statistics students to available – but difficult to access – large data sources. This section highlights pertinent literature, software, and data sources used in the development of the Shiny Database Sampler.

The goal of the tool is to allow students to treat large databases as populations from which to collect random samples through specified sampling schemes. The software also provides basic tools to numerically and visually assess the sampled data prior to download. `R` (R Core Team, 2013) was chosen as the computational engine to power the Shiny Database Sampler. `R` provides all of the sampling and computational functionality necessary many of the simple data operations, but two additional packages were integral to the construction of the web interface and database querying. The `shiny` package (RStudio and Inc., 2014) allows an `R` session running on a web server to dynamically generate the JavaScript for the user interface accessed via web browser. `shiny` operates on the principle of reactive programming, where certain changes to the user interface (editing fields, clicking buttons, etc.) are programmed to trigger `R` to reprocess data objects which are then incorporated into the JavaScript sent back to the browser for display. This allows for the strength of `R` for statistical computation to be applied on the server side of a point-and-click web interface. This strength is leveraged in the primary task of collecting random subsamples from a `MySQL` database (Oracle Corp., 2014). The accompanying query language to interacting with the databases is not simple; certainly no simple enough to expect proficiency from an introductory statistics student. The `RMySQL` package (James and DebRoy, 2012) allows for database queries to be executed within an `R` session. By coupling the functionality of `shiny` and `RMySQL` we are able to bypass the skill threshold, by allowing the student to set query specifications for random sampling through a simplified interface which automates the query behind the scenes.

The Shiny Database Sampler uses two of the large governmental databases mentioned in subsection 1.2.2.1: the 2001-2009 Fatality Analysis Recording System accident data from the National Highway Traffic Safety Administration (`www.nhtsa.gov/FARS`) and the Public Use Micro Sample data from the 2010 United States Census (`www.census.gov/`). The accidents database contains over one-millions records from fatal vehicle accidents, each included 29 variables pertaining to the accident. The census database has 26 variables from the census records of over three-million U.S. residents. Selecting a subset from a large database can be time intensive if the records are not properly indexed using *keys* (Schwartz et al., 2012a). The records in the accident and census databases were indexed using hierarchical keys in such a way that

prioritizes the speed of simple random sampling, while also greatly improving the efficiency of stratified random sampling.

Designing the user interface incorporated principles from the fields of software engineering, educational psychology and human computer interaction. *Quality characteristics* are used to define and evaluate the design attributes in software engineering. The ISO 9126 are popular set of definitions that include six quality characteristics: functionality, reliability, usability, efficiency, maintainability and portability (Bevan 1997; Berton and Vallencillo 2002). These elements provide design goals and an evaluation structure for general quality software construction. The design also incorporates *cognitive load theory* to improve the strength of the Shiny Database Sampler as a learning tool for sampling concepts. Cognitive load theory argues that learning is optimized by avoiding the expense of mental resources on tasks that are not helping to create new schema or integrate new knowledge into existing schema (Muller et al., 2008). *Human centered design* applies cognitive load theory to the development of educational software; simplifying interface operation to reduce extraneous cognitive load and improve intended learning outcomes (Oviatt, 2006).

After an educational software tool is constructed it is important to evaluate its performance. The ISO 9126 quality characteristics provide a basic framework for assess the software, however the evaluation process needs to then be tailored to the specific learning object. In the initial stages of development *heuristic evaluation* – a non-formal comparison of software to quality characteristics, and *pluralistic walkthroughs* – testing software with statistically proficient users to gain feedback, can be conducted to evaluate the usability of the software (Nielsen, 1994).

A student user survey using Likert scale responses was constructed as a more formal evaluation of the Shiny Database Sampler. The survey items are written to measure three important elements from human centered design: ease of use, connection to course concepts and engagement. Item response theory deals with the difficulties of eliciting latent characteristic using an ordinal rating scale. Ideally the items in the survey will be easy to answer in a way that reflects true belief (low *difficulty*), clearly differentiate examinees with different beliefs (high *discrimination*), and hold high agreement with other items measuring the same latent characteristic (high *reliability*) (DeMars, 2010).

Cronbach's α (often called *coefficient* α), is used to provide a measure the reliability of the items by summarizing the correlation structure between items (Cronbach, 1951). Higher values of Cronbach's α indicate higher agreement between item responses; George and Mallery (2003) and Nunnally and Bernstein (1978) provide commonly used scales for interpreting the reliability based on the value of Cronbach's α. Under Gaussian data with compound symmetry in the covariance structure, the distribution of Cronbach's α can be approximated with an F-distribution (Kistner and Muller, 2004).

It is also important to consider *acquiescence*, or a bias toward positive responses, that occurs in self-report data because of a human bias toward the social desirability of affirmation (Furnham, 1986). Negatively phrasing half of the items for each latent characteristic, then reversing the scoring, can be used in an attempt to counter-balance this bias. However the resulting double negation may create items that contra-intuitive as measures of positive responses (Friborg et al., 2006). The language of each item must be carefully constructed to attempt to navigate the inherent difficulties with eliciting honest and accurate human responses.

Many tools and concepts were leveraged in the development of the Shiny Database Sampler. Quality characteristics from software engineering, cognitive load theory from educational psychology and human centered design from human computer interaction were fundamental in the construction of the software using `R`, `shiny` and `RMySQL`. Item response theory from statistical survey methodology was used in formal evaluation of the Shiny Database Sampler.

1.5 Loss in Binned Scatterplots

The final body chapter of this dissertation shifts from educational tools incorporating large data sources to research on visualizing large data. The research focuses on quantifying the loss of information that occurs in the aggregation necessary for creating binned scatterplots. Literature on big data visualization, scatterplot adaptations, binning algorithms, and perceptual psychology is relevant to this pursuit.

Cleveland (1987) provides an overview of research in statistical graphics, identifying three primary areas of research: methods, computing, and perception. Research in graphical methodology pursues the development of new ways to encode information from data, data summaries,

or model elements into graphical aesthetics. Graphical computation research often investigates ways to optimize plot rendering and creation of user interfaces to support graphical exploration by data analysts. Research in graphical perception is necessary to evaluate the merit of new methods of visualization, because the information encoded in the graphical aesthetics is only valuable if human cognition is able to decode that information perceptually.

Methodology related to bivariate data displays and binned data graphics are relevant to work with binned scatterplots. The scatterplot is a widely used tool for display of bivariate quantitative data, that uses points on a Cartesian plane to visually display the coordinates of each pair. Friendly and Denis (2005) discuss the development of the scatterplot and state, "Indeed, among all the forms of statistical graphics, the humble scatterplot may be considered the most versatile, polymorphic, and generally useful invention in the entire history of statistical graphics." They find that many area plots and time series plots arrived in the eighteenth century (Playfair, 1786), but early forms of the scatterplot did not emerge until the nineteenth century (Herschel, 1833).

The scatterplot is very effective at visualizing the bivariate distribution of data pairs, but it often fails to overcome a fundamental challenge; scalability (Theus, 2006a). While the points do not have any area mathematically, they are visually displayed with round glyphs that occupy space on the graphic. *Over-plotting* occurs in scatterplots when two or more points share overlapping graphical space, thus obscuring part of the visual information. This problem becomes more pronounced as the number of points increases, substantially lowering the utility of the traditional scatterplot in ever growing modern data sizes. Many adaptations for the scatterplot have been suggested to overcome this problem, most either use changes to the glyphs representing each point (Tukey 1977,Few 2008, Keim et al. 2010, Hao et al. 2010, Janetzko et al. 2013) or through plotting binned aggregations of points (Cleveland and McGill 1984b, Carr et al. 1987, Wickham 2013, Liu et al. 2013).

Adaptations that alter the individual points may help in cases of mild over-plotting, but plots that employ binning have superior scalability for visualizing truly massive data. Liu et al. (2013) state that "Visualizing every data point can lead to over-plotting and may overwhelm users' perceptual and cognitive capacities. On the other hand, reducing the data through

sampling or filtering can elide interesting structures or outliers." The binned scatterplot is an aggregation based plot that is the two-dimensional analog to a histogram, which bins bivariate data and displays frequency through the shade of geometric tiles, as opposed to the height of a bar. Binned scatterplots are constructed by defining a two-dimensional tessellated grid over the range of the data, binning points based on the grid boundaries, then shading matching geometric tiles based on the frequency of points in the corresponding bin. Rectangular binned scatterplots are often referred to as *heatmaps*, which can be traced back to the end of the turn of the twentieth century (Wilkinson et al., 2000).

Specifying the binning algorithm is fundamental to the construction of a binned scatterplot. Binning algorithms in statistical graphics date back to Pearson (1895) who used univariate binning in the process of visualizing the binomial approximation to the normal distribution. Carr et al. (1987) introduced binned scatterplots using a hexagonal binning grid. Scott (1992) uses *mean integrated squared error* (MISE) – a loss measure of difference between points and bin centers – to evaluate bivariate binning strategies. He concluding that rectangular binning and hexagonal binning hold similar MISE, both superior to triangular binning. Wickham (2013) argues for fixed width rectangular binning for computational speed and ease.

Scott (1992) explains that the histogram convey frequency and relative frequency which are the essence of a density function. Similarly, the binned scatterplot can be thought of a visual estimate of the bivariate density. The binning and rendering of the plot encode the density information into the tiles through shading, therefore research on perception of color is important for understanding how that information is recovered visually. Several authors have tested graphical perception of colors in designed experiments. Cleveland and McGill (1984a) find that shading and color saturation are less accurate graphical aesthetics than position, size and angle. In repeating Cleveland and McGill's experiment using updated technology, Heer and Bostock (2010) similarly rate color as less precise for representing quantitative information than position, size and angle. Demiralp et al. (2014) find that people tend to judge color similarity primarily based on hue and secondarily based on shade. Healey and Enns (1999a) present evidence to suggest that up to seven distinct hues are simultaneously distinguishable for the average person, but there is more difficulty discerning different shades of the same

hue. This all points to the difficulty that is faced in extracting frequency information based on the tile shades in a binned scatterplot, therefore special consideration is paid in Chapter 4 to mapping frequency information to shade.

1.6 Literature Themes

The review of literature identifies important themes, knowledge, and wisdom to be incorporated in the research that follows. The literature on educational technologies has a distinct take-away for work in statistics education. The recursive message resonates clearly; technology presents the opportunity to create a rich learning environment and engage students on a deeper level, however the development and implementation of technology in education requires proper pedagogy and integration into curricula to effectively improve learning. This theme echoes throughout education literature from statistics, and also in the work of our peers in similar STEM disciplines. Great care must be taken when developing or implementing new educational technologies to ensure that students attentions are guided to learning new concepts, not distracted by the technology. These principles lend well to the pursuit of work on simulation-based learning and the construction of the Shiny Database Sampler. In statistical graphics, the literature commonly reminds us that we not only need to develop efficient ways to visualize information, but we also must critically assess the perceptual ability to accurately recover that information.

CHAPTER 2. COMPARISON OF LEARNING OUTCOMES FOR SIMULATION-BASED AND TRADITIONAL INFERENCE CURRICULA IN A DESIGNED EDUCATIONAL EXPERIMENT

Status: Submitted for Review to *Technology Innovations in Statistics Education* (TISE)

Authors

Karsten Maurer, Iowa State University, Primary Author

Dennis Lock, Iowa State University

Abstract

Conducting inference is a cornerstone upon which the practice of statistics is based. As such, a large portion of most introductory statistics courses is focused on teaching the fundamentals of statistical inference. The goal of this study is to make a formal comparison of learning outcomes under the traditional and simulation-based inference curricula. A randomized experiment was conducted to administer the two curricula to students in an introductory statistics course. The results indicate that students receiving the simulation-based curriculum have significantly higher learning outcomes for confidence interval related topics. While the results are not comprehensive in assessing the effect on all facets of learning, they indicate that learning outcomes for core concepts of statistical inference can be significantly improved with the simulation-based approach.

2.1 Introduction

Conducting inference is a cornerstone upon which the practice of statistics is based. As such, a large portion of most introductory statistics courses is focused on teaching the fundamentals

of statistical inference. In recent years the approach by which to teach inference in introductory statistics courses has been the topic of growing discussion. The traditional approach to inference curriculum is focused on distributional theory-based methodology, often characterized by use of distributional assumptions, formulas and tables. A modern alternative is a simulation-based approach to the inference curriculum. The simulation-based approach utilizes tactile and computational simulation to run inferential techniques such as bootstrapping for confidence intervals and simulation-based hypothesis testing. Many proponents of the simulation-based inference curriculum argue that this allows students to be exposed to the core concepts of the inference without first requiring the understanding of theoretical probability distributions.

The focus of the following study is to make a formal comparison of learning outcomes under the traditional and simulation-based inference curricula. The learning outcomes for concepts surrounding inference with confidence intervals and hypothesis testing are of primary interest. A randomized experiment was conducted to administer the two curricula to students in an introductory statistics course. The experimental design allows for causal inference to be drawn about the effect of curriculum type on the learning outcomes. The results indicate that significant improvement in learning outcomes for confidence interval related topics are achieved using the simulation-based teaching methods.

2.2 Literature Review

With the goal to make proper comparison of traditional versus simulation-based curricula for introductory statistical courses, we must first view where each approach stands within the constant evolution of statistics education. Using the term "traditional" to describe the current standard for introductory statistics course curriculum is relative to only the last two decades. Moore chronicled the reform movement of statistic education of the 1980's and 1990's as a period of drastic change in the introductory statistics classroom. The curriculum expanded greatly from a course dominated by theory-based inference methodology to the inclusion of the topics of data exploration, data production, model diagnostics and simulation. The content change indicated a shifting emphasis toward conceptual understanding and applied statistics. Moore also stated, "(w)hat is striking about the current reform movement is not only its

momentum but the fact that it centers on pedagogy as much as content" (Moore, 1997). The pedagogical push toward active learning was combined with content change and the increasing use of technology to form what may be referred to now as the traditional introductory statistics curriculum.

The tenets of the statistics education reform movement were formalized in the Guidelines for Assessment and Instruction in Statistics Education (GAISE) reports for pre-K-12 (Franklin et al., 2005) and introductory college courses (Aliaga et al., 2005). Six recommendations were made in the executive summary of the GAISE college report: emphasize statistical literacy and thinking, use real data, stress conceptual over procedural understanding, foster active learning, use technology for both learning and analysis, and use assessment as part of the learning process. In the past decade these principles have been widely adopted in statistics education with a noteworthy increase in technological integration. Technology in the statistics classroom now regularly takes the form of applets, graphing calculators, multimedia materials, and educational, analytical and graphical software (Chance et al. 2007; Rubin 2007). Technological proliferation in the statistics classroom came as a result of technologically receptive statistics educators taking advantage of computation that has become cheaper and more accessible. A large survey of introductory statistics instructors found that 76% of the instructors usually or always require students to use a computer program to explore and analyze data, and 90% of the instructors report a high level of comfort using computer applications to teach introductory statistics (Hassad, 2013).

Amidst the drastic increase in the use of technology in introductory statistics education there has been a growing group of educators who believe that the curriculum reform has stopped short of the possibilities that computation can provide. Cobb argues that statistics education has done well to adopt technology to displace tedious calculation but has not effectively changed the approach to teaching inference. Cobb strongly articulates a call for statistics instructors to use simulation-based methods for teaching inference to replace the traditional approach to inference using theory-based methodology. He states, "(o)ur curriculum is needlessly complicated because we put the normal distribution... at the center of our curriculum, instead of the core logic of inference at the center" (Cobb, 2007). If we view the introductory

statistics course as a constrained optimization problem with statistical literacy and conceptual understanding of inference as the items to maximize, then removing the burden of learning the normal distribution will present the opportunity for more time spent learning core concepts (Carver, 2011). In recent years, curricula for using a simulation-based approach to inference have been developed by a number of groups of statistics educators (Tintle et al. 2014; Lock et al. 2013; CATALST 2012; Carver 2011).

There has been research done on the efficacy of simulation-based inference curricula; however, due to the recency of the curricula development most of this preliminary research has been observational. Budgett, Pfannkuch, Regan & Wild conduct a case study on a small group of students receiving a simulation-based curriculum and found significant learning gains using pre and post testing based on the Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS). This study does not however attempt to make a comparison between the simulation-based approach and traditional approach to teaching inference (Budgett et al., 2013). Another pair of studies make comparisons on both learning outcomes and learning retention between the two types of curricula. Tintle, VanderStoep, Holmes, Quisenberry and Swanson found weak evidence for an overall improvement in learning outcomes and significant improvements within the topic of hypothesis testing for the cohort of students receiving the simulation-based curriculum, but the lack of random assignment of student to cohort obstructs the ability to draw any causal conclusions (Tintle et al., 2011). Tintle, Topliff, VanderStoep, Holmes and Swanson then found significant evidence for improvements to learning outcome retention after four months for students receiving the simulation-based inference curriculum, but again self-selection of students to cohort prevents establishing a causal link (Tintle et al., 2012).

The preliminary research shows promising results for the simulation-based approach to teaching statistical inference. A more rigorous experimental approach to comparing the traditional and simulation-based curricula has been taken in this study in order to establish a causal effect of curriculum on learning outcomes. Section 3 explains the structure and methodology implemented in the educational experiment and the measurement of student learning. Section 4 details the model based approach for assessing the effect of curriculum on specific learning

outcomes. Lastly, we discuss the study findings and explore the implications for designing future introductory statistics curricula.

2.3 Methodology

The subjects for this study were students enrolled in two sections of the Introduction to Statistics, Stat 104, course at Iowa State University in the spring semester of 2014. Stat 104 is an introductory statistics course tailored for students in the agricultural and biological sciences. Of the 112 students to complete the course, 101 students consented to the release of their course data for the purposes of this study. The students who did not consent were treated identically to those who consented, but their data was omitted from the analysis that follows. Students from both sections were randomly assigned to one of the two inference curriculum treatments, creating cohorts A,C and B,D, respectively. Cohorts A and B were exposed to the simulation-based curriculum; while the cohorts C and D were exposed to the traditional curriculum. Student cohorts were the basic units to which room assignments, instruction and curriculum treatments were applied.

The course was administered by the authors in a co-teaching setting for students from all cohorts. The course schedule involved two hours of lecture and two hours of lab per week. The co-teaching strategy was employed as an intentional attribute of the experimental design. The following subsections will detail the curriculum outline for each cohort of students, the experimental design for administering the curricula using the strengths of the co-teaching setup and the data collected for analysis.

2.3.1 Curricula Structures

To compare the learning outcomes for students receiving the traditional and simulation-based inference curricula we first needed to prepare a curriculum for each approach. Both curricula needed to satisfy the course guidelines set by the Department of Statistics at Iowa State University, covering the following topics: univariate and bivariate descriptive statistics, linear regression, experimental design, basic probability rules, the binomial distribution, the normal distribution, sampling distributions, and inference on means and proportions. Each

curriculum was composed of lecture, corresponding lecture notes, weekly lab assignments designed for groups of four to five students, weekly homework assignments, a midterm exam and a cumulative final exam. The curricula materials for the course did not require the use of a textbook, however specific textbooks that roughly follow the structure of each curriculum were recommended as supplementary study materials (Agresti and Franklin 2012; Lock et al. 2013).

Figure 2.1 outlines how these topics were structured within a weekly schedule for the sixteen week semester for each curriculum. Note that students from all cohorts were exposed to an identical curriculum for all non-inference related topics in the course. This includes identical lecture, course notes, homework assignments, lab assignments and midterm exam during the first half of the semester.

Starting at week 9 the curricula diverge into their respective approaches to inference. Cohorts A and B began the simulation-based inference curriculum in week 9 by first learning the concepts of sampling distributions then used computer simulation and sampling variability as a basis for exploring inference using bootstrap confidence intervals and simulation-based tests. To be specific, confidence intervals were constructed by estimating the standard error using the standard deviation of the bootstrap distribution, not through percentile-based bootstrap methods. Lectures, homework and labs for these cohorts utilized the StatKey software package (Lock et al., 2013) to conduct the simulation-based inference. The simulation-based curriculum then covered normal distributions and how they could be used to conduct inference on means and proportion. While many advocates for simulation-based methods may argue that the normal distribution should be pushed to a second course in statistics, course guidelines required that all students of this introductory statistics course be taught theory-based inference methodology.

Cohorts C and D progressed through the traditional approach by first learning the normal distribution and use of the normal tables. They were then introduced to applications of the normal approximation within inference. The traditional curriculum utilized simulation to display concepts, but only to the extent of demonstrating that sampling distributions can be approximated by normal distributions under certain conditions.

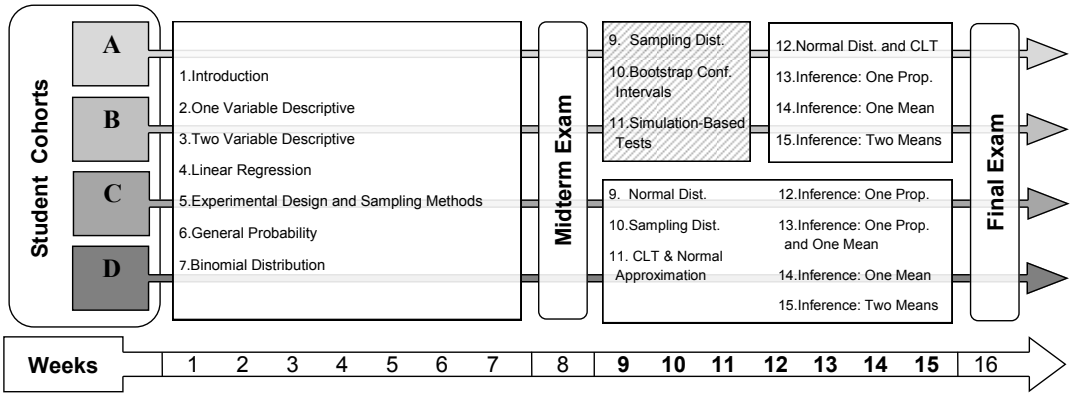


Figure 2.1: Curricula schedules.

During the second half of the semester the lectures, course note, homework and lab assignments differed between the two curricula. However, homework and lab assignments were kept similar when they covered similar topics. For example, all cohorts covered the topic of sampling distributions so the lab assignments were nearly identical between the two groups with the exception of a question pertaining to the normal approximation included for the traditional cohorts. By the end of the semester all cohorts covered how to conduct inference using normal theory; however cohorts A and B additionally learned the core concepts of inference using simulation-based methods prior to learning traditional theory-based inference methods.

2.3.2 Experimental Design

The logistics of administering a course with two distinct curricula and four cohorts of students required a well-structured design and creative scheduling on several fronts. The primary objectives for the experimental design were to eliminate differences in non-inference related curriculum administration to the extent possible, remove the confounding instructor effect on each curriculum and to mitigate the effect of unknown lurking variables through random assignment of students to curricula.

Students were randomly assigned to cohorts during the first week of the course. Of the 101 students who completed the course and consented to the release of their data there were 50 students in the traditional treatment group and 51 students in the simulation-based treatment group. It is also worthwhile to note that of the 4 students to drop the course, all did so prior

to week 9; thus, we can safely assume that the inference curriculum treatment did not play a role in the drop. All students who began the inference curricula completed the course.

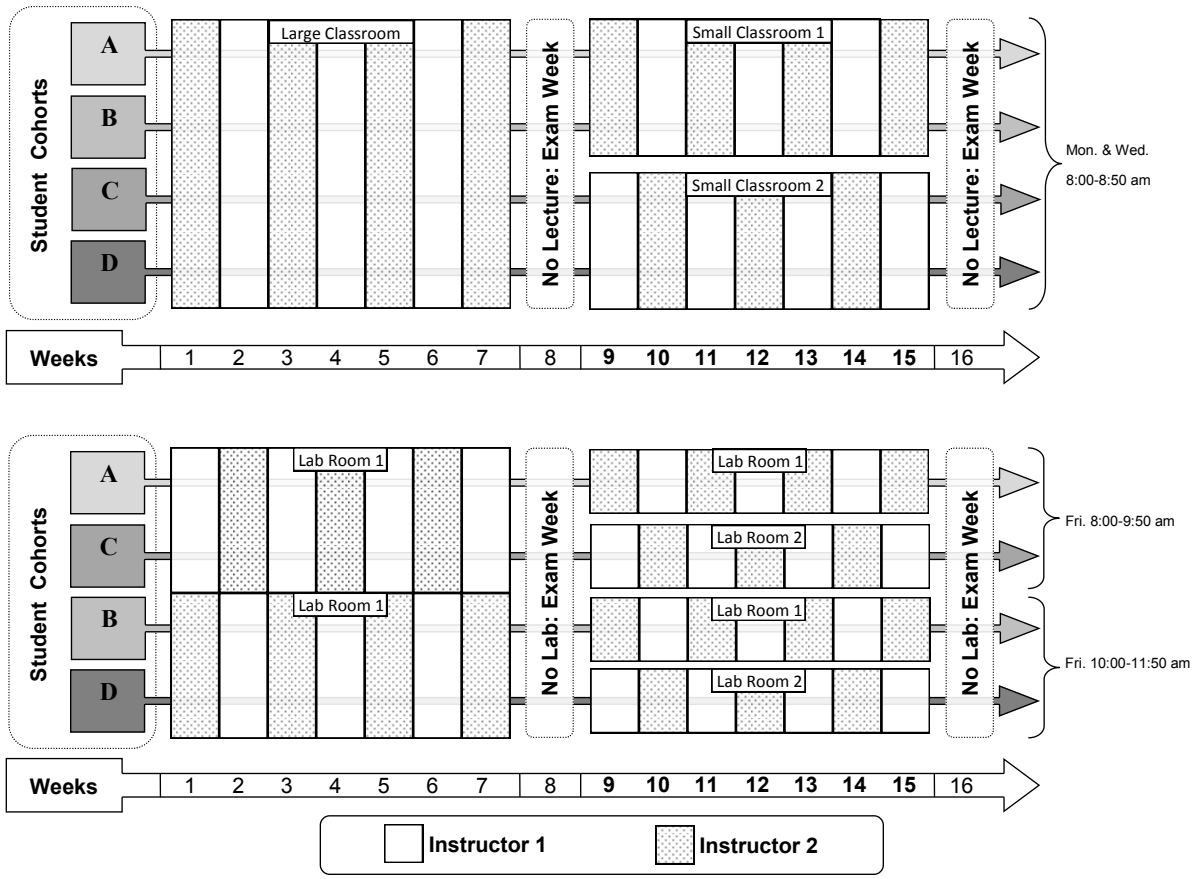


Figure 2.2: Instructor and room schedules.

Students were exposed to identical lecture and lab instruction for the first half of the semester and then diverge into two separate lecture and lab settings for the second half of the semester. This was done to make the experience as similar as possible such that both treatment groups would have the same exposure to terminology and ideas leading up to the inference topics. We could not reassign students to lecture and lab times different than the times for which they enrolled, which meant the logistics of the design required preemptive room scheduling and course time scheduling preparations. By working with the department chair and course coordinator before students enrolled into sections, we were able to schedule two sections of the course to have identical lecture times but separate lab times. Special room scheduling was required because all students needed to attend the same lecture and lab rooms for the first

half of the semester then split into separate lecture and lab rooms after the midterm. This room and course time scheduling allowed for students to be divided into cohorts and attend the lecture or lab specific to their curriculum. The lecture and lab room schedules for each cohort are displayed in Figure 2.2.

Assigning one instructor to each curriculum would confound the instructor effect and the curriculum effect. To avoid confounding, each treatment group would need to receive instruction from both instructors. An alternating weekly schedule was decided upon to spread out the instructor effects over both curricula. A coin was flipped to decide how to match the instructor to the curriculum when the alternation was initialized. The lecture and lab instruction schedules for each cohort can also be found below in Figure 2.2. Note that each figure has student cohort and times fixed across all weeks, reflecting the unchanged time structure that each student enrolled into. The instructors and room locations are what changed throughout the course.

2.3.3 Data Collection

In order to measure learning outcomes for specific inference concepts we utilized question sets from the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) for the topics of confidence intervals and hypothesis testing (ARTIST, 2006). The ARTIST scaled question sets each consist of 10 multiple choice questions that are geared toward critical thinking about the inference topic. These questions were administered as part of the written final exam for all students on the same day and time. The ARTIST scaled scores for the topics of confidence interval and hypothesis testing were recorded for each student. The multiple choice questions for the ARTIST scaled topics can be found in Appendix A.1.

The final exam also included two problems that tested the student's ability to conduct statistical inference in an applied setting using theory-based methodology. Each problem was based on a hypothetical scenario where data has been collected and inference needed to be conducted using the traditional approach using formulas and tables. The first problem provided data summaries and students needed to construct and interpret a confidence interval for a single population mean. The second problem required students to conduct a hypothesis test for a single proportion based on another set of data summaries. The applied inference problem scores

for each student are not used for the primary analysis on learning outcomes but are included for an interesting peripheral analysis on student ability to conduct inference using traditional theory-based methods. The applied inference problems and grading rubrics can also be found in Appendix A.1. The exams were graded blindly, with no identifying information of the student or treatment visible during the grading process.

In addition to the ARTIST and applied inference question scores, data were collected from the first eight weeks of the course – prior to student exposure to an inference curriculum. We have scores from homework assignments 1 to 7, lab assignments 1 to 7 and the midterm exam for each student. The midterm exam questions and grading rubrics can be found in Appendix A.2. Since all of these items were administered and graded equivalently for all students before being assigned to a curricula, the scores from these weeks will be referred to as the "pre-treatment measurements". Lastly, the data include the cohort to which each student belonged.

The research proposal approved by the Institutional Review Board specified that students' data would be entirely deidentified following the course, including all demographic information. At the conclusion of the semester the data for the 101 students who consented to the release of their data were saved, with names and identity information removed, to a spreadsheet. The deidentified student data was imported to `R` for the analysis described in Section 2.4 below.

2.3.4 Data Summary

The ARTIST and applied inference question scores from the final exam are the response variables on which we wish to compare the groups of students from the two inference curricula. Figure 2.3 displays the histogram, mean and standard deviation for each response variable, separated by curricula. Midterm exam scores as also included in order to provide a comparison of the curricula groups using a pre-treatment measurement.

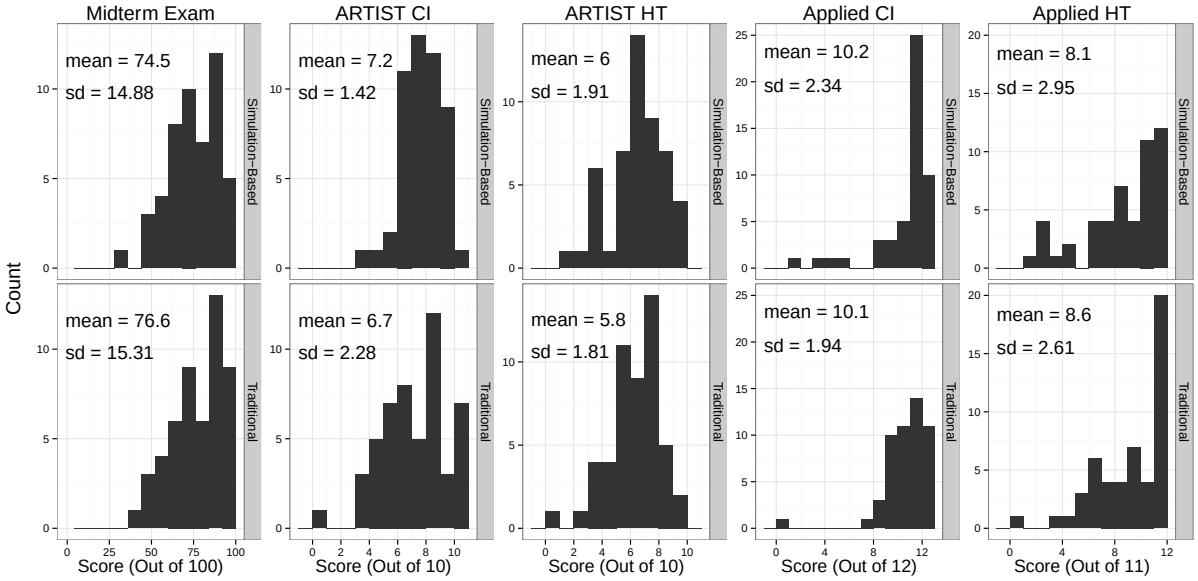


Figure 2.3: Histograms and summary statistics of scores by curricula group.

In Figure 2.3, we see that the midterm exam scores are very similarly distributed for each group; with the traditional curriculum group scoring only slightly higher on average than the simulation-based curriculum group. This similarity is expected – and desirable – because the midterm was conducted prior to the treatment being administered, and the class materials and instruction were designed to be identical at that stage of the course.

Comparing the distributions in Figure 2.3 we see that the simulation-based inference group had a higher average score than the traditional inference group on both of the ARTIST scaled question sets and on the applied confidence interval problem, but scored lower on average on the applied hypothesis testing problem. The simulation-based inference group had lower variability than the traditional inference group on the ARTIST question set for confidence intervals, but higher variability on all other scores. These data summaries are suggestive of differences in the inference learning outcomes of the two groups. In Section 2.4, we take a model-based approach to assess if these differences are statistically significant.

2.4 Analysis

The primary goal of the analysis is to investigate if there is a curricula effect on inference concept learning outcomes. Our data includes ARTIST scaled topic scores for confidence intervals and hypothesis tests which we use as the responses for the comparison of curricula. A model based approach is used to assess curricula effect while controlling for pre-treatment differences between students. With the two dimensional response and an assortment of covariates we employ a multivariate analysis of covariance (MANCOVA) model.

Both curricula groups were required to learn how to conduct normal-based inference. This leads to another question of interest. Does the added simulation-based material turn out to be detrimental to student's ability to use distributional theory-based methods to conduct inference? Two applied problems were included on the final exam that required students to use theory-based methods and formulas to conduct inference. These applied questions were used as the responses in a separate MANCOVA model to check for a curriculum effect. The bivariate MANCOVA models used for these two analyses are parameterized as

$$y_{i\ell} = \tau_\ell \mathbb{1}_{\{i \in T\}} + \beta_{\ell 0} + \sum_{p=1}^{P} x_{ip}\beta_{\ell p} + \epsilon_{i\ell}, \tag{2.1}$$

where

$y_{i\ell}$ is the ℓ^{th} response ($\ell \in \{1, 2\}$) from student i, $1 \leq i \leq n$,

τ_ℓ is the treatment effect of the simulation-based curriculum on response ℓ, and

$\mathbb{1}_{\{i \in T\}}$ is the indicator function for student i in the treatment group.

$\beta_{\ell 0}$ is the common intercept for response ℓ, and

$\beta_{\ell p}$, $1 \leq p \leq P$ are the model coefficients of the P covariates.

x_{ip} is the p^{th} pre-treatment covariate score of student i, and

$\epsilon_{i\ell}$ is the error for the ℓ^{th} response from the i^{th} student.

We assume that error pairs are independent and identically distributed:

$$\vec{\epsilon_i} = \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{bmatrix} \overset{iid}{\sim} \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix} \right)$$

With this parameterization, it is clear that the underlying structure of the MANCOVA model is a multivariate multiple linear regression that can include categorical and continuous covariates. Note that the paired error terms from each student are correlated but are specified as independent between students. The assumption of independence between student response scores is understood to be unrealistic for students from the same class. The repercussions of violating the assumption of independence between student responses will be explored through a simulation study in Section 2.4.3 following the analysis.

2.4.1 Modeling ARTIST Outcomes

We begin with the model for the ARTIST scaled topic scores. Many of the pre-treatment variables are highly correlated. To select a model with only the most predictive pre-treatment covariates, model selection was conducted by first running backward selection based on AIC then removing further covariates that posed collinearity issues. The model selected for final analysis included three covariates: an indicator variable for the curriculum treatment group, the lab 5 score and the midterm score. The midterm tested students on materials from weeks 1-7 and lab 5 assessed understanding of topics related to random selection techniques. We will refer to this selected model as the "ARTIST Model". Model fit for the ARTIST Model was assessed to be satisfactory; see Appendix A.3.1 for residual plots and other model diagnostics.

To test for overall covariate effects on the multivariate responses we use Pillai's Λ. Table 2.1 shows a weak overall effect of the curriculum treatment on the paired ARTIST scaled topic scores. This prompts us to investigate the treatment effect on the ARTIST scaled topic scores for confidence intervals and hypothesis tests separately, to see if the weak overall effect is driven by a significant effect on one of the two scores.

Table 2.1: Tests for bivariate effects on ARTIST question scores using Pillai's Λ.

	Pillai's Λ	Approx. F	Pr(>|F|)
Midterm	0.2109	12.8277	0.0000
Lab 5	0.0792	4.1261	0.0191
Treatment	0.0469	2.3605	0.0998

To investigate the effect of the curriculum treatment on each ARTIST scaled topic score, we analyze the two underlying linear models that comprise the overall MANCOVA model. Table 2.2 displays the coefficients of the linear model fit to the ARTIST scaled score for confidence interval learning outcomes along with covariate ranges to provide context to coefficient magnitudes. It should be noted that although the midterm and lab scores were recorded discretely, they were treated as continuous covariates when fitting the model. We find that midterm, lab 5 and the curriculum treatment effect are significant. Specifically, the simulation-based inference group scored significantly higher by 0.7149 out of a possible 10 points, a 7.146% improvement in confidence interval learning outcomes on the ARTIST scale while controlling for midterm and lab 5 scores.

Table 2.2: Coefficients for model fit to ARTIST score for confidence interval topic.

	Covariate Values	Estimate	95% Confidence Interval
Intercept	1	1.4648	(-0.5467 , 3.4763)
Midterm	{0,1,...100}	0.0477	(0.0253 , 0.0701)
Lab 5	{0,1,...100}	0.0183	(0.0057 , 0.0309)
Treatment	{0,1}	0.7146	(0.0435 , 1.3858)

Table 2.3 displays the coefficients of the linear model fit to the ARTIST scaled score for hypothesis test learning outcomes. We find that only the midterm score is significant for predicting learning outcomes for hypothesis testing. There was no significant curriculum treatment effect.

Table 2.3: Coefficients for model fit to ARTIST score for hypothesis test topic.

	Covariate Values	Estimate	95% Confidence Interval
Intercept	1	2.1053	(0.0046 , 4.2060)
Midterm	{0,1,...100}	0.0386	(0.0152 , 0.0620)
Lab 5	{0,1,...100}	0.0085	(-0.0046 , 0.0217)
Treatment	{0,1}	0.3050	(-0.3960 , 1.0059)

A final consideration in the comparison of learning outcomes using the ARTIST model is that we made two primary comparisons; the curriculum effect on each of the inference topics.

While several multiple comparisons adjustments have been developed for univariate response modeling, the Bonferroni method is the only traditional adjustment that is flexible enough for use in the MANCOVA setting. With the Bonferroni adjustment, if we wish to maintain an overall $\alpha = 0.05$ significance level then we hold each individual comparison to the $\alpha = 0.025$ level. After using the Bonferroni adjustment, the curriculum effect on learning outcomes for confidence interval topics would no longer considered significant (p-value $= 0.031 > 0.025$) at the overall $\alpha = 0.05$ level, but instead would be significant at the overall $\alpha = 0.1$ level. However, the Bonferroni method is well known for being overly conservative in its adjustment, and we are comfortable with maintaining the original interpretations.

2.4.2 Modeling Applied Theory-Based Inference Scores

As with the MANCOVA model for ARTIST scaled question scores, we consider all pre-treatment measurements in a new model for the two applied theory-based inference question scores. Backward stepwise selection was used to obtain a reduced MANCOVA model in a model selection process identical to that described in Subsection 2.4.1. We will refer to the selected model here as the "Applied Model". Residual plots and other model diagnostics for the Applied Model may be found in Appendix A.3.2.

Table 2.4 shows – based the Pillai's Λ – that there was no overall effect of curriculum treatment on the scores for the pair of applied inference problems. This is of particular interest because students receiving the simulation-based curriculum had three weeks less of coursework involving the use of normal distributions and normal tables. This implies that despite the increased complexity of the simulation-based material and the shortened exposure to theory-based inference concepts, there was no significant detriment to students' performance in conducting inference using theory-based methods. It should be noted that the applied questions from the final exam were written by the authors and have not been assessed as reliable metrics for learning outcomes. Thus, the results are reported as supplementary to the discussion on learning outcomes measured by the ARTIST scaled topics.

Table 2.4: Tests for bivariate effects on Applied question scores using Pillai's Λ.

	Pillai's Λ	Approx. F	Pr(>|F|)
Midterm	0.3238	22.9881	0.0000
Homework 2	0.0870	4.5750	0.0127
Treatment	0.0108	0.5217	0.5952

2.4.3 Model Assessment

The bivariate MANCOVA models that were employed in Sections 2.4.1 and 2.4.2 make the assumption of independent errors between students; an assumption which is very likely violated in practice because learning outcomes for students attending the same lectures and labs are very likely related. The assumption of independence is typically made for convenience and the lack of repetition on the lecture and lab levels prevents us from estimating a proper variance structure. It is therefore important to assess the consequences of this violation on the fitted MANCOVA model; specifically, the impact on the Type 1 error rates in tests for curriculum effects on learning outcomes. We elect to explore the consequences through a simulation study wherein the assumption of independence is knowingly violated and the effects on errors rates can be recorded.

We choose the ARTIST Model (2.1) from Sections 2.4.1 as the basis for our simulation study. Under the assumption of independence between students, we found weak evidence of a curriculum effect on the bivariate ARTIST learning outcomes using Pillai's test. Further inspection of the individual responses using t-tests revealed significant evidence of a curriculum effect on confidence interval learning outcomes, but no evidence of a curriculum effect on hypothesis test learning outcomes. In order to assess the trustworthiness of the results, we must first know how a violation of the assumption of independence impacts the Type 1 error rates for these tests. We simulate responses from a generative model without a curriculum effect on learning outcomes (i.e. $\tau_1 = \tau_2 = 0$) and where the assumption of independence between students is violated to a known degree. We then track the Type 1 error rate in curriculum effects when the ARTIST model, assuming independence, is fit to the simulated responses.

The generative model for the simulations is adapted from the MANCOVA model (2.1) by adding random effects to violate the independence between students. The generative model

includes fixed effects for lab 5 and midterm scores, and random effects for responses to each ARTIST topic, lab sections and lecture section. Recall that responses are nested within students, students are nested within lab sections, and labs are nested within lecture sections. Thus the generative model is defined as,

$$y_{ijk\ell} = [\beta_{0k} + x_{ijk1}\beta_{k1} + x_{ijk2}\beta_{k2}] + [\eta_\ell + \gamma_{k\ell} + \delta_{jk\ell} + \epsilon_{ijk\ell}]\,, \tag{2.2}$$

for $i \in \{1, \ldots, n\}$, $j \in \{1, 2, 3, 4\}$, $k \in \{1, 2\}$, and $\ell \in \{1, 2\}$. Where y_{ijkl} is the ℓ^{th} response from student i who is in lab section j and lecture section k. The fixed effects portion of the model, within the first square brackets, is defined identically to the original ARTIST model (2.1). In the generative model the β coefficients are set equal to the corresponding coefficient estimates from the original ARTIST model. The remaining terms, η_ℓ, $\gamma_{k\ell}$, $\delta_{jk\ell}$, and $\epsilon_{ijk\ell}$ are random effects for responses, lecture sections, lab sections and individual errors, respectively, distributed as follows

$$\vec{\epsilon}_{ijk} = \begin{bmatrix} \epsilon_{ijk1} \\ \epsilon_{ijk2} \end{bmatrix} \overset{iid}{\sim} \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix} \right),$$

$$\vec{\delta}_{jk} = \begin{bmatrix} \delta_{jk1} \\ \delta_{jk2} \end{bmatrix} \overset{iid}{\sim} \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, d\Sigma = \begin{bmatrix} d\sigma_{11}^2 & d\sigma_{12}^2 \\ d\sigma_{21}^2 & d\sigma_{22}^2 \end{bmatrix} \right),$$

$$\vec{\gamma}_{k} = \begin{bmatrix} \gamma_{k1} \\ \gamma_{k2} \end{bmatrix} \overset{iid}{\sim} \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, g\Sigma = \begin{bmatrix} g\sigma_{11}^2 & g\sigma_{12}^2 \\ g\sigma_{21}^2 & g\sigma_{22}^2 \end{bmatrix} \right),$$

$$\vec{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \overset{iid}{\sim} \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, z\Sigma = \begin{bmatrix} z\sigma_{11}^2 & z\sigma_{12}^2 \\ z\sigma_{21}^2 & z\sigma_{22}^2 \end{bmatrix} \right).$$

Thus the variance structure for the responses includes common variance components $\sigma_{\ell\ell'}$ and scaling parameters z, g and d. The variance components for $\sigma_{\ell\ell'}$ from the original ARTIST model are plugged in as variance parameters for the generative model. The scaling parameter z controls the variability that is common between all student responses; thus $z\sigma_{\ell\ell'}$ is the covariance of responses for ARTIST sets ℓ and ℓ' between students that do not share lecture or lab sections. The scaling parameter g controls the additive increase to covariance between students of the same lecture section, $g\sigma_{\ell\ell'}$, and d controls the additive increase to covariance

between students of the same lab section, $d\sigma_{\ell\ell'}$. If each of these variance scaling parameters is set to zero, there is no violation to the assumption of independence between students; thus making the variance structure of the generative model (2.2) match the error structure of the original ARTIST model (2.1). We will use these parameters to control for violations of the independence assumption to different degrees in the simulation.

The simulation procedure for assessing the Type 1 error rates under violations of independence between student responses is conducted through a five step process. We consider violations based on all combinations of $d \in \{0, 0.02, 0.04, 0.06, 0.08, 0.1\}$ and $g \in \{0, 0.02, 0.04, 0.06, 0.08, 0.1\}$. We set z=0 for all simulations because any random effect that is common to *all* students will not effect tests for curriculum effects (i.e. the difference in average responses between curricula groups remains constant). For each combination of variance scaling parameters we repeat the following simulation process for $m \in \{1, \ldots, 20000\}$:

1. Randomly permute the lecture and lab section labels in the data
2. Simulate the random effects by drawing $\vec{\epsilon_{ijk}}^{\,(m)}$, $\vec{\delta_{jk}}^{\,(m)}$,$\vec{\gamma_k}^{\,(m)}$, and $\vec{\eta}^{\,(m)}$ from the multivariate normal distributions defined above.
3. Compute the m^{th} simulated responses with generative model (2.2) as:

$$y_{ijk\ell}^{(m)} = \beta_{0k} + x_{ijk1}\beta_{k1} + x_{ijk2}\beta_{k2} + \eta_{\ell}^{(m)} + \gamma_{k\ell}^{(m)} + \delta_{jk\ell}^{(m)} + \epsilon_{ijk\ell}^{(m)}$$

4. Fit the ARTIST Model, assuming independence, to the simulated responses.
5. Conduct Pillai's test for overall curriculum effect, then t-tests for curriculum effect on each response individually. Record test statistics and p-values.

Recall that in the generative model (2.2) does not include a treatment effect, therefore any test where significant curriculum effects are found has incurred a Type 1 error. Figure 2.4 displays the observed Type 1 error rates in simulations under violations of independence between students. When the variance scaling parameters are set to zero there is no violation of independence and we see that tests hold at the nominal Type 1 error rate of α=0.05. However,

the error rates increase quickly when the variance scaling parameters, d and g, increase. This occurs more dramatically for the overall Pillai's test than for the individual t-tests.

To establish points of reference for interpreting Figure 2.4 we examine the error rates under a few specific parameter settings. When the between labmate covariance is 4% higher than for non-labmates (i.e. d=0.04 and g=0) the Type 1 error rates for individual t-tests are above 0.15 and the Pillai's test is above 0.20; over three and four times the nominal rate, respectively. Worse, when the between classmate covariance in 4% higher than for non-classmates (i.e. d=0 and g=0.04) the Type 1 error rates for individual t-tests are above 0.25 and the Pillai's test is above 0.35; over five and seven times the nominal rate, respectively. This error rate inflation occurs because the random effects based on lecture and lab sections are being misinterpreted in the ARTIST model as fixed effects of curriculum due to the assumption of independence between students; a misinterpretation made worse in the case of lecture section due to the complete confounding with curriculum.

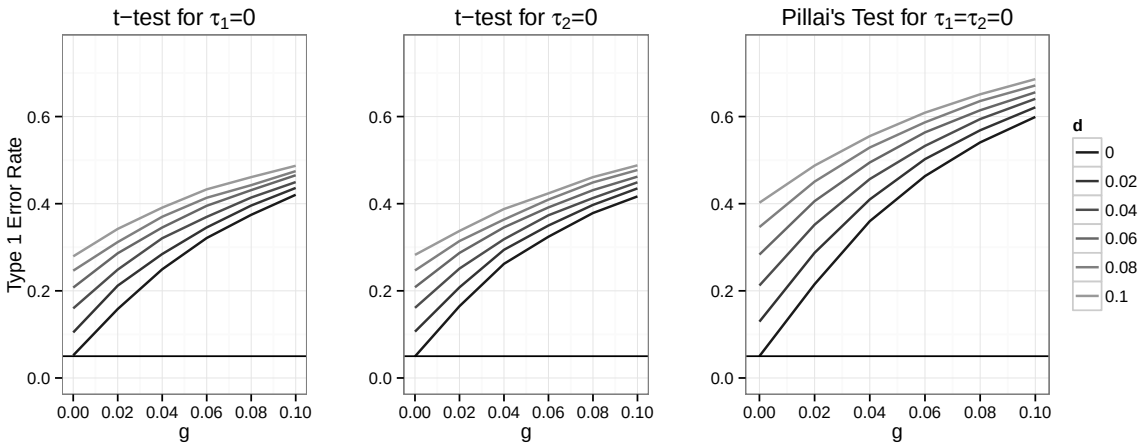


Figure 2.4: Type 1 error rates from 20,000 simulations under each combination of d and g for individual t-tests and Pillai's overall test for curriculum effects on the ARTIST response scores. The horizontal black line indicates the nominal Type 1 error rate of $\alpha = 0.05$.

Note that the magnitude of variance scaling parameters is necessarily attributed to non-curricular factors because the generative model is designed to carry no curriculum treatment effect. Great care was taken during the study design and administration to minimize all non-curricular differences that students encountered in lecture and lab sections; using the alternation of instruction, identical curricula administered with all students in the same room in weeks 1 to

8, and careful pedagogical preparation. However, the simulation study indicates that even in the case of very minor lecture or lab based variance structure the model suffers highly inflated Type 1 error rates and gives rise to major doubts about the results of tests for curricular effects discussed in Sections 2.4.1 and 2.4.2.

2.5 Discussion and Conclusions

The ARTIST model analysis indicates that students receiving the simulation-based curriculum have significantly higher learning outcomes for confidence interval related topics. The magnitude of the improvement was 7.146% on the ARTIST scale, after accounting for the midterm and lab 5 scores. There was no significant difference between traditional and simulation-based curricula on learning outcomes for hypothesis test topics. There are however several issues that merit serious concern and consideration about the validity of these results; including the model assumption of between student independence, the diverse population of interest, the replicability of treatments, and the measurement of learning outcomes.

The experimental design used randomization of individual students to curricula to aid in the creation of homogeneous groups receiving each the curricular treatment, but the treatments were administered on the lecture and lab section level. The bivariate MANCOVA model was fit under the assumption of independence between students because the lack of repetition in lecture and lab sections does not allow for proper estimation of lecture or lab based variance structure. Even when great efforts are made to control for non-curricular differences in the student experiences, as was done in this study, it is unreasonable to assume that students of the same lecture and/or lab sections would not share some degree of non-curricular connection in learning outcomes as a result of having shared the same physical learning environment. The simulation study indicates that even minor violations of this assumption leads to unacceptable inflation to Type 1 error rates in tests for curriculum effects.

This error rate inflation is a major problem not only for this study, but every educational study where curricular treatments are implemented in groups of students but comparisons are made between learning measurements from individual students. Alternative experimental designs could implement curricular treatments on the individual student level or have sufficient

group replication to make comparisons on the group level; each option suffering major logistical and resource demands. To overcome this issue, a deeper understanding of inter-student variability is needed to support the use of more appropriate covariance structures. Uniformity trials are an established method within agricultural statistics that examines variance structure through experiments with only a single treatment (see e.g. Richter and Kroschewski 2012). If applied in a classroom setting, this approach provides a potential avenue to identify variance structures that form due to student cohorts, which could then be used as a basis of plug-in estimates for variance components in studies without cohort repetition. This would require a widespread and concerted effort by many in the discipline to collect and catalog data on inter-student variance structure from many classroom scenarios.

We must also bear in mind the population for which the results of this study may be representative. The study was conducted with undergraduate students enrolled in an introductory statistics course at a large public Midwestern university. The course is required for students in agricultural and biological sciences. Students in the course are predominantly sophomores and juniors. The results are therefore only applicable to the extent to which these students represent the broader population of introductory statistics students.

Another important consideration is that the treatment itself was a half semester curriculum – a highly complex combination of lesson plans, lecture content, assignments and technology use. The treatment complexity poses a problem in identifying precisely which, if any, components of the curriculum might improve learning outcomes. Investigation of the efficacy of the individual components from the improved curriculum is an area for future research. There are also many possible ways to implement simulation-based inference within a course. One noteworthy characteristic was that the simulation-based curriculum that was employed in this study utilized bootstrapping to teach the concepts of confidence intervals as opposed to inverting a simulation-based hypothesis test. This study does not attempt to identify if all possible implementations of the simulation-based approach would achieve improvements in learning outcomes.

A surprising aspect of the results is that the curriculum effect was more pronounced for the learning outcomes of confidence interval topics than for hypothesis testing topics. It is

surprising because much of the literature on the simulation-based approach is focused on the theoretical benefits in simplifying the concepts of hypothesis testing. A potential explanation is that the benefits of the simulation-based methods were counterbalanced by the challenge faced when students were also required to learn theory-based methods; forcing them to mentally reconcile the differences between how each approach obtains a p-value. It is important to recall that due to departmental requirements for the course, the treatment group learned simulation-based inference in addition to an abbreviated unit on theory-based inference methods. However, based on the highly inflated Type 1 error rates in studies with cohort based variance structure, it is also very plausible that the disparity seen is purely random.

The evidence of improved learning outcomes is also contingent on the efficacy of the ARTIST scales to measure student learning. The ARTIST question sets have been criticized as being increasingly outdated and for lacking reliability and validity evidence (Ziegler, 2014). The Comprehensive Assessment of Outcomes in Statistics (CAOS; DelMas et al. 2007) was considered as alternative assessment of student learning because it is nationally normed and backed by a reliability study; however, the reliability was assessed for the CAOS test in its entirety (40 items) which was decided to be too extensive to be administered in addition to the other necessary components on the final exam. The Reasoning about P-values and Statistical Significance (RPASS; Lane-Getaz 2013) assessment was also considered, but not selected, because it does not assess learning outcomes for confidence interval related topics. The Basic Literacy in Statistics (BLIS; Ziegler 2014) and the Goals and Outcomes Associated with Learning Statistics (GOALS; Garfield et al. 2012) assessments were recently designed to better measure student learning in the contemporary statistics classroom, unfortunately these assessments were in development at the time of this study.

This study is unable to draw reliable conclusions on the efficacy of simulation-based methods for teaching statistical inference due to the volatility of the Type 1 error rates under minor violations to model assumptions. This is a fundamental problem for all comparative educational studies where pedagogical treatments are administered on the class level and measurements are taken the individual student level. Viewed as a case study of two curricula administered within extremely similar student groups, we found that simulation-based curriculum had a

noticeable improvement in learning outcomes associated with confidence intervals and a small improvement in learning outcomes associated with hypothesis testing. While the results are clearly not conclusive in assessing the curricular effect on learning outcomes, the findings do merit further consideration into the pedagogical benefits of a simulation-based curriculum.

CHAPTER 3. A SHINY NEW OPPORTUNITY FOR INTERACTION WITH BIG DATA IN UNDERGRADUATE EDUCATION

Status: In Preparation for Submission to *Technology Innovations in Statistics Education* (TISE)

Authors

Karsten Maurer, Iowa State University, Primary Author

Heike Hofmann, Iowa State University

Abstract

As the availability of truly massive data proliferates, it is enticing to incorporate these data sets into the curriculum of an undergraduate statistics course. Major barriers exist for interacting with big data due to the computationally intense nature of working with large databases. Difficulties include gaining access to databases, interacting with database management software, and obtaining summary statistics or manageable subsamples from the database for student use. This paper describes a web-based software application, the *Shiny Database Sampler*, which allows instructors and students to bypass these barriers using a simple point-and-click interface constructed through `R` and the `R` packages `shiny` and `RMySQL`. The Shiny Database Sampler allows instructors and students to obtain subsamples from databases, using a variety of random sampling schemes. Application and evaluation of the software indicate that students find the interface easy to use, well connected to course concepts, and engaging through access to real data.

3.1 Introduction

Statistics education has been rapidly evolving in the past decade with respect to undergraduate course curriculum and assessment. Technology has played the role as a catalyst for many of these major changes. An important change involves how data is accessed and analyzed in the classroom. The GAISE report (Aliaga et al., 2005) laid out six recommendations on how to improve the teaching of introductory statistics; two of which urge statistics instructors to "Use technology for developing conceptual understanding and analyzing data" and to "Use real data". There are many software tools and online repositories for instructors to access real data for use in the statistics classroom. These include the Data and Story Library (DASL Project, 1996) and its Australian counterpart, OzDASL (Smyth, 2011), the Data Archives of the Journal of Statistics Education (American Statistical Association, 2014), CAUSE Web Repository (CAUSE, 2014), and Many Eyes (IBM Corp., 2013b). These repositories are wonderful for accessing many real data sets but the majority of the data sets currently available are quite small in scale.

Finzer et al. (2007) argue that in curricula for introductory level statistics "(w)hat seems to us to be missing are data sets-especially large and highly multivariate data sets-that are ripe for exploration and conjecture driven by the students' intrigue, puzzlement and desire for discovery" (Finzer et al., 2007, p.1). Large, real data sources are becoming increasingly available, but tend to be less easily accessible. A major contributor to this trend is the Freedom of Information Act which to ensures that non-classified data from the United States government is publicly available (U.S.C., 1996). The online government resources at www.data.gov/ (U.S. General Services Administration, 2014), www.census.gov/ (U.S. Census Bureau, 2014), www.nhtsa.gov/ (National Highway Transportation Safety Administration, 2015) and www.cdc.gov/ (Center for Disease Control, 2014) are all locations of massive data stores. Governmental data sets contain rich information related to many socially relevant issues, making them prime candidates for engaging student interest.

The goal of connecting undergraduate statistics students with big data sources requires careful consideration to implement. Jacobs (2009) speaks of the difficulties associated with

using, "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time", including scaling computational tasks, avoiding sub-optimal storage schemes and parallel processing.

New data technologies are needed in order to allow introductory statistics students to interact with big data sources, such as the governmental databases. This paper discusses the construction and functionality of the Shiny Database Sampler; a web-based application that allows students to pull random samples from large databases. Then, details of the implementation within a lab assignment and course project for an introductory statistics course are discussed. Lastly, the survey responses from 265 introductory statistics students whom used the Shiny Database Sampler for the lab assignment are analyzed to evaluate the software.

3.2 Shiny Database Sampler

Exposing students to larger and larger data is tricky because it becomes increasingly unwieldy to transfer, store and access data on student's personal computers. Access through remote databases and database querying languages is outside the realm of comfort for both students and instructors in most undergraduate statistics courses. The Shiny Database Sampler is a simplified graphical user interface that is designed to allow students to take appropriately sized subsets from the databases to practice small sample methodology taught in most introductory statistics courses. Working with a small static subset from the large database would let the remainder of the database go to waste. Instead the tool allows students to specify a random sampling scheme that will pull the subset from database dynamically. This is done to emphasize the role of random sampling techniques in data collection, an important concept within an introductory statistics curriculum.

3.2.1 Layout and Design

The Shiny Database Sampler allows the user to randomly sample subsets from remotely stored SQL databases using a point-and-click graphical user interface. The tool is available online through the link at http://shiny.stat.iastate.edu/karstenm/ShinyDatabaseSampler/. A screenshot of the graphical user interface is shown in Figure 3.1.

The Shiny Database Sampler is a JavaScript based online application created using the `shiny` package (RStudio and Inc., 2014) in the statistical computing language `R` (R Core Team, 2013). The Shiny package uses `R` code files to generate the graphical user interface in a web browser that interacts with an `R` session running on the server. It is used in combination with the `RMySQL` package (James and DebRoy, 2012) to allow the `R` session on the server machine to query the database at the user's request via buttons in the graphical user interface.

The interface was designed with a focus on the quality of the software as an educational tool. The field of software development defines six attributes contributing to software quality: functionality, reliability, usability, efficiency, maintainability, efficiency and portability (Bevan 1997; Berton and Vallencillo 2002). The Shiny Database Sampler is highly portable, as it can be accessed online through any JavaScript enabled web browser. The reliability and maintainability of `shiny` applications depend upon proper implementation of `R` and `shiny` on the web server. The key consideration with respect to efficiency is to optimize the database querying using proper indexing so that sample retrieval occurs almost instantaneously (Schwartz et al., 2012b).

The graphical user interface contains two tabs, each broken into two panels. The layout is designed for the user to select actions in the side panel and view the results in the main panel. The functionality and usability of the *Sample and Summarize* and *Visualize* tabs are discussed in the following subsections.

3.2.1.1 The *Sample and Summarize* tab

A screenshot of the interface for the *Sample and Summarize* tab is shown in Figure 3.1 below. A sidebar panel which contains all the sampling options and controls, and a main panel which contains the data table and brief summary of the sampled data set.

The sidebar panel contains several fields and buttons for selecting and executing a sampling plan. At the top of the sidebar is a drop-down menu to select the database table from which the user may take a random subsample. The current version of the tool allows users to access the 2001-2009 Fatality Analysis Recording System accident data from the National Highway

Traffic Safety Administration (www.nhtsa.gov/FARS) and the Public Use Micro Sample data from the 2000 United States Census (www.census.gov/).

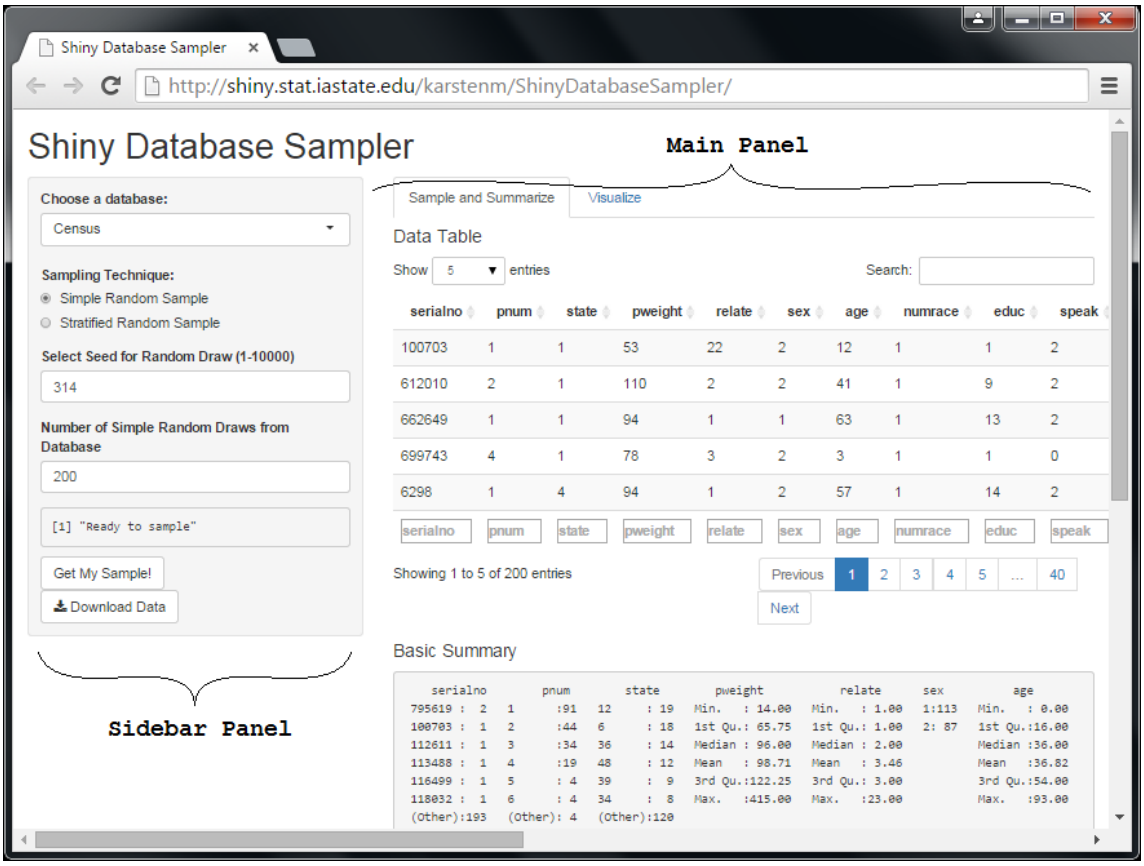


Figure 3.1: Shiny Database Sampler layout for *Sample and Summarize* tab.

After selecting the database, the user can choose between taking a simple or stratified random subsample of data from the database. If the user chooses simple random sampling, then they must specify a sample size; whereas if the user chooses stratified random sampling the strata variable and number of observations per stratum need to be specified. Lastly the user sets the seed for the selection algorithm. Setting the seed may seem to contradict the intention to draw random subsets, however it was included after careful consideration. For classroom settings, it is often desirable for students to work with identical data sets for consistency of class discussion and grading. This can be accomplished by setting the same seed and the same sampling specifications. Obtaining a random sample from the databases is still possible with the additional – but fairly trivial – step of first randomly generating a starting seed.

Once the sampling setup is specified, the user may click the *Get My Sample!* button and the randomly selected subsample of the database will be obtained and displayed in the main panel of the interface. Note that the interface only keeps track of the more recently selected subsample, which we will refer to as the *active data*. Lastly, the side panel contains the button to download the active data to a local drive on the user's computer. The data will be downloaded as comma separated values (csv) file to the default download folder on the user's computer.

The main panel of the Shiny Database Sampler interface displays a data table and a basic summary of each variable in the active data. When logging into the webpage a default sample is taken and displayed until a sample of the users choosing is selected. The data table is searchable, sortable and expandable which makes it easy for the user to take a quick peek at the variable names and values in the active data. The basic summary statistics for each variable are also displayed in the main panel below the data table; those familiar with *R* programming will quickly recognize this as the verbatim output of the *summary* function in *R*. In the case that stratified sampling was used to draw the data, the summary for each variable is broken down by strata. The displays in the main panel of the Shiny Database Sampler are not intended to be the location for any extensive analysis of the active data but instead a quick check that it was what the user intended to select.

3.2.1.2 The *Visualize* tab

The *Visualize* tab of the interface is designed to construct basic plots of the active data that was drawn in the *Sample and Summarize* tab. Figure 3.2 shows a screenshot of the layout of the *Visualize* tab. The sidebar panel contains fields for specifying variables and variable types that will be plotted. Univariate plots can be created by selecting the response variable from a drop-down menu containing a list of all variables and the variable type. Plots may also display bivariate relationships by additionally selecting an explanatory variable and it's variable type.

The variable types must be manually specified as numerical or categorical. Variable types are not automatically specified for variables in the database for the deliberate purpose of forcing student users to consider appropriate ways to display the data. Table 3.1 shows the possible plot types that can be created based on the options selected.

Once the plotting options are selected in the sidebar panel, the user may click the *Make my Plot!* button to generate and display the plot in the main panel. Since the visualizations are intended only for preliminary exploration, the plots have a default construction and labeling that are not able to be customized.

Table 3.1: Plot types supported in *Visualize* tab.

Response Variable	Explanatory Variable	Default Plot
Numerical	None	Histogram
Categorical	None	Barchart
Numerical	Numerical	Scatterplot
Categorical	Categorical	Stacked Barchart
Numerical	Categorical	Side-by-side Boxplots

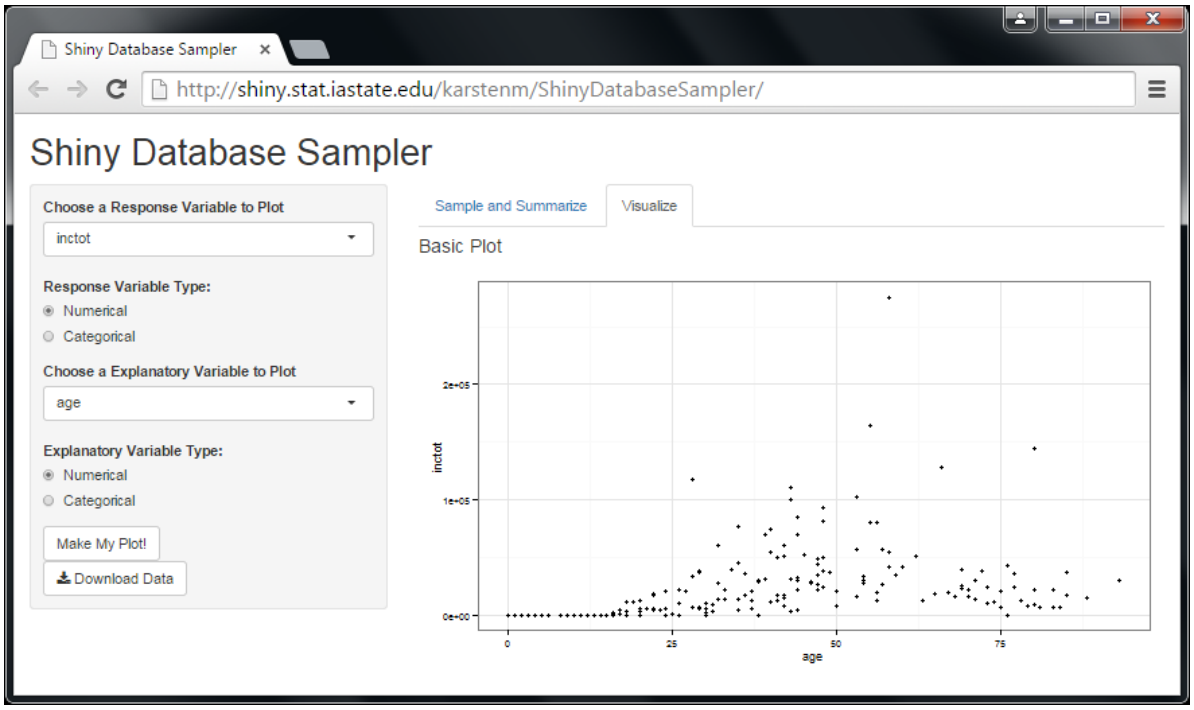


Figure 3.2: Shiny Database Sampler layout for *Visualize* tab.

3.2.2 Applications

The Shiny Database Sampler was developed for applications in statistics education and was implemented in the initial stages of development within the curriculum of an introductory statistics course. The alpha version of the software was used within one section of Statistics 104: *Introduction to Statistics* for a group lab assignment and a course project. Later, a beta version was extended to be used in a lab assignment for six more sections of the same course. These two applications demonstrate possible uses of the Shiny Database Sampler.

3.2.2.1 Lab Application Overview

The lab that utilized the Shiny Database Sampler was designed for students to think critically about sampling approaches, then the software allowed them to treat the large database as a population upon which to conduct their mock survey. Students were asked to consider the following pair of hypothetical situations:

1. Suppose that our goal is to estimate the mean age of all US residents. Similar to polling organizations we have a budget that allows us to survey around 1000 people. To collect our sample we decide to take a simple random sample of 1040 US residents.

2. Suppose now that our goal has changed. Now we wish to investigate the association between age and state of residency. We want to compare the median ages for different states. We still have a budget that allows us to survey of 1040 people. To collect our sample we decide to take a stratified random sample of 20 residents from each state in the United States plus the District of Columbia and Puerto Rico.

In each scenario students were asked to discuss the choice of sampling scheme and to identify potential problems or difficulties. The students used the Shiny Database Sampler tool to obtain a sample from the database containing a 1% microsample of the 2010 U.S. Census, from which they estimated mean and median age of U.S. residents. This lab was written to ensure that sampling concepts were the primary focus, with the Shiny Database Sampler acting in a supporting role. In order to avoid (sporadic) clicking of buttons to obtain samples without

ever stopping to consider why the sampling approach matters, we intentionally designed the assignment to invite students to carefully consider sampling options *before* using the tool. The complete lab assignment can be found in Appendix B.1.

3.2.2.2 Capstone Project Application Overview

A second application of the Shiny Database Sampler was as an optional data source for a capstone project for students of Stat 104. Students were required to work in small groups to complete a capstone project that took a statistical approach to answering a question of their choosing. In order to accommodate students' interests in the subject matter of the project, groups were allowed to pick their own data source. The only specific requirement for the project was a written report that explained the collection of bivariate data and an analysis of the association – included the appropriate plots, inference and interpretations to answer their question of interest.

Six of the groups chose to run a mock survey, using the databases available through the Shiny Database Sampler as a stand-in for a large population. These groups were required to demonstrate a strong argument for the sampling plan they used within the Shiny Database Sample. For instance, if a group wanted to know if the proportion of fatal accidents involving a drunk driver was higher in California than in Iowa; as part of the report they needed to argue why taking a stratified random sample of 100 fatal accidents per state would produce better information to answer their question than a simple random sample. Groups that gathered random samples from the Census and Accidents databases knew that they had gathered information from real life sources and they seemed genuinely engaged in the results of their projects.

3.2.3 User Survey for Software Evaluation

As discussed above, the Shiny Database Sampler was designed for student use in course assignments. The quality of the software was initially assessed using a pluralistic walk through (Nielsen, 1994), where the developer met iteratively with both statistical novices and experts to test the functionality of the early versions of the software. The software improvements that

followed this inspection created the beta version of the Shiny Database Sampler that was ready for student use.

The second stage of evaluation for the Shiny Database Sampler was a user survey to learn student opinions about their experience from using the software during the lab assignment described above. Specifically, we are interested in three aspects of the software: we want to know if students find the tool easy to operate, if they see the connection to sampling concepts and if they find the data engaging. Exploring the student responses on these three topics of interest helps us to assess the quality of the Shiny Database Sampler as an educational software.

The ease of use of the Shiny Database Sampler is an considered an important attribute from both an educational psychology and a software development viewpoint. Cognitive load theory postulates that human capacity to process information is limited and that learning is composed of *extraneous load*, effort to overcome obstructions to new knowledge, and *germane load*, effort used to form new schema and integrate ideas with existing knowledge. Muller et al. (2008) explain that "Finding ways to increase germane load and minimize extraneous load has been a central pursuit of researchers under this paradigm." Application of cognitive load theory is extended to the setting of software in *human centered design* (Oviatt, 2006). The goal is to make interfaces for educational technologies intuitive and easy to operate in order to minimize the extraneous load; thus allowing more mental resources to be devoted to developing and integrating new knowledge. The intuitive construction of the user interface is an important component to software development, where the usability of a tool is defined by its learnability, understandability and operability (Berton and Vallencillo, 2002).

Active learning requires a high cognitive load, and if prior knowledge is lacking more scaffolding is necessary to support learning (Muller et al., 2008). Since sampling concepts are typically new for students, it is important that students clearly recognize sampling concepts in the interface; allowing it to be used as part of an assignment that is scaffolded for active learning. If students are able to identify the role of the Shiny Database Sampler as a tool for learning about random sampling, it is more easily integrated into the learning process.

Finally, assessing student engagement with the data sources in the Shiny Database Sampler is important because higher student engagement is linked to higher academic performance and

learning (Carini et al., 2006). The GAISE guidelines recommend that technological tools should be used to help teach statistical concepts and that the use of real data is important for student engagement, hence we focus on these topics (Aliaga et al., 2005). Neumann et al. (2013) found that students consider real data more interesting and engaging. The hope is that students using the Shiny Database Sampler will find the real, nationally collected, governmental data sources engaging.

Student responses were collected in an anonymous survey following the group lab assignment – as described in Section 3.2.2.1 – that required students of an introductory statistics course, Stat 104, at Iowa State University to use the Shiny Database Sampler tool. Six sections of Stat 104 students were surveyed. The students were informed that the survey was not required and that no penalties or rewards were affiliated with its completion. Of the 320 students attending the lab, 265 completed the survey.

3.2.3.1 Survey Description

After completing the lab assignment, students were asked to respond to a survey consisting of twelve statements (referred to as *items* in the following, see Table 3.2 for an overview). For each statement, students were asked for feedback on their level of agreement on a Likert scale from strongly disagree to strongly agree. The twelve items were designed to assess student opinion within three topics of four items each: ease of use, connection to sampling concepts, and engagement with the census data. We will refer to these as the *Ease*, *Concept* and *Engagement* item sets. For each group of four items, two were worded positively and two were worded negatively. Introducing negation with half of the items was done to reduce the response bias associated with *acquiescence* – the tendency to respond positively irrespective of the item content due (Furnham, 1986). Responses were scored as -2 (strongly disagree), -1 (disagree), 0 (neutral), 1 (agree), 2 (strongly agree). Responses for negatively worded items were reverse-scored for the purposes of analysis.

From Table 3.2 we see that all response averages are positive after reverse-scoring. With the Ease items this indicates that students tend to find the tool relatively easy to operate.

Table 3.2: Survey items and response summaries *after* reverse-scoring (RS).

Topic Set	ID	Item	Polarity	Mean	SD
Ease	1	*I found the web tool easy to use*	+	0.84	0.76
	2	*The layout of the web tool was intuitive*	+	0.63	0.74
	3	*Using the web tool was difficult*	−	0.77	0.88
	4	*Learning to use the web tool was hard*	−	0.81	0.85
Concept	1	*The web tool helped me understand sampling concepts*	+	0.80	0.78
	2	*I understand sampling ideas less after using the web tool*	−	0.83	1.01
	3	*Sampling techniques are clearer after using the web tool*	+	0.58	0.73
	4	*The web tool made me less sure how to randomly sample*	−	0.89	0.87
Engagement	1	*I did not enjoy working with the Census data*	−	0.38	1.01
	2	*I thought the Census data was boring*	−	0.23	0.97
	3	*Knowing that the Census data was from real people made it more interesting*	+	0.82	0.84
	4	*I liked analyzing the Census data*	+	0.28	0.86

For frame of reference, we assume that students are comparing the difficulty of use with other educational technologies and webpages they have encountered in the past; in particular the `JMP` software used previously on their Stat 104 labs and homework. Students also tend to respond to Concept items in a manner that is affirmative that the tool connects them to sampling concepts. Students' responses are near to neutral for most items about engagement with the census data, with the exception of Engagement item 3. The phrasing of this question seems to have led students to reconsider their engagement level and led to a consistently more positive attitude.

3.2.3.2 Assessment of Internal Consistency for Item Topic Sets

The goal for this survey is to use the responses to sets of items to infer student opinions about the underlying topic of each set. It is reasonable to aggregate the responses over entire questions sets if we can show that items within each set are measuring the same latent topic. We use fluctuation diagrams and Cronbach's α (Cronbach, 1951) to assess this internal consistency.

A fluctuation diagram is the visual analog of the contingency table, displaying frequency of each unique responses combination as the area of blocks on the bivariate grid containing all possible pairs of response values. A fluctuation diagram with large blocks on the diagonal indicates strong agreement or *internal consistency* between responses of the two items. Figure 3.3 contains fluctuation diagrams for all item pairs within topic sets. We notice that most pairs of

responses fall heavily along the diagonal and are primarily in the upper right of each diagram. This indicates that most items within sets have strong agreement and that the response values are generally neutral to positive for all items (after reverse-scoring the negative statements). For the item pairs in the Concept topic set we see that fluctuation diagrams have slightly larger off-diagonal trends than items within the other two sets, which indicates a lower level of internal consistency for Concept items than in the other two topic sets.

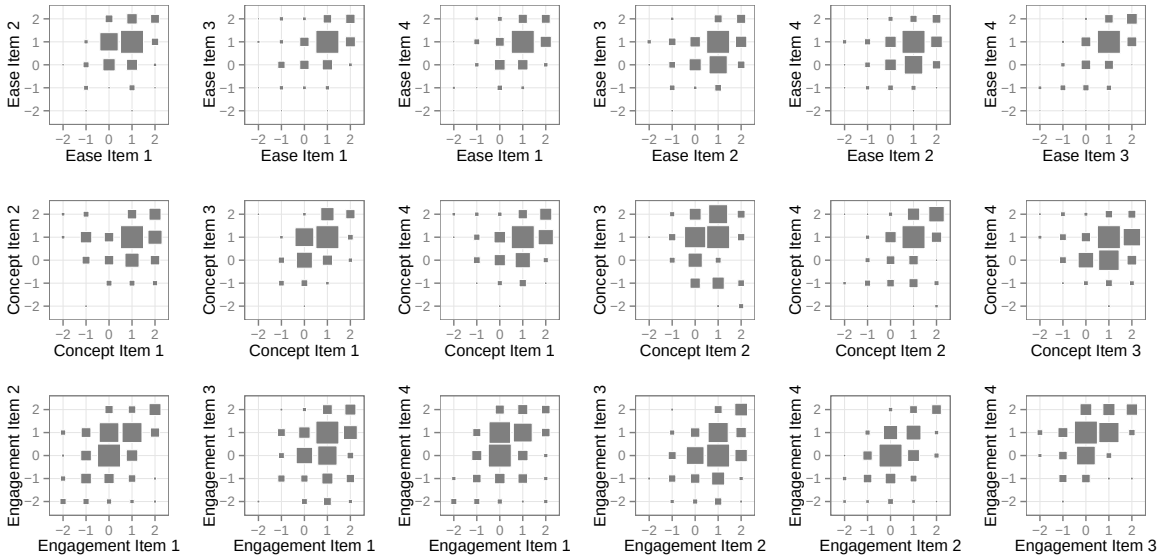


Figure 3.3: Fluctuation diagrams of all item pairs within topic sets.

Cronbach's α measures internal consistency within an item set by comparing the sum of individual response variances to the variance of the sum of the responses. It is defined as follows

$$\alpha = K/(K-1) \cdot \left(1 - \sum_{i=1}^{K} \mathrm{Var}\left(Y_i\right) \Big/ \mathrm{Var}\left(\sum_{j=1}^{K} Y_j \right) \right), \tag{3.1}$$

where Y_i denotes the response on the i^{th} survey item ($i = 1, ..., K$), and K is the number of survey items considered for internal consistency. Generally, $K = 4$ for the item sets of this survey. Cronbach's α reaches a maximal value of 1, if there is perfect agreement between items (i.e. all responses to the same item set are identical). In the case that items sets are independent, the internal consistency is measured as $\alpha = 0$. Cronbach's α can be negative in the situation of consistent *disagreement* between responses and will approach negative infinity

if there is perfect disagreement between items. See appendix B.2 for details on the bounds for Cronbach's α.

Nunnally and Bernstein (1978) propose that a Cronbach's α of 0.7 or above should be considered as an indication of "modest reliability". George and Mallery (2003) provide the commonly used extended scale, displayed in Table 3.3, for interpreting internal consistency based on Cronbach's α.

Table 3.3: Extended scale for Cronbach's α (George and Mallery, 2003).

Internal Consistency	Range
Excellent	$[0.9, 1.0]$
Good	$[0.8, 0.9)$
Acceptable	$[0.7, 0.8)$
Questionable	$[0.6, 0.7)$
Poor	$[0.5, 0.6)$
Unacceptable	$(-\infty, 0.5)$

Since Cronbach's α is a sample estimate for the internal consistency of an item set, it experiences sampling variability. Under the assumption of normally distributed responses, Cronbach's α follows approximately an F_{ν_1,ν_2}, where $\nu_1 = n - 1$ and ν_2 is based on a function of the eigenvalues from the quadratic linear combination of the roots of the variance matrix (Kistner and Muller, 2004). Assuming normality to construct confidence intervals for the true internal consistency of item sets would be questionable for the responses in this survey, so we have elected to bootstrap the intervals instead.

Table 3.4 displays the point estimates and 95% central bootstrap intervals for Cronbach's α for each item set from the student survey. The intervals were created using quantiles of the Cronbach's α values from 10,000 bootstrap resamples. The results indicate modest levels of internal consistency for Ease and Engagement item sets, and a lower level for the Concept item set. This is in agreement with the findings based on the fluctuation diagrams in Figure 3.3.

Table 3.4: Cronbach's α estimates for each item set with 95% central bootstrap confidence interval based on 10,000 bootstrap samples.

Set	Estimate	95% Confidence Interval
Ease	0.70	(0.613 , 0.759)
Concept	0.53	(0.410 , 0.637)
Engagement	0.72	(0.643 , 0.776)

3.2.3.3 Assessment of Polarity Issues

We next turn our attention to the polarity of the survey items; specifically we consider that positive and reverse-scored negative items may elicit a different responses. The survey contained six unique item pairs based on topic and polarity combinations. Figure 3.4 compares the distribution of responses from positive and negative item pairs within topics. We see strong similarity between positive and reverse-scored negative items in response distributions for the Ease and Engagement item sets. The Concept item set however displays a noticeable difference in response distributions from each polarity. In particular, we see that students are more neutral toward the positively worded questions. This polarity difference in student responses explains the lower internal consistency measured by Cronbach's α, and may be partly due to the problem that the negation of positive constructs can be linguistically counter-intuitive (Friborg et al., 2006). For instance, students may interpret the statement "It is not less clear" differently than the statement "It is more clear".

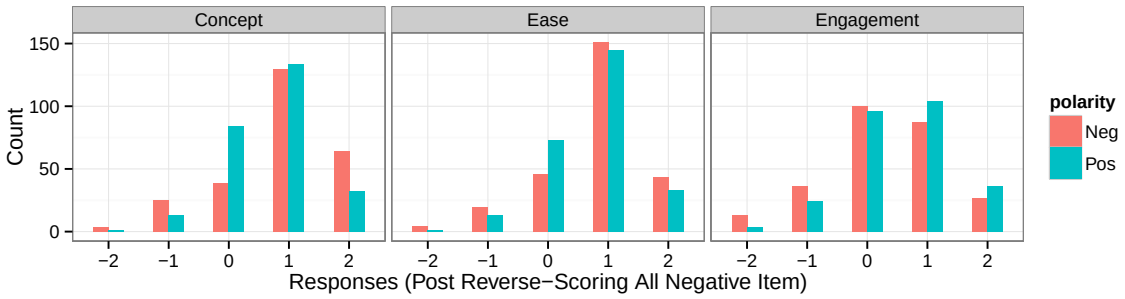


Figure 3.4: Item set response distributions by polarity.

To assess whether responses from positive and reverse-scored negative items can be reasonably grouped together within topic sets we turn to principal component analysis. We first combine the item pairs into averages for each of the six topic and polarity combinations, then

we decompose these six scores into principal components. The component variances and factor loadings from this decomposition are found in Table 3.5. We argue that the data could be reasonably reduced to four principal components because each of these components explains over 10% of the variance and together they explain 87.4% of the total variation. The uniformly aligned factor loadings for Component 1 reflect the general tendency toward student agreement to all items on the survey. The factor loadings for Components 2 and 3 displayed in Figure 3.5 show similar projections for positive and negative item scores for Ease and Engagement pairs but a dramatic separation in the positive and negative item scores for the Concept set.

Table 3.5: Summary statistics from principal component analysis with six topic/polarity item pairs.

	Principal Component	1	2	3	4	5	6
Variances	Prop. of Var	0.457	0.181	0.135	0.101	0.071	0.055
	Cumu. Prop. of Var	0.457	0.638	0.773	0.874	0.945	1.000
Loadings	Pos. Ease	-0.292	0.169	-0.509	-0.036	-0.055	0.789
	Neg. Ease	-0.444	-0.390	-0.464	0.429	0.372	-0.335
	Pos. Concept	-0.304	0.330	-0.389	-0.350	-0.537	-0.487
	Neg. Concept	-0.408	-0.641	0.250	-0.585	-0.067	0.116
	Pos. Engaged	-0.367	0.523	0.207	-0.321	0.663	-0.083
	Neg. Engaged	-0.569	0.164	0.519	0.496	-0.355	0.087

This principal component analysis, with all topic and polarity combinations, suggests that we can reduce the dimensionality by combining the positive items with the reverse-scored negative items for Ease and Engagement topics. This leaves only the Concept item set separated based on polarity for final analysis. The decision to combine the responses for Ease and Engagement items also aligns with the higher internal consistency for these item sets as displayed in Cronbach's α values and fluctuation diagrams in Figure 3.3. Thus, we will carry forward with the final analysis using four resulting item sets: Ease, Positive Concept, Negative Concept and Engagement.

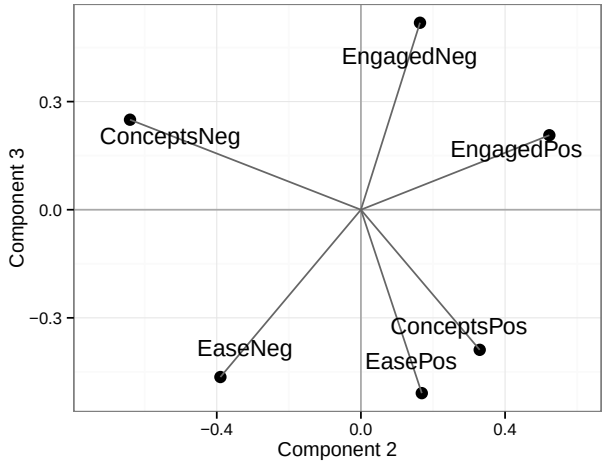


Figure 3.5: Item pair loadings on Components 2 and 3 from the principal component analysis on six topic/polarity item pairs.

3.2.3.4 Assessment of Orthogonality

The next major consideration is whether the item sets are truly measuring different latent topics, and therefore can be view as non-redundant. The ability of the survey to separately measure the topics of Ease, Concepts and Engagement can be assessed through the orthogonality of the responses from different item sets. To check the orthogonality of the sets we conduct another principal component analysis; this time on the average responses for each student from the four item sets – Ease, Positive Concept, Negative Concept and Engagement. Items sets will be considered highly orthogonal if the principle component analysis cannot reduce the dimensionality from the four sets.

Table 3.6: Principal component analysis with final four item sets.

	Principal Component	1	2	3	4
Variances	Prop. of Var	0.515	0.249	0.134	0.102
	Cumu. Prop. of Var	0.515	0.764	0.898	1.000
Loadings	Ease	-0.430	0.144	-0.307	0.837
	Pos. Concept	-0.408	0.562	-0.520	-0.497
	Neg. Concept	-0.619	-0.745	-0.101	-0.227
	Engagement	-0.515	0.330	0.790	-0.031

Table 3.6 displays the proportion of variance explained by each of the four principal components and factor loadings. The first principal component has similar loadings from all item sets, which we can interpret as the general tendency toward positively scored responses on all items. The second, third and fourth principal components create separation for mean responses of the Negative Concept item set, the Engagement item set and the Ease item sets, respectively. The variances in Table 3.6 reveal that over 10% of the variation is explained by the fourth component, thus it is necessary to retain all four principal components. This inability to reduce dimensionality implies that average student responses from the four item sets are largely orthogonal. Based on the separation in the loadings and the orthogonality of the principal components, we conclude that the average response scores from the four item set have interpretability as measurements of unique latent topics.

3.2.3.5 Survey Assessment Results

In the analysis of student responses, we find that the internal consistency, assessed with Cronbach's α and fluctuation diagrams, is acceptable for interpreting the combined item responses that measure ease of use and engagement with the census data. We do not have the same certainty with the responses to Concept items and therefore split the Concept items into two sets: the Positive and Negative Concept item sets. This split is supported by the initial principal component analysis of the six topic and polarity item pair scores. The follow-up principal component analysis on the combined responses for each of the four resulting item sets indicates that the factors were all fairly orthogonal. This ensures us that the survey was effective at eliciting unique characteristics of the user experience.

The barcharts found in Figure 3.4 show that the distribution for each item set is heavily skewed to the left, with the majority of students having neutral to positive responses. The small tail to the left in each distribution indicates that there was a small minority of students with expressedly negative views. The response distributions indicate that on average students found the Shiny Database Sampler easy to use, found that the tool connected them to sampling concepts and felt moderately engaged with the census data that was accessed with the application.

3.3 Conclusions and Future Work

The Shiny Database Sampler allows point-and-click access to large databases through the mechanism of random sampling. This approach adds pedagogical value over the use of static samples because it allows for course activities to highlight the concepts and process of collecting data through random sampling. The lab and project application of the Shiny Database Sampler within an introductory course were designed to emphasize thought about sampling concepts and the data, not about the software. Toward this scaffolded approach, the design aimed to make the interface easy to use, clearly display sampling concepts and provide engaging data. The student user survey indicates that these goals were met.

The Shiny Database Sampler could naturally be updated to access additional databases or provide more statistical analysis or visualization options directly within the interface. Also in future work, a similar interface could be developed where the user specifies an aggregation scheme instead of a sampling scheme. This approach would be more true to exploration techniques of big data sources than using random sampling; this would be a distinct – and perhaps exciting – departure from the content of a traditional introductory statistics curriculum. In such an interface, grouping variables or binning parameters could be used to direct dynamic data aggregation to then be explored; using display such as histograms or binned scatterplots. As the size and ubiquity of data in the world grows, students would be well served by attempts to thoughtfully incorporate big data into the undergraduate statistics curricula.

CHAPTER 4. BINNING STRATEGIES AND RELATED LOSS FOR BINNED SCATTERPLOTS

Status: In Preparation for Submission to *Journal of Computational and Graphical Statistics* (JCGS)

Authors

Karsten Maurer, Iowa State University, Primary Author

Heike Hofmann, Iowa State University

Susan Vanderplas, Iowa State University

Abstract

Dealing with the data deluge of the Big Data Age is both exciting and challenging. The demands of large data require us to re-think strategies of visualizing data. Plots employing binning methods have been suggested in the past as viable alternative to standard plots based on raw data, as the resulting area plots tend to be less affected by increases in data. This comes with the price of the loss of information inherent to any binning scheme. In this paper we discuss binning algorithms used in the construction of binned scatterplots. We define functions to quantify the loss of spatial and frequency information and discuss the effects of binning specification on loss in the framework of simulation and case studies. From this we provide several practical suggestions for binning strategies that lead to binned scatterplots with desirable visual properties.

4.1 Introduction

Technological advances have facilitated collection and dissemination of large data as records are digitized and our lives are increasingly lived online. According to an EMC report in 2014

"the digital universe is doubling in size every two years and will multiply 10-fold between 2013 and 2020 - from 4.4 trillion gigabytes to 44 trillion gigabytes" (http://www.emc.com/about/news/press/2014/20140409-01.htm). This "Data Deluge" of the Big Data Age (NY Times, Feb 2012) poses exciting challenges to data scientists everywhere: "It's a revolution ... The march of quantification, made possible by enormous new sources of data, will sweep through academia, business and government. There is no area that is going to be untouched"– Gary King, Harvard Institute.

Data sets with millions of records and thousands of variables are not uncommon. Friedman (1997) proposed in his paper on data mining and statistics that "Every time the amount of data increases by a factor of ten, we should totally rethink how we analyze it". Jacobs (2009) echoed the sentiment, stating that "big data should be defined at any point in time as *data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time*". The same holds true for visualizations. With a 100-1000 fold increase in the amount of data, the utility of some of our most commonly used graphical tools, such as scatterplots, deteriorates quickly (Unwin et al., 2006).

Area plots, such as histograms, do not tend to be as affected by increases in the amount of data because they display aggregations instead of raw data. By using binning strategies and the principles for displaying information in area plots, scatterplots can again become useful instruments for large data settings (Unwin et al., 2006).

In this paper we describe first the inadequacy of traditional scatterplots in large-data situations. We discuss different binning algorithms use in the construction of binned scatterplots and the *loss of information* inherent to binning. We will then explore the effects of binning specification on the properties of binned scatterplots through simulation and real-data case studies. We conclude with several practical suggestions for binning specifications for creating binned scatterplots that have desirable visual properties.

4.2 Scatterplots for Large Data Sets

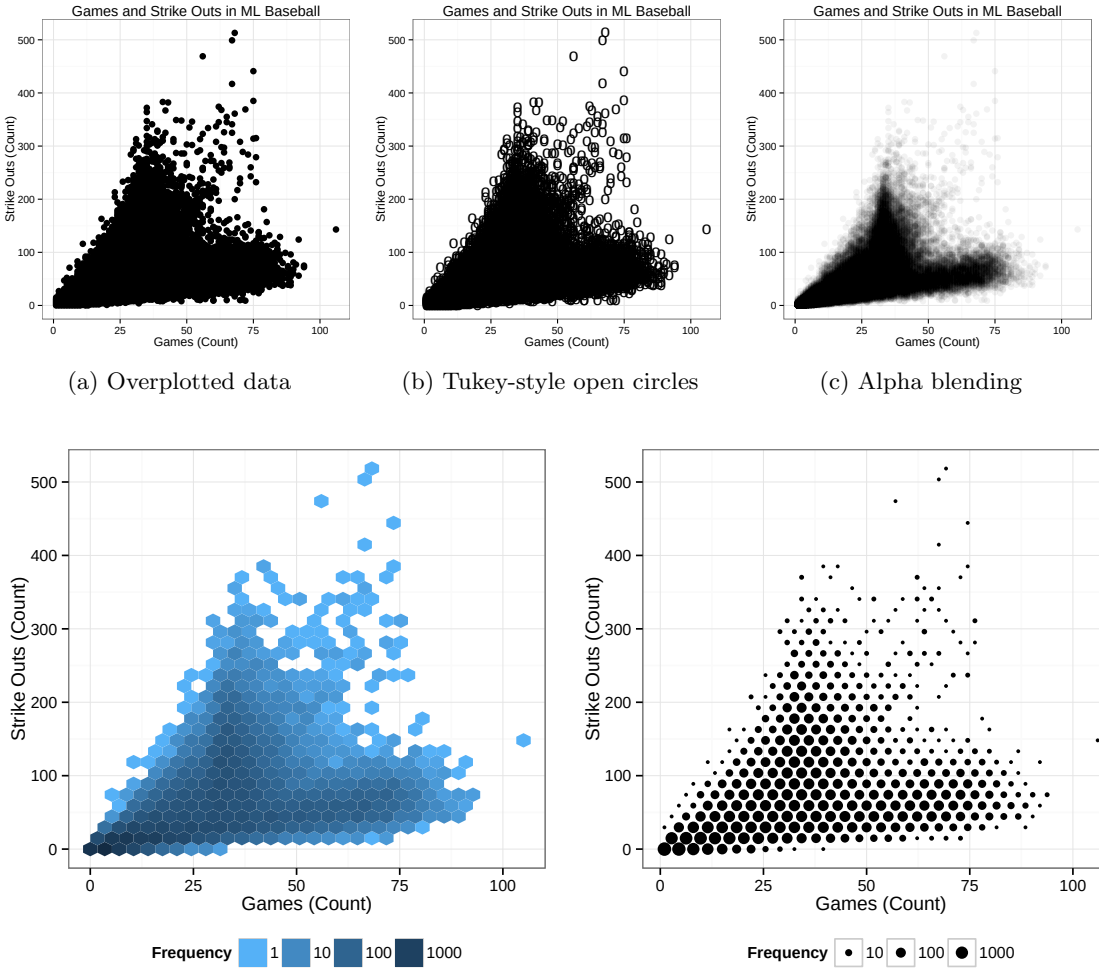
In the case of modestly sized data, scatterplots are great tools for showing bivariate data relationships. With large data, scatterplots suffer from over-plotting of points, which masks

relevant structure. Figure 4.1 shows an example taken from baseball statistics. The scatterplot shows 139 seasons (from the years of 1871 – 2009) of pitching statistics for every baseball pitcher as published in Sean Lahman's Baseball database (http://www.seanlahman.com/baseball-archive/). The number of games played in a season is plotted against the number of strikeouts a pitcher threw over the course of a season. While the data set is only medium sized with 42583 observations, it already shows some of the break-down patterns scatterplots experience with large data.

Figure 4.1(a) shows a traditional scatterplot with each observation is drawn with a filled circle. A triangular structure is apparent with some outliers at a medium number of games and high number of strikeouts; however the density within the triangular mass of points is indistinguishable. Tukey (1977) suggested the use of open circles (see Figure 4.1(b)) to mitigate the problem of over-plotting. Open circles make points that are close together more visually distinct; thus allowing for the perception of more density information than with filled points. A modern alternative to open circles is alpha blending (see Figure 4.1(c)). Alpha blending renders points as semi-transparent to provides more visibility of underlying points.

All of these methods fall short in the example. As can be seen in Figure 4.1, strategy (a) is the least effective, as it provides information about the outliers and range of the data but cannot provide any point density information. Tukey's open circles (b) help to a degree, but are also prone to over-plotting when the data set is very large. Alpha blending (c) highlights the structure, but minimizes the visual impact of outliers. The data set is large enough that neither alpha blending nor open circles are completely effective, and so we must pursue a different strategy which can provide better information about the relative density of points at a given location.

Other scatterplot adaptations have been introduced that avoid over-plotting by manipulating the display of the points by distorting the locations or the scales. Generalized scatterplots (Keim et al., 2010) display all individual observations, including those sharing identical coordinates, and use distortion of the point locations by having points repel one another to avoid overlapping. An extension of generalized scatterplots uses clustering and local principal components to allow ellipsoid oriented distortion to display local correlation structure in the data



(a) Overplotted data (b) Tukey-style open circles (c) Alpha blending

(d) Hexagonal binned scatterplot (e) Hexagonal binned bubble plot

Figure 4.1: Scatterplots of games versus strikeouts in major league baseball, using different strategies of dealing with the issue of over-plotting: (a) uses standard, opaque, filled circles, (b) uses Tukey's recommended open circles, and (c) uses filled circles with alpha blending (α=0.05). Plots (d) and (e) show hexagonal binning strategies with frequency mapped to color and area respectively.

(Janetzko et al., 2013). Variable-binned scatterplots (Hao et al., 2010) break the display into a non-uniform rectangular grid and resize the rows of cells according to density of points. This variable binning fragments the continuity of the axes into segments on different scales and also does not deal with points at identical coordinates. Generalized and variable-binned scatterplots make fine data structure more visible and allow color to be reserved for a third variable instead of frequency; however, the distortion of the point locations and/or axes warp the visual display of the association between the two primary variables.

Another approach is to reduce the graphical complexity by plotting binned aggregations of the data, namely frequencies, as opposed to plotting every observation as an individual point. This has the additional advantage of reducing the size of the stored data necessary for the construction of the plot, as only the bin centers and the bin frequencies must be stored. Wickham argues for a "bin-summarize-smooth" procedure to be applied to the visualization of big data and he notes that simple summary functions, such as counts, scale well with the size of data (Wickham, 2013). Liu, Jiang and Heer employ the computational benefits of binning for their interactive big data visualization program, *imMens* (Liu et al., 2013).

Methods commonly used to display binned variables include sunflower plots (Cleveland and McGill, 1984b), kernel density smoothing of tonal variation and binned scatterplots (Theus, 2006b). Sunflower plots are scatterplots of binned data, where the symbol used for the bin increases in complexity in proportion to the number of points in that bin. Sunflower plots are particularly useful when the number of points in each bin remains reasonably small. Kernel density smoothing can be used to vary α or color according to a smoothed density, providing features similar to binned scatterplots or alpha blended scatterplots in a more smooth, continuous fashion. However, these estimates require careful parameter tuning, as over-smoothing may hide gaps in the data while simultaneously de-emphasizes outlying points.

Histograms are a simple example of a plot that can be built using binned aggregations of the data; in their case the bin locations and bin counts act as a set of sufficient statistics necessary to reconstruct the plot. A natural extensions of histograms to higher dimensions is to form a tessellated grid on a two dimensional Cartesian plane using some other attribute, such as color or size or 3D renderings to provide joint density information within each grid cell, known as

a tile. A *bubble plot* is a binned data plot that scales the size of a filled circle in proportion to frequency. Bubble plots were first used by William Playfair (Playfair, 1786; Playfair et al., 2005). A *binned scatterplot* uses shading to provide frequency information, with tiles (rather than bars in a histogram) at the bin center, similar to a two-dimensional histogram viewed from above.

Figure 4.1 contains examples of a hexagonally binned scatterplot with frequency encoded as color (4.1(d)) and a bubble plot with frequency encoded as point size (4.1(e)). The hexagonally binned scatterplot and the bubble plot are more effective at displaying the shape of the joint density and preserving outliers than any of the scatterplots shown in Figure 4.1(a-c). The bubble plot is prone to suffer from the Hermann-grid illusion (Hermann, 1870), where the white spaces between circles on the evenly spaced grid appear shaded due to an optical illusion; whereas this will not occur for a binned scatterplot on a tessellated grid.

The inner structure of the baseball data is only apparent in the binned scatterplot and the bubble plot. The joint density consists two distinct ridges following two lines with very different slopes. The lower slope corresponds to the modern average strike out rate of pitchers of just under one strike-out per game. The other line has a slope of about four times that rate. This high rate is also associated with fewer games played. Closer investigation of other, related variables reveals that this high strike-out rate corresponds mainly to historic pitchers with much shorter seasons (in 1876 only 70 games were played in a season, as opposed to 162 in 2009), and qualitatively different balls and bats.

For extremely large data sets, binned scatterplots are a more useful visualization of two-dimensional density information than the scatterplot, and are less computationally demanding, as not every single point in the data set has to be rendered separately.

As with histograms, the width of bins (or the number of bins) is an important factor in the detail of the binned data and the resulting plot: if the bin width is too small in comparison to the amount of data available, there is little advantage to binning, but if the bin width is too large, interesting features of the joint distribution may be obscured by over-plotting.

4.3 Binning Algorithms

Binning algorithms used in making distributional approximations can be traced back to Pearson's work with the binomial approximation to the normal, where he mentions the need to define an origin and binwidth for segmenting the normal distribution (Pearson, 1895). More recently Scott has presented discussion on the importance of binning specification in the creation of histograms to appropriately display one dimensional density approximations (Scott, 1979). Scott (1992) extends to the properties of multivariate binning strategies.

Binning in dimensions X and Y provides us with a more condensed form of the data that ideally preserves both the joint distribution as well as the margins, while reducing the amount of information to a fraction of the original. Binning is a two-step procedure: we first assign each observation (x, y) to a bin center (x^*, y^*), and in a second step we count the number of observations assigned to each unique bin center; resulting in reduced data triples of the form (x^*, y^*, c), where c is the number of all observations assigned to bin center (x^*, y^*).

We will proceed with rectangular bins for simplicity, but other binning schemes, such as hexagonal bins (Carr et al., 1987) are also common. While hexagonal binning has been shown to have slightly better graphical properties (Scott, 1992); rectangular bins are advantageous because bins in x and y are orthogonal to each other, thus we can present the one-dimensional case which will easily generalize to two or more dimensions. We will however only consider binning in up to two dimensions, X and Y. The algorithms we discuss are immediately applicable to higher dimension, but we do not feel that the paper would benefit from a more general discussion.

For the univariate case with observations, x_i for $i \in \{1, \ldots, n\}$, binning algorithms require a set of bin centers x_j^* for $j \in \{1, \ldots, J\}$ and a binning function $b_X(.) : x_i \to x_j^*$ that maps observations to bin centers.

Rectangular binning accomplishes this by defining a sequence of J adjacent intervals, $(\beta_{j-1}, \beta_j]$ for $j \in \{1, \ldots, J\}$, which span over the range of the data. Note that half open intervals are used such that any observation falling on a bin boundary is assigned to a unique interval. Values x_i exactly equal to the lowest bin boundary β_0 are grouped into the first bin to close the leftmost

bound. Each observation is then mapped to a bin center, x_j^*; the midpoint for the interval to which the observation belongs.

This is expressed mathematically using the binning function $b_X(.) : x_i \to x_j^*$ defined as

$$b_X(x_i) = \begin{cases} x_1^* & \text{for all } x_i = \beta_0 \\ x_j^* & \text{for all } x_i \in (\beta_{j-1}, \beta_j] \end{cases} \tag{4.1}$$

Standard rectangular binning is a special cases of general rectangular binning that uses intervals of equal size for all bins; thus only the origin of the first bin, β_0, and a binwidth, ω_X, need to be specified. Standard rectangular binning is necessarily used in the construction of all histograms; the consistent binwidth makes the display of frequency proportional to density. Fixed width binning procedures are also highly computationally efficient (Wickham, 2013).

Note that this standard rectangular binning procedure utilizes intervals that are open to the left and closes the outer bound of the leftmost bin. These specifications are consistent with the binning procedure used in the `hist()` function for creating histograms in base `R` (R Core Team, 2013). These specifications were selected for this paper, but these choices are by no means consider universal for binning. For example, the `ggplot2` package creates histograms with intervals open to the right and does not close the outer bound of the rightmost bin (Wickham, 2009).

Quantile binning is another option that divides the range of the observations into bins each containing an equal number of points. The j^{th} bin interval takes the form $(Q_X((j-1)/J), Q_X((j)/J)]$, where $Q_X(p)$ is the the p^{th} empirical quantile using the inverse empirical distribution function. Note that this binning approach is *not* desirable for spatially visualizing density patterns, as it effectively balances the frequency counts in all bins; it does however have desirable properties for binned scatterplots that employ a second stage of binning to create discrete shade scheme for displaying grouped bin frequencies, which will be discussed in Section 4.4.2.

The bin boundaries and centers for each type of rectangular binning algorithm discussed above can be found in Table 4.1.

An alternative to the rectangular binning processes discussed above, is the random binning algorithm which utilizes a non-deterministic bin function $b_X^r(\cdot)$ to randomly assigns an obser-

Table 4.1: Rectangular and Random Binning Specifications

	Bin Boundaries	Bin Centers
General	$\{\beta_j \mid \beta_j > \beta_{j-1}\}$	$\{x_j^* \mid x_j^* = (\beta_{j-1} + \beta_j)/2\}$
Standard	$\{\beta_j \mid \beta_j = \beta_{j-1} + \omega_X\}$	$\{x_j^* \mid x_j^* = \beta_{j-1} + \omega_X/2\}$
Quantile	$\{\beta_j \mid \beta_j = Q_X(j/J)\}$	$\{x_j^* \mid x_j^* = Q_X((j-0.5)/J)\}$
Random	—	$\{x_j^* \mid x_j^* > x_{j-1}^*\}$

vation, x_i, to a bin center, x^*, from a set of possible bins. In this paper, we will consider the simplest case of just two bins, so that without loss of generality we can assume that x_i lies between bin centers x_j^* and x_{j+1}^*. The bin function assigns x_i to a bin center with a probability inversely proportional to the distance to that bin center; the closer a value is to a bin center, the higher the probability the value is assigned to that bin center. More formally,

$$b_X^r(x_i) = \begin{cases} x_j^* & \text{with probability } (x_{j+1}^* - x_i)/(x_{j+1}^* - x_j^*) \\ x_{j+1}^* & \text{with probability } (x_i - x_j^*)/(x_{j+1}^* - x_j^*) \end{cases} \tag{4.2}$$

for $x_i \in [x_{j+1}^*, x_j^*]$. In Table 4.1 we note that this random binning algorithm does not specify bin boundaries; only a sequence of bin centers. This method is easily extensible to also map x_i into more than two bins and can accommodate non-uniform distribution of bin centers.

The deterministic standard binning algorithm is an example of a "direct" binning algorithm, in which all points are assigned with weight one to the bin center. "Linear" binning (Theus, 2006b) is a *computationally intensive* alternative to direct binning in which adjacent bins are assigned a weight depending on the distance from the point to that bin, where all weights sum to one. With large data sets, the calculations required for linear binning become unwieldy, but the random binning algorithm can be considered an approximation to linear binning. Specifically, the expectation of the random binning algorithm is the same as for linear binning.

4.3.1 Extension to Two Dimensional Binning

The standard and random binning algorithms are easily extendable to higher dimension. For the purposes of creating binned scatterplots we will specify extension to rectangular binning in two dimensions. In this case we wish to assign data pairs (x_i, y_i) to bin centers of the form (x_j^*, y_k^*), with $j \in \{1, \ldots, J\}$ and $k \in \{1, \ldots, K\}$,where J and K are the number of bins in the

X and Y dimensions, respectively. The (j,k) pairs that index the bin centers can be linearized to a single index such that $\ell = j + J(k-1)$; thus making j the fast running index and k the slow running index. With this linearized index for all bins we now have a set of bin centers of the form (x^*_ℓ, y^*_ℓ), with $\ell \in \{1, \ldots, \mathscr{L}\}$, where $\mathscr{L} = J \cdot K$.

The standard rectangular binning function $b(.) : (x_i, y_i) \rightarrow (x^*_\ell, y^*_\ell)$ is defined as

$$b(x_i, y_i) = (b_X(x_i), b_Y(y_i)) \tag{4.3}$$

where $b_X(x_i)$ and $b_Y(y_i)$ are the univariate standard binning algorithms for the X and Y dimensions respectively. The random rectangular binning function, $b^r(\cdot) : (x_i, y_i) \rightarrow (x^*_\ell, y^*_\ell)$ is similarly defined as

$$b^r(x_i, y_i) = (b^r_X(x_i), b^r_Y(y_i)) \tag{4.4}$$

where $b^r_X(x_i)$ and $b^r_Y(y_i)$ are univariate random binning algorithms for each dimension. Figure 4.3 provides an illustration of each binning process extended to a two dimensional situation.

4.3.2 Binned Data Reduction

The second stage of binning requires a frequency breakdown of the number of observations associated with each bin center, forming reduced data triples, (x^*, y^*, c), where c is the number of all observations assigned to bin center (x^*, y^*). Table 4.2 makes use of a small set of simulated data to show the progression from the original data (a), to the binned data (b), to the reduced binned data (c). The reduced binned data is sufficient for constructing the binned scatterplot. In cases of large data, binning greatly reduces the storage size for the information and the computation time needed to construct a binned scatterplot.

Note that numerical attributes other than frequency of the binned data may also be recorded during binning, however only frequency is required to construct a binned scatterplot. Data reduction comes at the expense of spatial information of any of the individual points. After aggregation the original spatial locations cannot be recovered. The loss of information incurred from binning will be explored in following sections.

Table 4.2: Original, binned and reduced binned data tables, with data storage sizes. Binned using standard rectangular approach with origin $(\beta_{0,x}, \beta_{0,y})$ = (-10,-10) and binwidths $\omega_x = \omega_y = 10$.

(a) Original Data, 12 rows

x	y
-7.7325	-9.6340
-8.1176	-1.4529
-5.8996	-3.2033
-7.0375	-5.5563
-3.6354	-3.9315
-8.7639	0.9874
-2.9781	8.6802
0.8210	-8.6118
5.4477	-8.4555
4.6849	-5.6620
9.4785	1.1133
1.7579	5.3759

(b) Binned Data Centers, 12 rows

$b_X(x)$	$b_Y(y)$
-5	-5
-5	-5
-5	-5
-5	-5
-5	-5
-5	5
-5	5
5	-5
5	-5
5	-5
5	5
5	5

(c) Reduced Binned Data, 4 rows

x^*	y^*	c
-5	-5	5
-5	5	2
5	-5	3
5	5	2

4.4 Loss due to Binning

Problems with large data in scatterplots arise from over-plotting, which is a form of implicit data aggregation. In order to keep track of the number of observations near a given location, we switch to a weighted visual display which explicitly aggregates the data. The reduced binned data carries the sufficient information necessary to render the binned scatterplot. Making the data aggregation explicit allows us to calculate the loss we experience.

A traditional scatterplot is comparable to a *minimally binned scatterplot* – using small enough bins such that only a single unique coordinate pair exists within each bin – however, the "bins" of a traditional scatterplot are shaded in a binary manner with no indication of overlapping observations. Alpha blending as used in Figure 4.1(c) extends the binary shading of a standard scatterplot to an implicit shading according to frequency. The shading is implicit because the range of frequency information is not scaled to the range of shading values, so that maximum color saturation is usually reached well before the maximum frequency, truncating the perceivable frequency information. By explicitly shading bins according to frequency, more information is preserved than in a traditional scatter plot, as the frequency domain provides

visual weight to tiles which may represent more points. This generalization allows us to describe the plots in Figure 4.1(d) and (e) under the same framework as plots (a)-(c).

By additionally increasing the bin width, we provide increasingly higher-level summaries of the data by smoothing over local structures. Using a small number of large bins may mask the real signal in the data, while an extremely large number of small bins may not sufficiently smooth over *noise* inherent in any real data set. Figure 4.2 gives an overview of a data set and binned representations using different numbers of bins, demonstrating the loss of information with increasing bin size.

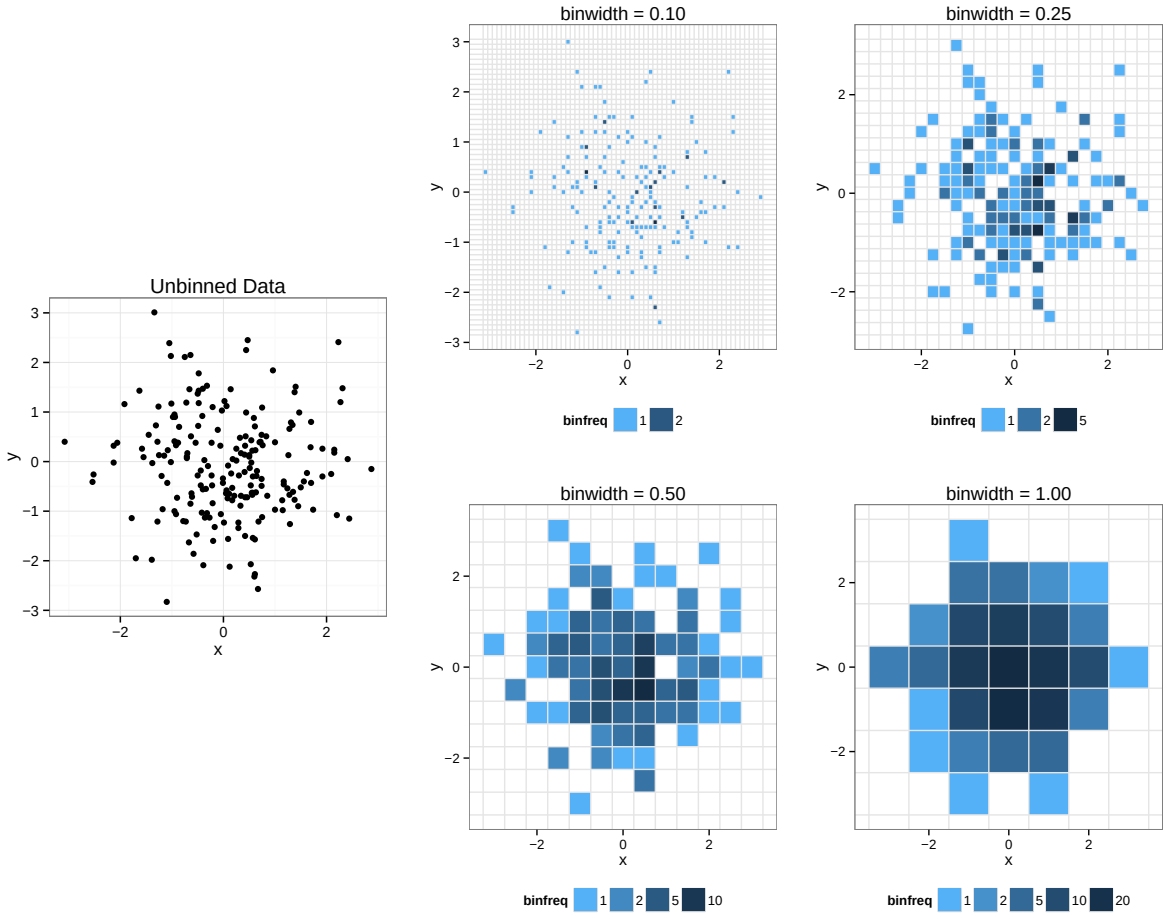


Figure 4.2: Series of scatterplots showing the original data (scatterplot, left), and versions of the binned data for different bin widths. The visual loss from binning at 0.1 is minimal, while a bin width of 1 gives a rough approximation.

In Figure 4.2 the minimally binned scatterplot, with bin width equal to 0.1, is visually very similar to the traditional scatterplot; but importantly the binned scatterplot contains

information about overlapping points. The second and third binned scatterplots, with bin width equal to 0.25 and 0.50 respectively, show higher-level summaries of the data but which may also provide more visually accessible information about the shape of the two-dimensional density distribution between x and y. The fourth binned scatterplot, with bin width equal to 1.0, provides only a rough bivariate density display due to over-smoothing from the large bins.

Loss of information occurs during the binning and rendering process. For the remainder of the paper we will assume that we are using shade in binned scatterplots to represent frequencies. We distinguish two sources of loss in the construction of a binned scatterplot:

- *Spatial Loss*, L^{S}, occurs when points (x_i, y_i) for observations $i \in \{1, \ldots, n\}$ in the data set are reduced to a set of tiles centered at (x^*_ℓ, y^*_ℓ) for bins $\ell \in \{1, \ldots, \mathscr{L}\}$. By displaying frequency information using shaded tiles instead of individual points there is a loss of information about the exact location of the points.

- *Frequency Loss*, L^{F}, occurs when bin counts, $c_\ell \in \{1, \ldots, \mathscr{L}\}$ are not mapped to a continuous shading scale. While shade can be *rendered* continuously in HSV color space, thus representing frequency exactly, a human reader can not *extract* this information at the same precision due to limitations of human cognition. In order to model these limitations we introduce a second stage of binning by using a discrete color scale for displaying binned frequencies, $b_C(c_\ell)$, $\ell \in \{1, \ldots, \mathscr{L}\}$.

Note that even though the losses from creating a binned scatterplot may turn out to be substantial, there is a huge gain with respect to an traditional scatterplot, where information can be masked in large data situations due to over-plotting of points. The idea of loss from one-dimensional binning was explored by Scott using mean integrated squared error as the loss function to be optimized by the choice of the number of bins in the construction of histograms (Scott, 1979). He later extended this discussion to two-dimensional binning, where he compared the mean integrated squared error loss for hexagonal, rectangular and triangular binning; finding that hexagonal and rectangular binning performed similarly, both far superior to triangular binning (Scott, 1992). We will take a similar approach to quantifying loss resulting from the binning algorithms detailed in Section 4.3 above.

4.4.1 Spatial Loss

When the individual points of a scatterplot are collapsed to bin centers to be displayed as tiles in a binned scatterplot there is a loss of the location information. This can be expressed as the Euclidean distance between points and the visual center of the tiles (i.e. the bin centers). Note that other distance metrics could be used, but the Euclidean distance has a desirable interpretability in $\mathbb{R}^2$. The *total spatial loss*, L^S, is defined as

$$L^S = \sum_{i=1}^{n} L_i^S = \sum_{i=1}^{n} \sqrt{(x_i - b_X(x_i))^2 + (y_i - b_Y(y_i))^2} \tag{4.5}$$

where L_i^S is the loss in the ith observation. Figure 4.3 visually displays the spatial loss for the data from Table 4.2 as a result of standard rectangular binning. Observations (x_i, y_i) and bin centers (x_ℓ^*, y_ℓ^*) are displayed as black points and gray crosses, respectively. The length of line segments connecting these represent L_i^S, the spatial loss for each observation; thus, the combined length of all line segments represents the total spatial loss, L^S.

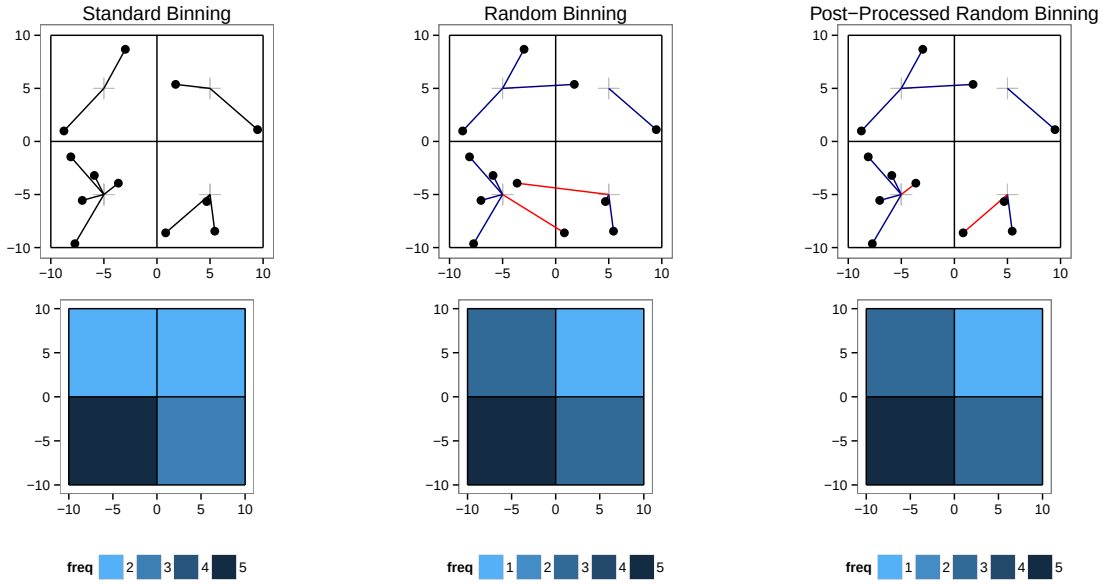


Figure 4.3: Visualization of spatial loss for same data using standard, random and post-processed random binning algorithms. Note that the total spatial loss for the standard algorithm is smaller than by random binning. The net spatial loss for the original random binned data is the total spatial loss of the post-processed assignment.

For random assignment the total spatial loss can be calculated by simply replacing the standard binning function with the random binning function in Equation 4.5. The total spatial

loss for randomly binned data is visualized in Figure 4.3. Special consideration should be paid to the two points with line segments highlighted in red in the panel for random binning. In standard binning these points are in neighboring bins but have been allocated to the opposite bin during random binning. We see that the binned scatterplot remains identical if these points are re-allocated back to the closer centers; however, the total spatial loss is smaller after the random allocation is post-processed.

The minimum total spatial loss from all binning allocations that result in the same reduced binned data, and thus the same binned scatterplot, will be referred to as the *net spatial loss*, L^S_{net}. The net spatial loss for standard binned data is always equivalent to the total spatial loss due to the deterministic bin allocations. For random binning algorithms the net spatial loss is achieved by first exchanging bin assignments for all pairs of points in neighboring bins that exist further from their own bin center than from their partner's bin center, then calculating the total spatial loss from this processed binned data.

The spatial loss is a Euclidean distance, but the units affiliated with this distance are based on the units on which each variable is recorded. If the two variables in the binned scatterplot share the same units this leads to direct interpretability of the spatial loss. However, if the two variables do not share the same units or the same magnitude of values it is advisable to standardize the variables prior to binning, thus making the spatial loss more universally interpretable as a distance in units of standard deviations.

4.4.2 Frequency Loss

It is important to note that there is no spatial loss for the traditional scatterplot, as the points are rendered at exact coordinates. The reason we are willing to sacrifice precision of location information is to alleviate the inherent loss of frequency information from over-plotting, which can be immense for large data sets. When a point is rendered in a traditional scatterplot there is no graphical change once any additional points are placed at the same coordinates; thus visually implying a frequency of one data value at the location. This implicit loss of frequency information is then exacerbated when points are rendered as a circle with a radius large enough

to also partially cover nearby points. In binned scatterplots, we have traded the exact locations for exact frequencies within bins.

Bin counts can be displayed using a continuous shading scale in HSV color space (Healey and Enns, 1999b) and thus we can theoretically map frequency to shade perfectly. While the tiles of the binned scatterplot would be *rendered* precisely, the ability of a human with average vision to extract that information by visually mapping the tile shade to a frequency scale in the plot legend is largely imprecise. Color perception of shade is extremely context sensitive allowing an inaccurate mapping of tile shade to the corresponding shade in the plot legend. It is therefore not realistic, to expect readers to be able to decode frequency from a continuous shade scheme, even though theoretically we can perceive shades continuously (see e.g. the discussion in http://hypertextbook.com/facts/2006/JenniferLeong.shtml). Figure 4.4 shows an example of the context sensitivity of colors. Even though there is a lot of spread between the colors in this example, most onlookers have trouble answering the question about the relative relationship of the shading of the darkest/lightest points in each cloud of points.

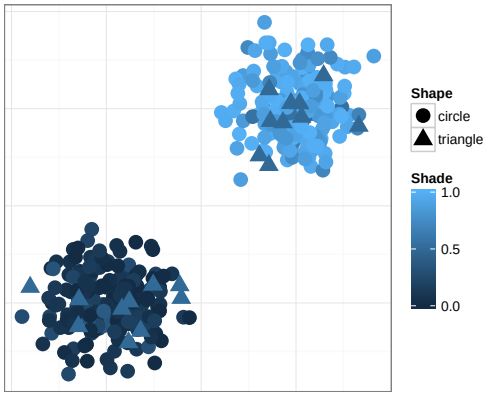


Figure 4.4: Assess the colors in this scatterplot. Are the triangles in the lighter cloud of points as dark as the triangles in the darker cloud of points? The answer is that all the triangles have the exact same shade!

Whenever the shading scheme for rendering the counts in each bin is discretized there is a loss of frequency information. We can model this as a second stage of binning; wherein the bin counts, c_ℓ, for bins $\ell \in \{1, \ldots, \mathscr{L}\}$ are placed into frequency bins using any of the previously

discussed univariate binning algorithms in Section 4.3. Figure 4.5 provides a visual example of a discrete color palette resulting from frequency binning.

Research suggests that even under optimal conditions, we can effectively compare only about seven color hues simultaneously, and that we are even more limited in terms of distinguishing shade (Healey and Enns, 1999b). This provides a physical upper limit on the amount of frequency variation we can perceive through color. As a result, a frequency binning which produces seven of fewer frequency categories is preferable.

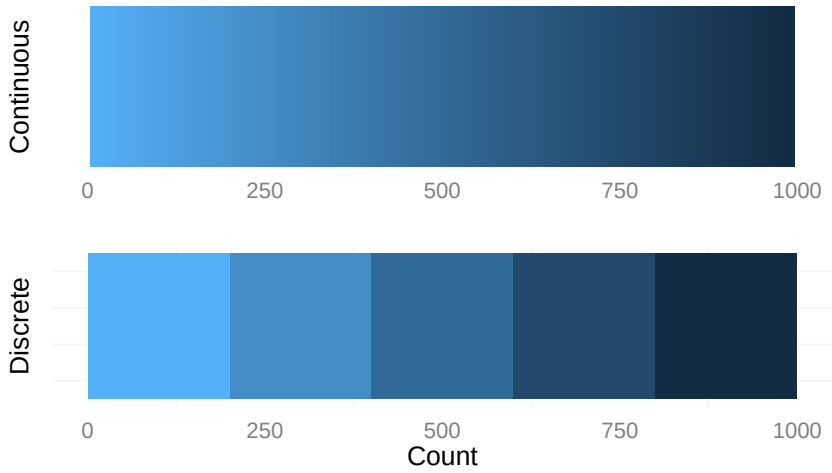


Figure 4.5: Continuous and discrete color scales for counts in the range 0 to 1000. Frequency binning done using standard rectangular binning with origin $\beta_0 = 0$ and binwidth $\omega_c = 200$.

The goal of binning the frequencies and using a ordinal shading scheme is to quantify the imprecision in visually extracting frequency information. It does so by using shade to display binned frequencies, $b_c(c_\ell)$ instead of the true frequencies, c_ℓ. The *total frequency loss*, L^F, is defined as

$$L^F = \sum_{\ell=1}^{\mathscr{L}} L_\ell^F = \sum_{\ell=1}^{\mathscr{L}} (c_\ell - b_c(c_\ell))^2 \tag{4.6}$$

where L_ℓ^F is the frequency loss for the ℓ^{th} bin. Note that this is effectively a sum of squared error from binning frequencies.

While this numerical assessment of frequency loss does not exactly account for limitations in human perceptual ability, it does provide a more realistic model for the loss in perception that does occur.

Frequency data consists of counts, which commonly exhibit skew densities, i.e. there are usually a lot of bins with small bin counts and a few bins with extremely large frequencies. The use of quantile binning is promising in the case of frequency binning because it seeks to place the same number of bins in each shaded group. An alternative is to use a log transformation which produces a more symmetric distribution of frequency information, increasing perceptual resolution. This is consistent with the Weber-Fechner law which suggests that increased stimulus intensity is perceptually mapped on the log scale (Goldstein, 2007). Using a logarithmic mapping of frequency to the shade aesthetic provides a more natural perceptual experience and simultaneously increases the perceptual resolution of the graph. The *log frequency loss*, $L^{\log F}$, is defined as

$$L^{\log F} = \sum_{\ell \in \mathscr{L}^*} L_\ell^{\log F} = \sum_{\ell \in \mathscr{L}^*} (\log(c_\ell) - b_c(\log(c_\ell)))^2 \tag{4.7}$$

where $\mathscr{L}^*$ is the index set for all non-empty bins, which is done to avoid asymptotic problems from log transforming bin counts of zero.

4.5 Exploring Properties of Loss

Binning data for the purpose of creating a binned scatterplot requires a choice of algorithm as well as a choice of parameters associated with that binning algorithm. This section aims to compare binning algorithms and identify the best parameter choices for minimizing loss under a number of scenarios. Some choices may be proven optimal through analytical properties, while other are data dependent and require empirical exploration of loss from binning. Whether analytical or empirical, data is needed to demonstrate how loss is impacted by binning choices.

Data sets were simulated from bivariate distributions to be used throughout this section: Exponential, Normal and Uniform. These distributions were selected for their variety in terms of shape, center and spread. 100,000 observation pairs were simulated for each data set from the following distributions:

- Set I: $x_i \sim$ iid Exp$(\lambda), y_i \sim$ iid Exp(λ) with $\lambda = 11$
- Set II: $x_i \sim$ iid Normal$(\mu, \sigma^2), y_i \sim$ iid Normal(μ, σ^2) with $\mu = 50$ and $\sigma = 11$

- Set III: $x_i \sim$ iid Uniform$(a, b), y_i \sim$ iid Uniform(a, b) with $a = 0$ and $b = 100$

The parameters were selected to have data values roughly span the region $[0, 100]^2$. The simulated data can be found in the top row of Figure 4.6.

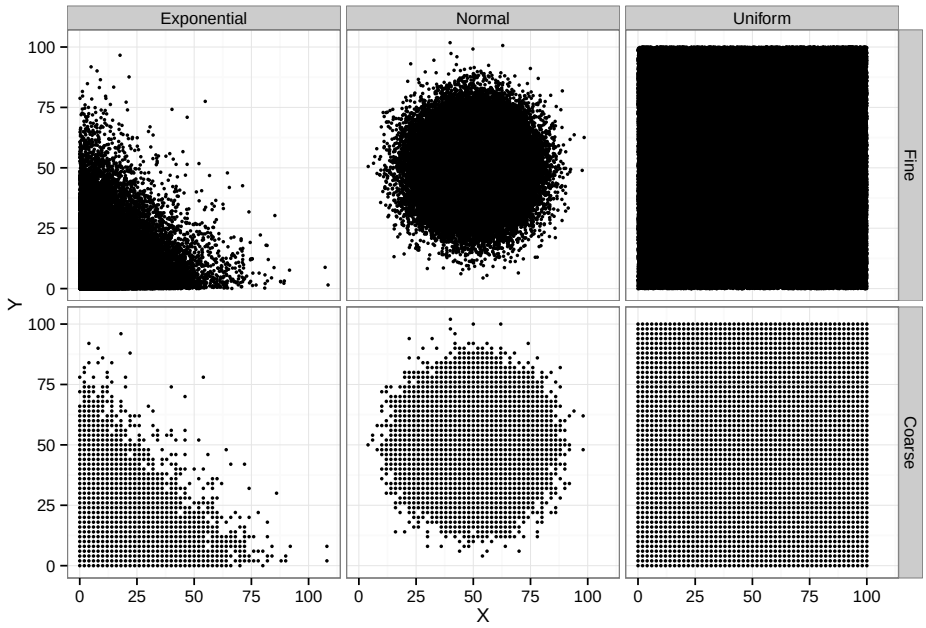


Figure 4.6: Scatterplots of fine and coarse versions of the simulated bivariate data. Density changes are impossible to make out in the solid black areas of the plots. 97.4%, 98.7% and 99.1% of the points are completely hidden behind at least one other point due to over-plotting in the coarse data for the uniform, normal and exponential simulations, respectively.

These simulated data sets are from continuous distributions, and thus the values are recorded to many decimal places; 6 decimal places in our simulate data. Real data is recorded to only the number of digits that measurement precision allows, and in many cases rounded even further.

The *data resolution* is defined as the smallest increment between successive data values. To observe loss from binning under more realistic conditions, we create three data sets by rounding the values from the originally simulated data sets to the nearest even number, thus a data resolution of 2 units in each dimension. This coarse version of the original simulated data is displayed in the bottom row of Figure 4.6. By exploring the loss properties for both the *fine* and *coarse* versions of the data, we identify which binning options are robust to the resolution at which data is recorded.

4.5.1 Rectangular Binning Specifications and Spatial Loss

For rectangular binning there are many specification options; including type of algorithm, location of the origin and binwidths for each dimension. To explore the spatial loss properties under different binning approaches we begin with a comparison of standard and random binning. For equal bin widths, the net spatial loss under standard binning is always less than or equal to the net spatial loss under random binning. This is because the minimal spatial loss for each data point under random binning is to allocate to the nearest bin center, which is how the point would be allocated in standard binning.

Figure 4.7 displays the net spatial loss from binning the fine resolution simulated data with standard and random binning algorithms using square bins with a sizes ranging from 2 units^2 to 20 units^2. It is apparent that the spatial loss for random binning is always higher than for standard binning and the loss grows as bin width increases. The net spatial loss from random binning increases with bin width at a faster rate than from standard binning, as indicated by the widening gap between the lines and the steeper slopes.

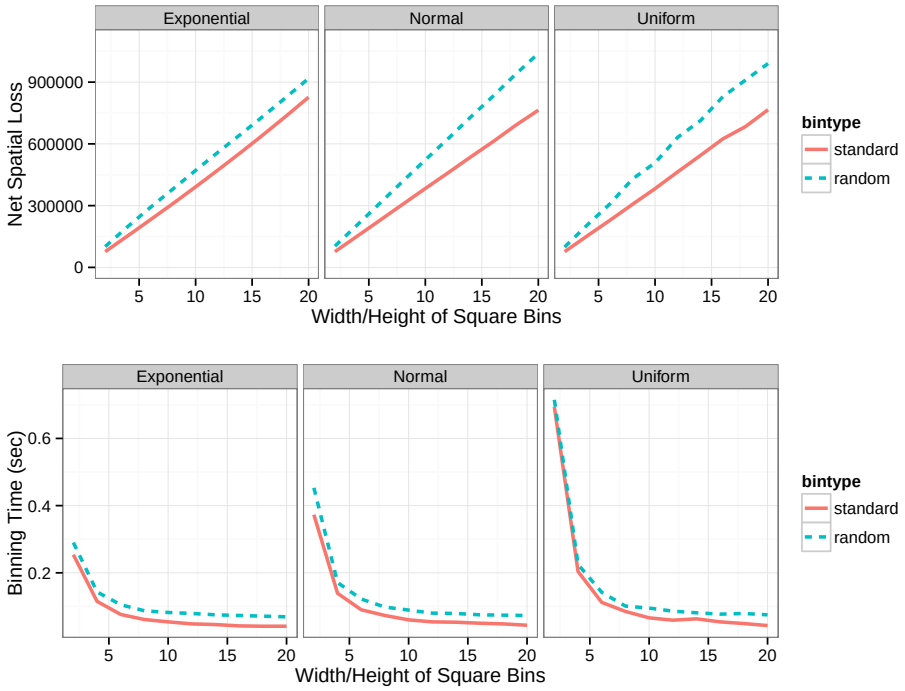


Figure 4.7: Lineplots for net spatial loss and computation times over a range of bin sizes for standard and random binning of the fine version of the simulated data from each bivariate distribution.

All rectangular binning algorithms require the specification of a bin width for each dimension and a binning origin. While using smaller bin widths leads to smaller spatial losses, reducing bin sizes comes at the cost of computation time; a potentially non-negligible consideration in settings with truly massive data. Figure 4.7 also shows that the computation time needed to bin the simulated data sets is a decreasing function of the bin size and that random binning is marginally slower than standard binning across all bin sizes. Computation times for binning were collected using a commercial Asus laptop with an Intel Core i7 processor running at 2.80 GHz. Another pertinent argument for using larger bins is that it will smooth over larger regions if we are primary interested visualizing the large scale density structure.

If we view the binned scatterplot as a visual estimator of the bivariate density then we may consider a few desirable properties of estimators: unbiasedness, consistency and efficiency. Binned scatterplots are inherently biased displays of spatial information as they shift visual emphasis from the true location of individual points to the geometric centers of tiles. However, as bin sizes become increasingly fine, the bias decreases as the plot shows more precise spatial locations. When minimal binning occurs – with one unique coordinate pair at the center of each bin – the density estimate is spatially exact. Also, the density estimation more perfectly reflect the bivariate density as the sample size increases; thus making the binned scatterplot a consistent visual estimator. We may also consider minimally binning as a spatially efficient estimator because it minimizes spatial loss in the visual density estimator. It is worth noting that a density estimate requires the combination of spatial and frequency information and that the estimation properties were considered only through the lens of spatial loss. This makes the assumption that frequency information is rendered through a precise continuous mapping of bin frequencies.

While bin widths can be chosen as any positive real value, the selection should be restricted to an integer multiple of the resolution of the data to avoid undesirable visual artifacts in the binned scatterplot. For example, if the bin dimensions are smaller than the resolution of the data, there will be empty rows or columns of bins in the binned scatterplot. These gaps of white space are undesirable because they create visual discontinuity that interferes with the interpretation of bivariate density. The grid of white spaces also creates an optical illusion, the

Hermann-grid illusion (Hermann, 1870; Spillmann, 1994), which makes dark spots appear in the crossings of the lines and additionally interferes with visual evaluation.

There is also a disruptive visual consequence if bin dimensions are not integer multiples of the bivariate data resolution; where bins have systematically different numbers of possible data values that can cause *artificial striping* – an oscillating density pattern imposed by the binning that does not exist in the raw data – to appear in the binned scatterplot.

To demonstrate the importance of properly selecting binwidths we consider the coarse version of the simulated uniform data which is recorded to a resolution of two units in each dimension. Figure 4.8 displays the binned scatterplots under several scenarios. Under standard binning we see the white-space gaps with one-by-one unit bins, vertical and horizontal artificial stripes with five-by-five unit bins, and the appropriate view of an evenly spread density with four-by-four unit bins. Note that random binning is effective at smoothing out the artificial striping patterns when non-integer multiples of the data resolution are selected.

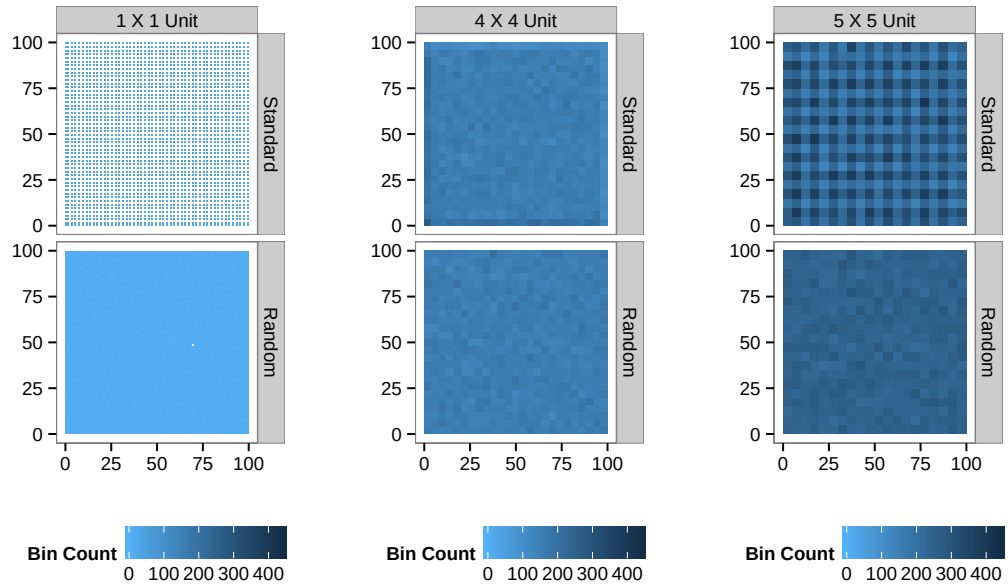


Figure 4.8: Binned scatterplots of coarse uniform data with 1X1, 4X4 and 5X5 square bins. For standard binning, the 1X1 bins leaves white-space gaps between bins, and 5X5 bins cause artificial density stripes; whereas random binning smoothes density over these poor choices in bin dimensions.

The binning origin can also influence the spatial loss in the binned scatterplot. For data with fine resolution compared to the bin dimensions, the origin is only largely consequential for distributions with high density in outermost bins. Figure 4.9 displays the binned scatterplots of ten-by-ten unit bins for the fine exponential data with two different origin choices: (0,0) and (-9,-9). The binning beginning at the (-9,-9) origin suffers visually from the illusion that the density drops off near the lower bounds due to the fact that these lowest bins in each dimension are largely empty due to the negative regions. The binning origin at (0,0) is also superior in terms of the net spatial loss, which is about seven percent lower than for the (-9,-9) origin. For data with very fine resolution, it is a good idea to set the origin at the minimum value for each dimension.

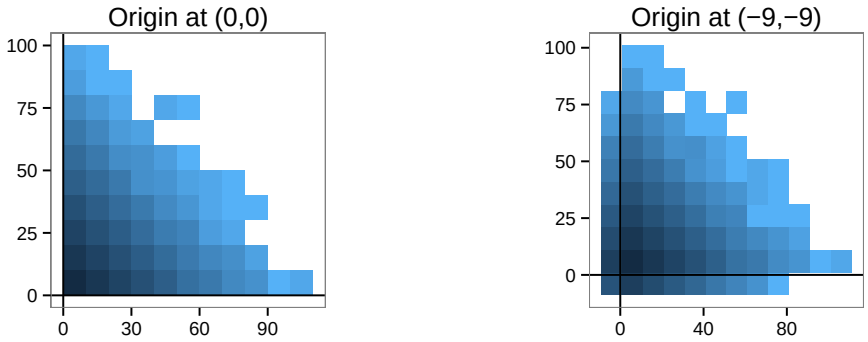


Figure 4.9: Binned scatterplots for the fine exponential data using standard binning with 10X10 square bins with origins at (0,0) and (-9,-9). The bold lines at x=0 and y=0 denote the lower bounds of the data.

For data with a coarse resolution the location of the origin is also important because the origin controls the proximity of possible data values to bin centers. We define the *origin offset* for each dimension, as the tuple, (o_x, o_y), by which we shift the bivariate bin origin $(\beta_{0,x}, \beta_{0,y})$ from the minimal values in x and y:

$$(\beta_{0,x}, \beta_{0,y}) = (x_{(1)}, y_{(1)}) - (o_x, o_y), \tag{4.8}$$

where $x_{(1)}$ and $y_{(1)}$ are the minimal data values in x and y. Thus the offset indicates the number of units in each dimension to shift the binning origin below the origin naturally encouraged by the data.

It can be shown analytically that an origin offset of $(\alpha_x/2,\alpha_y/2)$ units minimizes the net spatial loss in the situation with the following three properties: (i) data are recorded to a resolution of α_x units in the X dimension and α_y units in the Y dimension, (ii) points are symmetric distributed within rectangular bins, (iii) the bin dimensions are integer multiples of α_x and α_y, respectively (see proof in Appendix C.1). In practice the $(\alpha_x/2,\alpha_y/2)$ origin offset is found to be a reasonable choice for lowering spatial loss for symmetric bivariate data at a coarse resolution while using bin dimensions that are integer multiples of α_x and α_y.

Figure 4.10 shows how the net spatial loss changes as the origin offset is shifted while using standard rectangular binning for the coarse simulated data sets. Note that for simplicity, changes to the origin offset are made equally in each dimension, thus shifting the origin at a 45 degree angle. Since the coarse data has a 2X2 unit resolution, we pay special attention to an origin offset of (1,1) in order to assess how well the theoretically optimal origin offset at $(\alpha_x/2,\alpha_y/2)$ works for data that violate the theoretical assumptions to differing degrees. The round glyphs indicate the origin offset where the net loss reaches an absolute minimum in the simulation. For the two symmetrically distributed data sets, normal and uniform, the origin offset of (1,1) was found to either minimize the net spatial loss or achieve a local minimum very near to the overall minimum (within a 0.2% increase from the minimum spatial loss) for each considered bin size. For the bivariately skewed exponential data, the origin offset of (1,1) minimized net loss for the smaller intervals but was not optimal for the largest intervals; 2.5% and 7% above the minimum spatial losses for the 8X8 and 10X10 unit bins, respectively.

4.5.2 Frequency Binning Specifications and Frequency Loss

The reduced binned data from spatial binning contains the center and count information for all bins. The bin frequencies may be mapped continuously to a precisely rendered shade, however it is naive to believe that human perception will be able to perfectly extract that information. There is implicitly loss of frequency information occurring when the shade of

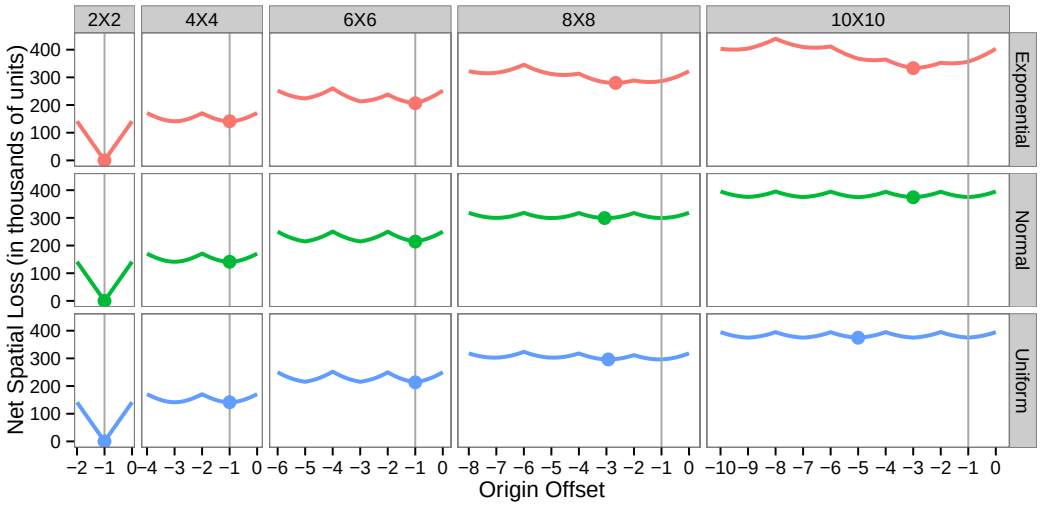


Figure 4.10: Net spatial loss for coarse simulated data at a 2X2 unit resolution using various sized square bins over the range of possible origin offsets. The vertical gray line in each facet indicates the origin offset of (1,1).

a tile is visually mapped back to a frequency through the use of a shading scale index. Bin frequencies may themselves be binned in order to discretize the color scale for the binned scatterplot, thus making the loss explicit.

Figure 4.11 displays binned scatterplots with varying numbers of standard binned frequency groups. If we attempt to discern differences between similarly shaded tiles: it is trivial when only four shades exist, it becomes much more difficult at seven bins, and at ten frequency bins we are hardly able to discriminate better than in continuous shading. This aligns with Healey and Enn's theory on the number of discernible colors (Healey and Enns, 1999b). Our exploration of frequency loss will focus on frequency binning with at most ten bins because above this we experience implicit frequency from perceptual bounds that are not well reflected in the explicitly defined frequency loss.

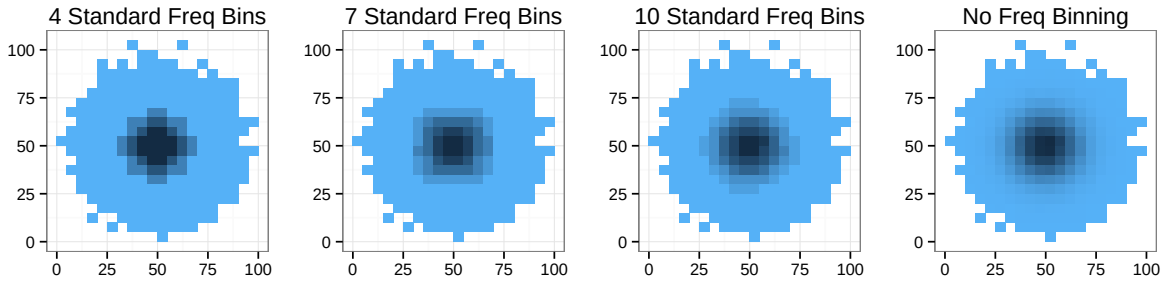


Figure 4.11: Binned scatterplots for the simulated bivarate normal data with varying numbers of standard binned frequency groups. At 10 frequency bins the shades are difficult to distinguish, with little difference from a binned scatterplot with continuous frequency shading.

The loss of frequency information in frequency binning is dependent on the selection of a discrete color mapping, where the binning algorithm and number of frequency bins must be specified. The top row in Figure 4.12 displays the frequency loss from using standard and quantile frequency binning algorithms, with between one and ten frequency bins, from each set of simulated data. We first notice the large difference in the magnitude of frequency losses based on the bivariate distribution; frequency losses are highest for the exponential data, lower for the normal data, and lowest for the uniform data. We also note that the frequency loss is a decreasing function of the number of frequency bins for both binning algorithms. This relationship largely flattens out after the fourth frequency bin, each subsequent bin reducing the frequency loss less than the previous. Thus we should consider using between four and seven bin shades in order to reduce loss while also allowing for easy perception of frequency groups in the binned scatterplot. Lastly, the frequency loss with the standard frequency binning is be higher than for the quantile-based algorithm when a minimal number of bins are used. But in all three simulated data sets, the losses drop to comparable levels for each algorithm when four or more frequency bins are used. Therefore, it is strongly recommended to use the quantile-based algorithm to bin untransformed frequencies.

Due to the difference in scales between counts and log counts, the frequency loss and log frequency loss can not be directly compared. The bottom row of Figure 4.12 displays the

log frequency loss from using standard and quantile algorithms for binning the *log* counts for bins from the same sets of simulated data. Log frequency binning behaved very similarly to standard frequency binning, where log frequency loss decreased as the number of frequency bins increased. The same advice to use between four and seven frequency bins also holds for shading a binned scatterplot based on log counts.

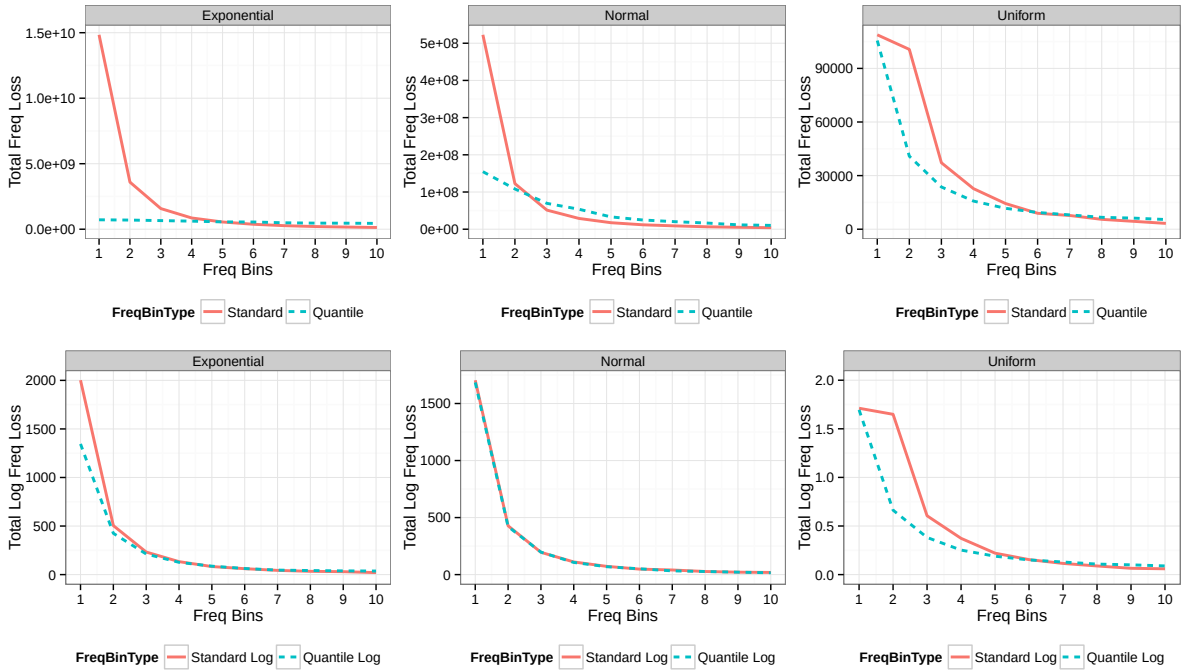


Figure 4.12: Lineplots for total frequency loss (top row) and total log frequency loss (bottom row) from standard and quantile binning of bin counts and log bin counts, respectively, using reduced binned data from each bivariate distribution.

Log transforming the frequencies prior to binning and using quantile based binning on the raw frequencies are two methods for dealing with the same problem for binned scatterplots: heavily right skewed bin counts where dense bins visually overshadow any structure in low density bins. Frequency groups for quantile binning are invariant to the log transformation because a monotone transformation does not affect groupings based on quantiles, thus it is preferable leave the frequencies untransformed before quantile binning for better contextual interpretability. In the common situation where bin frequencies are heavily skewed, the choice is between quantile frequency binning and standard log frequency binning.

Since the loss scales are not comparable, the choice is guided by desired interpretation. Figure 4.13 shows binned scatterplots for the simulated normal data with quantile frequency binning and standard log frequency binning. For standard log frequency binning, the shade is to be interpreted as an ordinal indicator based on equally spaced groupings of log bin frequencies; whereas for quantile frequency binning the shade denotes groups based on frequency quantiles. This is analogous to the difference in interpreting a histogram of log transformed data and a boxplot of untransformed data in univariate visualization. Both shading schemes are effective at reducing the visual impact of the highest density bins near the center of each plot, allowing for the differences in the surrounding bin frequencies to be emphasized. It should be noted that both frequency binning algorithms produced very similar looking binned scatterplots, but this will not always be the case. It occurred in this scenario because the distribution of log bin frequencies from the bivariate normal data is nearly uniform, thus the standard log frequency binning aligns closely with groupings from the quantile frequency binning.

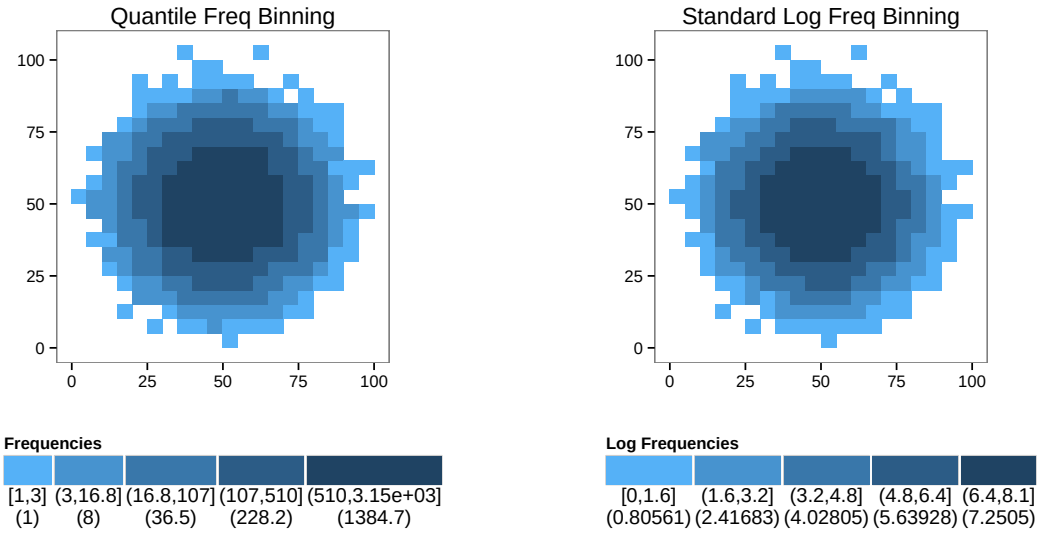


Figure 4.13: Simulated bivariate normal data spatially binned using standard algorithm with 5X5 unit bins and origin at (0,0). Binned scatterplots with five frequency bins from quantile frequency binning (left) and standard log frequency binning (right).

4.6 Discussion and Examples

The exploration of spatial and frequency loss in the process of creating a binned scatterplot yields a number of important properties and practical recommendations for their construction. These recommendations and loss properties will now be demonstrated through two examples using real bivariate data. First we further explore the baseball pitching data. In a second example, we will investigate the relationship between scheduled and actual departure times for American commercial airline data.

4.6.1 Binning Loss in Baseball Data: Strikeout and Game Counts

We now revisit the baseball data used earlier and construct binned scatterplots to display the relationship between strikeouts and game counts. We begin by specifying the bin dimensions and origin for standard rectangular binning. The data are counts, each variable is therefore recorded as an integer, corresponding to a one game by one strike-out data resolution. We use integer bin widths in each dimension of spatial bins to avoid artificial density stripes. The data has only 42,583 observations, so we can bin on a fine grid without taking an inordinate amount of computation time. We use approximately 50 bins in each dimension; the data range from 1 to 106 games and from 0 to 513 strike-outs, thus we use bins that are two games wide and ten strikeouts high. The origin will be set to a half unit below the minimum value in each dimension, at $(0.5, -0.5)$.

The leftmost plot in Figure 4.14 displays the binned scatterplot with a continuous shading of raw bin frequencies. The most striking feature in the frequency distribution is the dark spot in the bottom right of the plot representing a large number of pitchers that played very few games and had very few strikeouts. It is nearly impossible to distinguish the density structure across the remaining bins, representing better pitchers who played many games and earned many strikeouts. In this case there are two frequency binning options that we can employ to deal with this skewness of the distribution of bin counts. Quantile frequency binning using four shade groups (center plot of Figure 4.14) allows us to visualizes the quartiles of the bin densities. Alternatively, standard log frequency binning using four shade groups (rightmost

plot of Figure 4.14) uses the log transformation to diminish the visual impact of high density bins. Each frequency binning approach adds a layer of complexity to interpreting the shade, however they both effectively emphasize the forked ridges in the density structure.

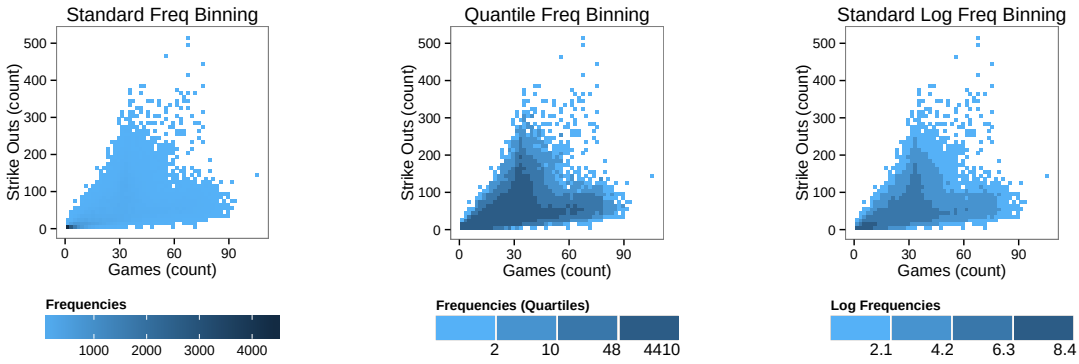


Figure 4.14: Binned scatterplots for games versus strikeouts.

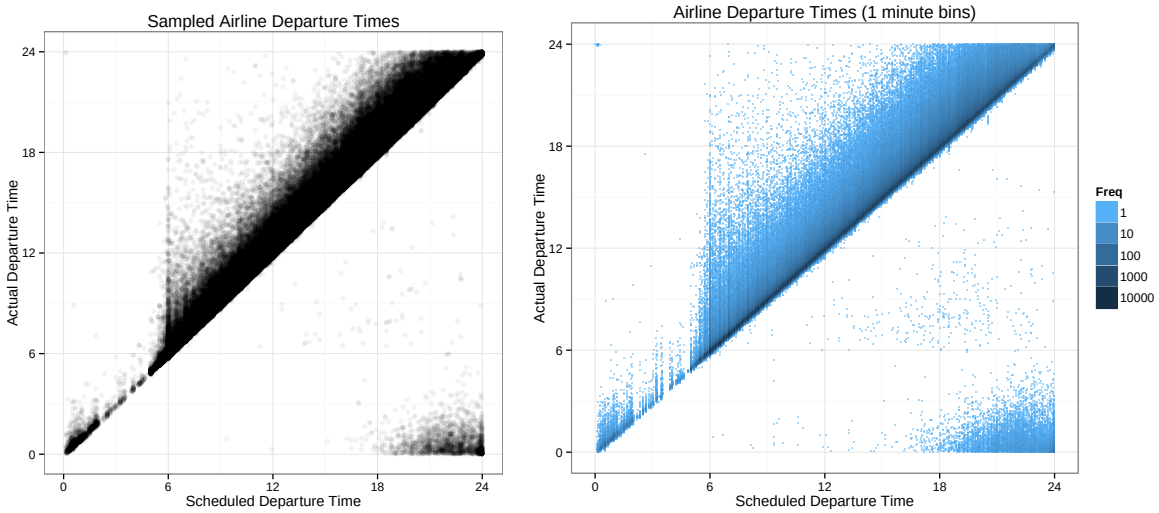
The interpretability of the net spatial loss in this scenario suffers because the units and scales of the two variables differ. To remedy this, the two variables can be standardized then binned equivalently by rescaling the bin dimension, where the net spatial loss is interpreted as a distance in units of standard deviations. The net spatial loss for the standardized data is the same for each plot in Figure 4.14 because they do not differ in location, only in frequency shading. The net spatial loss is 2595.33 standard deviations, an average of 0.0609 standard deviation per data point. We also find that we have approximately 3% lower net spatial loss, on the standardized scale, for our binning specification using the recommended origin offset than if we were to naively set the binning origin at the minimum data values of one game and zero strikeouts.

4.6.2 Big Data: Airline Departure Times

The Federal Aviation Association (FAA) requires all airlines based in the United States to report details for every single flight. These are published online by the Bureau of Transportation Services at `http://www.transtats.bts.gov/DataIndex.asp`. Every day there are about 16,000 flights across the United States adding up to almost 6 Million flights a year. Scheduled and actual departure times for all flights in 2011 make up –in uncompressed form–

a file of about 450 MB. A comparison of scheduled and actual departure times allows us an investigation of on-time performance of air carriers.

Figure 4.15 shows two plots of the relationship between scheduled versus actual departure times. The plot on the left shows the scatterplot from a sample of one million of those records. Even while using alpha blending this results in a severely over-plotted graph. On the right is a minimally binned scatterplot of the reduced data from standard binning at 1-minute intervals. If the origin is offset by a half minute for both departures and arrivals, this binning does not have any spatial loss, as all observations will be centered within bins. The 1-minute standard binning also reduces the data to 176,384 reduced data triple, less than 3% of the original data.



(a) Million-point sample of Airline Data

(b) Minimal binning of Airline Data

Figure 4.15: Scheduled and actual departure times of flights across the United States in 2011. The plot on the left is based on a sample of the data, the plot on the right shows all flights. The large scale distributional patterns are visible in both plots, but the plot on the left misses some of the finer level details in scheduling that is visible in the plot on the right.

Both plots show the same large scale distributional patterns: scheduled and actual arrival times are highly correlated, recognizable from the conglomeration of points along the identity line. Scheduled departure times past 6:00 am in the morning are much more common than earlier flights. It is much more likely for a flight to be delayed than to leave early, leading to the wash-out effect above the line, that is getting thinner with increasing delays. The range of delays on usual days starts at about one hour at 6:00 am and increases during the day to

about a two hour delay. The small number of flights before 6:00 am are also visible in both plots. The triangle of observations on the bottom right in both plots is nothing but an artifact of the data collection consisting of flights that are scheduled before midnight, but are delayed to departures past midnight. The cloud of outliers halfway between the two main structures is potentially interesting, since no immediate explanation comes to mind, and would be worthy of a follow-up investigation.

What is not apparent in the alpha-blended scatterplot, is some fine-level structure that the minimally binned scatterplot shows. Note that because we have bin widths equal to the resolution to which the data is recorded in each dimension, we know that the spatial binning algorithm will not cause any *artificial* density stripes. A close inspection of the plot on the right hand side reveals darker colored vertical lines at 30 minute intervals. It is obvious that more flights are scheduled with departures on the hour and at 30 minutes past the hour.

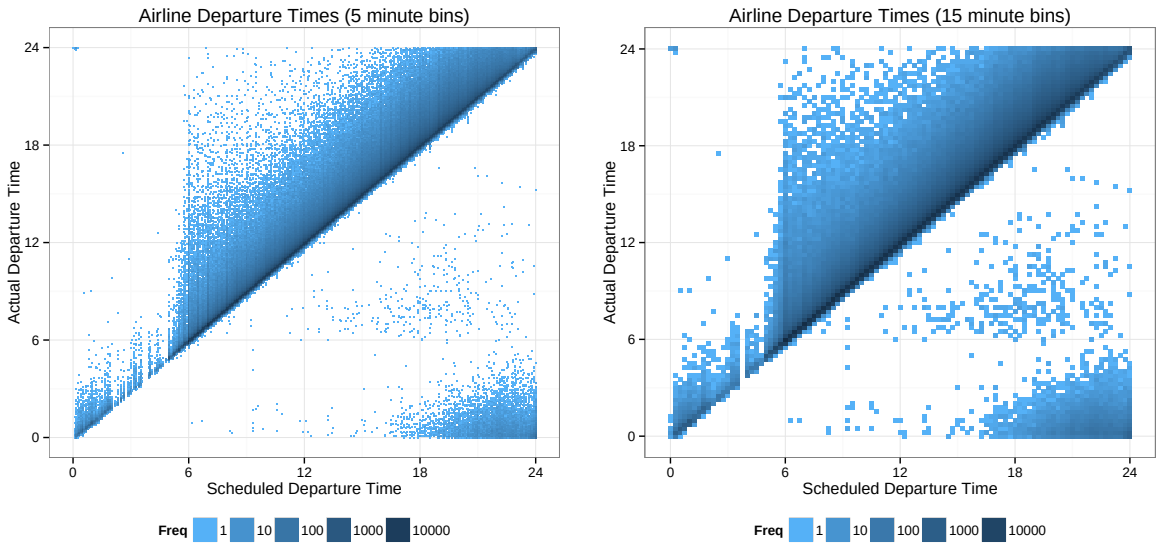


Figure 4.16: 5-minute bins produce a higher-level summary of the data than shown in Figure 4.15b. 15-minute bins produce an even more coarse summary of the data.

Figure 4.16 displays the binned scatterplots for the flight departure data with larger bins. Binning data by five-minute intervals produces a more high-level summary of the relationship between actual and scheduled departure time, though it necessarily obscures some of the finer details. In addition, binning data by five minute intervals reduces the size of the data set to a

much more manageable 19,787 reduced binned data triples, which can be easily manipulated on probably any modern computer. Binning by 15-minute intervals reduces the data set to a nearly-trivial 3,575 reduced binned data triples, but the graphical summary becomes granular and less appealing at that resolution.

4.7 Conclusions and Future Work

Large bivariate data sets are very difficult to visualize in raw form, due to over-plotting of points. Binning allows for the visualization and manipulation of large data sets, and easily translates into binned scatterplots which are more appropriate for the human visual system. Reducing the data for binned scatterplots has distinct computational and visual advantages, however the aggregation comes at the cost of losing precision in spatial and frequency information.

We have presented two algorithms for spatially binning data points; standard and random rectangular binning algorithms. The random binning algorithm displayed strong advantage of avoiding the problem of artificial stripes that occur when data recorded to a coarse resolution are binned using a bin width that was a non-integer multiple of the data resolution. However, the standard binning algorithm is superior due to lower spatial loss. For data with a coarse resolution (α_x units in the X dimension and α_y units in the Y dimension) artificial stripes in the standard binning process can be avoided, if bin dimensions are chosen as integer multiples of α_x and α_y. We were also able to show through simulation that a reasonable default for the binning uses an origin offset of $(\alpha_x/2, \alpha_y/2)$ because it resulted in minimal or near minimal spatial loss for symmetric data and performed well even for spatially skewed data.

Spatial binning with smaller bin dimensions will lead to lower spatial losses; however, finer binning requires more processing time and does not highlight large scale density structure. It is left to the plot designer to decide how much spatial information they are willing to sacrifice in order to simplify the display of density structure.

If we elect to use frequency binning to shade tiles ordinally, it is recommended to use between four and seven distinct shades. This is done to minimize the frequency loss within the bounds of human perceptual ability to distinguish multiple shades simultaneously. Using quantile

frequency binning or standard log frequency binning are shown to be reasonable methods – with slight differences of interpretability – for handling situations with heavily skewed bin counts.

Future implementations of software for constructing binned scatterplots would be well served to allow for choices in specification of binning algorithms. The findings of this research provide suggestions for reasonable default settings of binning parameters that maintain spatial and frequency information and lead to desirable visual properties in the binned scatterplot.

CHAPTER 5. CONCLUSIONS

The goal at the onset of this dissertation was to overcome challenges and grasp opportunities associated with large data and emerging technology to advance the fields of statistics education and statistical graphics. The preceding chapters have demonstrated a successful pursuit of this goal through the review of existing literature to guide relevant research topics, the work to address difficult challenges in both fields, and the application of advanced statistical methods and analysis to draw actionable insights on each topic. To conclude, we review the primary results from each of the studies.

The study on traditional and simulation-based statistical inference curricula took an ambitious and statistically rigorous approach to comparing learning outcomes through a designed experiment. The randomization of students to curricula helped to isolate the curricular effect by creating homogeneous cohorts of students which received each curriculum. The student learning outcomes, measured using the ARTIST scaled questions sets for confidence intervals and hypothesis tests, were analyzed using a bivariate MANCOVA model which controls for pre-treatment factors, including scores from a midterm and lab. The model found no significant curriculum effect for learning outcomes related to hypothesis tests; despite the added complexity in the simulation-based curriculum. The model also indicates a statistically significant 7% improvement on the ARTIST scale for the simulation-based curriculum over the traditional curriculum for learning outcomes related to confidence intervals, while accounting for pre-treatment covariates. However, the validity of these results is highly questionable because the model assumption of independence between students is likely untrue. The simulation study demonstrated that the Type I error rates in tests for curricular effects inflate dramatically under minor violations that have lecture or lab section based variance structure. Therefore,

caution must be taken in interpreting the results of this study which are suggestive, but not conclusive, of a simulation-based curriculum benefit.

The Shiny Database Sampler stands as a successful proof of concept that big data sources can be incorporated into an introductory statistics curriculum. The web-based interface is designed to allow students to easily obtain samples from large databases, using random sampling methodology to infuse the software with pedagogical value. Course activities for an introductory statistics course were also developed to incorporate and leverage the Shiny Databases Sampler for learning about sampling concepts. The software was assessed through a student user survey designed to measure the ease of use, connection to sampling concepts and the engagement with the data. The survey indicates that students find the Shiny Database Sampler easy to use, notice the connection to sampling concepts and are moderately engaged with the data context.

The research on loss related to the construction of binned scatterplots yielded interesting exploration of loss properties and several recommendations for binning specification. Traditional scatterplots are fundamentally unable to scale for large data due to over-plotting which obscures frequency information. Binned scatterplots trade individual points for shaded tiles, sacrificing exact data locations to allow frequency information to be displayed. The spatial loss is measured as the sum of Euclidean distances from points to the bin centers. Bin frequency information can be rendered precisely through continuous shading, however the perceptual ability to extract that information is imperfect. We propose the use of frequency binning to make this imprecision explicit, thus allowing the measurement of frequency loss. The exploration of loss properties provides a few main recommendations for binning specifications: bin dimensions should be integer multiples of the data resolution, it is often beneficial to offset the binning origin by half the data resolution in each dimension, quantile and standard log frequency binning are two viable strategies for dealing with heavily skewed bin frequencies, and four to seven frequency groups should be used when frequency binning is employed.

This research contributes to methodology in statistics education and statistical graphics. While the research goal of this dissertation was successfully pursued, there remain many challenges and opportunities within these fields that provide abundant avenues for future work.

Bibliography

Agresti, A. and Franklin, C. (2012), *Statistics: The Art and Science of Learning From Data*, Upper Saddle River, NJ: Pearson.

Albert, J. (2009), "Discrete Bayes in R," *Technology Innovations in Statistics Education*, 3, 1–16.

— (2014), *LearnBayes: Functions for Learning Bayesian Inference*, r package version 2.15.

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., and Witmer, J. (2005), "Guidelines for Assessment and Instruction in Statistics Education: College Report," .

American Statistical Association (2014), "Journal of Statistics Education (JSE) Data Archives," http://www.amstat.org/publications/jse/jse_data_archive.htm, accessed: 09/03/2014.

ARTIST (2006), apps3.cehd.umn.edu/artist/tests/index.html, accessed: 09/08/2014.

ATLAS (2014), "Applied Technologies for Learning in the Arts & Sciences (ATLAS) Data Repositories," http://www.atlas.illinois.edu/services/stats/archive/repositories/, accessed: 09/04/2014.

Baglin, J., Bedford, A., and Bulmer, M. (2013), "Students' Experiences and Perceptions of Using a Virtual Environment for Project-Based Assessment in an Online Introductory Statistics Course," *Technology Innovations in Statistics Education*, 7, 1–15.

Baglin, J. and DaCosta, C. (2013), "Comparing Training Approaches for Technological Skill Development in Introductory Statistics Courses," *Technology Innovations in Statistics Education*, 7, 1–23.

Berton, M. F. and Vallencillo, A. (2002), "Quality Attributes for COTS Components," *I+ D Computacion, 1 (2)*, 128–143.

Bevan, N. (1997), "Quality and usability: a new framework," *Achieving software product quality, van Veenendaal, E, and McMullan, J (eds)*.

Blackboard Inc. (2014), http://www.blackboard.com, accessed: 09/03/2014.

Bokhove, C. and Drijvers, P. (2010), "Digital tools for algebra education: Criteria and evaluation," *International Journal of Computers for Mathematical Learning*, 15, 45–62.

Boon, P. (2009), "A designer speaks: Designing educational software for 3D geometry," *Educational Designer*, 1.

Brown, D. (2015), *Tracker Video Analysis and Modeling Tool Version 4.87*.

Bryan, J. (2006), "Technology for physics instruction," *Contemporary Issues in Technology and Teacher Education*, 6, 230–245.

Budgett, S., Pfannkuch, M., Regan, M., and Wild, C. (2013), "Dynamic Visualisation and the Randomization Test," *Technology Innovations in Statistics Education*, 7, 1–21.

Bulmer, M. and Haladyn, K. J. (2011), "Life on an Island: a Simulated Population to Support Student Projects in Statistics," *Technology Innovations in Statistics Education*, 5, 1–20.

Burrill, G., Allison, J., Breaux, G., Kastberg, S., Leatham, K., and Sanchez, W. (2002), *Handheld graphing technology in secondary mathematics*, Texas Instruments.

Carini, R. M., Kuh, G. D., and Klein, S. P. (2006), "Student engagement and student learning: Testing the linkages*," *Research in higher education*, 47, 1–32.

Carlson, K. and Winquist, J. (2011), "Evaluating an Active Learning Approach to Teaching Introductory Statistics: A Classroom Workbook Approach," *Journal of Statistical Education*, 19, 1–23.

Carr, D., Littlefield, R., Nicholson, W., and Littlefield, J. (1987), "Scatterplot Matrix Techniques for Large N," *Journal of the American Statistical Association*, 82, 424–436.

Carver, R. (2011), "Introductory Statistics Unconstrained by Computability: A New Cobb Salad," *Technology Innovations in Statistics Education*, 5, 1–14.

CATALST (2012), "Change Agents for Teaching and Learning Statistics (CATALST) Materials," www.tc.umn.edu/~catalst/materials, accessed: 09/06/2014.

CAUSE (2014), "Consortium for the Advancement of Undergraduate Statistics Education Resourses," https://www.causeweb.org/resources/, accessed: 09/03/2014.

Center for Disease Control (2014), "Center for Disease Control (CDC)," www.cdc.gov, accessed: 09/04/2014.

Chance, B., Ben-Zvi, D., Garfield, J., and Medina, E. (2007), "The Role of Technology in Improving Student Learning of Statistics," *Technology Innovations in Statistics Education*, 1, 1–26.

Cleveland, W. (1987), "Research in Statistical Graphics," *Journal of the American Statistical Association*, 82, 419–423.

Cleveland, W. S. and McGill, R. (1984a), "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American statistical association*, 79, 531–554.

— (1984b), "The Many Faces of a Scatterplot," *Journal of the American Statistical Association*, 79, 807–822.

Cobb, G. (2007), "The Introductory Statistics Course: A Ptolemaic Curriculum," *Technology Innovations in Statistics Education*, 1, 1–15.

Cook, T. (2002), "Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for not Doing Them," *Educational Evaluation and Policy Analysis*, 24, 175–199.

Cronbach, L. (1951), "Coefficient alpha and the internal structure of tests," *Psychometrika*, 16, 297–334.

DASL Project (1996), "The Data and Story Library (DASL)," http://lib.stat.cmu.edu/DASL/, accessed: 09/03/2014.

DelMas, R., Garfield, J., Ooms, A., and Chance, B. (2007), "Assessing Students' Conceptual Understanding After a First Course in Statistics," *Statistics Education Research Journal*, 6, 28–58.

DeMars, C. (2010), *Item response theory*, Oxford University Press.

Demiralp, C., Bernstein, M., and Heer, J. (2014), "Learning perceptual kernels for visualization design," .

Desire2Learn (2014), http://www.brightspace.com/, accessed: 09/03/2014.

Drijvers, P. et al. (2012), "Digital technology in mathematics education: Why it works (or doesn't)," in *12th International Congress on Mathematical Education, Seoul*.

Everson, M. G. and Garfield, J. (2008), "An Innovation Approach to Teaching Online Statistics Courses," *Technology Innovations in Statistics Education*, 2, 1–18.

Few, S. (2008), "Solutions to the Problem of Over-plotting in Graphs," *Visual Business Intelligence Newsletter*.

Finzer, W., Erickson, T., Swenson, K., and Litwin, M. (2007), "On Getting More and Better Data into the Classroom," *Technology Innovations in Statistics Education*, 1, 1–10.

Flick, L. and Bell, R. (2000), "Preparing tomorrow's science teachers to use technology: Guidelines for science educators," *Contemporary issues in technology and teacher education*, 1, 39–60.

Forbes, S. (2012), "Data Visualization: A motivational and Teaching Tool in Official Statistics," *Technology Innovations in Statistics Education*, 6, 1–19.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., and Schaffer, R. (2005), "Guidelines for Assessment and Instruction in Statistics Education: A Pre-K-12 Curriculum Framework," .

Freiman, V. (2014), "Technology Design in Mathematics Education," in *Encyclopedia of Mathematics Education*, Springer, pp. 605–610.

Freudenthal Institute (2014), "Digital Math Environment," http://www.fi.uu.nl/dwo/en/, accessed: 5/1/2015.

Friborg, O., Martinussen, M., and Rosenvinge, J. (2006), "Likert-based vs. semantic differential-based scoring of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience," *Personality and Individual Differences*, 40, 873–884.

Friedman, J. H. (1997), "Data mining and statistics: What's the connection," in *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics*.

Friendly, M. and Denis, D. (2005), "The early origins and development of the scatterplot," *Journal of the History of the Behavioral Sciences*, 41, 103–130.

Furnham, A. (1986), "Response Bias, Social Desirability and Dissimulation," *Personality and Individual Differences*, 7, 385–499.

Gage, M., Pizer, A., and Roth, V. (2002), "WeBWorK: Generating, delivering, and checking math homework via the Internet," in *ICTM2 international congress for teaching of mathematics at the undergraduate level, Hersonissos, Crete, Greece. http://www. math. uoc. gr/˜ ictm2/Proceedings/pap189. pdf.*

Gang, X. (2015), "WWW Interactive Multipurpose Server," http://wims.unice.fr/wims/en_home.html, accessed: 5/1/2015.

Garfield, J., delMas, R., and Zieffler, A. (2012), "Developing statistical modelers and thinking in an introductory, tertiary-level statistics course," *The International Journal on Mathematics Education*, 44, 883–898.

George, D. and Mallery, P. (2003), *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update (4th Edition)*, Boston, MA: Allyn & Bacon.

Goldstein, E. B. (2007), *Sensation & Perception*, Belmont, CA: Thomason Wadsworth.

Green, K. (2013), "The Campus Computing Project: A National Survey of Computation and Information Technology," .

Hakenholz, E. (2010), "Compass and Ruler Metal Geometry," http://db-maths.nuxit.net/CaRMetal/index_en.html, accessed: 5/1/15.

Hao, M. C., Dayal, U., Sharma, R. K., Keim, D. A., and Janetzko, H. (2010), "Visual Analytics of Large Multidimensional Data Using Variable Binned Scatter Plots," in *IS&T/SPIE Electronic Imaging*, International Society for Optics and Photonics, pp. 753006–753006.

Harlow, J., Ashman, B., and Sainudiin, R. (2009), "Extending Galton's Binomial Quincunx to the Trinomial Septcunx," *Technology Innovations in Statistics Education*, 3, 1–16.

Harraway, J. (2012), "Learning Statistics Using Motivational Videos, Real Data and Free Software," *Technology Innovations in Statistics Education*, 6, 1–21.

Hassad, R. A. (2013), "Faculty Attitude towards Technology-Assisted Instruction for Introductory Statistics in the Context of Educational Reform," *Technology Innovations in Statistics Education*, 7, 1–17.

Healey, C. G. and Enns, J. T. (1999a), "Large datasets at a glance: Combining textures and colors in scientific visualization," *Visualization and Computer Graphics, IEEE Transactions on*, 5, 145–167.

— (1999b), "Large Datasets at a glance: combining textures and colors in scientific visualization," *IEEE Transactions on Visualization and Computer Graphics*, 5, 145–167.

Heer, J. and Bostock, M. (2010), "Crowdsourcing graphical perception: using mechanical turk to assess visualization design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 203–212.

Hermann, L. (1870), "Eine Erscheinung simultanen Contrastes," *Archiv für die gesamte Physiologie des Menschen und der Tiere*, 3, 13–15.

Herschel, J. (1833), "On the investigation of the orbits of revolving double stars," *Memoirs of the Royal Astronomical Society*, 5, 171–222.

Hoff, S., Heiny, R., and Perrett, J. (2012), "An Exploration of the Exact Distribution and Probability for Sample Means for Small Random Samples," *Technology Innovations in Statistics Education*, 6, 1–19.

Howe, K. (2004), "Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for not Doing Them," *Educational Evaluation and Policy Analysis*, 10, 42–61.

IBM Corp. (2013a), *IBM SPSS Statistics for Windows, Version 22.0*, Armonk, NY.

— (2013b), "Many Eyes," www.ibm.com/manyeyes, accessed: 09/03/2014.

ICOTS9 (2014), "International Conference on Teaching Statistics 9: Program," http://icots.info/9/programme.php, accessed: 09/06/2014.

ICPSR (2014), "Inter-university Consortium for Political and Social Research," www.icpsr.umich.edu, accessed: 09/04/2014.

Jackiw, N. (2002), "The Geometer's Sketchpad v4. 0," .

Jacobs, A. (2009), "The Pathologies of Big Data," *Communications of the ACM*, 52, 36–44.

James, D. A. and DebRoy, S. (2012), *RMySQL: R interface to the MySQL database*, r package version 0.9-3.

Janetzko, H., Hao, M., Mittelstadt, S., Dayal, U., and Keim, D. (2013), "Enhancing Scatter Plots Using Ellipsoid Pixel Placement and Shading," in *2013 46th Hawaii International Conference on System Sciences (HICSS)*, pp. 1–10.

Kaplan, J. (2011), "Innovative Activities: How Clickers can Facilitate the Use of Simulation in Large Lecture Classes," *Technology Innovations in Statistics Education*, 5, 1–14.

Kaput, J. J. and Thompson, P. W. (1994), "Technology in mathematics education research: The first 25 years in the JRME," *Journal for research in mathematics education*, 676–684.

Keim, D., Hao, M., Dayal, U., Janetzko, H., and Bak, P. (2010), "Generalized Scatter Plots," *Information Visualization*, 9, 301–311.

Kistner, E. and Muller, K. (2004), "Exant Distributions of Intraclass Correlation and Cronbach's Alpha with Gaussian Data and General Covariance," *Psychometrika*, 69, 459–474.

Konold, C. and Kazak, S. (2008), "Reconnecting Data and Chance," *Technology Innovations in Statistics Education*, 2, 1–37.

Lane-Getaz, S. (2013), "Development of a Reliable Measure of Students' Inferential Reasoning Ability," *Statistics Education Research Journal*, 12, 20–35.

Lee, H. and Lee, T. (2009), "Reasoning about Probabilistic Phenomia: Lessons Learned and Applied in Software Design," *Technology Innovations in Statistics Education*, 3, 1–21.

Liu, Z., Jiang, B., and Heer, J. (2013), "*imMens*: Real-Time Visual Querying of Big Data," in *Eurographics Conference on Visualization (EuroVis)*, International Society for Optics and Photonics, vol. 32.

Lock, R., Lock, P., K.L.Morgan, Lock, E., and D.F.Lock (2013), *Statistics: Unlocking the Power of Data*, Hoboken, NJ: Wiley.

MATLAB (2014), *version 8.3.0 (R2014a)*, Natick, Massachusetts.

McDaniel, S. and Green, L. (2012a), "Independent Interactive Inquiry-Based Leaning Modules Using Audio-Visual Instruction in Statistics," *Technology Innovations in Statistics Education*, 6, 1–20.

— (2012b), "Using Applets and Video Instruction to Foster Students' Understanding of Sampling Variability," *Technology Innovations in Statistics Education*, 6, 1–17.

McGowan, H. (2011), "Planning a Comparative Experiment in Educational Settings," *Journal of Statistics Education*, 19, 1–19.

McGowan, H. and Gunderson, B. (2010), "A Randomized Experiment Exploring how Certain Features of Clicker Use Effect Undergraduate Student's Engagement and Learning in Statistics," *Technology Innovations in Statistics Education*, 4, 1–29.

Microsoft Corp. (2013), *Microsoft Excel*, Redmond, WA.

Minitab, Inc. (2010), *Minitab 17 Statistical Software*, State College, PA.

Moodle (2014), https://moodle.org/, accessed: 09/03/2014.

Moore, D. S. (1997), "New Pedagogy and New Content: The Case of Statistics," *International Statistical Review*, 65, 123–165.

Muller, D. A., Sharma, M. D., and Reimann, P. (2008), "Raising cognitive load with linear multimedia to promote conceptual change," *Science Education*, 92, 278–296.

National Highway Transportation Safety Administration (2015), "National Highway Transportation Safety Administration (NHTSA)," www.nhtsa.gov, accessed: 04/27/2014.

NCTM, P. (2000), "Standards for School Math," *Nat'l Council of Teachers of Math.*

Neumann, D., Hood, M., and Neumann, M. (2013), "Using Real-Life Data When Teaching Statistics: Student Perceptions of this Strategy in an Introductory Statistics Course," *Statistics Education Research Journal*, 12, 59–70.

Nicholson, J., Ridgeway, J., and McCluster, S. (2013), "Getting Real Statistics into all Curriculum Subject Areas: Can Technology Make this a Reality," *Technology Innovations in Statistics Education*, 7, 1–16.

Nielsen, J. (1994), "Usability inspection methods," in *Conference companion on Human factors in computing systems*, ACM, pp. 413–414.

NORC (2014), "General Social Survey," http://www3.norc.org/GSS+Website/, accessed: 09/04/2014.

Nunnally, J. and Bernstein, I. (1978), *Psychometric Theory (Third Edition)*, New York, New York: McGraw Hill.

Ooms, A. and Garfield, J. (2008), "A Model to Evaluate Online Educational Resourses in Statistics," *Technology Innovations in Statistics Education*, 2, 1–17.

Oracle Corp. (2014), *MySQL*, Oracle Corp.

Oviatt, S. (2006), "Human-centered design meets cognitive load theory: designing interfaces that help people think," in *Proceedings of the 14th annual ACM international conference on Multimedia*, ACM, pp. 871–880.

PASCO (2015), "PASPORT Probeware," .

Pearson, K. (1895), "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material," *Philosophical Transactions of the Royal Society of London*, 186, 343–414.

PhET (2015), "PhET Interactive Simulations," https://phet.colorado.edu/, accessed: 5/3/2015.

Playfair, W. (1786), *Commercial and Political Atlas*, London.

Playfair, W., Wainer, H., and Spence, I. (2005), *Playfair's Commercial and Political Atlas and Statistical Breviary*, Cambridge University Press.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ragasa, C. (2008), "A Comparison of Computer-Assisted Instruction and the Traditional Method of Teaching Basic Statistics," *Journal of Statistics Education*, 16, 1–10.

Richter, C. and Kroschewski, B. (2012), "Geostatistical models in agricultural field experiments: investigations based on uniformity trials," *Agronomy Journal*, 104, 91–105.

Richter-Gebert, J. and Kortenkamp, U. (2013), "The Interactive Geometry Software Cinderella Version 2.8," http://www.cinderella.de/, accessed: 5/4/2015.

Ridgeway, J., Nicholson, J., and McCluster, S. (2013), "'Open Data' and the Semantic Web Required a Rethnk on Statistics Teaching," *Technology Innovations in Statistics Education*, 7, 1–12.

Rios, J. M. and Madhavan, S. (2000), "Guide to adopting technology in the physics classroom," *The Physics Teacher*, 38, 94–97.

Roper Center (2014), "Roper Center Public Opinion Archives," www.ropercenter.uconn.edu/, accessed: 09/04/2014.

RStudio and Inc. (2014), *shiny: Web Application Framework for R*, r package version 0.9.1.

Rubin, A. (2007), "Much Has Changed; Little Has Changed: Revisiting the Role of Technology in Statistics Education 1992-2007," *Technology Innovations in Statistics Education*, 1, 1–33.

Sage Developement Team (2015), *Sage Version 6.6*.

SAS Institute Inc. (2014), *SAS/STAT Software, Version 9.1*, Cary, NC.

Schmidt, K. K. (2013), "Virtual Discussion for Real Understanding: The Use of an Online Discussion Board in and Introductory Biostatistics Course," *Technology Innovations in Statistics Education*, 7, 1–14.

Schwartz, B., Zaitsev, P., and Tkachenko, V. (2012a), *High Performance MySQL*, Sebastopol, CA: O'Reilly.

— (2012b), *High performance MySQL: Optimization, backups, and replication*, " O'Reilly Media, Inc.".

Scott, D. (1979), "On Optimal and Data-Based Histograms," *Biometrika*, 66, 605–610.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, John Wiley & Sons.

Shaltayev, D., Hodges, H., and Hasbrouck, R. (2010), "VISA: Reducing Technological Impact on Student Learning in an Introductory Statistics Course," *Technology Innovations in Statistics Education*, 4, 1–20.

Shutes, K. (2009), "A Note on Using Individualised Data Sets for Statistics Coursework," *Technology Innovations in Statistics Education*, 3, 1–9.

Smyth, G. (2011), "Australasian Data and Story Library (OzDASL)," http://www.statsci.org/data, accessed: 09/03/2014.

— (2014), "Statistical Science Web Data Sets," http://www.statsci.org/datasets.html, accessed: 09/04/2014.

Spillmann, L. (1994), "The Hermann Grid Illusion: a Tool for Studying Human Perceptive Field Organization," *Perception*, 23, 691–708.

StataCorp LP (2013), *Stata Statistical Software: Release 13*, College Station, TX.

Stohl Drier, H., Harper, S., Timmerman, M. A., Garofalo, J., and Shockey, T. (2000), "Promoting appropriate uses of technology in mathematics teacher preparation," .

Texas Instruments (2015), "Texas Instruments Calculator-Based Laboratory 2," .

The College Board (2014), "AP Statistics: Calculator Policy," apstudent.collegeboard.org/apcourse/ap-statistics/calculator-policy, accessed: 09/05/2014.

Theus, M. (2006a), "Scaling Up Graphics," in *Graphics of Large Datasets*, New York: Springer.

— (2006b), "Scaling Up Graphics," in *Graphics of Large Datasets*, New York: Springer.

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., and Vanderstoep, J. (2014), *Introduction to Statistical Investigations*, Hoboken, NJ: Wiley.

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V., and Swanson, T. (2012), "Retention of Statistical Concepts in a Preliminary Randomization-Based Introductory Statistics Curriculum," *Statistics Education Research Journal*, 11, 21–40.

Tintle, N., VanderStoep, J., Holmes, V., Quisenberry, B., and Swanson, T. (2011), "Development and Assessment of Preliminary Randomization-Based Introductory Statistics Curriculum," *Journal of Statistics Education*, 19, 1–25.

Trouche, L. and Drijvers, P. (2010), "Handheld technology for mathematics education: Flashback into the future," *ZDM*, 42, 667–681.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Lebanon, IN: Addison Wesley.

Unwin, A., Theus, M., and Hofmann, H. (2006), *Graphics of Large Datasets*, New York: Springer.

U.S. Census Bureau (2014), "U.S. Census Bureau," www.census.gov, accessed: 09/04/2014.

U.S. General Services Administration (2014), "data.gov," www.data.gov, accessed: 09/04/2014.

U.S.C. (1996), "FOIA UPDATE: THE FREEDOM OF INFORMATION ACT, 5 U.S.C. SECT. 552, AS AMENDED BY PUBLIC LAW NO. 104-231, 110 STAT. 3048," .

Vernier Software and Technology (2015), *LoggerPro Version 3.9*.

Waterloo Maple Inc. (2015a), *Maple 2015*, Waterloo, ON.

— (2015b), *Maple T.A. 10*, Waterloo, ON.

WHO (2014), "World Health Organization Multi-Country Studies Data Archive," http://apps.who.int/healthinfo/systems/surveydata/index.php/catalog, accessed: 09/04/2014.

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, New York: Springer.

— (2013), "Bin-Summarize-Smooth: A Framework for Visualising Large Data," Tech. rep.

Wijers, M., Jonker, V., and Drijvers, P. (2010), "MobileMath: exploring mathematics outside the classroom," *ZDM*, 42, 789–799.

Wilkinson, L., Rope, D., Carr, D., and Rubin, M. (2000), "The Language of Graphics," *Journal of Computational and Graphical Statistics*, 9, 530–543.

Williams, A. (2012), "Online Homework vs. Traditional Homework; Statistics Anxiety and Self-Efficacy in an Educational Statistics Course," *Technology Innovations in Statistics Education*, 6, 1–19.

Wolfram (2015), *Mathematica Version 10.1*, Champaign, IL.

Wolfram Alpha LLC (2015), "Wolfram Alpha," http://www.wolframalpha.com/, accessed: 5/1/2015.

World Bank (2014), "World Bank Open Data," data.worldbank.org/, accessed: 09/04/2014.

Xanthopoulos, D. (2010), *Physion Simulation Software Version 1.01*.

Xie, Y. (2013), *Dynamic Documents with R and knitr*, Boca Raton, FL: Chapman and Hall/CRC.

Ziegler, L. (2014), "Reconceptualizing Statistical Literacy: Developing an Assessment for the Modern Statistics Course," Ph.D. thesis, University of Minnesota.

APPENDIX A.

A.1 Appendix: Final Exam Used in Curricula Study

The following appendix contains the questions and point rubrics for the ARTIST scaled questions and the applied theory-based inference problems as they appeared on the final exam.

A.1.1 ARTIST Scaled Multiple Choice Question Set for Confidence Intervals

1. Answer the following general multiple choice questions regarding confidence intervals. There is only one correct answer for each (circle the best option).

i. Two different samples will be taken from the same population of test scores where the population mean and standard deviation are unknown. The first sample will have 25 data values, and the second sample will have 64 data values. A 95% confidence interval will be constructed for each sample to estimate the population mean. Which confidence interval would you expect to have greater precision (a smaller width) for estimating the population mean?

a. I expect the confidence interval based on the sample of 64 data values to be more precise.

b. I expect both confidence intervals to have the same precision.

c. I expect the confidence interval based on the sample of 25 data values to be more precise.

ii. A 95% confidence interval is computed to estimate the mean household income for a city. Which of the following values will definitely be within the limits of this confidence interval?

a. The population mean

b. The sample mean

c. The standard deviation of the sample mean

d. None of the above

iii. Each of the 110 students in a statistics class selects a different random sample of 35 Quiz scores from a population of 5000 scores they are given. Using their data, each student constructs a 90% confidence interval for μ the average Quiz score of the 5000 students. Which of the following conclusions is correct?

a. About 10% of the sample means will not be included in the confidence intervals.

b. About 90% of the confidence intervals will contain μ.

c. It is probable that 90% of the confidence intervals will be identical.

d. About 10% of the raw scores in the samples will not be found in these confidence intervals

iv. A 95% confidence interval for the mean reading achievement score for a population of third grade students is (43, 49). The margin of error of this interval is:

a. 5

b. 3

c. 6

v. Justin and Hayley conducted a mission to a new planet, Planet X, to study arm length. They took a random sample of 100 Planet X residents and calculated a 95% confidence interval for the mean arm length. What does a 95% confidence interval for arm length tell us in this case? Select the best answer:

a. I am 95% confident that this interval includes the sample mean (x) arm length.

b. I am confident that most (95%) of all Planet X residents will have an arm length within this interval.

c. I am 95% confident that most Planet X residents will have arm lengths within this interval.

d. I am 95% confident that this interval includes the population mean arm length.

vi. Suppose that a random sample of 41 state college students is asked to measure the length of their right foot in centimeters. A 95% confidence interval for the mean foot length for students at this university turns out to be (21.709, 25.091). If instead a 90% confidence interval was calculated, how would it differ from the 95% confidence interval?

a. The 90% confidence interval would be narrower.

b. The 90% confidence interval would be wider.

c. The 90% confidence interval would be the same as the 95% confidence interval.

vii. A pollster took a random sample of 100 students from a large university and computed a confidence interval to estimate the percentage of students who were planning to vote in the upcoming election. The pollster felt that the confidence interval was too wide to provide a precise estimate of the population parameter. What could the pollster have done to produce a narrower confidence interval that would produce a more precise estimate of the percentage of all university students who plan to vote in the upcoming election?

a. Increase the sample size to 150.

b. Increase the confidence level to 99%.

c. Both a and b

d. None of the above

viii. A newspaper article states with 95% confidence that 55% to 65% of all high school students in the United States claim that they could get a hand gun if they wanted one. This confidence interval is based on a poll of 2000 high school students in Detroit. How would you interpret the confidence interval from this newspaper article?

a. 95% of large urban cities in the United States have 55% to 65% high school students who could get a hand gun.

b. If we took many samples of high school students from different urban cities, 95% of the samples would have between 55% and 65% high school students who could get hand guns.

c. You cannot use this confidence interval to generalize to all teenagers in the United States because of the way the sample was taken.

d. We can be 95% confident that between 55% and 65% of all United States high school students could get a hand gun.

ix. The Gallup poll (August 23, 2002) reported that 53% of Americans said they would favor sending American ground troops to the Persian Gulf area in an attempt to remove Hussein from power. The poll also reported that the margin of error for this poll was 4%. What does the margin of error of 4% indicate?

a. There is a 4% chance that the estimate of 53% is wrong.

b. The percent of Americans who are in favor is probably higher than 53% and closer to 57%.

c. The percent of Americans who are in favor is estimated to be between 49% and 57%.

x. Suppose two researchers want to estimate the proportion of American college students who favor abolishing the penny. They both want to have about the same margin of error to estimate this proportion. However, Researcher 1 wants to estimate with 99% confidence and Researcher 2 wants to estimate with 95% confidence. Which researcher would need more students for her study in order to obtain the desired margin of error?

a. Researcher 1.

b. Researcher 2.

c. Both researchers would need the same number of subjects.

d. It is impossible to obtain the same margin of error with the two different confidence levels.

A.1.2 ARTIST Scaled Multiple Choice Question Set for Hypothesis Testing

2. Answer the following general multiple choice questions regarding hypothesis testing. There is only one correct answer for each (circle the best option). As a helpful note the term "statistically significant" means that you reject the null hypothesis.

i. The makers of Mini-Oats cereal have an automated packaging machine that is set to fill boxes with 24 ounces of cereal. At various times in the packaging process, a random sample of

100 boxes is taken to see if the machine is filling the boxes with an average of 24 ounces of cereal. Which of the following is a statement of the null hypothesis being tested?

a. The machine is filling the boxes with the proper amount of cereal.

b. The machine is not filling the boxes with the proper amount of cereal.

c. The machine is not putting enough cereal in the boxes.

ii. A research article gives a p-value of .001 in the analysis section. Which definition of a p-value is the most accurate?

a. the probability that the observed outcome will occur again.

b. the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true.

c. the value that an observed outcome must reach in order to be considered significant under the null hypothesis.

d. the probability that the null hypothesis is true.

iii. If a researcher was hoping to show that the results of an experiment were statistically significant they would prefer:

a. a large p-value

b. a small p-value

c. p-values are not related to statistical significance

iv. A researcher compares men and women on 100 different variables using a difference in means t-test. He sets the level of significance at 0.05 and then carries out 100 independent t-tests (one for each variable) on these data. If, for each test, the null hypothesis is actually true, about how many "statistically significant" results will be produced?

a. 0

b. 5

c. 10

d. none of the above

Problems (v) and (vi) refer to the following situation: Food inspectors inspect samples of food products to see if they are safe. This can be thought of as a hypothesis test where Null: the food is safe, and Alternative: the food is not safe. Identify each of the following statements as a Type I or a Type II error.

v. The inspector says the food is safe but it actually is not safe.

a. Type I

b. Type II

vi. The inspector says the food is not safe but is actually safe.

a. Type I

b. Type II

vii. A newspaper article claims that the average age for people who receive food stamps is 40 years. You believe that the average age is less than that. You take a random sample of 100 people who receive food stamps, and find their average age to be 39.2 years. You find that this is significantly lower than the age of 40 stated in the article (p-value ¡ .05). What would be an appropriate interpretation of this result?

a. The statistically significant result indicates that the majority of people who receive food stamps is younger than 40.

b. Although the result is statistically significant, the difference in age is not of practical importance.

c. An error must have been made. This difference is too small to be statistically significant.

viii. A newspaper article stated that the US Supreme Court received 812 letters from around the country on the subject of whether to ban cameras from the courtroom. Of these

812 letters, 800 expressed the opinion that cameras should be banned. A statistics student was going to use this sample information to conduct a test of significance of whether more than 95% of all American adults feel that cameras should be banned from the courtroom. What would you tell this student?

a. This is a large enough sample to provide an accurate estimate of the American public's opinion on the issue.

b. The necessary conditions for a test of significance are not satisfied, so no statistical test should be performed.

c. With such a large number of people favoring the notion that cameras be banned, there is no need for a statistical test.

ix. A researcher conducts an experiment on human memory and recruits 15 people to participate in her study. She performs the experiment and analyzes the results. She obtains a p-value of .17. Which of the following is a reasonable interpretation of her results?

a. This proves that her experimental treatment has no effect on memory.

b. There could be a treatment effect, but the sample size was too small to detect it.

c. She should reject the null hypothesis.

d. There is evidence of a small effect on memory by her experimental treatment.

x. It is reported that scores on a particular test of historical trivia given to high school students are approximately normally distributed with a mean of 85. Mrs. Rose believes that her 5 classes of high school seniors will score significantly better than the national average on this test. At the end of the semester, Mrs. Rose administers the historical trivia test to her students. The students score an average of 89 on this test. After conducting the appropriate statistical test, Mrs. Rose finds that the p-value is .0025. Which of the following is the best interpretation of the p-value?

a. A p-value of .0025 provides strong evidence that Mrs. Rose's class outperformed high school students across the nation.

b. A p-value of .0025 indicates that there is a very small chance that Mrs. Rose's class outperformed high school students across the nation.

c. A p-value of .0025 provides evidence that Mrs. Rose is an exceptional teacher who was able to prepare her students well for this national test.

d. None of the above.

A.1.3 Applied Theory-Based Confidence Interval Question

3. Farmer Cindy is in charge of the chickens on her family's farm, and is curious about the average number of eggs the entire flock produces in a month. Observing the entire flock would be time consuming, so she approaches you asking what her options are. You inform her that 30 chickens should be selected at random to be observed for a month. After the month she observes the average number of eggs her sample of 30 chickens produced is 25 eggs with a standard deviation of 6 eggs.

1. (5 pts) Construct a 95% confidence interval for the average number of eggs chickens from her entire population produce in a month. (You don't need to check any conditions here)

2. (3 pts) Interpret the 95% confidence interval constructed in the previous part:

3. (2 pts) Cindy is disappointed with the width of the interval you provide for her, suggest to her two ways she could obtain a narrower confidence interval.

A.1.4 Applied Theory-Based Hypothesis Testing Question

4. Cindy becomes concerned about a disease some of her chickens are catching that causes a decrease in the chicken's egg production. She is interested in the proportion of her entire flock that has the disease, but detecting the disease requires taking blood from the chicken which is expensive and time consuming. After working with you in the past, she understands that she can estimate this proportion by taking just a sample of her chickens! Assume she selects a sample of 100 chickens in the best possible way and observed that 15 of the chickens had the disease.

Cindy's pessimistic guess is that 25% of the flock has the disease. Complete the following steps to test if the proportion of her entire population diseased is less than 0.25.

1. (2 pts) State the Null and Alternative hypothesis:
2. (2 pts) Check the conditions for a hypothesis test:
3. (2 pts) Calculate the test statistic:
4. (2 pts) Find the p-value:
5. (3 pts) Make a decision about your hypothesis and state your conclusion in context of the problem.

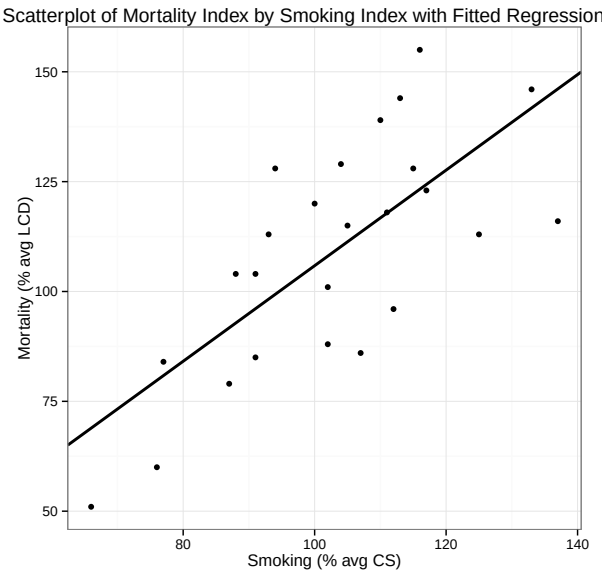
A.2 Appendix: Midterm Exam Used in Curricula Study

The following appendix is the midterm exam in its entirety, along with point designations for each question. Note that Problem 6 for the midterm is the ARTIST scaled question set for assessing topics related to "data collection". This ARTIST set was included with the goal to strengthen the midterm as a effective covariate for controlling for pre-treatment differences in the curricula groups in the model comparing learning outcomes using the ARTIST scores.

1. [28 points] Linear Regression: A study was conducted in England in 1989 investigating the relationship between smoking and lung cancer mortality within different occupational groups of males. The data include a smoking index and a lung cancer mortality index for men in 25 occupational groups in England.

The smoking index measures the percentage of the number of cigarettes smoked per day by men in the particular occupational group compared to the average number of cigarettes smoked per day by all men. The units associated with the smoking index are percent of average cigarettes smoked (% avg CS). The mortality index measures the percentage of the rate of deaths from lung cancer among men in the particular occupational group compared to the rate of deaths from lung cancer among all men. The units associated with the mortality index are percent of average lung cancer deaths (% avg LCD).

The data for the 25 occupations are displayed in the scatterplot below. Use the information from the scatterplot, the sample statistics and linear regression equation listed below to answer all parts of this problem.



$\bar{x} = 102.88\%$ avg CS	$\bar{y} = 109\%$ avg LCD	r=0.7162
$S_x = 17.1982$ % avg CS	$S_y = 26.1135\%$ avg LCD	$R^2 = 0.5129$

Prediction Equation: $\hat{y} = -2.8853189 + x(1.0875323)$

(a) [2 points] What is the response variable?

(b) [2 points] What is the explanatory variable?

(c) [4 points] The correlation coefficient, r=0.7162. Using this value, what can be said about the relationship between the smoking index and the lung cancer mortality index?

(d) [4 points] The slope of the linear regression model is 1.0875323. Show the calculations to prove that this is the value for the slope. Then, interpret this value in the context of the problem.

(e) [4 points] The intercept of the linear regression model is −2.8853189. First, show the calculations to prove that this is the value for the intercept. If it is appropriate to interpret this value in the context of the problem, then do so. Otherwise, specify why it is inappropriate to interpret this value.

(f) [3 points] In the prediction equation for the regression line explain that the symbols $\hat{y}$ and x represent.

(g) [3 points] Using the prediction line given above, predict the morality index (in % avg LCD) for an occupational group that has a smoking index of 110% avg CS (round your answer to 2 decimal places).

(h) [3 points] Suppose that there was an occupational group in our data set with a smoking index of 110% avg CS and a mortality index of 140% avg LCD. Calculate the residual for this occupational group (round your answer to 2 decimal places).

(i) [3 points] Interpret the R^2 value in terms of the context of the problem.

2. [10 points] A Gallup poll conducted on February 26, 2014 investigated the wellbeing of American adults. The survey used the Gallup-Healthways Well-Being Index to classify the person's wellbeing as "thriving", "struggling" or "suffering". The telephone survey found that 55% of 1500 randomly selected American adults were classified as thriving.

For this study, identify the following:

(a) [2 points] Identify the population of interest.

(b) [2 points] Identify the sample.

(c) [2 points] Identify the population parameter.

(d) [2 points] Identify the sample statistic.

(e) [2 points] Identify the variable.

3. [10 points] A study investigated the relationship between heights of husbands and wives. Heights were recorded as groupings of "Tall", "Medium" and "Short". The researcher recorded height pairings for 205 married couples and wants to know how the height of the husband associated the height of the wife. Use the information found in the contingency table below to answer the questions that follow.

(a) [2 points] What is the probability that a randomly selected married couple will have a short husband?

		Wife			
		Tall	Medium	Short	Total
Husband	Tall	18	28	14	60
	Medium	20	51	28	99
	Short	12	25	9	46
	Total	50	104	51	205

(b) [2 points] What is the probability that a randomly selected married couple will have a tall wife *and* a short husband?

(c) [2 points] What is the probability that a randomly selected married couple will have a tall wife *or* a short husband?

(d) [2 points] What is the probability that a randomly selected married couple will have a short husband *given* that we know the wife is tall?

(e) [2 points] Are the events of short husband and tall wife *independent*? Justify your answer mathematically.

4. [8 points] Suppose that we have a box containing 23 marbles of equal size and shape so that each marble is equally likely to be selected from the box. There are 10 red, 8 green, 3 yellow marbles and 2 black marbles. Suppose we assign points to each color marble: 2 points for a red, 5 point for a green, 10 points for a yellow and 15 points for a black. Let X be the discrete random variable that counts the number of points that a randomly selected marble is worth.

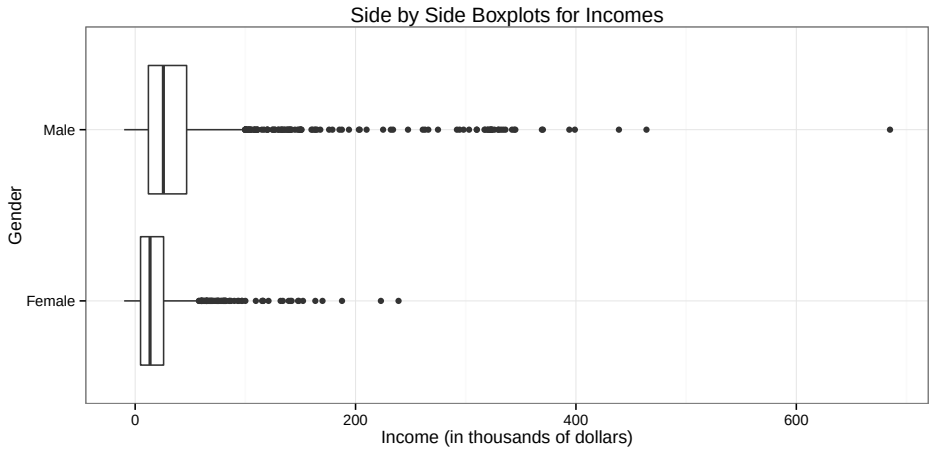
(a) [2 points] Find the probability distribution of X? (fill out table and round all values to 2 decimal places)

x	P(x)

(b) [4 points] What 2 properties must be true for the values in the P(x) column of the probability distribution in part 4(a)? Note that X is a discrete random variable.

(c) [2 points] What is the mean of X? (If you were unable to complete part 4(a), describe how you would find the mean)

5. [10 points] A survey on gender and income was conducted and a random sample of 3700 Adult U.S. residents was gathered. In the figure below are side-by-side boxplots of the income (in thousands of dollars) for male and female U.S. residents. Use this graphical display to answer the following questions.



Gender	mean	sd	min	Q1	median	Q3	max
Male	39319	54242	-10000	12000	25600	46700	685000
Female	18786	21570	-10000	4880	13500	25800	239000

(a) [5 points] Using full sentences, describe the distribution of income for *only* the male residents of the U.S.

(b) [5 points] Using full sentences, compare the distribution of male and female distributions of income. A complete answer will compare shapes, centers and spreads. Be specific about which measures of center and spread are being used to make these comparisons. (ie writing "the spread is bigger" will not earn full credit)

6. [18 points] Multiple Choice: *Clearly* circle the selected answer.

(i) [2 points] In a survey people are asked 'Which brand of toothpaste do you prefer?' The data gathered from this question would be what type of data?

a. continuous

b. categorical

c. quantitative

Items ii and iii refer to the following situation: A student is gathering data on the driving experiences of other college students. One of the variables measured is the type of car the student drives. These data are coded using the following method: 1 = subcompact, 2 = compact, 3 = standard size, 4 = full size, 5 = premium, 6 = mini van, 7 = SUV, and 8 = truck.

(ii) [2 points] What type of variable is this?

a. categorical

b. quantitative

c. continuous

(iii) [2 points] The student plans to see if there is a relationship between the number of speeding tickets a student gets in a year and the type of vehicle he or she drives. Identify the response variable in this study.

a. college students

b. type of car

c. number of speeding tickets

d. average number of speeding tickets last year

(iv) [2 points] A researcher is studying the relationship between a vitamin supplement and cholesterol level. What type of study needs to be done in order to establish that the amount of vitamin supplement causes a change in cholesterol level?

a. Survey

b. Randomized experiment

c. Time Series Study

d. Survey

(v) [2 points] An instructor is going to model an experiment in his statistics class by comparing the effect of 4 different treatments on student responses. There are 40 students in the class. is the best way for the instructor to distribute the students to the 4 treatments for this experiment?

a. Assign the first treatment to the first 10 students on the class list, the second treatment to the next 10 students, and so on.

b. Assign a unique number to each student, then use random numbers to assign 10 students to the first treatment, 10 students to the second treatment, and so on.

c. Assign the treatment as students walk into class, giving the first treatment to the first 10 students and the second treatment to the next 10 student, and so on.

d. All of these are equally appropriate methods.

e. None of these is an appropriate method.

Items vi and vii refer to the following situation:

Suppose two researchers wanted to determine if aspirin reduces the chance of a heart attack.

(vi) [2 points] Researcher 1 studied the medical records of 500 randomly selected patients. For each patient, he recorded whether the person took aspirin every day and if the person had ever had a heart attack. Then he reported the percentage of heart attacks for the patients who took aspirin every day and for those who did not take aspirin every day. What type of study did Researcher 1 conduct?

a. Observational

b. Experimental

c. Survey

d. None of the above

(vii) [2 points] Researcher 2 also studied 500 patients that visited a regional hospital in the last year. He randomly assigned half (250) of the patients to take aspirin every day and the other half to take a placebo everyday. Then after a certain length of time he reported the percentage of heart attacks for the patients who took aspirin every day and for those who did not take aspirin every day. What type of study did Researcher 2 conduct?

a. Observational

b. Experimental

c. Survey

d. None of the above

(vii) [2 points] The dean of a college would like to determine the feelings of students concerning a new registration fee that would be used to upgrade the recreational facilities on campus. All registered students would pay the fee each term. Which of the following data collection plans would provide the best representation of students' opinions at the school?

a. Survey every 10th student who enters the current recreational facilities between the hours of 1:00 and 5:00 pm until 100 students have been asked.

b. Randomly sample fifty student ID numbers and send a survey to all students in the sample.

c. Place an ad in the campus newspaper inviting students to complete an online survey. Collect the responses of the first 200 students who respond.

d. All of the above would be equally effective.

(ix) [2 points] A team in the Department of Institutional Review at a large university wanted to study the relationship between completing an internship during college and students' future earning potential. From the same graduating class, they selected a random sample of 80 students who completed an internship and 100 students who did not complete an internship and examined their salaries 5 years past graduation. They found that there was a statistically higher mean salary for the internship group than for the non-internship group. Which of the following interpretations do you think is the most appropriate?

a. More students should take internships because having an internship produces a higher salary.

b. There could be a confounding variable, such as student major, that explains the difference in mean salary between the internship and no internship groups.

c. You cannot draw any valid conclusions because the samples are not the same size.

7. [**8 points**] When playing the game Settlers of Catan**TM** there is a game piece known as the robber that blocks resources from being obtained. On each player's turn, that player rolls a pair of 6 sided dice. If the sum of the dice roll equals 7 then that player may move

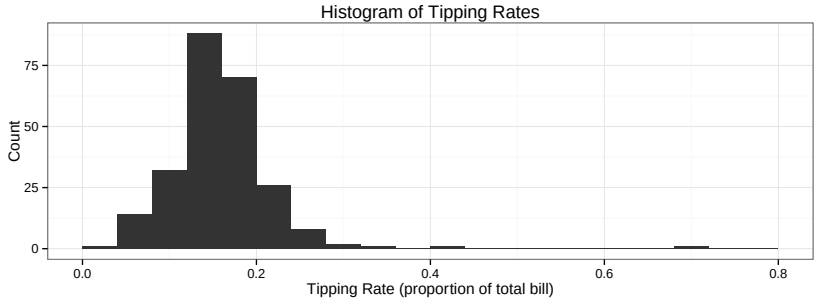
the robber. This means that there is a 6/36 chance of moving the robber with every roll of the dice, because there are 6 of the 36 possible combinations that sum to 7. Let X count the number of times that one player moves the robber in 12 turns.

(a) [4 points] Check the 4 conditions to argue that X is a Binomial Random Variable

(b) [2 points] What is the probability that a player gets to move the robber 2 times during the 12 turns?

(c) [2 points] What is the mean number of times during 12 turns that a player will be able to move the robber?

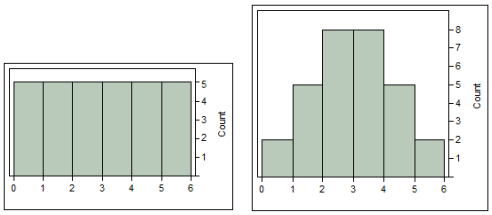
8. [4 points] A waiter from a restaurant keeps track of the information about tips he receives from dinning parties. Part of this data includes the tipping rate, as a proportion of the total bill, that was tipped. Below is a histogram of tipping rates from the 244 dinning parties served by the waiter. Use this histogram to answer the following question.



(a) [2 points] What is the best measure of center to describe the distribution of tipping rates? Explain your answer. (Note: no calculations or estimated values are needed to answer this problem)

(b) [2 points] What is the best measure of spread to describe the distribution of tipping rates? Explain your answer. (Note: no calculations or estimated values are needed to answer this problem)

9. **[4 points]** Below are histograms for two separate random samples of 30 quiz scores. Which quiz will have the larger sample standard deviation for quiz scores? **Explain your choice briefly**. (Note: You do not, and should not, do any calculations to answer this question)



Quiz 1 Scores Quiz 2 Scores

A.3 Appendix: MANCOVA Model Diagnostics

For the ARTIST Model described in Subsection 2.4.1 and the Applied Model described in Subsection 2.4.2 a set of model diagnostics were conducted. Each of these MANCOVA models are parameterized as specified in Section 2.4; as such the assessment of model assumptions will be very similar for each model. Residual plots will be used to assess the assumptions of linearity and constant variance. The univariate and bivariate normality of the error terms will be assessed visually using normal quantile plots and a scatterplot of the paired residuals from the model. Additionally, although it is not a modeling assumption, MANCOVA models are best behaved with low to moderate correlation between the response variables, because then then model is able to capture variance unique to each response.

A.3.1 ARTIST Model Diagnostics

The ARTIST Model is assessed to satisfactorily meet the modeling assumptions. The correlation between the ARTIST scores for confidence intervals and hypothesis tests is acceptable for modeling with MANCOVA with a correlation of 0.288 between the responses. The normality of the errors is upheld by the plots in Figure A.1. The normal quantile plots only display a slight bend for the residuals from the Hypothesis Test topic scores, and the bivariate distribution of the residuals are visually consistent with bivariate normality.

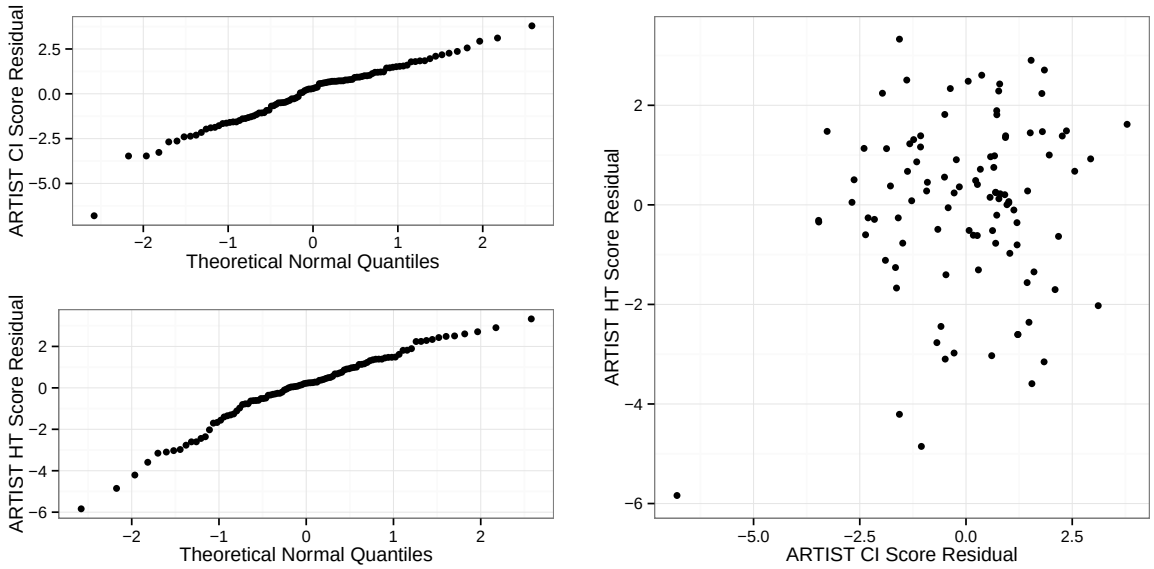


Figure A.1: Normal quantile plots (left) and bivariate scatterplot (right) for residuals of each response from ARTIST Model.

The residual plots in Figure A.2 display stripped bands of points that run shallowly downward, creating the optical illusion of a trend in the points. This is because the model fits the discretely recorded ARTIST scores as continuous response variables, thus there is a discrete set of residuals possible for any particular fitted value. Note that, as with all least-squares regression models, the residuals are uncorrelated with the fitted values.

The residual plots in Figure A.2 are overlain with Loess smoothers – and corresponding 95% confidence envelopes – to check for violations of linearity. There is no issue with the assumption of linearity in the prediction of the confidence interval score, but there is a slight significant dip in the pattern for hypothesis interval residuals. The assumption of homoscedasticity appear to hold with the residuals spread fairly evenly at all levels of the fitted values. There are however a few outliers on the residual plots, specifically a few students who scored abnormally lower than predicted. These outliers were investigated and found to be low leverage and non-influential.

A.3.2 Applied Model Diagnostics

A correlation of 0.546 between the pair of applied problem scores for confidence intervals and hypothesis tests is acceptable for modeling with MANCOVA. The assumption of normality

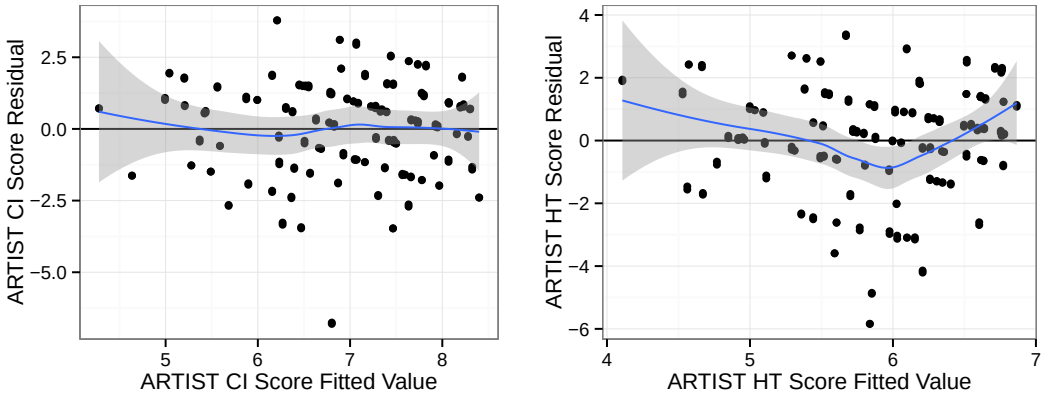


Figure A.2: ARTIST Model residual plots overlaid with Loess smoother and corresponding 95% confidence envelopes.

of errors in the Applied Model is potentially problematic. The normal quantile plot for the residuals from the applied confidence interval score in Figure A.3 show a distinct curve. The scatterplot of the residual pairs from the Applied Model also appear to have a non-normal bivariate distribution.

The Loess smoothers do not significantly departing from a the horizontal lines at zero for the residual plots in Figure A.4 and thus no departures from the assumption of linearity. The residuals do however show signs of changing variance over the range of the fitted values. This appears to be driven by the upper bound on student scores for each question.

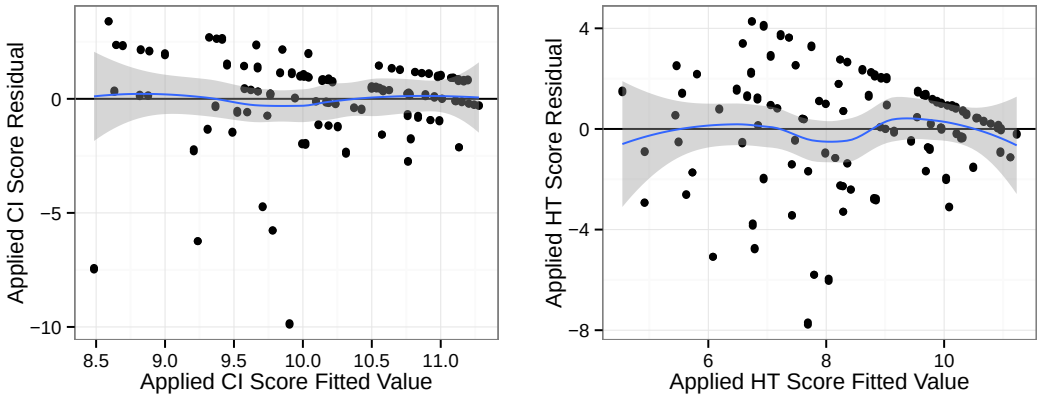


Figure A.4: Applied Model residual plots overlaid with Loess smoother and corresponding 95% confidence envelopes.

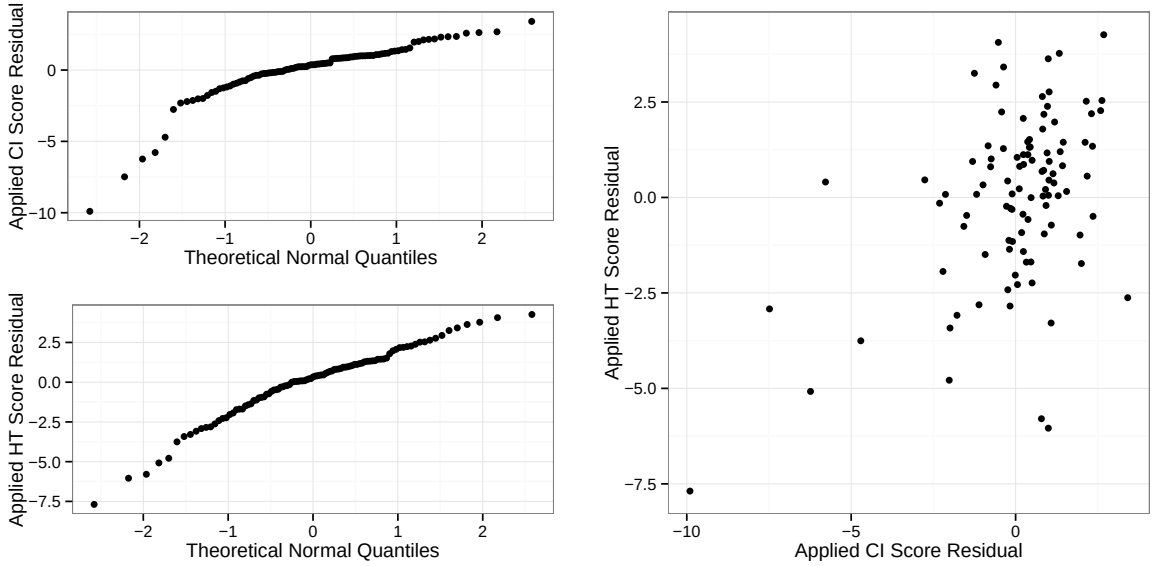


Figure A.3: Normal quantile plots (left) and bivariate scatterplot (right) for residuals of each response from Applied Model.

APPENDIX B.

B.1 Appendix: Lab Assignment Using Shiny Database Sampler

For this activity you will be using a tool called the Shiny Database Sampler to take a random sample of United States residents from US census data. The census data is the Public Use Microdata Sample (PUMS) which is a 3 million person subset of the entire Census data. For this activity we treat our samples as though they are selected from the full census records.

We are going to explore how these random sampling plans relate to the goals of a sample survey. The tool will allow you to define either a simple random sampling plan or a stratified random sampling plan. In the following two scenarios we will explore the advantages and disadvantages of these two sampling plans. Access the tool at http://shiny.stat.iastate.edu/karstenm/ShinyDatabaseSampler.

Scenario 1: Suppose that our goal is to estimate the mean age of all US residents. Similar to polling organizations we have a budget that allows us to survey around 1000 people. To collect our sample we decide to take a simple random sample of 1040 US residents.

(a) Is this study and example of an experiment or an observational study? Explain your answer.

(b) Your colleague Bob claims that we are wasting our budget to get only 1040 people using random sampling. He says that we could get 20000 responses to the survey if we invested that money into a mailing campaign in Minneapolis. Explain why the random selection is important.

(c) Another colleague, Jill, asks why we do not stratify by state when we take the sample so that we get 20 people from each of the 50 states along with Puerto Rico and the District

of Columbia. Explain why this idea would not create a representative sample to pursue our goal.

Now that we have decided on our sampling plan, let's go collect our data. The Shiny Database Sampler needs to be told 4 pieces of information in order to collect census records the way you want. (1) Choose the database called "Census", (2) select the "simple random sample" option, (3) enter a random seed, any number between 1 and 10000, you can do this by rolling a 10-sided die 4 times and (4) lastly tell it that we want "1040" random draws. Once you have drawn your samples the page will display basic summary statistics for the variables in the census.

(d) Report the 5-number summary and sample mean age.

(e) Use the 5-number summary to construct a box plot of age.

(f) Go to the "Visualize" tab. Choose age as your Response Variable to Plot. What type of variable is this? By clicking on Make My Plot? a histogram of the sample of ages will be displayed. Describe the shape of the data distribution of age.

(g) Is the relationship between the sample mean and sample median consistent with your description of shape? Explain briefly.

(h) If our goal was to not only estimate the mean age of all the U.S. residents but also come up with estimates of the median age of all residents in each of the 50 states, plus the District of Columbia and Puerto Rico what is a drawback of using the simple random sample of 1040? Hint: Set the Data Table to display 100 records per page and go to the page that has "states" 10 and 11 (Delaware and the District of Columbia).

Scenario 2: Suppose now that our goal has changed. Now we wish to investigate the association between age and state of residency. We want to compare the median ages for different states. We still have a budget that allows us to survey around 1040 people. To collect our sample we decide to take a stratified random sample of 20 residents from each state in the United States plus the District of Columbia and Puerto Rico.

(i) Explain in general why collecting a stratified random sample is a better plan than a simple random sample for answering this question.

Now that we have decided on our new sampling plan, let's go collect our data. The Shiny Database Sampler will need to be told 5 pieces of information in order to collect census records the way you want this time. (1) Choose the database called "Census", (2) select the "stratified random sample" option, (3) enter a random seed, any number between 1 and 10000, you can do this by rolling a 10-sided die 4 times, (4) select "state" as strata variable and (5) lastly tell it that we want "20" random draws from each state, plus the District of Columbia and Puerto Rico.

It will take a minute or two to collect these data. It is sifting through millions of records and randomly selecting them from within state groups after all! Once you have drawn your samples you can take a peek at your data set in the main panel of the webpage. You will be able to answer the following questions using the summaries provided on the webpage.

You will notice that the summaries are all broken down by state, but the states are not given names, they are given a code number. This is done on the census to save computer storage space (saving a "19" is much smaller than "Iowa"). A list of all the state codes is available at https://www.census.gov/geo/reference/ansi_statetables.html (Click on FIPS Codes for the States and the District of Columbia).

(j) Report the mean and 5-number summary for the age of the sample from the state of Iowa (`state = 19`).

(k) Report the mean and 5-number summary for the age of the sample from the state of Alaska (`state = 2`).

(l) Compare the distribution of ages in Alaska and Iowa using the values from parts j and k.

(m) Making comparisons as we have done above would become tedious if we wanted to compare ages between all pairs of states in the country. What would be a good way to visually display this information so aid in making these comparisons? Explain your answer.

B.2 Appendix: Cronbach's α Properties

Recall the form of Cronbach's α from equation (3.1):

$$\alpha = K/(K-1) \cdot \left(1 - \sum_{i=1}^{K} \mathrm{Var}\left(Y_i\right) \Big/ \mathrm{Var}\left(\sum_{j=1}^{K} Y_j\right)\right)$$

Claim 1: Perfect agreement in items leads to $\alpha = 1$

Proof: Let $Y = Y_1 = Y_2 = ... = Y_k$, thus having perfect agreement.

$$\Rightarrow \mathrm{Cov}(Y_i, Y_j) = \mathrm{Var}\left(Y\right) = \sigma_y^2 \quad \forall i \neq j$$

$$\Rightarrow \mathrm{Var}\left(\sum_{j=1}^{K} Y_j\right) = \sum_{i=1}^{K} \mathrm{Var}\left(Y_i\right) + \sum_{i \neq j} \mathrm{Cov}(Y_i, Y_j) = K\sigma_y^2 + K(K-1)\sigma_y^2$$

$$\Rightarrow \alpha = (K/(K-1)) \left(1 - \sum_{i=1}^{K} \mathrm{Var}\left(Y_i\right) \Big/ \mathrm{Var}\left(\sum_{j=1}^{K} Y_j\right)\right) =$$
$$(K/(K-1)) \left(1 - K\sigma_y^2 \Big/ K\sigma_y^2 + K(K-1)\sigma_y^2\right) =$$
$$(K/(K-1)) \left(1 - 1/K\right) = (K/(K-1)) \left((K-1)/K\right) = 1$$

Claim 2: For independent items $\alpha = 0$

Proof: Let $Y_1 = Y_2 = ... = Y_k$ be independent

$$\Rightarrow \sum_{i=1}^{K} \mathrm{Var}\left(Y_i\right) = \mathrm{Var}\left(\sum_{j=1}^{K} Y_j\right)$$
$$\Rightarrow \alpha = (K/(K-1)) \left(1 - \sum_{i=1}^{K} \mathrm{Var}\left(Y_i\right) \Big/ \mathrm{Var}\left(\sum_{j=1}^{K} Y_j\right)\right) =$$
$$\alpha = (K/(K-1)) \left(1 - \mathrm{Var}\left(\sum_{j=1}^{K} Y_j\right) \Big/ \mathrm{Var}\left(\sum_{j=1}^{K} Y_j\right)\right) =$$
$$\alpha = (K/(K-1)) \left(1 - 1\right) = 0$$

Claim 3: Perfect disagreement in items leads to $\alpha = -\infty$

Proof: Let $K = 2$ and $Y_1 = -Y_2$, thus having perfect disagreement.

$$\Rightarrow \mathrm{Var}\left(Y_1 + Y_2\right) = \mathrm{Var}\left(Y_1 - Y_1\right) = \mathrm{Var}\left(0\right) = 0$$
$$\Rightarrow \alpha = (K/(K-1)) \left(1 - \sum_{i=1}^{K} \mathrm{Var}\left(Y_i\right) \Big/ \mathrm{Var}\left(\sum_{j=1}^{K} Y_j\right)\right) = (2/1)(1 - 2\sigma_y^2/0) = -\infty$$

APPENDIX C.

C.1 Appendix: Optimal Offset for Symmetric Data Recorded to Resolution α_x

The following proof assumes three conditions for standard rectangular binning of a set univariate data hold: (i) data are recorded to a resolution of α_x units in the X dimension, (ii) points are symmetric distributed within standard rectangular bins, (iii) the bin width, ω_x, is an integer multiples of α_x (i.e. $\omega_x = k\alpha_x$ for some $k \in \{1, 2, \dots\}$). Under these conditions it will be shown that spatial loss is minimized by setting the binning origin, β_x, to $\alpha_x/2$ units below the minimum data value (i.e. $\beta_x = x_{(1)} - \alpha_x/2$).

Let $x_1, x_2, \dots, x_k \in \mathbb{R}$ represent the values in a single bin such that $x_{i+1} = x_i + \alpha_x$ for some constant $\alpha_x \in \mathbb{R}$. Thus $x_j = x_1 + (j-1)\alpha_x$.

Suppose then that we bin the data using standard rectangular binning with origin, $\beta_x = x_1 - \theta$, and binwidth ω; where θ is the *origin offset* from the data. Thus

$b(x_j) = \beta_x + \omega/2 = (x_1 - \theta) + (k\alpha_x/2)$

Spatial Loss, $L^S = \sum_{i=1}^{k} ||x_i - b_(x_i)||$ is definitionally minimized when $b_(x_i)$ is the *geometric median*. The geometric median for $x_1, \dots, x_k = Q_x(.5) = (x_{\lceil \frac{k+1}{2} \rceil} + x_{\lfloor \frac{k+1}{2} \rfloor})/2$, where $Q_x(\cdot)$ is the empirical quantile function.

Thus the optimal offset is the θ such that

$b(x_i) = Q_x(.5)$

$\Rightarrow (x_1 - \theta) + (k\alpha_x/2) = (x_{\lceil \frac{k+1}{2} \rceil} + x_{\lfloor \frac{k+1}{2} \rfloor})/2$

$\Rightarrow 2x_1 - 2\theta + k\alpha_x = (x_1 + (\lceil \frac{k+1}{2} \rceil - 1)\alpha_x) + (x_1 + (\lfloor \frac{k+1}{2} \rfloor - 1)\alpha_x)$

$\Rightarrow -2\theta + k\alpha_x = (\lceil \frac{k+1}{2} \rceil - 1)\alpha_x + (\lfloor \frac{k+1}{2} \rfloor - 1)\alpha_x$

$\Rightarrow -2\theta + k\alpha_x = ((k+1) - 2)\alpha_x$

$\Rightarrow -2\theta = -\alpha_x$

$\Rightarrow \theta = \alpha_x/2$

Thus the optimal offset for reducing spatial loss in this scenario is $\theta = \alpha_x/2$. This result holds for data that is symmetrically distributed within the bin since the median will not change. It extends to multiple contiguous bins with resolution α_x data that has symmetrically distributed data withing each bin.

If the same conditions are extended to the two dimensional case, then the origin for minimal spatial loss is at $(x_{(1)} - \alpha_x/2, y_{(1)} - \alpha_y/2)$ where α_x and α_y are the data resolution for each dimension, respectively.