Statistical Computation Using Student Collaborative Work

John D. Emerson
Middlebury College, Middlebury, Vermont 05753

Abstract
Undergraduate statistics courses, even some courses at the beginning level, have turned to **R** for their computational needs. The open-source availability of **R**, and its strengths for replicable statistical research and analysis, brings to students important advantages for their future work. The adoption of **RStudio**, an integrated development environment (IDE) for **R**, and of student-friendly **R** packages like **Mosaic** help to make the **R** environment a realistic choice for students taking their first few statistics courses. Nonetheless, using **R** is not without challenges for many students who lack programming experience. The adoption of strategies based on student collaboration can aid in addressing these challenges. This paper outlines experiences at a liberal arts college with student collaboration, including team projects, to strengthen students' facility with statistical computation. These experiences may also turn challenges that can intimidate students into rewarding learning experiences that deepen conceptual understanding.

Key Words: undergraduate curriculum, collaborative learning, team projects, simulation

1. Introduction

The field of statistics is in the midst of change; it faces growing challenges and opportunities. Statistics education and training is evolving as it keeps pace with these challenges. In particular, undergraduate statistics education is undergoing rapid expansion and substantial revision. A 2015 study found that statistics is the fastest growing STEM major, with a 95% increase in the number of bachelor's degrees between 2010 and 2013 (*Amstat News*, April 2015 Issue #454, p. 14). At least 130 universities have awarded bachelor's degrees in statistics. In the decade ending in 2013, the number of undergraduate degrees in statistics conferred grew by more than 140%, from 673 to 1656 (IPEDS 2013 as cited in *Curricular Guidelines for Undergraduate Programs in Statistical Science*, 2014).

With expanded enrollments have come evolving methods of teaching and using statistics. In November 2014, the American Statistical Association endorsed new curriculum guidelines for undergraduate programs in statistical science. The executive summary of the report identified four key points – 1. Increased importance of data science; 2. Real applications; 3. More diverse models and approaches; and 4. Ability to communicate – that reflect changes in curriculum and suggested pedagogy since the previous guidelines were disseminated in 2000. The course development described in this report responds to the new *Guidelines*.

The increased importance of **data science** points to the need for extensive computing skills. The *Guidelines* suggest that undergraduate students in statistics "need facility with professional statistical analysis software, the ability to access and manipulate data in various ways, and the ability to perform algorithmic problem solving." The report urges that students should be fluent in higher-level programming languages and facile with database systems.

An emphasis on **real applications** anticipates that students should work with data sets that have been collected to "solve an authentic and relevant problem" and, in particular, to analyze non-textbook data. One implication is that students will begin to see the inevitable need for data-cleaning and dealing with missing data.

The introduction of **more diverse models and approaches** includes a use of computer-intensive resampling methods as an alternative to the traditional methods associated with "normal theory". New curricular materials are appearing, in particular at the level of the introductory statistics course, that provide students with first-hand exposure to sampling distributions. These materials may be used before students are introduced to core concepts in formal statistical inference (Lock et al., 2013; Tintle, et al., 2015).

A greater focus on developing **communication skills** is essential for students entering a world where collaboration, consulting, and addressing statistical questions in varied disciplinary and multi-disciplinary settings are nearly routine. The *Guidelines* (2014) recommend that students be given multiple opportunities to practice and refine their communication skills.

Two courses offered at a highly selective liberal arts college were designed to address the four key recommendations of the *Guidelines* (2014). The first course, *Understanding Uncertainty: Exploring Data Using Randomization*, was offered in a concentrated one-month term in which students take a single course on an intensive basis. Although this course did not list prerequisites, a perception that the course might be especially demanding meant that the enrolling students brought at least one of three particular strengths to the course: a prior course in statistics; experience with computer programming; or study of mathematics at a level beyond first-year calculus. Enrollment was limited to twelve students because the course satisfied a College writing requirement by focusing on expository writing about science and technology.

The first course was not a traditional beginning statistics course; rather, it introduced the use of randomization techniques and resampling to carry out computer-intensive experiments on both real data sets and on simulated data. Some illustrations and exercises were drawn from a preliminary version of the text by Tintle et al. (2015). I assigned students to four teams of three students, with special attention given to balancing the strengths of each team. I designed much of the student work around team projects. Writing assignments, both individual and team-based, reported on the results of computer analyses or statistical experiments. I gave detailed feedback on papers, and in several instances I asked that the papers be revised and resubmitted.

The second course, the principal subject of this report, was a revision of a long-standing existing course, *Introduction to Statistical Science.* This course serves students from all four classes, and most students come from the physical sciences, environmental studies, and geography. It has no prerequisites, although most students have studied some calculus. The course enrolled 24 students in two sections of 12, so that each course section had four student teams with three members. (One student dropped the course near the end of the term for medical reasons.) The 12-week course met three times a week for an hour, and it also had a weekly 1½ hour lab. Each class meeting had assigned work, either problems or projects, and nearly all of this work used the **R** statistical software and **RStudio**. I collected the assignments twice per week. The course had two 2-hour exams, two quizzes, and a comprehensive final examination.

Over the years the course had been based on either of two widely-used textbooks, *Stats: Data and Models* (3rd ed. De Veaux, Velleman, and Bock, 2012), and *Introduction to the Practice of Statistics* (5 th ed. Moore and McCabe, 2014). The newly-revised course made use of web-based materials from ***OpenIntro*** ***(https://www.openintro.org/stat/textbook.php)***, and the ***OpenIntro*** curriculum: *Introductory Statistics with Randomization and Simulation* (1st ed. Diez, Barr, and Cetinkaya-Rundel, 2014), a free download and also available in print at under $10.

2. Incorporating the 2014 *Curricular Guidelines*

The course *Introduction to Statistical Science (ISS)* began with an introduction to the **R** and **RStudio** environments, data frames, exploratory data analysis, and plotting data. I provided many **R** scripts and the students experimented with these, making modest changes and extensions. As recommended in the *Guidelines* (2014), nearly all data sets were "real data". We devoted the second week to a brief but intense study of probability, with an emphasis on random sampling and simple probability experiments illustrating coin tossing, die rolling, and urn models. We always examined results of simulations using plots, and we generated and explored sampling distributions.

In the third week we tackled binomial proportions and we addressed questions of inference using random sampling under a null hypothesis. Thus simulation and sampling played a central role in developing the concepts of hypothesis testing and p-values. Only after thinking deeply about displays of sampling distributions did we introduce **R**'s built-in procedures for analyzing proportions and eventually for using normal approximations for the various sampling distributions.

Students gradually learned to use **R** for analyses by working with lots of illustrative scripts, by making modest changes to the scripts, and by discussing them and experimenting with them in their teams. I avoided the temptation to let computer programming take over a significant part of our agenda; learning about statistical science and all that it entails is a large enough challenge for a first course in statistics. But not

surprisingly, some of the stronger students got "hooked" by some of the programming issues that naturally arise in using **R** for simulation and other analyses.

Lab exercises introduced students to the convergence of sampling distributions to normal distributions, so it was natural to discuss normal distributions and the Central Limit Theorem early in the course. After a brief detour with chi-squared distributions for tables of counts, we moved to a systematic study of measured data. We tried to lean heavily on what we learned from the study of proportions as we developed the inference of one sample, matched pairs, and then two samples of measurements. I believe that our seeing sampling distributions again and again helped to solidify the parallels between the two types of problems. Of course special care was needed in introducing t-distributions (as opposed to t-statistics, where the parallels seem natural).

I planned the course with a limited use of traditional lecturing. The students usually had a reading assignment from the ***OpenIntro*** materials, and they seemed to enjoy that more than traditional textbooks. I sometimes began class by inviting questions, or by framing questions myself and inviting the teams to spend a few minutes collaborating before responding in class. When I didn't get an initial response to a question, I would ask the teams to "convene" and then I left the room for a few minutes; the students understood that I would ask for team reports when we reconvened. This approach seemed quite effective; and it ensured that no one fell asleep even in the 8:00 a.m. section.

I gave a two-hour exam following the fourth week of the course. One exam problem, adapted from an exercise in the ***OpenIntro*** textbook (ISRS 2014, p. 114), developed a one-binomial hypothesis test relating to the effectiveness of Assisted Reproductive Technology at a specific clinic, as compared with a benchmark given by the Center for Disease Control. This problem generated considerable student interest, and I decided to build on it for use in our first team project.

3. Student Team Projects

I delayed forming student teams and introducing team projects until the fifth week of the course, following the first examination. The late start let me use both student biographical data and the results of the examination to constitute the teams in a way that was "balanced". I invited student preferences for team membership, and then worked with those to form teams that included at least one member having stronger preparation and who had done well on the first examination. This approach got little if any student "pushback". Each of the two sections of the course had four teams of three students; I believe that this size is optimal in terms of logistics. Larger teams pose greater challenges in arranging meetings of all team members, including occasional meetings with their instructor.

Project 1: Assisted Reproductive Technology The first team project used a modified version of the data set on Assisted Reproductive Technology, introduced on the first

examination. I described results from a (hypothetical) second clinic and asked the students to test the hypothesis that the two independent sets of binomial outcomes came from a distribution with the same success rate. Although they had previously encountered a "normal-theory" solution to the two-binomial problem, I expected them to do the project using only a simulation approach – a permutation test. They referred to my solution using simulation for the one-binomial problem introduced on the exam, and their task was to extend the ideas to the two-binomial problem and to generate the null sampling distribution for a binomial difference. Each team was to prepare a carefully-written report on its findings, and I provided guidelines for how I would evaluate their reports. Meanwhile the course was continuing and we were learning about inferences for a single group and for two independent groups of measurements. In doing that, I tried to emphasize the parallels for measured data with the binomial data problems which they had already studied extensively.

I gave the teams about a week and a half to complete the project. Still, the results were disappointing in all but a handful of cases. I learned yet again a lesson that I've had to learn many times over years of teaching: the transferal and extension of basic concepts to new settings does not always come naturally. For example, some teams treated the proportion of successes from the first clinic as known and fixed, and treated only the results from the second clinic as being random. Students seemed not to appreciate the importance and value of taking advantage of their teammates to think deeply about the problem and its structure. I suspected that some teams communicated only using e-mail or other electronic communication, and did not sit down together and engage in a real team effort. The quality of writing was often poor, indicating clearly that students invested little effort in proofreading, revising drafts, and taking advantage of the multiplicity of "authors". My feedback to the teams was both thorough and frank, and when I returned the projects I told them that I had not recorded the grades. I required a substantial revision and resubmission. I believe that this response helped establish the expectations for their subsequent work on team projects; the final revisions of the first project were very substantially improved over the first submission.

Project 2: Voter Attitudes on Banning Large Sodas The teams worked on the second team project in parallel with their revision for the first report. This project asked students to use t-tests and related confidence intervals for measurements. The questions addressed one sample, paired comparisons, and two independent samples. The problem culminated in a paired analysis of "differences of differences". Along the way, students were asked to assess the conditions for using normal theory; the data set had just 26 records for voters in 26 cities. Normal theory assumptions were generally satisfied, and simulation was not called for in the problem. Although one or two teams failed to correctly address the appropriate uses of paired comparison versus two independent sample tests, the projects were generally well-done and my assessments compared favorably with those for the (revised) first project.

Project 3: Gender Bias in Promotion Decisions The third team project used a data set from ***OpenIntro*** (*ISRS* 2014, pp. 61-65) that records results from a randomized

experiment done at a management institute. The participants, who were managers, evaluated dossiers of candidates for promotion at a bank, and recommended whether or not to award the promotion. The explanatory factor being investigated was gender of those who were reviewed; two versions of the same dossier, with only the gender of the candidate distinguishing them, were randomly assigned to the managers. My students had already encountered this data set in the context of a hypothesis test, and the project asked for a confidence interval for the gender-based differences in rates of promotion. A simulation approach makes use of the bootstrap, a method I had only recently introduced them to.

The project asked the teams to begin by using a permutation test to examine the hypothesis of no gender difference. A plot of the null sampling distribution supported a finding that there was a statistically-significant gender difference in rates of promotion, and students reported their simulated two-sided p-values. The problem then asked the students whether their work could provide a confidence interval for the true difference in promotions rates (it cannot). Next I asked them to write about how a bootstrap study of the data would differ from the simulation analysis used in the hypothesis test. Finally I asked students to carry out a bootstrap analysis, give a 95% confidence interval for the difference, and discuss their findings.

Following the recommendation by Tim Hesterberg (2014), I had demonstrated two elementary methods for building a confidence interval from the bootstrap distribution of differences in proportions: (i) using the standard deviation of the bootstrap distribution, and using the quantiles of the distribution. Although I had warned that the quantile method is less reliable in small data sets, the majority of teams used that approach, perhaps because it felt more natural and intuitive when one is staring at a plot of the bootstrap distribution. A couple of teams reported both intervals and compared the two.

The third team project was clearly the most challenging so far. Some students likely had little understanding of the difference between a simulated permutation distribution and a bootstrap distribution. I believe that the use of the team structure helped address this challenge for most but not all students. The average grade on this project was between the averages of the first two projects (averages in order: 83, 89, 87), but the standard deviation was the largest of any team project. The teams seemed either to “get it” with the bootstrap, or not. The bootstrap concept is an initial challenge, and it does generate considerable interest and discussion among the students.

Project 4: The Earth’s Solar System The fourth team project, done in the 11th week of the course, used a data set on the planets in the earth’s solar system that came from the textbook by De Veaux, Velleman, and Bock (2012, pp. 255-6). Students used three variables; (i) the position order of the planets, (ii) the distance from the sun, and (iii) the length of a year on each planet, given in earth-years. We had encountered uses of re-expression in the context of simple linear regression, and this project asked students to explore relationships among the variables and to try and find expressions of the variables that straightened the relationship. The initial explorations relied largely on scatter plots,

and after students found suitable data transformations they used regression fits and related diagnostics to examine their results in detail. The project statement did not specify what re-expression to try or what variable(s) to try it on.

A regression of log(distance) against position gave an R^2 of 97.9%, and was an unambiguous improvement over both analyses of the original variables and of log-log data. The more interesting exploration was one for length of year against distance from the sun. A log-log analysis yielded a scatter plot that appeared perfectly straight, and simple linear regression using these scales confirmed that impression by giving an R^2 of 100%. The project than asked the teams to, "Give some regression diagnostics for the regression above. Discuss what you discover from this examination. Is it worthwhile to look carefully at the residuals?" The students who followed the implied analyses were in for a surprise.

A scatter plot of the residuals from the simple linear regression of log(length-of-year) against log(distance from sun) revealed that the residual from one planet, Pluto, was a clear (negative) outlier. According to De Veaux, Velleman, and Bock (2012), this finding contributed to the discovery in 2006 that Pluto may not be a planet but rather a large icy object in the Kuiper belt of icy objects. The members of several of the teams were puzzled by the finding, given the R^2 value of 100%. Although I had defined the project in a rather open-ended way, six of the eight teams arrived at the key discovery and five teams presented nicely-written reports that earned an A- or A grade.

After evaluating these team reports, I sensed that most teams were finally working as I had envisioned and hoped. I decided to survey the students on their course experiences, and 22 students completed the survey using a Likert scale. Two survey items addressed student attitudes about their team work. Fifteen of 22 students agreed or strongly agreed that "small team work aided my understanding", whereas a single respondent strongly disagree with the statement. But 9 of 22 students agreed with a statement "the team projects were more trouble than they were worth", and 10 students disagreed with the statement. In private conversations with a few students, I learned that unfavorable attitudes about teams derived largely from the logistical issues in their scheduling meetings of the teams, and from getting all members to participate fully and to "sign off"" on the team report.

4. Team Component of Final Examination

I devoted the last week of the course to a review of the various computer-intensive methods for testing hypotheses and building confidence intervals, and we delved more deeply into this material as it relates to core concepts of statistical inference. I assigned a three-part simulation team project as a component (25%) of the final examination. Our goal was to use simulation to explore the roles of sample size and classical distributional assumptions in making valid inferences. In particular, the project provided a context for comparing classical normal-theory methods and computer-intensive methods, where the

"right answer" is known. The price paid for this convenience was of course that we were no longer beginning the analysis with "real" data.

Final Team Projects The project began with two "populations", normal and exponential, with the same mean and standard deviation for both ($\mu = \sigma = 10$). We examined samples with n=4, 16, 64, 256, 1024, and 4096, although for some parts we did not use all of these sample sizes. We always used 10,000 replications. The first simulation problem examined the sampling distribution of the sample means, and compared results for the normal and exponential data at each sample size. I wanted the focus to be on statistical concepts rather than on technical challenges in the **R** language, so I gave students code that generated the two populations and the simulations of the sampling distribution for the mean at n=16. It was then easy for them to extend to other values of n. Their task was to look closely at the simulated results, to make sense of them, and to find ways to present comparisons that convey the results for the two population distributions and the various sample sizes. This portion of the project extended work we had done in one of the early labs in the course, so the observed behavior of the histograms and normal probability plots were not entirely new to the students.

The problem gave students a chance to "see" the standard error shrinking as n grows, to see the central limit theorem in action, and to learn how large an n is needed for the sample means from exponential data to appear just as "normal" as those for means of data drawn from the normal population. In all cases we used both histograms and normal probability plots, but we did not use a formal test for normality. Most students reported that they detected skewness in the sample means for the exponential data when n was 64, but not for n=256. We also recorded the coverages of the true mean by 95% confidence intervals constructed using the classical normal-theory methods; perhaps not surprisingly the intervals using exponential data displayed under-coverage when n was 64 (or smaller) and not for n=256 (or greater).

Table 1. Standard Errors and Confidence Intervals for the Mean; Normal and Exponential

	n=4	n=16	n=64	n=256	n=1024
Normal.SE	5.01	2.50	1.24	0.620	0.311
Skewed.SE	5.01	2.53	1.26	0.632	0.316
Normal.Coverage	94.76	94.85	95.08	95.10	94.75
Skewed.Coverage	88.41	91.80	93.67	94.64	94.62

Notes.
a. Both "populations" have mu = 10 and sigma = 10
b. Analysis of sample means from Normal and Exponential (Skewed) populations
c. Simulations use 10,000 replications throughout, as suggested by Hesterberg (2014)
d. Confidence intervals based on normal theory, using t-distributions
e. "Coverage" refers to the percent of simulated confidence intervals containing the true mean

The second part of the project, done in the same context, addressed an issue that was entirely new for the students; we looked at the sampling distributions of a quantile (the 99th percentile) from the right tail of the distribution. A motivating question for doing this was to explore the variability of the endpoints of a confidence interval that is based on quantiles of a sampling distribution. This exploration opened a new world to the students, in part because the sampling distribution of the quantile is **not** symmetric for either normal data or for exponential data.

Table 2. Estimating Means and 99% Quantiles; Normal and Exponential

	n=4	n=16	n=64	n=256	n=1024	n=4096
Normal.SE	**5.07**	**2.49**	**1.23**	**0.622**	**0.309**	**0.154**
Skewed.SE	**5.04**	**2.49**	**1.23**	**0.617**	**0.313**	**0.152**
Normal.Q99.SE	**7.04**	**5.08**	**3.36**	**2.065**	**1.143**	**0.566**
Skewed.Q99.SE	**11.68**	**11.59**	**8.68**	**5.379**	**2.930**	**1.473**
Normal.Q99.Mean	**20.19**	**27.03**	**31.06**	**32.60**	**33.13**	**33.23**
Skewed.Q99.Mean	**20.40**	**32.34**	**40.84**	**44.08**	**45.18**	**45.48**

Notes. a. The first four rows of Table 2 give standard errors from the sampling distributions of the sample means and of the 99% quantiles, for both Normal and Exponential data.

b. The true 99% quantiles are: 33.3 (Normal, $\mu=\sigma=10$) and 46.1 (Exponential, $\mu=\sigma=10$)

The simulations confirmed that the standard error for the sample mean is the same for normal and exponential data, and of course is always proportional to $\frac{1}{\sqrt{n}}$. Other noteworthy findings are:

- The standard error of the 99% quantile is always greater than that for the mean;
- The standard error of the 99% quantile is, for given sample size, always substantially greater for the exponential distribution than for the normal distribution. For large n these standard errors differ by a factor of around 2.8.
- As n increases, the 99% quantile grows monotonically toward the true population values of 33.3 (for normal) and 46.1 (exponential). The sample quantiles are substantially biased estimators for the population values.
- With normal data, skewness is evident in the sampling distribution of the 99% quantile even for n=1024. The skewness is more pronounced with exponential data than for normal data, and only the sampling distributions of the quantile for samples of size 4096 appeared to be symmetrical. (We omit the plots used to support these claims.)

The third component of the project used resampling to compare the bootstrap methodology to replicated random sampling as used in the first two parts of the project. The bootstrap cannot improve on the estimation of a parameter (the mean, or a 99% quantile), but how well does it estimate the variability of a statistic of interest? We continued to use data drawn from the same normal and exponential populations, and to examine sample means and sample 99% quantiles. I provided code for doing a reasonably

extensive analysis using only sample size n = 256. I gave students the option of looking at other sample sizes, and a few of the teams did pursue that. The primary focus in this final part of the team project was to compare bootstrap distributions with ordinary sampling distributions for the same parameter and the same sample size.

Figure 1 show the sampling distributions and the bootstrap distributions for the sample mean and for the 99% quantile, when the sample consists of 256 observations from an exponential distribution with mean 10. Clearly the variability is greater for the 99% quantile, and the distributions are not symmetric. Two of the student teams were especially interested in the observed behavior, notably the graininess of the bootstrap distribution. We discussed possible reasons for this behavior, and one team explored it further using larger sample sizes. I did my own investigation right along with the student teams, including a look at the coverages of two versions of bootstrap confidence intervals.

Figure 2 shows a bootstrap distribution and its normal probability plot for a bootstrap sample of 10,000 99% quantiles for samples drawn from the normal distribution (μ=10, σ=10). The graininess of the distribution is striking and in sharp contrast to the corresponding bootstrap distribution of the sample mean. We explored the same two plots for sample size n=4096, and the graininess was still evident.

Table 3 gives coverages for three confidence interval constructions: traditional normal-theory intervals using the t-distribution; bootstrap confidence interval that uses the standard error of the bootstrap distribution of the sample mean; and bootstrap confidence interval using the percentiles of the bootstrap distribution. I provide here my own results for sample sizes 16, 64, 256, 1024, and 4096, with data drawn from the normal distribution and from the exponential distribution. The work grew out of conversations I had with one of the student teams.

Table 3. Coverages for 95% Confidence Intervals Using Classical t-Distribution, Bootstrap Standard Error, Percentile Bootstrap (PB); Normal & Exponential Data

	n=16	n=64	n=256	n=1024	n=4096
Normal.Coverage.t	**94.92**	**94.90**	**94.84**	**95.42**	**95.31**
Normal.Coverage.B	**94.15**	**94.78**	**94.77**	**95.43**	**95.33**
Normal.Coverage.PB	**91.95**	**94.36**	**94.69**	**95.37**	**95.27**
Skewed.Coverage.t	**91.63**	**94.17**	**94.57**	**95.24**	**95.36**
Skewed.Coverage.B	**90.93**	**93.97**	**94.55**	**95.20**	**95.37**
Skewed.Coverage.PB	**89.67**	**93.81**	**94.41**	**95.21**	**95.35**

The results for the simulated coverages indicate that:

- Coverages by intervals that use samples of normally-distributed data are generally close to the nominal 95% level. However the bootstrap percentile

interval yields under-coverage for n=16 (91.95%) and possibly for n=64 (94.36%).

- When samples are from the exponential distribution, all three intervals show under-coverage at n=16, and modest under-coverage at n=64.
- The bootstrap interval that used the observed standard error of the bootstrap distribution generally outperforms the bootstrap percentile interval.
- The simulations do not give evidence that bootstrap intervals used here offer advantages in their coverage over the classical t-based confidence interval.

5. Discussion and Findings for Students of Statistics

This report describes and reviews two somewhat experimental undergraduate statistics courses without statistics prerequisites. Both courses relied heavily on computer simulation and resampling in the **R/RStudio** environments. Because the second course was an alternative version of an introductory statistics course that has been taught over many years, we have reported in much greater detail on the changes made in that course and on the results. This report is largely descriptive, and to the extent that it reports on student outcomes it can only be considered as "impressionistic"; we have no control group.

I view the introductory course as being "experimental" in two distinguishable ways: (1) It made heavy use of computer-intensive methods for doing statistical inference; and (2) It expected students to wrestle with much of the more challenging material in the course in teams of three students. To assess student attitudes about each of these areas, I administered a 17-item survey on a Likert scale on the last class day of the course. I coded the 22 responses to each item on a scale of 1 (strongly disagree with statement) to 5 (strongly agree), and calculated average responses.

I can summarize some key student attitudes about the computer-intensive methods as follows:

- Students feel they learned much from a focus on sampling distributions (x=4.2)
- Student would like to have a text that used more R code (x=4.3)
- Students do NOT support using more simulation earlier in the course (x=2.6)
- Students are ambivalent about using less computer simulation in order to add coverage of more statistical topics (x=3.2)
- Students would welcome more resources to support their R programming (x=3.9)
- Students report they gradually learned to use R scripts on their own (x=3.7)
- Students understand the use of the bootstrap for estimating standard error (x=3.9)

Only two items addressed student attitudes toward the team projects:

- Students believe the small team work aided their statistical understanding (x=3.8)
- Students are ambivalent about whether the teams were more trouble than they were worth (x=3.0)

Students also routinely complete College course response forms that assess each of their courses, their own work in them, and the effectiveness of their instructors. The two-page form includes seven items on which students are asked to write "considered, candid responses" to the questions posed. In several instances the student responses proved useful in interpreting the assessments of my own course-specific items referred to above. A careful reading of the student responses provides a reminder of the variety of student views and their educational needs; I am no longer surprised when successive forms give what seem to be directly opposing views of a particular aspect of a course.

Many students gave positive assessments of the computer-intensive aspects of the course, and they seemed to appreciate that their extensive use of **R/RStudio** enhanced their statistics education. But other students complained that the heavy reliance on **R** meant that they were taking two courses at once – in statistics and in computer science. A few students would have liked more class time devoted to systematic instruction in programming in **R.**

The commentary about the use of student teams seemed more strongly divided. Some student praised the use of team projects and they identified them as being the aspect of the course from which they learned most. Other students commented explicitly that the team assignments required a lot of time, and that little if any additional learning resulted from the time and effort invested. I maintain an open mind about the use of teams and I suspect that both of these student perspectives have validity. The success of team collaboration seems to hinge largely on a team having at least one member who is a leader invested in the successful engagement of a problem by all members of the team. But to the extent that team projects contribute to active student engagement, I think that they contribute more to deep learning than do traditional text exercises.

I end this report with a summary of some of my own impressions of this teaching experience. I believe that an introductory course that relies heavily on simulation and computer-based sampling, coupled with a healthy dose of student collaboration, can provide the following:

- A more direct and immediate understanding of randomness in general, and, in particular, of the random variability of a statistic used for making inferences;
- A hands-on acquaintance with a wide variety of sampling distributions;
- A deeper appreciation of the role of traditional assumptions about distributions of statistics and how those assumptions can fail with real-world data sets;
- A strengthened ability to distinguish the random variation inherent is a statistical model (often with an assumed hypothesis) from systematic departures beyond such variation;
- A basic (though sometimes superficial) appreciation of such computer-intensive concepts as permutation tests and bootstrap distributions;
- A beginning appreciation of some of the (surprising) issues relating to sample size that have emerged from the fascinating insights given in the work of Tim Hesterberg (2014);

- An acquaintance, though not expertise, with a statistical software environment (**R/RStudio**) that has become a world-wide *de facto* standard for managing and exploring data, providing graphics, carrying out statistical analyses, and performing statistical simulations;
- An experience with the kinds of skills that are essential in tackling challenging questions as a participant in a collaborative team project;
- A realization that most important questions do not always have unambiguous "textbook" or "cookbook" answers; ambiguities abound and statistics does involve some "art" as well as sound statistical science.

Although the results of my own first effort to create a substantially revised introduction to statistics were surely mixed, I remain an optimist and I will continue the "experiment" with some adjustments and further revisions. Most importantly, I had a lot of fun as I worked through the material arm-in-arm with my students. I know that more fun is in store.

Acknowledgments

Joe Chang first made me aware of the ***OpenIntro*** curriculum and thus of the web-based course materials that I used in the introductory course. While on a sabbatical at Yale University, Jay Emerson introduced me to more about modern statistical computing than I ever thought possible; at a general level this background proved exceedingly useful for my work in this course. Finally, Tim Hesterberg's unusual, inspiring, and often surprising overview of computer-intensive methods, especially the bootstrap, motivated the focus of the team projects that ended the course.

References

American Statistical Association (2014). *Curriculum Guidelines for Undergraduate Programs in Statistical Science*, www.amstat.org/education/curriculumguidelines.cfm (accessed August 27, 2015).

De Veaux, R.D., Velleman, P.F., and Bock, D.E. (2012). *Stats: Data and Models* (3rd ed.), Boston: Pearson Addison-Wesley.

Diez, D.M., Barr, C.D., Cetinkaya-Rundel, M. (2014). *Introductory Statistics with Randomization and Simulation* (1st ed.), Text is free and available under the Creative Commons Attribution-ShareAlike License, https://www.openintro.org/stat/ (accessed August 27, 2015): OpenIntro.

Hesterberg, T. (2014). *What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Curriculum,* available at http://arxiv.org/abs/1411.5279 (accessed August 27, 2015).

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

Tintle, N., Chance, B.L., Cobb, G.W., Rossman, A.J., Roy, S., Swanson, T., VanderStoep, J., (2015). *Introduction to Statistical Investigations* (Preliminary Edition), Hoboken, NJ: Wiley.

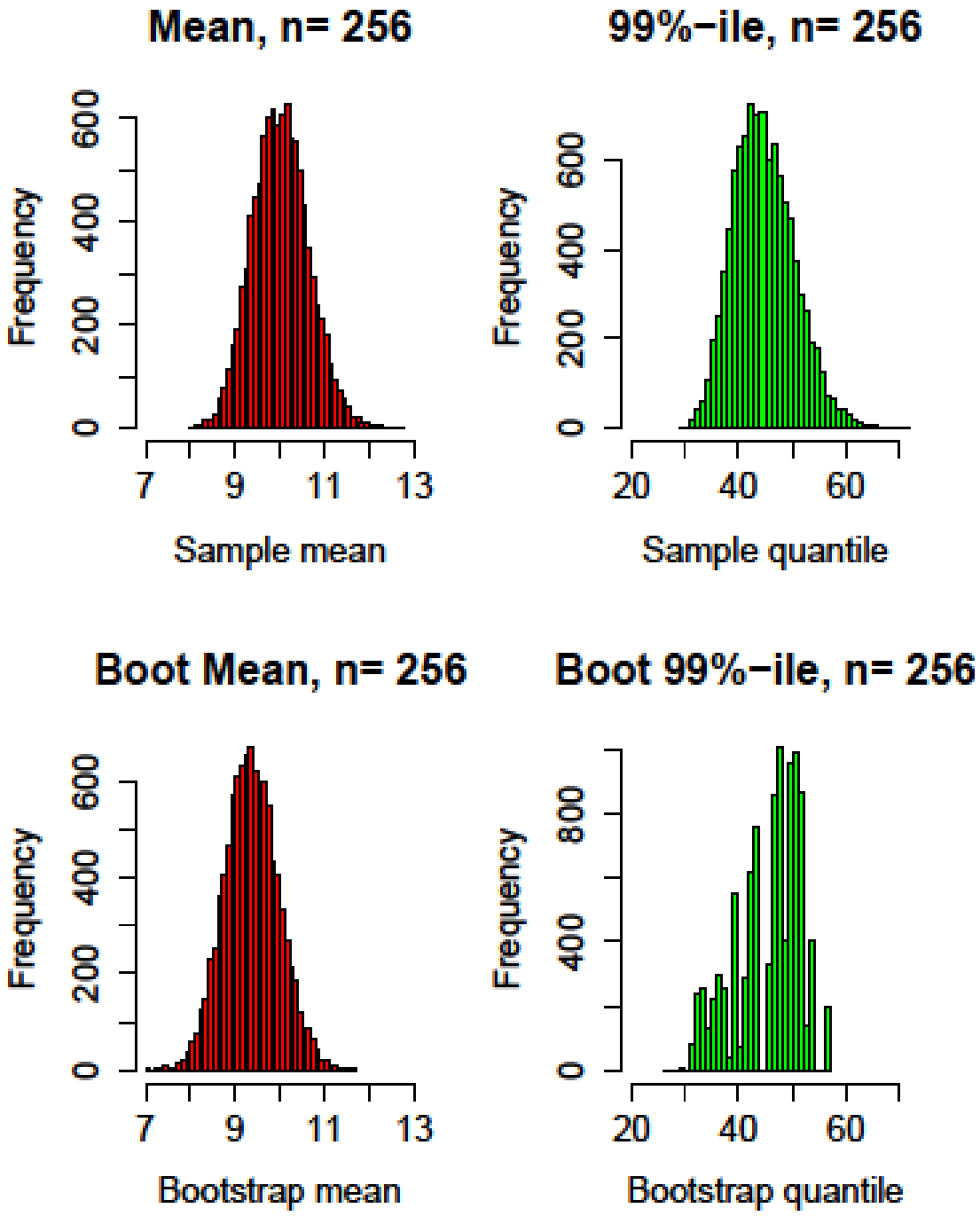

Figure 1. Sampling Distributions and Bootstrap Distributions of the Sample Mean and the Sample 99% Quantile from Exponential Data

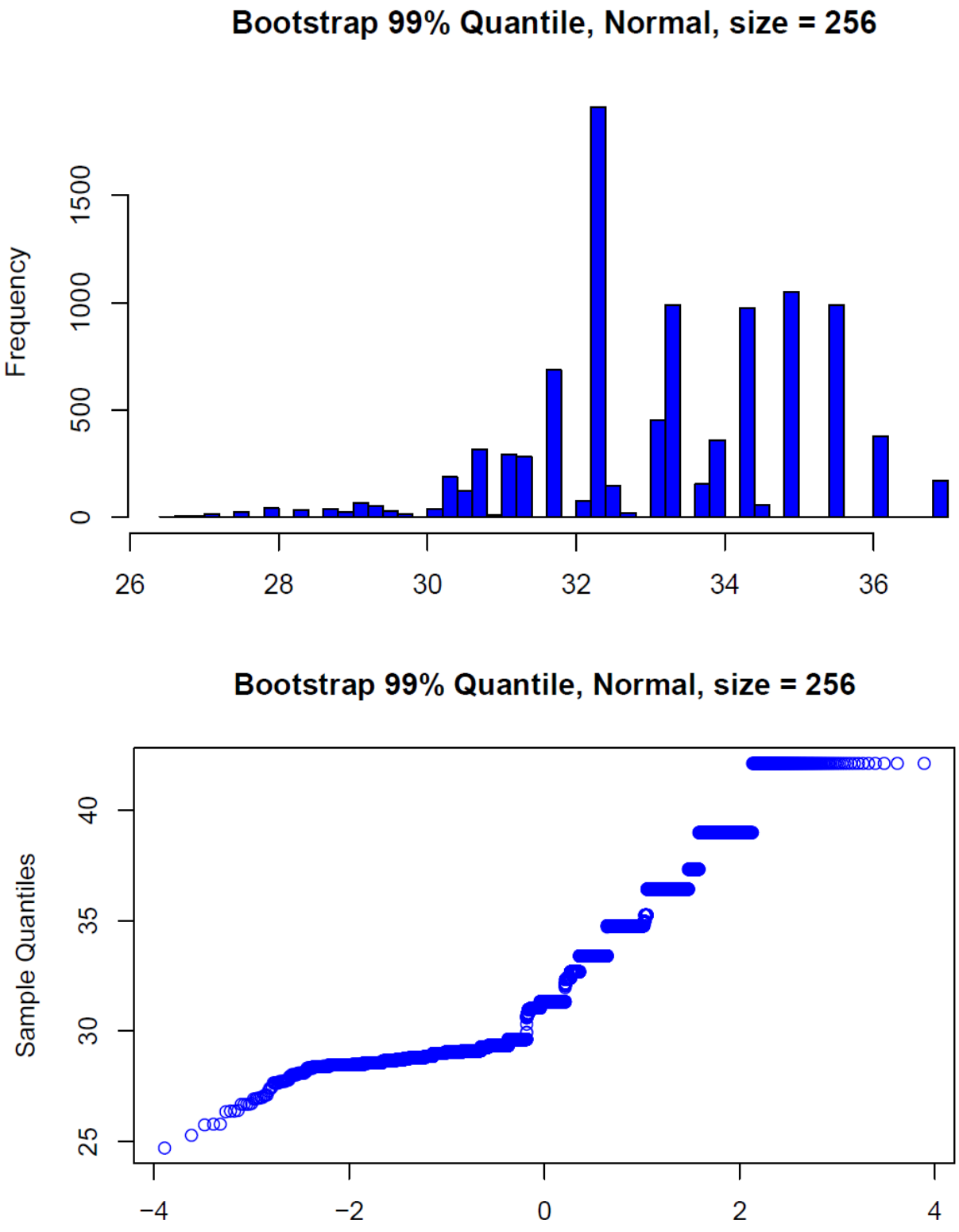

Figure 2. Bootstrap Distribution and Quantile Plot for Sample 99% Quantile, from Normal Distribution with n=256 and 10,000 bootstrap samples