Applied Probability at the Sophomore Level:

An Opportunity and a Challenge

Carlton, Matthew A.
California Polytechnic State University, Statistics Department
1 Grand Avenue
San Luis Obispo, CA 93407, USA
E-mail: mcarlton@calpoly.edu

Five years ago, we replaced our traditional introductory sequence for statistics majors with courses that emphasize statistical reasoning and de-emphasize both probability and formal mathematics. At the same time, our students floundered in their senior-level theory courses. To address both issues, we created a sophomore-level course in applied probability. This presented some special challenges, such as selecting a textbook and appropriate software, as well as "pitching" the material at the right level. But this also gave us a chance to introduce several applied probability methods at an accessible level for younger students, including some topics that appear nowhere else in our curriculum.

Background.

California Polytechnic State University ("Cal Poly") is a primarily undergraduate institution located in San Luis Obispo, California. We are part of the 23-campus California State University system, though our admission standards are comparable to many of the University of California schools. The Statistics Department gained its independence from the Mathematics Department about 25 years ago, and in that time the faculty have carved out a niche as national leaders in statistics education. (Four of the fifteen full-time faculty are fellows of the American Statistical Association, all for contributions to statistics education and/or undergraduate textbook writing.) A typical one-year statistics major cohort now consists of 12-20 students, though many statistics major courses are also taken by mathematics and economics majors.

Until 2005, statistics majors took a traditional two-term introductory sequence in statistics and probability, taught from Devore's *Probability and Statistics for Engineering and the Sciences*. This course followed the path many of us experienced in our first statistics courses: descriptive stats, probability and random variables, sampling distributions, and finally inference techniques. Students did not see much of probability *per se* again until their senior year, when they took a one-year sequence in probability and mathematical statistics, taught from DeGroot & Schervish's *Probability and Statistics*. Except for the very strongest students, statistics majors struggled in that sequence, due in part to the confluence of high mathematical rigor and lack of familiarity with (or, perhaps more precisely, lack of memory of) the core content.

Beginning in 2005, the two-term intro sequence was completely revamped, based upon statistics education research by Joan Garfield, Bob delMas, George Cobb, Roxy Peck, Beth Chance, Allan Rossman, and many others (Peck, Chance, and Rossman are all faculty at Cal Poly). The modern introductory sequence strips away probability and a topic unto itself and integrates descriptive and inferential techniques. This so-

called "randomization-based curriculum" still relies implicitly on probability, of course, to drive the ideas of random sampling/assignment, sampling variability, and *P*-values. But set-theory-style probability, along with formal probability rules and random variables, are not a core part of the randomization-based curriculum.

While this change seemed to improve students' understanding of basic statistical principles, our department was faced with a challenge. Statistics majors now saw less probability than they had previously (which many, though not all, of us argue is intrinsically bad); students' first exposure to serious probability occurred in the senior year, meaning junior-level courses could not leverage notions of random variables, expectation, and variance; and students performed even worse than before in the senior-level theory sequence. To address these issues, I developed a new course in applied probability, targeted for students at the end of their sophomore year. This course is what I'll describe in detail in what follows.

Course goals, content, and prerequisites.

"STAT 325: Introduction to Probability Models" was taught for the first time at the end of the 2007-08 academic year. STAT 325 recovers the topics lost from the "traditional" introductory statistics course, and supplements that material with several applied probability topics. Briefly, the 10-week (40-hour) course covers the following:

- basic notions of event probability (union/intersection, Venn diagrams, addition rule)
- simulation*
- conditional probability and independence (tree diagrams, Bayes' rule)
- counting rules
- discrete random variables (pmf and cdf, expectation and variance)
- discrete models (e.g., binomial, geometric, Poisson)
- continuous random variables (pdf and cdf, expectation and variance)
- continuous models (e.g., normal, exponential)
- jointly distributed discrete random variables (joint pmf, covariance and correlation)
- rules of expectation and variance
- central limit theorem
- Poisson processes*
- reliability theory (survival and hazard functions)*
- Markov chains*

The starred topics (*) were not part of the probability content of our previous "traditional" introductory statistics course. Simulation is integrated throughout the course.

To take STAT 325, students must have completed (a) one year of calculus, (b) a course in linear algebra (for Markov chains), and (c) a course in computer programming. We do not require students to have multivariate calculus under their belts, since this would delay the course for many students until their junior year; hence, we limit discussion of joint distributions to the discrete case. (Statistics majors see joint continuous distributions in the first quarter of their year-long, senior-level sequence.) In the first year of STAT 325, computer programming was not a prerequisite; this caused serious problems, which are discussed later.

When I developed this course, naturally I leveraged a lot of the material already available from the now-defunct traditional intro course. However, even that "old" material was re-tuned to address several goals for the course, set out by the statistics department:

1. Students should be able to use ideas from set theory, combinatorics, and conditional probability to solve problems.
2. Students should be able to calculate and understand expectation, variance, and standard deviation of random variables.
3. Students should be able to apply discrete and continuous models where appropriate, and assess the reasonability of those models for a given situation.
4. Students should be able to simulate random phenomena, writing original code as necessary, to estimate the probabilities of events and the distributions of random variables.
5. Students should be able to use a variety of applied probability models, including Markov chains, Poisson processes, and reliability models, to solve problems.

Pedagogical methods and examples.

We now discuss our approach to each one of these goals.

Goal 1: Use ideas from set theory, combinatorics, and conditional probability to solve problems.

Statistics majors will see a rigorous, theorem-proof version of this material in their senior year. For STAT 325, in contrast, we want to emphasize recognizing when to use each of these basic tools and translating "word problems" into a more mathematical structure.

Example: Suppose that 14% of all Cal Poly students have gotten a speeding ticket and 32% of all Cal Poly students have gotten a parking ticket.

(a) What is the sample space?
(b) Define two events appropriate to this example.
(c) What can we say about the probability a randomly selected student has earned at least one ticket?
(d) Suppose further that 5% of students have gotten both a speeding ticket and a parking ticket. Calculate the probability a randomly selected student has earned at least one ticket.

While this example looks very simple (and it is), it allows us to explore several ideas that trip students up the first time they study probability. First, model-building includes basic steps like properly defining the sample space and naming appropriate events. Second, just because 14% + 32% < 100% doesn't mean the two events are disjoint (thus, part (c) doesn't equal 46%). Third and most vitally, a model can be *underspecified*; here, the model isn't completely specified until part (d).

We look at three approaches to this problem: formulas, a Venn diagram, and a contingency table. With the last, students are able to more easily see why the original information is incomplete (we even discuss the "degrees of freedom" notion). With all three, students see that we'll always need three pieces of information to completely specify a model in this setting (two root events), but that the information can be rendered in multiple ways.

In particular, it's vital that students recognize that 14% does not represent the probability a student has a speeding ticket *and only a speeding ticket* (similarly for the meaning of 32%). By contrasting different tools, we reinforce the difference between $P(A)$ and $P(A \cap B^c)$, which students can distinguish in our mathematical language but often not as well in plain English.

Finally, students see that a contingency table takes a little more time to build, but makes a variety of follow-up questions trivial to answer (e.g., What's the probability a student has exactly one type of ticket?).

Example: A survey was made of television viewing habits for people in a metropolitan area. The survey specifically focused on which of three major networks — ABC, Fox, or CBS — people watch on Wednesday nights. The survey found that 42% of people in this area watch ABC, 57% watch Fox, 28% watch CBS, 15% watch ABC and Fox, 9% watch ABC and CBS, 19% watch Fox and CBS, and 4% watch all three networks.

(a) Use a Venn diagram and the information above to partition the sample space into 8 disjoint categories, along with their probabilities. [Hint: Work your way out from the center!]
(b) Report your findings in terms of these 8 disjoint categories. In other words, what do the 8 percentages in your diagram mean?
(c) What's the probability a randomly selected computer monitor has no defects?
(d) What's the probability a randomly selected computer monitor has at least one of these three types of defects?
(e) What's the probability a randomly selected computer monitor has exactly one of these three types of defects?
(f) What's the probability a randomly selected computer monitor has either a minor or major defect (or both), but not a critical defect?
(g) Come up with a general formula for $P(A \cup B \cup C)$ that relies on the seven probabilities given at the beginning of this example. Then use your formula to verify your answer to part (d).

Through this example, students again must distinguish the given information from the 8 pieces of a partition created by a Venn diagram. Even with the hint in part (a), some students assign the seven given probabilities to the seven disjoint pieces formed by three overlapping circles in a Venn diagram. As seen in part (g), the students ultimately derive the addition rule for three events by themselves. These lead into a deeper discussion of degrees of freedom in a model and the more general inclusion-exclusion principle.

Goal 2: Calculate and understand expectation, variance, and standard deviation of random variables.

Our statistics majors already know about mean and standard deviation from other courses. They will see these tools used repeatedly in their junior- and senior-level courses to examine the bias and variability of numerous estimators, so they need experience with those sorts of calculations. Of course, simple computational examples (finding the mean and variance of a given discrete or continuous distribution) are worth knowing, but they don't exemplify the skills we want our stats majors to carry with them into upper-division classes.

Example: (adapted from the 2008 Advanced Placement Statistics Exam) An experiment will be conducted to study the effect of temperature on an electronic device used in an undersea communications system to be installed at the Cal Poly pier. The experiment will be conducted at multiple temperatures. At

each temperature, devices will be submerged for 5000 hours, and the researchers will record how many of these devices are still working at the end of the experiment.

(a) A sample of 25 devices will be tested at 40°F. Write out the mean and standard deviation for $\hat{p}_{40}$, the sample proportion of devices that still work after 5000 hours, assuming the true proportion is p_{40}.

(b) A sample of 75 devices will be tested at 50°F. Write out the mean and standard deviation for $\hat{p}_{50}$, the sample proportion of devices that still work after 5000 hours, assuming the true proportion is p_{50}. (This should be easy if you got part (a).)

(c) The researchers believe there is a linear relationship between temperature and the probability of working 5000 hours. So, a logical estimate for the proportion of devices that work 5000 hours at 45°F is the average of the estimates at 40°F and 50°F:

$$\hat{p} = \frac{\hat{p}_{40} + \hat{p}_{50}}{2}$$

Find the expected value and the standard deviation of this estimator.

(d) Suppose that 17 of the 25 devices tested at 40°F work for 5000 hours and that 65 of the 75 devices tested at 50°F work for 5000 hours. Use this information and your answers to part (c) to come up with a point estimate for the true proportion of devices that work for 5000 hours at 45°F, as well as the estimated standard deviation of that proportion.

Not surprisingly, most students can find the expected value of the estimator in (c) and plug in the values from (d), but few correctly derive the standard deviation on the first attempt. This, coupled with a discussion of standard deviation formulas they already know from other statistics courses (e.g., for the difference of two sample means) reinforces the Pythagorean Theorem far better than examples of the flavor "sd(X) = 4, sd(Y) = 7, X and Y are independent, find sd(X+Y)."

Goal 3: Apply discrete and continuous models where appropriate, and assess the reasonability of those models for a given situation.

Despite the applied flavor of STAT 325, we actually derive the probability mass functions of the binomial, geometric, negative binomial, and hypergeometric distributions. The reason is obvious: the derivations of these formulas help students understand the reasoning behind each model, how the model assumptions are relevant, and the distinctions between similar models (especially binomial and negative binomial).

Continuous models (normal, exponential, etc.) are presented more simply, since there is no "natural" derivation of their pdfs. Instead, in the early going I emphasize correct calculation using continuous models, so students will be ready to tackle the central limit theorem and Poisson processes, which of course leverage these continuous models.

Example: Blocks on a computer disk are being scanned for errors. Each block has a 9.3% chance of being bad, independent of the status of all other blocks.

(a) On average, how many good blocks will be observed before the fifth bad block is found, and what is the corresponding standard deviation?

(b) What is the probability that exactly 4 of the first 25 scanned blocks will be bad?

(c) What is the probability that the fourth bad block will be found on the 25th scan?

(d) Write an expression for the probability that at most 8 scans are required to uncover the first bad block.

This example requires students to master the interplay between the binomial and negative binomial models, to distinguish "number of trials" from "number of failures" in the latter case, and to build expressions for "at least"/ "at most" probabilities as in part (d).

Example: (from our department assessment item bank) A civil engineer is interested in X = the number of overflow conditions that occur at a dam during a year. Suppose that the dam has an average of 12 overflow conditions during a year.

(a) Does it seem reasonable that the number of overflow conditions that occur at a dam during a year would follow a Poisson distribution? Explain.

Regardless of your answer from (a), suppose that X does follow a Poisson distribution.

(b) Give the probability mass function for X = the number of overflow conditions that occur at a dam during a year.

(c) Find the probability of at least two overflow conditions during a year.

(d) Find the probability of exactly three overflow conditions during a 4-month period.

Part (a) tests whether students understand that the Poisson model requires independent and identical behavior across non-overlapping time intervals, which is not realistic here. The other parts are simply Poisson calculations, though (d) requires pro-rating the average.

Goal 4: Students should be able to simulate random phenomena, writing original code as necessary, to estimate the probabilities of complicated events and the distributions of complicated random variables.

Thanks to our department's revamped curriculum, our students encounter the notion of simulation repeatedly before they take STAT 325, primarily as a way to estimate a P-value for a hypothesis test. Most simulations involve online Java applets or pre-fabricated code provided by the instructor (typically in Minitab). Now students must combine that background, their prerequisite basic programming course, and notions of probability.

We typically spend 1-2 days acclimating students to software (Matlab or R): how to open the program, how to write and save scripts and functions, syntactic peculiarities of the software, pseudo-random numbers, and several examples of basic simulations (using both for loops and vectorized operations).

Example: Write a program that simulates a 162-game baseball season and determines the team's longest winning streak during that season. Your program should have two inputs: p = the probability your team wins any particular game, and N = the number of iterations for your program (i.e., the number of 162-game seasons to be simulated). Your program should have one output: an N-by-1 vector whose ith entry is the longest winning streak during the ith simulated 162-game season. Use your program to answer the following questions.

(a) Assume p = .5 (you're rooting for a .500 team). Create a histogram of the variable "longest winning streak" based on your N-by-1 vector. N should be at least 10,000.

(b) Give a 95% confidence interval for the true expected longest winning streak in a 162-game season under the assumption p = .5.

(c) Give a 95% confidence interval for the probability that your team has a winning streak of at least 7 games during the 162-game season under the assumption $p = .5$.

(d) Now assume $p = .6$ (your team is really good). Repeat (a)-(c).

(e) Still assuming your team has a 60% success rate, repeat (a)-(c) but for the variable "longest losing streak." *Hint:* You shouldn't have to rewrite your program; just be clever about your inputs.

At this point in the course, students already understand the use of a binary vector to represent wins and losses; now, they must think about how to determine the longest streak of 1s within a vector.

Example: (from Baron, *Probability and Statistics for Computer Scientists*, pp. 128-129) A supercomputer is shared by 250 independent subscribers. Each day, the probability any particular subscriber uses the computer is 0.3. The number of tasks sent by a subscriber to the computer can be modeled by the Geometric distribution with a mean of 6.5. The amount of time that it takes the supercomputer to complete one task can be modeled by an Exponential distribution with mean 3 minutes.

Your assignment is to simulate one day's worth of tasks (many, many times), in order to answer the questions below. Notice that the model has three random variables, and that they are layered: the number of tasks will depend on the number of users that day, and the time to complete all tasks will depend on the number of tasks sent.

(a) Estimate the mean number of tasks submitted per day. Give a 95% confidence interval for the true mean.

(b) Estimate the probability that all tasks get completed within 24 hours. Give a 95% confidence interval for the true probability.

(c) Estimate the mean completion time for one day's worth of tasks. Give a 95% confidence interval for the true mean.

(d) Estimate the standard deviation of the completion time for one day's worth of tasks. Give a 95% confidence interval for the true standard deviation.

This exercise requires students to integrate several skills. They must recognize a probability model (here, the binomial distribution for number of users), use software's built-in probability distributions as part of a simulation, and write code to manage two different levels of random variables (here, the number of tasks and the times for those random number of tasks). Notice that, unlike in the previous example, I give them no guidance about the nature of the inputs or outputs.

Goal 5: Students should be able to use a variety of applied probability models, including Markov chains, Poisson processes, and reliability models, to solve problems.

Only STAT 325 covers these topics at Cal Poly; students will not see them anywhere else in the curriculum, except for some discussion of reliability theory in our survival analysis course (where the emphasis, quite rightly, is on model and parameter estimation from data). Once again, the emphasis is on application rather than theory. For example, we discuss and exploit the Poisson-Gamma relationship (an excellent review of discrete versus continuous variables), but we don't derive it. Similarly, I motivate several properties of Markov chains by example and/or simulation (e.g., the Chapman-Kolmogorov equations, steady-state probabilities), but I don't formally prove these.

Example: Consider three identical electrical devices, whose lifetimes can be modeled by an Exponential distribution with mean 5 (units = hundreds of hours). For each of the following structures, find the reliability function, mean time to failure, probability density function, and hazard rate.

(a) a series design

(b) a parallel design

(c) two devices in series, with that series connected in parallel to the third device

In class, students must solve these problems mathematically and provide a computer simulation. This requires them to think about each question in two ways: first, how reliability and design interplay with rules of intersection and union; second, how system lifetime relates to the minimum/maximum of the device lifetimes. (For example, the reliability of the series system can be expressed as $P(T_1 > t \cap T_2 > t \cap T_3 > t)$, but the lifetime itself — needed for simulation — is equal to $\min(T_1, T_2, T_3)$.)

The final example, below, spreads across several lessons and ties together many of the key ideas relating to Markov chains. We only discuss finite-state discrete-time homogeneous chains, with an emphasis on matrix representation of transition probabilities. Students learn how to manipulate the transition matrix to calculate various items requested in the example, but again we don't formally derive/prove anything.

Example: A taxi driver conducts business in three different towns. Whenever he is in Town 1, the probability his next passenger is going to Town 1 is 0.3, to Town 2 is 0.2, and to Town 3 is 0.5. Whenever he is in Town 2, the probability his next passenger is going to Town 1 is 0.1, to Town 2 is 0.8, and to Town 3 is 0.1. Finally, whenever he is in Town 3, the probability his next passenger is going to Town 1 is 0.4, to Town 2 is 0.4, and to Town 3 is 0.2.

(a) What are the "states" in this example, and how are they labeled?

(b) How should we define $X(n)$ in this chain?

(c) Write out all of the transition probabilities, and draw the corresponding *state diagram*.

The driver resides in Town 3, so $X(0) = 3$.

(d) What is the probability his second fare takes him to Town 1?

(e) What is the probability his second fare takes him to Town 2?

(f) What is the probability his second fare takes him to Town 3?

(g) If the taxi driver is currently in Town 2, what is the probability he will be in Town 1 two fares later?

(h) If the taxi driver has 10 fares today, what is the probability he ends up back "at home" (i.e. Town 3)?

Suppose he sleeps in his taxi, so each new day starts in a "random" town. Specifically, let's suppose for now that he has a 20% chance of waking up in Town 1, a 50% chance of waking up in Town 2, and a 30% chance of waking up in Town 3.

(i) What is the probability his first fare wants to go somewhere in Town 3?

(j) What is the probability his first fare wants to go somewhere in Town 2?

(k) Write down the pmf for $X(1)$, the destination of his first fare.

(l) Find the steady-state probabilities for this Markov chain using matrix algebra.

(m) Interpret the steady-state probabilities.

(n) The driver lives in Town 3. On average, how many fares does he handle before his next return to Town 3?

Other considerations: textbook and software.

When I first developed STAT 325, I established three criteria for a suitable textbook: (1) topic coverage, including the applied models; (2) explicit and repeated use of simulation; (3) the correct level of mathematical rigor/sophistication for my sophomore audience. Textbooks for the "traditional" probability and statistics sequence fail on criteria (1) and (2), while most probability texts are too advanced to meet criterion (3).

As an example, I am very fond of the books by Sheldon Ross, and his text *Introduction to Probability Models* certainly covers all the topics we need. However, the pace of the early chapters is too fast for students seeing semi-formal probability and random variables for the first time, and the discussions of Poisson processes and Markov chains are too advanced for younger undergraduates. (Ross' exposition is meant to be thorough, e.g. considering both finite- and infinite-state Markov chains, but sophomore students need a gentler introduction even at the sacrifice of completeness.)

I eventually settled on *Concepts in Probability and Stochastic Modeling* by Higgins and Keller-McNulty. The text's exposition is easy-to-read for undergraduates and rich in its usage of simulation. Many sections have a nice mixture of elementary and advanced exercises, including simulations; those few sections which are a little lacking can easily be supplemented. The treatment of Poisson processes is rather light, but otherwise the text fully covers the topics for STAT 325. My only complaint is the book's availability: written in 1992, the book is out of print, and Duxbury/Cengage has not been amenable to our creating photocopies at a reduced price. As a result, students and the campus bookstore must hunt down copies, which often prove to be very expensive.

Equally important to the smooth operation of the course is software selection. Two of us (Allan Rossman and myself) have taught STAT 325, I with Matlab and Allan with R. Both provide a natural programming language for simulation and expose students to software they're likely to see in industry (though I would argue most students, especially mathematics majors, are more likely to encounter Matlab than R). At Cal Poly, every campus computer lab has Matlab installed, and students can download the freeware version, Octave, to their home computers. Obviously, R is free to all students. Students with a previous computer programming course can pick up both languages quickly, though matrix manipulations (critical for Markov chains) are much more natural in Matlab than in R. In any case, students need to experience a proper computer language in a code-driven, command line structure; menu-driven software such as Minitab, while sufficient in other courses, does not meet the course goals for applied probability.

As mentioned previously, the inaugural STAT 325 class did not require computer programming as a prerequisite; this was a dire mistake. Half of the class had little to no experience with if/else statements, for and while loops, or even basic notions of writing, saving, and executing code. My two-day Matlab tutorial sufficed for the experience programmers in the group, but the inexperienced half lagged behind all quarter long. At the end of the course, evaluations were uniformly positive from those with prior programming experience and uniformly negative from the rest.

Conclusions.

An early course in applied probability models serves several purposes in the undergraduate statistics curriculum. The core material fills the hole left by the removal of probability from traditional introductory

sequences; instructors in upper-division courses can leverage knowledge of expectation and variance to discuss properties of estimators; students apply their freshman calculus and linear algebra knowledge earlier; students experience computer programming in a probabilistic setting (simulation) rather than just statistical settings (data management and analysis); and students are better prepared for senior-level theory courses by having earlier exposure to key concepts and formulas. In our department, we have seen a marked improvement in statistics majors' performance in the senior-level theory course; we have not yet evaluated whether our students handle theoretical arguments in other upper-level classes better.

Before developing an applied probability course, it's critical to think about your audience, what level of mathematical maturity they have (not just what courses they've taken), and what outcomes you'd like to see in later courses. Textbook and software selection are of course critical to the success of any class; for a sophomore-level probability course, probability textbook options are limited, though one could forego some applied topics and use any of several "math/stat" texts (e.g., Larsen & Marx). Command-line languages give students a more genuine modeling experience than menu-driven software, with Matlab and R being the best fit for an applied probability course.