Redesign of a First-Year Theory Course Sequence in Biostatistics

Jesse D. Troy[1,*], Kara McCormack[1], Steven C. Grambow,[1] Gina-Maria Pomann[1] & Gregory P. Samsa[1]

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA

*Correspondence: Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA. Tel: 1-919-668-2932. E-mail: jesse.troy@duke.edu

Received: February 22, 2022 Accepted: August 26, 2022 Online Published: October 19, 2022

doi:10.5430/jct.v11n8p1 URL: https://doi.org/10.5430/jct.v11n8p1

Abstract

This communication describes the process and results of a curriculum review of a first-year sequence of courses in statistical inference within a Master of Biostatistics program. Our primary aim was to develop an innovative course in statistical theory that meets the needs of a diverse student audience, the majority of whom are seeking a terminal master's degree while a minority will pursue PhD training in biostatistics. The main results were (1) different course paths for job-bound and PhD-bound students; and (2) the development of an innovative first course in statistical inference, which is a computationally-aided self-discovery of a salient (albeit not comprehensive) set of key concepts and techniques pertaining to statistical inference. The redesign process addressed a key conceptual barrier: namely, the unexamined assumption that deductive proofs are a necessary condition for rigorous presentation. Consistent with the principles of constructivism, we navigated this barrier by redefining the task to which pedagogic rigor should be applied: namely, to help students to develop a sound mental map of statistical inference. We believe that the approach we used to accomplish this redefined task could be generalized to additional aspects of statistical education, among others.

Keywords: constructivism, curriculum design, statistical inference

1. Introduction

Our context is a 2-year Master of Biostatistics (MB) program. Approximately 30% of our students proceed directly to doctoral work after graduation, with the remainder entering the workforce, typically at pharmaceutical companies, contract research organizations, or academic medical centers. In their first year all students take a theory sequence, a data analysis sequence, a programming sequence, and a practice of biostatistics sequence (G. P. Samsa, 2020; Troy et al., 2021). This communication describes the results of a curriculum review of the theory sequence, and the accompanying redesign and plan for evaluation.

Our intended audience extends beyond those who develop curricula in statistical inference, and includes instructors in mathematics, statistics and related disciplines who are attempting to navigate the disconnect between the intuitive way that mathematics is performed (and conceptualized by its practitioners) and how it is explained *via* deductive proofs (Ben-Zeev & Star, 2001). Simply stated: while deductive proofs are an essential component for building the intellectual edifice of mathematics, they aren't necessarily an effective way to teach mathematics to others. Or, perhaps: in mathematics, symbol manipulation without sufficient context isn't an effective path to deep understanding and mastery (Wilkerson-Jerde & Wilensky, 2011).

Here, we discuss both the content and the process around curriculum redesign in a mathematically based discipline such as biostatistics, including a plan for evaluation. As it turned out, a key conceptual barrier was the unexamined assumption that deductive proofs are a necessary condition for rigorous presentation. We navigated this barrier by redefining the task to which pedagogic rigor should be applied: namely, to help students to develop a sound mental map of statistical inference. We believe that the approach we used to accomplish this redefined task could be generalized to additional aspects of statistical education, among others. Indeed, it is our hope that our experience will be of interest to those in less mathematical disciplines, as an invitation to identify and address similar barriers to effective instruction within their fields of study.

2. Current Structure

Our theory sequence is relatively traditional. Indeed, its structure was initially derived from programs which offer the same courses to all graduate students during their first year, followed by a qualifying examination, at which point they differentiate into masters and doctoral. In such programs the theory sequence is effectively designed with PhD students in mind, as these students will take additional courses in statistical inference and related topics and must master the prerequisite information for those courses, with the ultimate goal of completing a dissertation which, of course, will typically involve derivations, proofs, and the like.

At present, the MB program does not offer separate inference courses for job-bound students (JBS) and PhD-bound students (PBS). One rationale is the desire not to prematurely place students on a job-bound track, as some begin with this expectation only to discover that doctoral training is a better match for their interests. Another rationale is that all students, including PBS, could benefit from a foundational review, especially if that review adopts a different and more general perspective than they have previously encountered.

3. Philosophical Foundation

Originally, an additional rationale for providing the same training to all first-year students was the view that "you can't properly apply statistical techniques unless you understand how they were derived", and thus that it was important that all students encounter an inferential sequence which, even though less complex than measure-theoretic advanced inference courses, was nevertheless rigorous and comprehensive. This view was not universally shared by our teaching faculty, yet was especially prevalent among our more methodologically-oriented instructors, who had a significant voice in the original design of the theory sequence.

4. Critique

The notion that you can't properly apply statistical techniques unless you understand how they were derived was sufficiently prevalent and strongly held as to perhaps constitute a meme. (Indeed, for simplicity of exposition this will henceforth be referred to as "the meme".) During the initial development of the theory sequence the meme was not critically examined. In its favor, the meme certainly seems plausible. Moreover, when presented with the choice of having a student understand how a particular statistical technique was derived versus not having that understanding, anyone would prefer the former. This is not even to mention that, as a practical matter, advocating that students need to know less rather than more isn't necessarily a recipe for persuasion.

Nevertheless, the meme has two problematic elements. The first is the general principle of educational pedagogy that goals and objectives are best stated in terms of actionable skills -- in other words, "to do" rather than only "to know" (Anderson et al., 2001). The second is the presence of numerous counterexamples to the meme. For example, the previous iteration of our theory sequence included a demonstration that the density for the standard normal distribution integrates to 1 (i.e., a requirement for proper distributions). This demonstration requires a page of algebra, including a critical juncture involving transformation to polar coordinates. Would a JBS who is using the normal distribution in an applied setting be harmed if this demonstration was included within the optional supplemental materials, or if they simply accepted the value of this integral on faith? We think not.

As another counterexample, the previous iteration of our theory sequence had as a stated goal the ability to understand not just the statement of the Central Limit Theorem but its derivation, which requires distinguishing between various types of convergence. Would a JBS who is applying the Central Limit Theorem be harmed if, rather than encountering nuances around convergence, they were trained to distinguish between large-sample and small-sample problems, and recognize that for the Central Limit Theorem to apply the quantity in question must be expressible as a mean? Again, we think not, especially if the student has sufficient skills in creating simulations to empirically explore the implications of large-sample approximations when needed.

If this critique of the meme is accepted, an implication is that, by delivering PhD-focused course content to JBS, we were imposing a gratuitous requirement. Clearly, if progress was to be made the question at hand needed reframing. An initial step in this reframing was to consider whether the meme, which was a core belief of some very knowledgeable and respected colleagues, might be true, but just not as literally stated. (Or, perhaps, that some elements of the meme are literally true but others would benefit from restatement.) In other words, we asked what underlying constructs the meme might actually represent.

5. Underlying Constructs

We believe that the meme relies upon two underlying constructs. The first construct is that some of the information in the theory sequence is prerequisite knowledge upon which later instruction will build. For PBS this prerequisite knowledge includes, although it isn't limited to, techniques for accomplishing derivations and proofs, and so it can be argued in this case that "understanding how statistical techniques were derived" actually means "mastering techniques of derivations and proofs for effective application of these techniques later in the program (e.g., when developing proofs as part of their dissertations)". This restatement has the additional advantage of describing things that students are expected to do rather than simply to know.

Restating the first construct as above also serves to illustrate that it doesn't describe a skill which is critical for JBS. Nevertheless, this latter group of students will take additional courses, and these courses will require an effective working knowledge of some of the concepts and techniques encountered in the theory sequence. As a simple example: they will encounter likelihood functions throughout the curriculum, and should have a basic mastery of the use of calculus to find parameter values which maximize those functions. Accordingly, we considered a key task to be identifying those concepts and techniques which will be used by both PBD and JBS later in the curriculum, those tasks which will be only required of PBS, and then differentiating between the two.

The second embedded meme construct is the notion that some sort of prerequisite knowledge is required in order to correctly apply statistical techniques, which is a skill that all students must master. The above counterexamples suggest that this prerequisite knowledge doesn't pertain to derivations. Upon reflection, we argue that this prerequisite knowledge actually pertains to the structure of statistical inference -- that is, understanding how everything fits together and being able to act accordingly. In the language of constructivism: what is needed is an explicit and actionable mental map of how experienced practitioners understand and apply the core concepts of statistical inference (Biggs, 1996). Some implications of this observation are discussed later.

7. An Unintended Barrier to Progress

During discussions around the theory sequence, an additional implication of the meme emerged -- namely, that since "you can't properly apply statistical techniques unless you know how they were derived", and since we want our students to properly apply statistical techniques, then "we shouldn't reduce the amount of information about derivations which this sequence covers and, more generally, we shouldn't drop any content". Indeed, in the extreme case this suggests an unexamined assumption that any proposal to drop content reflects an insufficient commitment to properly applying statistical techniques.

As with the meme itself, we argue that although the assertion that we shouldn't drop any content isn't literally true it embeds an important insight -- namely, that what JBS shouldn't receive is a watered-down version of a course for PBS. The sound educational instinct underpinning this insight is that, for example, removing some topics makes it more difficult for the student to discern how the remaining ones fit together, presenting results without adequate background makes it more difficult for the student to correctly apply them in new situations, etc.

8. The Impasse

We were seemingly at an impasse. Our curriculum was imposing gratuitous mathematical and statistical requirements for JBS, but we worried that removing those requirements might make matters worse. Moreover, our instructors had consistently provided feedback that students (including PBS) appeared to struggle in applying the principles of inference later in the curriculum, and that at times preoccupation with symbol manipulation seemed to have replaced deep understanding. This feedback suggested to us that not only should we consider doing things differently for JBS, but also for PBS.

Indeed, this insight ultimately provided a way to break the impasse. We recognized that the task for JBS was not to do less of the same thing, but instead to do something different and more effectively. Ideally, whatever changes were made would also enrich the experience for PBS, and satisfy our instructors that all our students were gaining the ability to better apply the theory sequence's content.

9. Curriculum Changes

Different programs would likely make different changes in response to the above considerations. The structural changes that we chose to make replaced a two-course sequence offered to all students that tried to serve both JBS and PBS with the following:

- For all students, the first course covers general concepts of statistical inference
- For PBS, the second course is an enhanced version of the previous second-semester inference course
- For JBS, the second course covers general concepts in causal inference

This approach allows core concepts of inference that serve both JBS and PBS to be taught to all first-year students. Subsequently, because it is only serving the PBS audience, the second-semester inference course moves at a quicker pace, is a comprehensive treatment, and can pay greater attention to regularity conditions, derivations, and other mathematical niceties. It only required modest changes, as the previous version was generally acknowledged to be excellent.

The course in causal inference was previously a second-year elective. It covers concepts such as difference in strength of inference between randomized trials and observational studies, confounding, interaction, the use of directed acyclic graphs (Hernan & Robins, 2020) to represent conceptual models of the underlying science, all of which are of direct utility to the practice of biostatistics within a job setting.

The class schedule is arranged so that JBS can "sit in on" (i.e., essentially, informally audit) the second-semester inference course and the PBS can sit in on the causal inference course, if desired. PBS can take (or sit in on) the causal inference course as an elective, if desired. The first-semester course on general concepts of statistical inference is new, we believe to be innovative, and is described next.

10. First-semester Theory Course

10.1 Organization

The first-semester theory course is intended to be a computationally-aided self-discovery of a salient (albeit not comprehensive) set of key concepts and techniques pertaining to statistical inference. The classroom is flipped, and significant time is devoted to working through detailed exercises and communicating the results. Although each module has a theme (e.g., joint and conditional probability distributions), ideas are interleaved throughout the course rather than being presented entirely in isolation. The primary textbook is In All Likelihood (Pawitan, 2013), and more traditional texts are used as supplemental resources.

In contrast to axiomatic approaches, our approach is "illustrate first, then loosely define" rather than "precisely define, then illustrate". As an illustration of the choice of topics, we concluded that both JBS and PBS would benefit from exposure to a standard set of statistical distributions (one of these being the binomial). For these distributions, the content which was modified pertained to calculating moments (especially for continuous distributions), whether directly or via moment-generating functions, and, algebra-intensive exercises such as working with conditional distributions of continuous variables. The content which was added pertained to describing how the distributions in question arise in actual practice, for example, how a set of independent Bernoulli trials can be identified. Also added because of their practical relevance were limiting distributions, for example, the Poisson distribution as the limit of many Bernoulli trials each with a small probability of success. However, students are not required to master the relevant algebraic manipulations (e.g., they aren't required to master the various ways which "e" arises as a limiting case). Simulation exercises illustrate, for example, the implications of replacing N Bernoulli trials, all with the same probability of success θ with N Bernoulli trials whose average probability of success is θ, but with different values of θ from trial to trial. Finally, while some traditional content was modified, like manipulation of moment generating functions, the course still retains coverage of these topics, e.g., what moments are and how they arise in practice, while the more traditional mathematical presentation is provided as supplementary material for PBS.

10.2 Example Module on Joint and Conditional Probabilities

As an example, Appendix 1 presents the module on joint and conditional probabilities. It uses the simplest possible case: namely, a 2x2 table describing the operating characteristics of a diagnostic test. The context is intended to be relevant to statistical practice for both JBS and PBS and, indeed, other first-semester courses use this example as well.

To begin, joint and conditional probabilities are illustrated and defined by example. Marginal probabilities are defined intuitively by first defining a subset of the original population as a new sample space, and then illustrating that the same results can be obtained using the original sample space and the law of total probability, which is introduced by example rather than definition. Conditional probabilities are treated similarly. Probabilities and sample spaces aren't described axiomatically, although many students (especially PBS) will have encountered the relevant axioms in an undergraduate course, and an axiomatic treatment is available in the supplemental materials.

The above probabilities are then used to produce Bayes theorem, and students are asked to produce Venn diagrams as an active learning exercise to solidify their understanding.

Each module includes a section on "how this is used in statistics". One of the goals of this section is to help students develop a mental map around inference. Sometimes "how this is used" is quite specific. For example, in the module on joint and conditional probability distributions this section includes a practical application of Bayes theorem to calculate the positive predictive value of a diagnostic test, including asking the student to generate a plain English explanation of their results.

The "how this is used in statistics" section also illustrates how ideas are interleaved rather than being treated in isolation. For example, the idea that some probability calculations can be simplified by applying the notion of conditional independence is first illustrated by standard card problems, and then extended to illustrate why sampling without replacement can be similar to sampling with replacement, but only for when the sampling fraction is small. This isn't quite a "derivation" in the traditional sense of the term, but is intended to illustrate a core concept that students can apply in their future work.

The "how this is used in statistics" section includes other similar content, such as active learning exercises which combine calculation and explanation.

10.3 Example Module on Introduction to Likelihood Functions

As another example, Appendix 2 presents the module on introduction to likelihood functions, using the binomial distribution as a running example. It builds upon, among others, a module on Bernoulli trials which also serves to illustrate the definition and straightforward calculation of its mean and the variance. (Indeed, we "derive" these moments of the binomial distribution not from its probability mass function directly, nor by a moment generating function, but by simply noting that a binomial distribution is the result of summing independent Bernoulli random variables, thus eliminating unnecessary algebra. The standard algebraic derivation is included in the supplemental materials.) Its content is subsequently used in multiple modules -- for example, hypothesis testing is illustrated within the context of a likelihood ratio test.

This exercise begins by making the distinction between a probability distribution (mass) function and a likelihood function, and involves students calculating the values of functions and then graphing the results. In part, a reason for emphasizing this point is feedback from instructors that students weren't universally clear about the nature of functions. Programming in R is used to ensure that students work through each step in the argument and thus solidify their understanding.

In this module, because likelihoods are so ubiquitous in statistics, the "how this is used in statistics" section is mostly just a segue into the next module. In that spirit, it introduces a geometric interpretation of the likelihood function, and links this to concepts such as signal versus noise, Fisher information, and the interpretation of hypothesis tests and confidence intervals. In passing, we note that even though subsequent modules simplify the maximization problem in the usual way by dropping constants and working with the log of the likelihood, for this introductory module it is pedagogically more straightforward to use the actual likelihood function, thus keeping intact the exact linkage between the probability distribution function and the likelihood function.

10.4 Evaluation Plan

Our evaluation of the revised curriculum will be discussed in a separate communication, once sufficient experience is accumulated. Briefly, evaluation will be performed at multiple levels, ranging from a micro-level to a macro-level. For example, a micro-level assessment would compare two different approaches to teaching how to integrate a probability density function to find the cumulative distribution function within a single course module using a randomized design. A macro-level assessment asks how well students can apply fundamental concepts of statistical inference within the context of significant and integrative work products such as the Masters Qualifying Examination (a 1-week take-home examination) (G. Samsa, 2021), internships and the master’s thesis. The macro-level assessment is the more challenging of the two, both because of the broader nature of the skills to be assessed, and because of the time required to asses (e.g., students complete their thesis 3 semesters after taking the theory course). In addition, the macro level assessment is not amenable to randomized experiments for several practical reasons, chiefly: 1) the unit of intervention is large, i.e., a semester-long course, and there are practical and financial constraints on conducting a concurrent randomized comparison; and 2) the manifestation of the construct we are attempting to address is most likely to appear later in the curriculum, such as during performance on the qualifying exam or thesis project (in the language of clinical trialists, the outcomes we want to measure are distal rather than proximal). Thus, for the macro level assessment will rely primarily on comparing qualifying exam performance pre-

and post-implementation of our new curriculum. We believe these plans are consistent with contemporary guidance for evaluation of education programs (What Works Clearninghouse, n.d.).

11. Discussion

This communication describes the process and results of a curriculum review of a first-year sequence of courses in statistical inference within a 2-year masters' program in biostatistics. One major change is that JBS and PBS now take a different sequence of courses: identical during the first semester but different in the second. Another major change is that the initial course has been redesigned to be primarily conceptual -- a more detailed rendering being "a computationally-aided self-discovery of a salient (albeit not comprehensive) set of key concepts and techniques pertaining to statistical inference". A distinguishing feature of our approach, consistent with principles of constructivism, is attention to explicitly describing the intellectual edifice (i.e., mental map) of statistical inference: in essence, we describe and then practice "how to properly think about statistical inference".

These efforts are consistent with Bloom's Taxonomy of learning objectives (i.e., in ascending order: remember, understand, apply, analyze, evaluate, create) (Anderson et al., 2001). At a minimum, both JBS and PBS need to reach "apply" and, indeed, part of the impetus for the redesign was feedback from instructors that they were having difficulty in consistently doing so. The new first-semester course is calibrated to this level of the taxonomy. For JBS, the second-semester course is also calibrated to "apply", whereas for PBS the second course is calibrated to "create".

The previous iteration of the theory sequence implicitly recognized that JBS need not create derivations and proofs, which received relatively less emphasis on examinations, but the problem that the content was in effect organized with PBS in mind remained. In essence, by trying to simultaneously serve two audiences we were slowing down the PBS while still imposing gratuitous requirements on JBS. One might reasonably ask whether we were simply exchanging one set of problems with its opposite. We believe not. Although technically adept, our PBS were not universally clear about how thinking about inference is actually structured, hadn't yet fully mastered "apply", and so a course at an application level with a conceptual focus actually provides new information to them. We also utilize an extensive archive of more traditional instructional materials as supplementals and so, for example, a PBS who wishes to encounter a more axiomatic treatment of these topics (e.g., as preparation for their next theory course) can do so.

Implementing this approach requires close coordination with other courses in the curriculum. For example, students are assumed to be relatively facile with R on day 1, with R Markdown being ideal, as many of the exercises combine coding, results and explanation. Our programming sequence is designed with this in mind and, indeed, students are exposed to the basics of R in a "preorientation curriculum" taken before beginning the first semester (Neely et al., 2022; G. P. Samsa, 2020). In the other direction, for example, the data analysis course benefits from distributions being introduced in a specific order. Evaluation is ongoing.

12. Comment

Although our intention is not to overly focus on the sociology of academia, we believe that a short commentary about our deliberations around curriculum redesign might be helpful to others. Like many other programs, we began with a theory sequence which was implicitly designed with PBS in mind. This sequence was quite similar to what our instructors, all of whom have doctorates, had experienced as PhD students, and which served them well. As one of the most difficult challenges in teaching is to put oneself in the place of a new learner, and perhaps a new learner with different educational goals and background than you, it isn't necessarily intuitive for an instructor to recognize that a course which worked well for them might not be appropriate for JBS. Central to this argument is the seldom discussed disconnect between the typical axiomatic approach to teaching statistics and the way that it is practiced.

We are unaware of a literature which discusses the disconnect between pedagogy and practice in statistics, and so will cite the literature in the related discipline of mathematics. Mathematics is described as an axiomatic system (Snapper, 1979) but expert practitioners don't learn new information by following proofs step-by-step. Instead, practitioners focus on examples, constructions and prototypes, essentially by applying what we've called a mental map (Wilkerson-Jerde & Wilensky, 2011; Wilkerson, 2008). Thus, our approach is consistent with the recommendation of the National Council of Teachers of Mathematics to "build procedural fluency from conceptual understanding" and which states that "effective teaching of mathematics builds fluency with procedures on a foundation of conceptual understanding so that students, over time, become skillful in using procedures flexibly as they solve contextual and mathematical problems" (Leinwand et al., 2014). Therefore, we feel that our approach is consistent with modern contemporary approaches to pedagogy in mathematics and related fields.

Theoretical arguments aside, other more pragmatic aspects of academia also tend to support a pedagogical status quo. Namely, in many academic departments the masters' program funds the PhD program, with the latter being of greater interest to faculty members because PhD students can assist them with their research whereas masters' students often cannot. The risk is that the masters' program will be treated mostly as a funding source, with insufficient attention to quality of instruction. We believe differently: since in our field of biostatistics there is very high demand from employers, which enables our students to have prosperous careers without pursuing a PhD, it is our responsibility to send our graduates into the workforce with the skills they need to be successful in their careers. Therefore, it is critical that we determine when JBS should receive different instruction than PBS, so that the learning experience for both groups of students is as effective as possible. In the present context, an additional benefit of determining that JBS should receive different instruction than PBS is in eliminating gratuitous mathematical pre-requisites. In our experience, some faculty members wrongly conflate doing so with reducing the level of programmatic rigor. In this case, it can be helpful to recall that such gratuitous requirements often disproportionately impact underrepresented minorities (Freedle, 2003), and thus their elimination also serves to increase the level of equity within the program.

In reflecting upon our experience, a key step toward progress was recognizing an unexamined assumption / conceptual barrier which had unnecessarily constrained our curriculum design. It is our hope that this will be of interest to those in less mathematical disciplines, serving as an invitation to identify and address similar barriers to effective instruction within their fields of study.

References

Anderson, L., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., PR, P., Raths, J., & Wittrock, M. (2001). *Taxonomy for Learning, Teaching, and Assessing, A: A Revision of Bloom's Taxonomy of Educational Objectives* (L. Anderson, B. Bloom, & D. Krathwohl (eds.)). Longman.

Ben-Zeev, T., & Star, J. (2001). Intuitive mathematics: Theoretical and educational implications. In B. Torff & R. Sternberg (Eds.), *Understanding and teaching the intuitive mind: Student and teacher learning* (pp. 29-56). Lawrence Erlbaum Associates.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347-364. https://doi.org/10.1007/BF00138871

Freedle, R. (2003). Correcting the SAT's Ethnic and Social-Class Bias: A Method for Reestimating SAT Scores. *Harvard Educational Review*, *73*(1), 1-43. https://doi.org/10.17763/haer.73.1.8465k88616hn4757

Hernan, M., & Robins, J. (2020). *Causal Inference: What If*. Chapman & Hall/CRC. Retrieved from https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Leinwand, S., Brahier, D., Huinker, D., Berry, R. I., Dillon, F., Larson, M., Leiva, M., Martin, W., & Smith, M. (2014). Principles to Actions: Ensuring Mathematical Success for All. In *Principles to Actions: Ensuring Mathetmatical Success for All*. Retrieved from http://www.nctm.org/principlestoactions

Neely, M. L., Troy, J. D., Gschwind, G., Pomann, G.-M., Grambow, S. C., & Samsa, G. P. (2022). Preorientation Curriculum: An Approach for Preparing Students with Heterogenous Backgrounds for Training in a Master of Biostatistics Program. *Journal of Curriculum and Teaching*, In Press.

Pawitan, Y. (2013). *In All Likelihood*. Clarendon Press.

Samsa, G. (2021). Evolution of a Qualifying Examination from a Timed Closed-Book Format to an Open-Book Collaborative Take-Home Format: A Case Study and Commentary. *Journal of Curriculum and Teaching*, *10*(1), 47. https://doi.org/10.5430/jct.v10n1p47

Samsa, G. P. (2020). Using Coding Interviews as an Organizational and Evaluative Framework for a Graduate Course in Programming. *Journal of Curriculum and Teaching*, *9*(3), 107-140.

Snapper, E. (1979). The Three Crises in Mathematics: Logicism, Intuitionism and Formalism. *Mathematics Magazine*, *52*(4), 207-216. https://doi.org/10.2307/2689412

Troy, J. D., Neely, M. L., Grambow, S. C., & Samsa, G. P. (2021). The Biomedical Research Pyramid: A Model for the Practice of Biostatistics. *Journal of Curriculum and Teaching*, *10*(1), 10-17. https://doi.org/10.5430/jct.v10n1p10

What Works Clearninghouse. (n.d.). *Standards Handbook, Version 4.1*. Retrieved from

https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

Wilkerson-Jerde, M. H., & Wilensky, U. J. (2011). How do mathematicians learn math?: Resources and acts for constructing and understanding mathematics. *Educational Studies in Mathematics*, *78*(1), 21-43. https://doi.org/10.1007/s10649-011-9306-5

Wilkerson, M. (2008). How do mathematicians learn mathematics? In *Proceedings of the Joint Meeting of PME 32 and PME-NA XXX* (Vol. 4).

Appendix

1. Joint and Conditional Probability, Discrete Case

Consider the following table, which cross-classifies 300 patients according to the results of a diagnostic test (X) and the presence of a disease (Y). The observed data are highlighted in italic text, other quantities are derived by summation.

	Y=0: disease absent	Y=1: disease present	Marginal distribution of test results
X=0: test is negative	*90*	*40*	130
X=1: test is positive	*10*	*160*	170
Marginal distribution of disease	100	200	Total=300

For example, if a patient is randomly selected from this population the probability that the disease is present is 200/300 = 0.67.

"Joint" events consider both X and Y, whereas "marginal" events consider one but not the other.

The probabilities for the 4 joint events are:

Pr{X=0 & Y=0} = 90/300.

Pr{X=0 & Y=1} = 40/300.

Pr{X=1 & Y=0} = 10/300.

Pr{X=1 & Y=1} = 160/300.

These 4 joint events are mutually exclusive and exhaustive, and thus constitute a sample space (i.e., the set of all possible outcomes).

The probabilities for the 4 marginal events are:

Pr{X=0} = 130/300.

Pr{X=1} = 170/300.

Pr{Y=0} = 100/300.

Pr{Y=1} = 200/300.

The probabilities for the marginal events are most simply obtained by using the information in the margins of the table (i.e., the numbers without highlighting). However, they can also be constructed using the joint probabilities. For example:

Pr{X=1} = Pr{X=1 & Y=0} + Pr{X=1 and Y=1} = 10/300 + 160/300 = 170/300, as before.

This illustrates the law of total probability, which holds when the events Y=0 and Y=1 are disjoint.

Conditional probabilities such as Pr{Y=1|X=1} can be calculated in two ways. One approach is to treat those individuals with X=1 as a new population of 170 individuals and notice that 160 of them have Y=1, implying that the desired probability is 160/170. Another approach uses the joint probabilities and the formula Pr{Y=1|X=1} = Pr{X=1 and Y=1} / Pr{X=1}, which in turn equals 160/300 / 170/300, which is identical to 160/170 after cancellation. The second approach illustrates the law of conditional probability.

We could apply the law of total probability to the denominator and obtain:

Pr{Y=1|X=1} = Pr{X=1 & Y=1} / Pr{X=1} = Pr{X=1 & Y=1} / (Pr{Y=0 & X=1} + Pr{Y=1 & X=1}),

which equals 160/300 / (10/300 + 160/300), with the same answer as before.

Finally, we can rearrange Pr{Y=1|X=1} = Pr{X=1 & Y=1}/Pr{X=1}, and plug this into the formula:

Pr{X=1|Y=1} = Pr{Y=1}*Pr{X=1|Y=1} / (Pr{Y=0}*Pr{X=1|Y=0} + Pr{Y=1}*Pr{X=1|Y=1}, which equals 200/300*160/200 / (100/300 *10/100 + 200/300*160/200), which comes out to the same thing after cancellation. This final version of the formula is termed Bayes theorem.

Exercise: Create a Venn diagram or other graphic illustrating the law of total probability. Do the same for the law of conditional probability. Do the same for Bayes theorem.

How this is used in statistics

Having multiple versions of the same formula might seem to be more trouble than its worth, but the advantage is that you might have one set of probabilities but want the other. For example, a patient has a positive test and wants to know the probability that they have the disease. The information we actually have pertains to the disease's prevalence (i.e., Pr{Y=1}) and the operating characteristics of the test: namely, Pr{X=1|Y=1} and Pr{X=1|Y=0}. These operating characteristics might have been derived from a separate experiment whereby, for example, a sample of patient with the disease are given the test and a sample of patients without the disease are given the test.

Bayes' theorem allows us to derive what we want from what we have, and we could tell the patient with a positive test that their chance of having disease is 160/170. Indeed, one way of quantifying the impact of a positive test is that it changed the physician's estimate of the probability of disease from the baseline value of 200/300 = 0.67 to 160/170 = 0.94. Similarly, a negative test would change the estimate of probability of disease from 0.67 to 40/130 = 0.31.

Bayes theorem also allows to make the same calculation for a hypothetical population with a different prevalence of disease, in similar fashion.

Exercise: assume that Pr{X=1|Y=0} and Pr{X=1|Y=1} are as before, but that the disease prevalence is changed to 5%. Calculate Pr{X=1|Y=1} and Pr{X=1|Y=0}. You should find that Pr{X=1|Y=1} is smaller than before. In plain English, explain why.

When calculating probabilities, it can be useful to structure the calculation as a chain of conditional events, each of them independent. For example, to draw 5 diamonds in a row from a deck of 52 cards, the first card must be a diamond, with probability 13/52 (i.e., 13 diamonds / 52 cards), then the second card must be a diamond, with probability 12/51 (i.e., the deck now contains 51 cards, 12 of which are diamonds, the particular diamond which was initially selected doesn't matter), then the third card must be a diamond, with probability 11/50, then the fourth card must be a diamond, with probability 10/49, and then the fifth card must be a diamond, with probability 10/48. These events are conditionally independent, the probabilities multiply, and so the answer is (13/52)*(12/51)*(11/50)*(10/49)*(9/48). This approach is much simpler than enumerating all the possible hands and then counting the number of hands with 5 diamonds.

Exercise: In the above example, we sampled without replacement -- in other words, once a card was selected it wasn't returned back to the deck. What would be the probability of selecting 5 consecutive diamonds if you sampled with replacement instead? Which probability is larger -- sampling with or without replacement?

Now suppose that you are sampling without replacement from a large deck of 52,000 cards, 13,000 of which are diamonds. What is the probability of selecting 5 consecutive diamonds? You should find that it is very similar to the probability you obtained using sampling with replacement. Sampling with replacement leads to the binomial distribution, sampling without replacement leads to the hypergeometric distribution, and as the size of the population increases the hypergeometric distribution approaches the binomial distribution. Results such as this are the mathematical justification for ignoring sampling with replacement, which simplifies matters, and should also serve

as a warning of the dangers of making this simplification when the sample sizes are sufficiently small.

So far, we've illustrated joint and conditional probabilities -- that is, probabilities associated with conditional events. This information will be used in, among others, the causal inference course.

In the module on the Bernoulli distribution, we will discover that when the possible values of a random variable X are 0 and 1, then Pr{X=1} = E(X) and so (1) in this special case, the above probabilities can be replaced with expected values; and (2) the resulting formulae hold for expected values in general. For continuous random variables, summation will be replaced by integration. JBS should be familiar with the underlying logic, and able to apply the resulting formulae in straightforward cases. (Note: we currently require mastery -- exposure is sufficient.)

The idea of conditioning occurs throughout statistics.

One application of this idea is stratification -- for example, in a randomized trial we might want to analyze the data separately for males and females, which is rather like the above.

Another application involves using conditioning to create random variables with a particular distribution. For example, if you have a Poisson process and observe 1 event between time 0 and time T, the waiting time until the event is exponential. These examples are covered elsewhere.

As an illustration of the basic idea, suppose that you toss a fair coin 4 times, and obtain exactly 1 head. Label the results of the tosses X1-X4 (all having values of 0 or 1), and label their sum X.

What are the 4 possible ways (i.e., joint events) to obtain X=1? (Hint: one way is {X1=0 & X2=1 & X3=0 & X4=0}?

What probabilities are associated with the 4 joint events?

Conditional on X=1, what is the probability that X1=1? What distribution does this represent?

Conditional on X=1, what is the probability that X2=1? What distribution does this represent?

Suppose that the probability of a head is 0.80. Does Pr{X1=1 | X=1} change?

What general principle about conditional distributions does this example illustrate?

2. Likelihood Functions, an Introduction

Consider 7 independent Bernoulli trials, each with success probability θ=0.6.

Exercise: use R to generate the pdf of X. In other words, fill in the table below. For example, the entry for X=0 should be 0.4^7

You should find the following:

X	Pr{X=x|θ=0.6}}
0	.00164
1	.01720
2	.07741
3	.19354
4	.29030
5	.26127
6	.13064
7	.02799

Notice that the experiment has not yet been performed, and so we do not know the value of X which will actually be observed. X is random, θ is fixed, as illustrated by the following notation:

$$\Pr\{X|\theta)\} = {}_nC_x\, \theta^X\, (1-\theta)^{(n-x)}$$

Now suppose that we observe X=4, and ask whether the data are consistent with the value θ=0.6. The value of X=4

will be observed almost 30% of the time, which provides informal support to the notion that θ might in fact be 0.6.

Exercise: use R to fill in the above table with θ=0.5. You should find Pr{X=4|θ=0.5} = .27344. The value of X=4 will be observed almost 30% of the time, which provides information support to the notion that θ might be 0.5, although there is marginally less support for θ=0.5 in comparison with θ=0.6.

Now, let's consider the same information from a different perspective -- in particular, we will treat X as fixed and vary the value of θ.

Exercise: use R to fill in the above table with Pr{X=4|θ}, for values of θ from 0 to 1 by 0.01. X=4 is impossible for θ=0 and also for θ=1 -- in plain English, explain why.

θ	Pr{X=4|θ}
0	
.01	
.02	
...	
.99	
1	

You should find that the maximum probability occurs where θ=0.57. We are using the same formula as before, but with 2 differences. First, the value of X is now fixed to be what was observed in the data. Second, the formula is interpreted to be a function of θ, and describes how Pr{X=4} varies as a function of θ.

Exercise: plot the above function (with θ on the x-axis and Pr{X=4|θ} on the y-axis).

The function that you have just created and plotted can be used to make inferences about the true (but unknown) value of θ. Indeed, we've already done so informally: the data seem consistent with θ=0.60, the data seem consistent with θ=.50, albeit slightly less so, and the value of θ which is the most consistent with the data is 0.57. "Most consistent with the data" can also be rendered as "the most likely value of θ given the data", and so the function in question is termed the "likelihood function". We denote it by L(θ|X} to emphasis that this is a function of θ, conditional on a fixed value of X. In our example, we might even use the notation L(θ|X=4}.

The value of θ that maximizes the likelihood function is called the maximum likelihood estimator (MLE) of θ, denoted by $\hat{\theta}$.

Exercise: plot the likelihood function L(θ|X) for 3 different scenarios: 4 successes out of 7 Bernoulli trials, 40 successes out of 70 Bernoulli trials, and 400 successes out of 700 Bernoulli trials. The value of the MLE should be identical.

How this is used in statistics

Much of statistical inference is based on the notion of generating parameter estimates using maximum likelihood. What follows is not intended to be a comprehensive treatment, but instead as an introduction to the topic to help you get your bearings.

Consider the likelihood function associated with 400 successes out of 700 Bernoulli trials. The slope of the function at the value of the MLE is 0 (this follows from a basic principle of calculus, covered in the next module). This function has a much steeper peak than, for example, the likelihood function for 4 successes out of 7 Bernoulli trials.

Exercise: Fill in the following table.

N	X	L(X|θ=0.57)	L(X|θ=.50)
7	4		
70	40		
700	400		

The third column is the value of the likelihood function at the MLE. The fourth column is the value of the likelihood function at another value of θ. We ask how much less likely θ=0.50 is than θ=0.57. To put the comparisons on the same scale, we divide L(X|θ=0.50) by L(X|θ=0.57). Consistent with the plots of the various functions, you should notice that within the range of θ=(0.50-0.57) the likelihood function for X=4 (n=7) is relatively flat, whereas the likelihood function for X=400 (n=700) is relatively steep. As an implication, verify that the ratio of these two likelihoods is near 1 for n=7 and far from 1 when n=700.

The next module will (among others) demonstrate the following:

- The "signal" (i.e., the most plausible value of θ) is the MLE.
- The "noise" in the estimate of θ (i.e., the imprecision in the signal) is related to the curvature of the likelihood function at the MLE.
- The "noise" depends on the inverse of the likelihood function's second derivative.
- The larger the absolute value of the second derivative, the steeper the likelihood function, and the more "information" we have about the actual value of θ.
- One way to quantify how much more likely the MLE of θ=0.57 is than θ=0.50 is to take the ratio of the two likelihoods. This idea underpins both hypothesis tests and confidence intervals.
- A confidence interval can be defined using the rule "take all values of θ for which the value of the likelihood function is close to the value of the likelihood function for the MLE".
- A confidence interval can also be created from the MLE plus or minus a multiple of the magnitude of noise.