Applied Statistics and Data Analysis

Rick Cleary, Babson College
John Gabrosek, Grand Valley State University
Patti Frazer Lock, Saint Lawrence University, chair
Mary Parker, Austin Community College
Allan Rossman, California Polytechnic University

I. The Course

CUPM recommends, and the MAA Board of Governors agrees, that every student majoring in the mathematical sciences take an introductory course in Applied Statistics, with a clear focus on data analysis. We recommend that this course be taken during the first two years of the undergraduate program and that it be focused squarely on applied data analysis. This is a course quite distinct from the usual upper-level sequence in probability and mathematical statistics that is offered as an elective in most undergraduate mathematics programs and also quite distinct from the low-level procedural course or quantitative literacy course taught at many institutions.

II. Student Audience

Although the course may serve a much broader audience, the audience we focus on in this report includes all students majoring in the mathematical sciences, including programs in mathematics, applied mathematics, mathematics education, operations research, actuarial science, and combined mathematics majors (combined with economics or biology, for example). We believe that an applied data analysis course, taken relatively early in the program, is a necessary component in all of these mathematical sciences programs.

III. History and Background

We are guided by our CUPM predecessors in this process. Every Curriculum Guide for at least the past 30 years has demonstrated to us the need to increase understanding of statistics in students majoring in the mathematical sciences and to do so in a course clearly focused on data analysis.

- **From the 2004 *Curriculum Guide***: "The CUPM Guide 2004 supports the 1991 CUPM recommendation that every mathematical sciences major should study statistics or probability with an emphasis on data analysis."
- **From the 1991 *Curriculum Guide***: "Every mathematical sciences major should include at least one semester of study of probability and statistics … The major focus of this course should be on data and on the skills and mathematical tools motivated by problems of collecting and analyzing data."
- **From the 1981 *Curriculum Guide***: "The Statistics Subpanel believes that an introductory course in probability and statistics should concentrate on data and on skills and

mathematical tools motivated by the problems of collecting and analyzing data."

We reiterate and endorse, as strongly as we can, the recommendation that every mathematical sciences major include a course in applied statistics, focused on data analysis.

IV. Current Status

The MAA Curriculum Guides have been recommending for more than 30 years, and with increasing emphasis, that every student majoring in the mathematical sciences take a course in statistical data analysis. How are we doing at meeting this recommendation? In an effort to discover the answer to that question, we did some of our own data collection and analysis. The results we found were quite discouraging and even worse than we had expected.

In our sample of fifty-five undergraduate programs in mathematics, selected from a wide variety of different types of schools, only 4 (about 7%) currently require a course in applied statistics. Furthermore, only 12 (about 22%) even *allow* a course in applied statistics (which is distinct from the upper-level probability and mathematical statistics courses) to count toward the major program. Even the 22% estimate is probably quite high, since many of these electives in "applied statistics" appear to be weighted more heavily toward probability rather than data analysis. Despite the strong recommendations to the contrary, we are making very little progress in producing mathematics graduates with a sound knowledge of statistical data analysis.

The situation in undergraduate programs in Applied Mathematics and in Mathematics Education appears to be slightly better but still not good. In our relatively small sample of these programs, we found that about 50% of programs in Applied Mathematics allow an applied statistics course to count toward the major, while about 63% of Mathematics Education programs do so. The results in Mathematics Education are a particular concern, since prospective secondary math teachers need a thorough preparation to teach the modern data collection and interpretation content of the Common Core standards. All students going on to K-12 math education should be required to take at least one and preferably two modern data-driven courses in applied statistics. See Appendix C.

We are living in a world in which the ability to analyze data is increasingly important, across almost all disciplines. Graduates of undergraduate programs in the mathematical sciences go on to a wide range of careers in education and business, graduate and professional programs in a wide range of areas, and doctoral programs across the range of mathematical sciences. In every single case, our graduates would be well-served with solid knowledge and skills of statistical data analysis. Indeed, we believe such a course would serve the vast majority of our students far better than one additional theoretical math elective.

The obvious question, of course, is why departments continue to discourage students from taking such a course as a part of the major program. The second obvious question to ask is what can we do to change this pattern.

We provide several possible answers to the first question. One is largely historical: Introductory statistics courses at many institutions have been viewed as low-level service courses with not enough mathematical content to warrant credit for a student majoring in mathematics. One is based on resources: There are not enough faculty members to provide such a course for majors, and there are not enough statisticians around to teach it. One is based on the tendency of mathematicians to view all courses through the lens of theoretical mathematics and to therefore evaluate the courses on how they might prepare students for doctoral programs in pure mathematics. Answers to the second question on what we can do to change the pattern are harder to come by. We address some of these challenges in Appendix B.

V. Description of the Recommended Course

Our Statistics Area Study Group is fortunate to have a widely respected set of guidelines from which to start. The GAISE Guidelines (Guidelines for Assessment and Instruction in Statistics Education) were written in 2005 and endorsed by the American Statistical Association. They have since also been endorsed by the American Mathematical Association of Two-Year Colleges. The GAISE Guidelines include "Goals for Students in an Introductory Course: What it Means to be Statistically Educated" as well as a list of six specific recommendations to help students attain the learning goals. These goals and recommendations are available online and are included in Appendix A of this report. The GAISE Guidelines are currently being updated and the revised guidelines are expected to be available in 2015. We unanimously support the goals and recommendations of the GAISE report, and these guidelines strongly inform our work.

For students majoring in the mathematical sciences, we recommend a course focused on applied data analysis and driven by real data. The course should stress conceptual understanding, foster active learning, and introduce students to statistical technology. The focus should be on the effective collection and analysis of data, along with appropriate interpretation and communication of results.

Just as Mathematics Departments routinely do with calculus courses, such a course in Applied Statistics can serve a wide audience. Also as with calculus, some institutions will have one level of the course while other larger institutions might have different courses for different audiences. In every case, however, the focus should be on understanding effective data analysis rather than on the underlying mathematical theory. The concepts involved in statistical inference are notoriously challenging for students to master, and a course focusing on these concepts provides an intellectually rigorous course, even without teaching these concepts from a theoretical mathematics perspective. There are significant differences between statistical and mathematical thinking (see, for example, Cobb & Moore, "Mathematics, Statistics, and Teaching", *The American Mathematical Monthly*, November 1997) and this course should focus explicitly on statistical thinking.

Even though we have used the singular "course" in this section, we believe that many different courses could achieve the goal of introducing mathematics students to effective data analysis. We provide syllabi for some such courses in Appendix D.

Cognitive Goals:
Applied Statistics is an outstanding course for helping students meet the cognitive goals set out in this *Guide*. Specifically, in the process of working with real data, students have to read with understanding, recognize patterns, identify essential features of a complex situation, and apply appropriate methodologies. All of these enhance critical thinking skills. Communication skills are also emphasized in such a course, as students learn to effectively interpret and justify their conclusions. Learning to use technology intelligently as an effective tool is an integral part of a good data analysis course.

Applied Statistics is also an outstanding subject to promote awareness of connections to other subjects. A strong course in applied statistics will illustrate applications to a wide variety of different areas and the importance of interacting across disciplines. Such a course can also serve to enhance student perceptions of the vitality and importance of mathematics and statistics in the modern world.

Additional cognitive goals of an applied statistics course include dealing with randomness and uncertainty, understanding the distinction between exact answers and models/approximations, and working with data visualization.

An introductory course in Applied Statistics should be taught using all the current best thinking about how people learn. Classes should be interactive with regular active participation by students. Statistics lends itself well to student projects, to experiential learning, and to team explorations, and we strongly encourage the use of these interactive pedagogies in statistics classes.

Mathematical Outcomes:
In addition to the outcomes listed in the cognitive goals, we offer the following goals for an introductory course in Applied Statistics:

- An understanding of the process by which statistical investigations are performed, from formulating questions to collecting data, then analyzing data and drawing inferences, and finally interpreting results and communicating conclusions.
- A solid conceptual understanding of the key concepts of statistical inference: estimation with intervals and testing for significance.
- The ability to perform statistical inference procedures, using traditional methods and/or modern resampling and permutation methods.
- Experience using technology to explore statistical concepts and to analyze data graphically, numerically, and inferentially.
- An understanding of the importance of data collection, the ability to recognize limitations in data collection methods, and an awareness of the role that data collection plays in determining the scope of conclusions to be drawn.

- The knowledge of deciding which statistical methods to use in which situations and the ability to check necessary conditions for those methods to be valid.
- Extensive experience with interpreting results of statistical analyses and communicating conclusions effectively, all in the context of the research question at hand.
- An awareness of the power and scope of statistical thinking for addressing research questions in a variety of scientific disciplines and in everyday life.

Prerequisites:
Basic proficiency in algebra is all that is required, combined with a bit of analytical maturity. Some data analysis courses could have calculus as a prerequisite.

Sample Syllabi:
Sample syllabi and course outlines are provided in Appendix D. We also include in Appendix C a recommended two-course sequence for future mathematics teachers, shared with us by the authors of the MET2 report (Math Education of Teachers). With the increased emphasis on statistics in the Common Core State Standards in Mathematics, and the dramatic and record-breaking rise in students enrolling in AP Statistics courses, this recommendation for future mathematics teachers is particularly important.

VI. What About Other Courses in Statistics?
Most statistical analyses involve the analysis and modeling of relationships between many variables. While a first course in applied statistics is likely to focus mainly on univariate and bivariate methods of data analysis, the course can serve as a bridge to and introduction of data analysis situations involving many variables. Statistics is much more than can be covered in just this one semester course! There is much to say about other courses in statistics for students majoring in mathematical sciences, including discussions of excellent second-level courses in applied statistics, the theoretical probability and mathematical statistics sequence, and other options for introductory courses for other majors and at other levels. These other courses, however, are both more accepted already within mathematics programs and less controversial as courses for math majors. For these reasons, we have opted to focus on the most important change we think undergraduate mathematics programs need to make as we prepare 21st century mathematics majors: to provide and actively encourage all students majoring in the mathematical sciences to include at least one course in applied data analysis.

Appendix A: GAISE Guidelines

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) were developed in 2010 and the full details are available online.

Recommendations:

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 concepts and analyzing data.
6. Use assessments to improve and evaluate student learning.

Goals:
Students should believe and understand why:

- Data beat anecdotes
- Variability is natural, predictable, and quantifiable
- Random sampling allows results of surveys and experiments to be extended to the population from which the sample was taken
- Random assignment in comparative experiments allows cause-and-effect conclusions to be drawn
- Association is not causation
- Statistical significance does not necessarily imply practical importance, especially for studies with large sample sizes
- Finding no statistically significant difference or relationship does not necessarily mean there is no difference or no relationship in the population, especially for studies with small sample

Students should recognize:

- Common sources of bias in surveys and experiments
- How to determine the population to which the results of statistical inference can be extended, if any, based on how the data were collected
- How to determine when a cause-and-effect inference can be drawn from an association based on how the data were collected (e.g., the design of the study)
- That words such as "normal," "random," and "correlation" have specific meanings in statistics that may differ from common usage

Students should understand the parts of the process through which statistics works to answer questions, namely:

- How to obtain or generate data
- How to graph the data as a first step in analyzing data, and how to know when that's enough to answer the question of interest
- How to interpret numerical summaries and graphical displays of data—both to answer questions and to check conditions (to use statistical procedures correctly)
- How to make appropriate use of statistical inference
- How to communicate the results of a statistical analysis

Students should understand the basic ideas of statistical inference, including:

- The concept of a sampling distribution and how it applies to making statistical inferences based on samples of data (including the idea of standard error)
- The concept of statistical significance, including significance levels and p-values
- The concept of confidence interval, including the interpretation of confidence level and margin of error

Finally, students should know:

- How to interpret statistical results in context
- How to critique news stories and journal articles that include statistical information, including identifying what's missing in the presentation and the flaws in the studies or methods used to generate the information
- When to call for help from a statistician

Appendix B: Challenges

Statisticians:
This is a bit of a Catch-22. In order to attract more quantitatively-inclined students into statistics, we need to expose more of them to the subject earlier in their college careers. However, in order to offer these courses, we need to recruit more statisticians as faculty in Mathematics Departments. We don't have a ready solution for this one, but a recent ASA/MAA Joint Report, "Qualifications for Teaching an Introductory Statistics Course," offers some guidance.

Resources:
How can departments, already stretched for resources, afford the resources to offer courses such as the one proposed? One solution is to reconfigure the current introductory statistics course offered at many schools. As a low-level probability and statistics course with a focus on procedures and formulas, the course is designed for non-math majors. By redesigning the course, however, with a focus on concepts, technology, and real data analysis, the course remains viable (and better) for the current audience while also becoming an appropriate course for students majoring in the mathematical sciences.

Connecting with Other Courses:
Students should be encouraged to take an applied data analysis course at any time in the first two years, which could be before, concurrent with, or after calculus. Just as students take Calculus and then are exposed to the theoretical foundations in Advanced Calculus and Real Analysis, students taking this course could then be exposed to the theoretical underpinnings in the probability and mathematical statistics sequence (as well as, if resources allow, more advanced courses in applied data analysis.) We see no problem with interactions between this course and existing courses.

Attitude:
One of the biggest challenges to the goal of having all mathematical sciences students complete a semester of data analysis is the belief of many mathematicians that such a course is not an appropriate course for a student majoring in mathematics. Some hold a strong belief that all true mathematics courses should follow a theorem and proof model. However, departments are starting to embrace the idea of offering a mathematical modeling course that asks students to

deal with complex real situations and that is project-based with a heavy emphasis on communication skills. In the same way, we hope that departments will embrace the idea of having their students explore the field of statistics, so that they are prepared for a world full of data and are exposed to more of the richness of the mathematical sciences.

Appendix C: Recommendations for Mathematics Education of Teachers

The MET2 report (Math Education of Teachers) recommends that prospective high school teachers obtain a thorough preparation to teach the modern data collection and interpretation content of the Common Core standards. Chris Franklin and Dick Schaeffer recommend the following two semester lower level sequence for future teachers, shared with us by the authors of the MET2 report. The Statistics Education of Teachers (SET) report currently being produced by the American Statistical Association (ASA) will provide more details about the content of these two courses. The report should be available online through the ASA in early 2015. We strongly recommend that schools refer to this document. This would be an outstanding sequence for future teachers, for schools that can support a two-semester sequence.

a. One-semester Introductory Statistics course emphasizing data analysis:

- formulation of statistical questions
- exploration of univariate data sets and comparisons among multiple univariate data sets
- introduction to the use of randomization in data production and inferential reasoning
- inference for means and proportions and differences of means or proportions; notions of p-value and margin of error
- introduction to probability from relative frequency perspective; additive and multiplicative rules, conditional probability and independence

b. One-semester Statistical Methods course

- bivariate categorical data: two-way tables, association, chi-square test
- bivariate measurement data: scatterplots, association, simple linear regression, correlation
- exponential and quadratic models; transformations of data (logs, powers)
- introduction to study design: surveys, experiments and observational studies
- randomization procedures for data production and inference
- introduction to one-way ANOVA

Appendix D: Sample Syllabus #1

Course Area: Statistics

Title of Course: Introductory Applied Statistics

Credit hours/semester: 3 or 4

Description of the target student audience: Similar to the range of students taking calculus.

Students majoring in the mathematical sciences as well as those in a variety of other fields (which will vary depending on the institution)

Course Description: An introduction to applied data analysis, designed to enable students to effectively collect data, describe data, and make appropriate inferences from data. Students are expected to communicate effectively about statistical results and to use a statistical software package for data analysis.

Proposed prerequisites: Basic algebra

How the course fits into a program of study: The course should be taken during the first two years of an undergraduate program in the mathematical sciences. It can be taken concurrently with calculus or any other sophomore-level courses in mathematics. Those students wishing to continue on in statistics are urged to take the probability and mathematical statistics sequence as well as any additional courses in advanced data analysis offered at the institution.

Course Outline:

- Data collection, including random sampling and design of experiments (2 weeks)
- Data description, including graphs and summary statistics for categorical and quantitative variables and relationships between variables (2 weeks)
- Introduction to the key ideas of estimation and testing, using modern resampling methods to build conceptual understanding (3 weeks)
- More on confidence intervals and hypothesis tests, using the normal and t distributions (3 weeks)
- Advanced tests, as time permits, such as chi-square tests, ANOVA, regression tests, multiple regression (3 weeks)

1. Students complete three data-analysis projects during the semester, each using a statistical software package and culminating in a written report (and, if possible, an oral report).
2. The course uses active learning, with regular in-class activities and group projects.
3. The course uses real data of interest to the students and emphasizes the connections of the subject to a wide variety of other fields.
4. The focus is on deep understanding of concepts such as variability of sample statistics, understanding random chance, estimation, and the meaning of the p-value, rather than on memorizing procedural methods.
5. Students are regularly presented with data in a real context, so that they experience the problem-solving and multiple approaches often necessary to move from a question of interest to reaching a conclusion.
6. Students gain extensive experience with effectively interpreting and communicating the results of data analysis.

Appendix D: Sample Syllabus #2

Course Area: Statistics

Title of Course: Introduction to Statistics

Credit hours/semester: 3 or 4

Description of the target student audience: Students majoring in mathematical sciences as well as those in other mathematically-related fields such as biology and economics

Course Description: Introduction to statistics for mathematically inclined students, focused on reasoning process of statistical investigations from asking question and collecting data to analyzing data and drawing inferences. Substantial use of statistical software.

Proposed prerequisites: Calculus I

How the course fits into a program of study: The course should be taken during the first two years of an undergraduate program in the mathematical sciences. Students wanting to continue in statistics should take a second course that introduces more advanced concepts and methods such as regression techniques and analysis of variance.

Course Outline:

Unit 1: Analyzing single binary variable

- Simulation, null model, statistical significance, p-value, binomial probabilities
- Two-sided test, significance level, rejection region, test decision, types of error, power
- Normal probability model, normal probability calculations, z-score, test statistic
- Standard error, critical value z^*, confidence interval, sample size determination
- Sampling, sampling bias, simple random sampling, precision

Unit 2: Comparing two groups on binary variable

- Two-way table, conditional proportions, segmented bar graph
- Binomial simulation analysis for comparing two proportions
- Normal approximation, standard error, two-proportion z-test, z-interval
- Observational studies, confounding variables, randomized comparative experiment
- Simulating randomization test for assessing statistical significance with 2×2 tables
- Hypergeometric probabilities, Fisher's exact test, relative risk, odds ratio

Unit 3: Comparing two groups with quantitative response

- Simulating randomization test for comparing two groups with quantitative response
- Histogram, measures of center and variability, five-number summary, boxplot
- Two-sample t-test, t-interval for comparing means
- Randomization test for paired data, paired t-procedures, prediction interval

Course Principles:

- Simulation-based inference is introduced throughout, prior to exact probability methods and theory-based techniques based on normal approximations.
- Students analyze genuine data from scientific research studies throughout.
- Students work through investigation activities to discover statistical concepts.
- Students complete course projects applying all aspects of statistical investigation process.