Math 300R: Additional improvement of Math 300

Daniel Kaplan

Sept. 24, 2022

This site holds the proposal for the Spring 2023 version of Math 300.

Background

Up through Spring 2022, Math 300 was organized around the Moore and Notz textbook: *Statistics: Concepts and controversies* 10/e. This book was designed for a non-technical audience of "consumers of statistics" but is dramatically outdated. For instance, it has no data science content and introduces only primitive statistical methods. A course with such shortcomings seems inappropriate for cadets going on to be officers who will inevitably have to work with modern data and methods.

In Fall 2022, Math 300 switched to a very different book, Ismay and Kim, *Statistical Inference via Data Science: A ModernDive into R and the Tidyverse*. The *ModernDive* book introduces computing on data in an accessible but modern way. It is the only well-known statistics text based on a data-science perspective. Nonetheless, the statistical inference portions of the book regress to the same sort of primitive statistical methods from *Concepts and Controversies*.

To support the Fall 2022 course using *ModernDive*, a complete set of roughly 35 Notes to Instructors (NTI) was written by Prof. Bradley Warner along with problem sets and other needed materials and deployed for the course.

This proposal is for additional improvements to Math 300, building on the Fall 2022 course but replacing the statistical inference portions of the course with more contemporary and general-purpose inference techniques and support for concepts and methods relevant to decision-making.

In the following, I refer to three different versions of Math 300:

- The Fall 2022 version of the course, using the *ModernDive* book, will be called *Math 300*.
- The previous version of the course, as taught for several years before Fall 2022, will be called *300CC*, which refers to the textbook then used, *Concepts & Controversies*.
- The course proposed in this document, a revision of part of *Math 300*, will be called *Math 300R*. The R stands for "revised."

Overall goals of Math 300R

The design of a course revision needs to take into account several factors:

- The target audience's anticipated technical ability and motivation and, therefore, the appropriate pedagogy and the balance between theory and practice to use in the course.
- Institutional goals that inform the prioritization of the topics included in the course.

- Constraints of class time and internal coherence of the course, that is, using later topics to reinforce student learning of the earlier topics.

Later, in the rationale section section of this document, I describe how I came to the following conclusions, but for now, a simple statement will suffice.

1. The target audience is humanities and social science majors, many of whom will not be confident in the use of calculus but all of whom have had previous exposure to R in the core calculus course.
2. Institutional goals (as revealed by discussions with humanities and social science departments and the wording of the catalog description of Math 300) include a substantial emphasis on *data science techniques* (data wrangling and visualization) and the use of statistical concepts and methods to support *decision-making*.

Statistical topics and framework

Transitioning from Math 300CC to Math 300 has already accomplished many data science goals. This proposal centers on the statistical topics/methods to be covered and the path through them.

The class-time demands of the new emphasis on data science techniques in Math 300 (and retained in Math 300R) dictate that the statistical concepts and methods be taught more compactly than in Math 300CC. Low-priority, legacy topics from Math 300CC should be dropped. (The GAISE report provides some guidance here.) We can use to advantage that students in Math 300 already see many modeling-related topics in Math 141Z/142Z. Since students already have a background in model-building and computing, we can choose statistical topics that relate well to decision-making.

A traditional path for statistical methods starts with descriptive statistics (e.g., standard deviation) and then presents "1-sample" statistics (e.g., mean, proportion) and inferential techniques (confidence interval, hypothesis test) in that context. Next comes the inferential techniques for the analogous "2-sample" statistics (difference in mean, difference in proportion), followed by inference techniques for regression.

This path is unnecessarily long for our students since regression encompasses all the traditional methods.[1] Framing statistical inference in the context of regression avoids the need to teach method-specific calculations or cover the variety of formats for non-regression test results. Regression is part of the data scientist's standard toolbox and relates well to more advanced data techniques such as machine learning. The *ModernDive* textbook uses regression as the segue from the first block (about data wrangling and visualization) to the third block (about inference).

[1]The chi-square test is an exception, but the actual statistical setting is better served by logistic regression.

Additional streamlining comes from motivating statistical inference using a simulation approach. Simulation draws on two conceptually simple data operations: resampling and permutation (shuffling). This approach is well established in the statistics community and is considered by many (including the *ModernDive* authors) to be a better pedagogy than the traditional formula-and-distribution presentation of statistical methods. Since Math 300 (and 300R) students will already have worked with wrangling and visualization, they will be well prepared to work with the data generated by repeated simulation trials.

The focus on decision-making in 300R appears in the addition of new concepts and techniques treated minimally in traditional statistics courses. These include *risk*, *prediction* (and its close cousin *classification*), *causality*, and *confounding*. Introductory epidemiology courses provide a model for teaching about risk, causation, and confounding. The pedagogy for these topics in Math 300R comes from the epidemiology course I introduced at Macalester. In addition, Math 300R draws on my decade of experience teaching causality as part of an introductory statistics course. (See the causation chapter of my *Statistical Modeling* text.)

The *Statistical Modeling* pedagogy for causality uses directed acyclic graphs (DAGs) and causal simulations based on them. Unlike resampling and permutation, which re-arrange existing data, the DAG simulations generate synthetic data with specified properties (such as effect sizes). Simulations allow a concrete demonstration of the extent to which regression techniques can and cannot recover causal information from data.

The DAG-simulation approach lends itself naturally to the demonstration of statistical phenomena such as sampling variation and estimation of prediction error. As an example, consider the statistical fallacy of regression to the mean, as with Galton's finding about comparing children's heights to their parents'. The natural hypothesis that heights are determined by genetic and other factors is represented by this DAG:

$$\epsilon \rightarrow parent \longleftarrow GENES \longrightarrow child \leftarrow \nu$$

In this DAG there is no causal mechanism included for "regression to the mean." However, Galton's empirical finding is replicated by data from the DAG-simulation.

Scope of the proposed changes

Math 300R will retain the first 17 lessons of Math 300. All teaching materials for this part of the course will be used unaltered. (Exception: revisions to Math 300 the Fall 2022 teaching team deems appropriate. Such revisions are not part of this proposal.)

The following 19 lessons are entirely refactored and based on new readings, NTIs, exams, and other materials. Objectives for each of these 19 lessons are itemized here.

- The corresponding *ModernDive* chapters are not used in Math 300R.

- The software used is the same as that used in the first half of the *ModernDive* book, specifically the `ggplot2` graphics package and the `tidyverse` data wrangling packages. However, the `infer` package used in the second half of *ModernDive* is dropped.

The theme of the refactored 19 lessons is "informing decisions with data." Statistical approaches that can inform decision-making include anticipating the impact of interventions, predicting individual outcomes, and the quantification of risk. These are all included in Math 300R.

Topics to be de-emphasized are the algebra of computing confidence intervals and p-values and the (controversial) role of p-values as a guide to practical "significance." About half of a traditional course is about the construction of confidence intervals in various settings and, more or less equivalently, the conversion of data into p-values. However, in the contemporary era, when "observational" data are collected *en masse*, p-values can become very small ("significant") even when the relationship under study is slight and insubstantial.

Confounding and methods for dealing with it (statistical adjustment, experiment) are treated substantially in Math 300R. Decision-making about interventions often relies on understanding causal effects. The possibility of confounding is a major source of skepticism about making causal judgments. In a world where much data is observational, the sweeping principles that "correlation is not causation" and "no causation without experimentation" do not support making responsible conclusions about causal connections and the need to make decisions even when data cannot provide a definitive answer. Decision-makers need this support.

Rationale for course revisions

Relationship to Math 357 and Math 377

DFMS offers three courses satisfying the statistics component of the Academy's core requirements: Math 300, Math 357, and Math 377. In designing 300R, attention should be paid to the reasons for supporting three distinct courses. The catalog copy lays out the differences in terms of intended student major, software, mathematical background, and orientation to data science.

Intended student major: The catalog says, "Math 300 is designed primarily for majors in the Social Sciences and Humanities." while "Math 356 is primarily designed for cadets in engineering, science, or other technical disciplines. Math majors and Operations Research majors will take Math 377." Math 377 is also the intended course for prospective Data Science majors, although this is not in the catalog.

Software: The catalog does not describe any software component for either Math 300 or Math 357, but states that, in Math 377, "modern software appropriate for data analysis will be used." In reality, as of Fall 2022, much the same software is used in all three courses: R with

the `dplyr` package for data wrangling, `ggplot2` for data visualization, and "R/Markdown" for creating computationally active documents.

One difference between Math 300 and 357/377 relates to computer programming. Both 357 and 377 include content about the underlying structure of the R language, object types, the construction of functions, and arrays and iteration. In contrast, Math 300 is based on a small set of command patterns using data frames. Students see R in Math 300 more or less as an extension of what they learned in 141Z/142Z; what's added is a few statistical and data-wrangling functions and a handful of new graphics types.

Students' mathematical background: Math 377 explicitly refers to "calculus-based probability." Math 300 and 357 share identical catalog copy, though in reality Math 357 and Math 377 use the same textbook. Calculus is indeed necessary for the probability topics in Math 357 and 377. My interpretation is that Math 300 should serve as a safe haven for those who lack confidence in their calculus skills. Both the Fall 2022 edition of Math 300 and the proposed Math 300R serve this role as safe haven.

Orientation to Data Science: Starting with the Fall 2022 edition, Math 300 develops and draws on data-science skills for wrangling and visualization. In this, the new Math 300 is in line with both Math 357 and 377.

The above analysis indicates that Math 300 and 300R should diverge from Math 357/377 in these ways:

1. Math 300R should make little or no use of calculus operations.
2. Math 300R should include little consideration of probability distributions or (non-automated) calculations with any but the simplest.
3. Math 300R should be computational, but should not draw heavily on computer programming skills such as types of objects, arrays, indexing, and loop-style iteration. Use of R/Markdown documents should be considered as a pedagogical choice, and retained or discontinued based on how it contributes to student success in the other areas of the course.

In addition, I suggest that …

4. Math 300R include some work with assembling/curating data using spreadsheets and basic data cleaning with spreadsheets. Awareness of the ubiquity of data errors and a basic understanding of how to deal with such errors is an important component of working with data. (This is not to suggest that data *analysis*, *modeling*, and *graphical* depiction be taught using spreadsheets, which are notoriously unreliable, difficult, and limiting for such purposes. Spreadsheets are, however, appropriate for the phase where non-tabular data is transcribed into a tabular arrangement.)

Institutional goals

It can be difficult to translate broadly stated institutional goals to apply them to a single course. However, catalog descriptions of programs and individual courses provide some assistance. Here is the catalog copy for Math 300 (which is identical to the catalog description of Math 357).

> Math 300. Introduction to Statistics. An introduction in probability and statistics **for decision-makers**. Topics include basic probability, statistical inference, prediction, data visualization, and data management. This course emphasizes critical thinking among **decision-makers**, preparing future officers to be **critical consumers of data**. (*Emphasis added.*)

I interpret the final sentence as a description of the overall objective of the course:

> **Overall objective**: *Prepare officers to* ***use data*** *to* ***inform decisions****.*

Returning to the idea that the topics listed in the catalog copy ought to be interpreted as serving the overall objective of the course, let's consider those topics one at a time:

1. data management
2. data visualization
3. prediction
4. statistical inference
5. basic probability

(1) Strictly speaking, as a term of art the phrase "data management" is business jargon describing enterprise-level activities that are unrelated to the other items on the list. It would be unheard of to include it, in this strict sense, in a statistics course. I believe the intent of the phrase to be better served by terms like "data wrangling," "data cleaning," "database querying," and such which make up an important part of "data science." Data wrangling is a major feature of Math 300 launched and is covered using professional level computing tools well suited to both small and large data. But whatever "data management" might reasonably be taken to mean, it was utterly ignored in Math 300CC.

(2) "Data visualization" is generally taken to be the process of using graphics to discover and highlight patterns shown in data. Math 300CC included only statistical graphics such as histograms, box-and-whisker plots, and basic "scatter plots." Math 300 adds to this modern modes of graphics such as transparency, color, and faceting that make it possible to display relationships among multiple variables. The software used in Math 300 is the professional-level `ggplot2` which provides the ability to increase the sophistication and generality of data display, using for example density graphics such as violin plots. As such, Math 300 is a big step on the road to rich data visualization. Some of these will be introduced in Math 300R in the second half of the course.

(3) "Prediction" is a central paradigm used in the important area of "machine learning." It is also an often used method used to inform decision making and characterize risk, for instance, by indicating the distribution of plausible outcomes. Math 300CC emphasized paradigms such as hypothesis testing and confidence intervals that are not aligned with making and interpreting predictions. Math 300 focuses on these same paradigms. Math 300R will treat prediction as a central statistical path, as well as highlighting its proper use, interpretation, and evaluation.

(4) "Statistical inference" is traditionally taken to mean the calculation and interpretation of hypothesis tests and confidence intervals in various simple settings. Such settings include the "difference between two means," the "correlation coefficient," and the "slope of a regression line." Math 300CC introduced a handful of such settings, providing distinct formulas for each of them. The "controversies" referred to in the title *Concepts and controversies* includes the problematic interpretation of "p-values" and the need to use random sampling and/or random assignment in data collection to get "correct" results. Math 300 retains the emphasis on confidence intervals and p-values in the simple settings, but emphasizes a more general and accessible methodology based on bootstrapping and permutation tests.

Unfortunately, appealing to random sampling/assignment is often whistling past the graveyard, since these idealized data collection processes are rarely available. Instead, professionals include "covariates" in their data collection in order to "adjust" for the factors that would have been scrambled into insignificance by random sampling/assignment if it had been available. Math 300R incorporates covariate methods and highlights the importance of identifying appropriate covariates.

(5) "Basic probability" can mean different things to different people. In most introductory statistics courses it refers to the construction, calculation, and study of named distributions such as the binomial, normal, chi-squared, t, etc. Such distributions play an important role in the statistical theory of confidence intervals and hypothesis testing. That is, they are support for statistical inference. Math 300CC followed the traditional pattern of having students memorize which distribution is relevant to which setting and using printed tables for calculation. As described earlier, Math 300 provides a much more natural route to inference through bootstrapping and permutation tests.

What's left out in this conception of basic probability is the support for decision making. Essential to this is the proper use of "conditional probability." Math 300R emphasizes appropriate use and interpretation of conditional probability, seen most clearly in the "classifiers" part of the course.

Faculty opinions

Insofar as faculty internalize the goals of the institution, their views can point to ways that existing courses do and do not reflect those goals.

Within DFMS and other departments, there is a general discontent that Math 300CC was not doing what it ought to. Reasons for this can be seen by examining the textbook used in Math 300CC. The book has clear deficiencies, among which are:

- the material is out of date and does not reflect any of the consensus recommendations (such as GAISE) developed in the last 30 years.
- it does not use data at any level beyond hand calculation.
- it does not deal with decision making at any serious level. (The only decision formally supported is whether or not to reject the Null hypothesis.)

The opinions of faculty *outside* DFMS can also be an important guide to institutional priorities. In AY 2021-22 I contacted the departments in the social sciences and humanities. Three of these—history, political science, economics—responded with interest. Discussion with groups of faculty from these departments elucidated a number of points:

- The faculty most highly valued the development of data-science skills such as computing for data wrangling and data visualization.
- The then-current version of Math 300 did not contribute to the development of such skills.
- Math 357 is not seen as an appropriate alterative to Math 300, both because of perceived difficulty of 357 and because faculty do not value the emphasis on probability distributions seen in 357.

From my experience at Macalester and in conducting reviews at many colleges, I am often wary of the motivation of faculty in other departments. These can represent a desire for service courses like Math 300 to cover discipline-specific techniques. However, the faculty I spoke to also had an eye on what their students will need for their post-graduation jobs. Particularly the USAF officers drew on their field experience in areas such as military intelligence.

Based on these findings, the group of faculty planning for revisions to Math 300 made an easy decision: replace the textbook with one oriented to data science. We selected the *ModernDive* book, which is unique among introductory statistics textbooks in starting out with data wrangling and visualization. This change of textbook addresses the "use data" part of the course objective stipulated above.

The other part of the objective—**inform decisions**—remains problematic even with the switch in the Math 300 textbook. Discussions I had with the *ModernDive* authors made clear that their purpose in writing the book was to provide a way to introduce data science into introductory statistics, but that they stuck to the traditional hypothesis-testing/confidence-interval framework in order not to make the change too daunting for instructors thinking of adopting

the text. In other words, they were not trying to turn the topic toward decision-making with data, the motivation of the ideas presented in this proposal for Math 300R.

Plan of work

1. Early October 2022: Preliminary approval, with appropriate modifications, of the proposed objectives.
2. October 2022: DTK will draft new day-by-day NTIs for the second half of the course in the same style as the existing NTIs for the first half of the course. In the process of drafting, there will likely be some re-arrangement and modification of the objectives in (1).
3. November 2022: With the draft NTIs in hand, a faculty team will make a more detailed examination of the proposed objectives. I recommend that this examination be structured as a set of hour-long discussions, one for each of the five divisions described in **?@sec-topics**.
4. November/December 2022: DTK (and others, as interested) will assemble student readings to replace the second half of *ModernDive*. Much of the content already exists in the form of a draft textbook by DTK. These will be re-arranged to correspond to the day-to-day objectives as determined in (3).
5. January/February 2023: The first 18 lessons of 300R will be taught as a repeat of those lessons from Math 300 Fall 2022. DTK will participate mainly as an observer.
6. January/February 2023: Revision and refinement will be made of the readings and NTIs in (3) and (4) above.
7. March/April 2023: Teaching the new lessons. DTK will participate as an instructor for these lessons.