Strategies for Teaching an Enduring Knowledge of Statistics

Larry Weldon
Simon Fraser University

Introduction

The futility of teaching students to master calculations without conceptual understanding is widely acknowledged. Unfortunately, there are incentives for both instructors and students to focus on calculations, and the move to teaching concepts is difficult. In this paper strategies are suggested for encouraging students to gain a conceptual understanding of statistics: open book tests and exams, elimination of the traditional textbook, emphasis on verbalization and visualization, use of open-ended case studies, and emphasis on process rather than product. It is argued that these strategies are likely to result in a more enduring knowledge of statistics than more traditional approaches. Some evidence from the author's experience is provided.

My main message is that the undergraduate statistics curriculum is a choice made by statistics faculty members, and I suspect many statistics faculty have not really thought seriously about what they would choose to teach if they could break away from tradition. There is an attitude in the statistics education community that any proposals for change need to be backed up by data-based research. But I think this is a recipe for inadequate adaptation to what is needed to provide students with a useful introduction to statistics. The forces of inertia in the academic world will keep things reflecting the statistics faculty choices of 1960. In education, data-based research is only useful on the edges, where change is not too threatening to the status quo, and my contention is that we have big problems with the core.

First Course

Let us first consider what should be in a first course in statistics. The vast majority of students will only take one course in the discipline, and so we need to accomplish something in one course. If we only provide a basis for future courses, we will waste the time of the many students who will not take future courses. A first course should give students an idea of what the practice of statistics is about, and why it is important, interesting, and useful. Those few students who take a first course with the intention of taking many statistics courses must be quite rare, and while one could argue for a specialized program for these few, it is likely that they also need the overview exposure that would be included in a terminal first course. Without the overview of the field, how will the future statisticians make good choices of courses and projects during their university education? If all students need this overview introduction, it could be compulsory even for the future statisticians and scientists.

However, there could still be a difference in the choice of application material of interest to the major student groups: life sciences, natural sciences, social sciences, or humanities. But the techniques and concepts would be similar in each group. So what is proposed for a first year course is a "terminal" course for all students. The idea is that students requiring more expertise in statistics require more than an introduction to the discipline, but not necessarily a different introduction.

Statistics and Mathematical Statistics

But what is wrong with providing a specialist introduction in situations where the numbers requiring a specialist knowledge warrant it? The problem that arises is the tendency to teach mathematical statistics to this group. There are some likely dangers to this approach:

1. Students with little interest in data analysis but a facility in mathematics will view this course as entry to a job-related discipline, and may in the end find applied statistics uncomfortable.
2. The range of topics that can be covered in a mathematically rigorous way in a single course is rather limited: an introduction to statistics should not stop at estimation and hypothesis testing.

To elaborate on point 2. The intro course should introduce things like multiple regression, simulation, smoothing, design of experiments and sampling surveys, statistical software, graphical analysis and summary, and categorical data analysis at the very least. Moreover, should not the student be exposed to the most successful applications of statistics like clinical trials, epidemiology, quality assurance, weather analysis, and data mining? This wide range of topics can be introduced in a single course if the mathematical rigour is eliminated. There is still much in the way of logical rigour that does not require the machinery of mathematics, and can be introduced while portraying the broad list of topics suggested here.

The danger in a broad course is that it is too shallow to be useful. I would like to suggest ways in which a broad course can be made both useful and memorable. It requires a rethink about what aspects of the discipline are the most useful to all students. My argument is that there are many basics that are often omitted from the first course yet are more broadly useful than the more traditional material.

Higher Level Courses in Statistics

Before outlining suggestions for such a first course in statistics, I would suggest that the ideas can be extended to higher-level courses. Traditional curricula in statistics introduce increasing levels of mathematics for these higher-levels: more probability models, more estimation criteria, more frameworks for hypothesis testing, more variables in regression and multivariate analysis, more design of experiments, Bayesian and nonparametric approaches, stochastic processes, and time series analysis. Certainly mathematical sophistication is needed for the clear development of these topics. But mathematics alone is not enough. There is at least as much time needed to be spent of the non-mathematical aspects of these more advanced subjects. Here are some examples of these non-mathematical aspects:

- Probability models: What parameter values are appropriate for the most common applications? What is the connection between simulating a probability model and sampling from a population? When is an empirical probability model preferred to a parametric probability model?
- Estimation Criteria: Under what circumstances is unbiasedness desirable? When is least squares to be preferred to nonparametric smoothing? How should outliers be handled?
- Hypothesis Testing: Under what circumstances is resampling useful? When is it advisable to take account of prior information in the analysis? How is exploratory data analysis different from confirmatory data analysis? Is data mining different from data analysis?
- Regression & Multivariate: How are higher-order relationships graphed or described in words? What are the dangers of step-wise regression? What is the difference between univariate outliers and multivariate outliers?
- Design of Experiments: How is causality established? Why are most studies Observational Studies? What common circumstance does Simpson's paradox warn us against?
- Bayesian and Nonparametric Approaches: When are these approaches needed and/or advisable? How is decision-making difference from Investigation? When are parametric models needed for data summaries, instead of graphical summaries?
- Stochastic Processes: If the model is always "wrong", what use is a model for forecasting? How prevalent are chaotic systems?
- Time Series Analysis: What sort of sample paths are consistent with stationarity?

To a mathematician, these issues are "applications" issues. But actually, they are issues of statistical theory and practice, and they need to be taught in statistics courses. Of course, anyone wanting to be expert in statistics needs to learn all the associated mathematics as well: matrix theory, optimization methods, algorithm strategies and software, computer graphics, differential equations, complex analysis. The difficult question is how the mathematical and non-mathematical aspects of statistics can be integrated into an undergraduate program. The lack of textbooks that span the two approaches is one problem, but the lack of instructors willing and able to span the approaches is an even greater problem. And without textbooks and willing instructors, the curriculum planners cannot expect much success.

Separation of Mathematics and Statistics

One solution is the have the math department teach the maths, and the stat department teach the stats. This might seem a bit regressive, but actually it would allow the stat department to get away from thinking they are teaching a certain brand of mathematics. For a statistics instructor to teach multivariate analysis, it would be wonderful to assume a familiarity with eigen-analysis, optimization methods, projections in hyperspace, and pseudo-inverses. My own experience with teaching multivariate analysis was that I had to teach all the math first, and then there was not enough time to explain how multivariate analysis can be used to solve data analysis problems: strategies of data mining, classification problems, index construction, outlier identification, multivariate distributions, etc.

These problems at advanced levels stem from the basic intellectual bias at the introductory level: that the important part of statistics is mathematical statistics. Scholars who believe this tend to design two kinds of intro-stat courses: i) formula courses that amateurs can handle but which provide very little of practical use or intellectual stimulation, and ii) mathematical courses which are intellectually rigorous but cover only a tiny portion of statistical strategies, and almost none of the practical problems that a statistical practitioner faces. The realization that statistics is not mathematics can result in a huge change to all undergraduate statistics courses, not only the introductory one. Statistics considers data context crucial, whereas mathematics tries to eliminate context with its abstractions. For example, beginning lectures in statistics deal with types of measurement: ratio, interval, ordinal, categorical. Beginning mathematics usually does not deal with this distinction, since it is a context issue. At a more advanced level, statistics deals with fractional factorial experiments since they are necessary for practical reasons, whereas mathematics deals with fractional factorial designs because they have interesting group structure: the motivation is quite different, and the instruction has to be motivated appropriately.

Implications for the Intro-stat course

To return to the first course, I have argued that,

- i) all students need the same kind of first course
- ii) it should teach concepts and tools of statistics with as little mathematics as possible
- iii) to make the course material as useful as possible to a wide range of student careers, and for a long time.

With these three objectives in mind, I will outline some strategies I have used in a first course, and some anecdotal information about the student response to them.

Strategies for a useful first course in statistics

I will list these strategies and explain my rationale in each case in the following notes.

1. Avoidance of a technique-oriented textbook
2. Experiential Presentation of Techniques in areas of interest
3. Use of Computing by Instructor for Simulation and Graphics
4. No student computing, few formulas
5. Open Book & Notes for tests and exams
6. Use of graphics for explanations
7. Require verbalization of why, what, when
8. Sample tests and exams to impart objectives
9. Application material that is an important part of a general education
10. List of Concepts, Contexts and Techniques as an aid to exam preparation
11. Math as a simplification technique, to summarize technique structure

1. Avoidance of a technique-oriented textbook

Almost all statistics textbooks are technique-oriented. Chapter headings are things like "descriptive statistics", "inference with one variable", "regression and correlation", "study design", "hypothesis testing", etc. But techniques are viewed as abstractions to most students in an introductory course, and the general utility of the techniques will not likely be appreciated – they will seem like context-free abstractions that need to be absorbed in order to pass the course, but otherwise have little to offer their careers or even intellectual development. Conscientious instructors try to demonstrate each new technique with motivating examples and interesting applications, but the pedagogical goal with these examples is to make clear the technique itself. A consequence is that the actual applications used are viewed as ancillary and not to be taken too seriously, and so utility of the technique is inadvertently downplayed.

One alternative is to use an "experiential learning" approach: case studies with important results are introduced and the techniques required for those studies are described as needed. The focus in these modules is on the useful result, so multiple techniques will usually be used. The techniques that are used in many case studies will be explained many times so that the repetition so necessary for in-depth learning can be achieved. Towards the end of the course the traditional structure of the technique list can be clarified: one variable-two variables-multiple variables, description-inference, estimation-hypothesis testing, etc. For such a course, a "reader" is useful for providing background to the case studies, but almost any statistics textbook (perhaps from the used-book store) can supply definitions and formulas in a technique-ordered way. Clearly, the instructor's notes have to be clear to present the focus of each case study and to identify the techniques required.

The experiential approach has the potential to convey statistical strategies in contexts relevant to the student, and memorable as a result.

2. Experiential Presentation of Techniques in areas of interest

The choice of case-study contexts should reflect the interests of the student group addressed, when this is known. Unselected groups should be given case studies that have important lessons for everyday life: evaluation of new drugs or diets, randomness in investment returns and sport results, rationales of insurance and lotteries, monitoring of social networks, etc. For students within specific major categories, motivating case studies can be drawn from the general areas of life science, or social sciences, or natural sciences.

3. Use of Computing by Instructor for Simulation and Graphics

The importance of computing for statistics can be conveyed in first courses by the instructor's use of computing for demonstrations. Simulations of random samples can be portrayed as dotplots, and the associated probability models can be simultaneously portrayed graphically. Difficult concepts like the sampling distribution of the sample mean, the linearity of the correlation coefficient, the modeling assistance of the residual plot, and the power of a two-sample means test, can be experienced second hand by students with this device. Of course, the assignments and tests must reinforce the importance of understanding these computer-based strategies in order to engage students in owning the ideas.

It may not be advisable to engage students in the frustrations of statistical software in a first course. In subsequent courses, while the market will suggest certain packages such as SAS, it is a good idea to inform students of the free and world-wide support of a general purpose package like R. Of course, optional use of software by students can always be encouraged, if the tutorial system is prepared to support it.

4. No student computing – few formulas

There is no question that statistical practice requires a very good knowledge of statistical software. However, students with only one course will not be prepared for independent statistical practice, no matter what the course content. The student who spends the required time mastering Minitab, JMP, or R will not have time to learn enough to avoid to stay out of trouble using these software packages. My suggestion is that in a first course, do not attempt to have the students use software – most of the students will have more use for the big ideas of statistics. Although one of the big ideas is that simulation is a very useful tool, this can be conveyed through demonstrations rather than practice.

5. Open Book&Notes for tests and exams

Statistical software has matured to the point where very good software is now available at very low cost. The practitioner relies on software for calculation details, and the skill in data analysis depends more on when, why and how to use the software rather than the calculation detail the software achieves. Of course there is a crucial need for users to be able to judge the reasonable-ness of answers but even this does not require a knowledge of calculation details. One way to focus students' attention on the logic of various analysis methods is to convince them that memory-work has little value unless it is accompanied by conceptual understanding. And one way to do this is to allow students to take whatever books and notes they would like into tests and exams. Of course this also changes the nature of the examination: one cannot ask students for things that merely require a copy from notes to the test paper. Students should realize that this also implies that they try hard to understand the "why" and "when" of techniques and not spend so much effort on memorizing the procedural details. This kind of knowledge will be useful to a wide variety of students, and moreover, it is likely to stick. As Einstein said "Education is what remains after one has forgotten everything he learned in school."

6. Use of graphics for explanations

Many subjects have their graph paradigm: in economics it is the demand-supply curve, in computing it is the flow chart, and in statistics it is the scatter plot. So many phenomena of statistics are described in terms of one- and two-dimensional scatter plot (perhaps the dotplot in one dimension). These are important for making clear the difference between a population (or probability model) and sample data. Students learn that dotplots from a normal population do not look normal, regression lines from a bivariate normal scatter have a lesser slope than the major axis of the elliptical contour, and symmetric random walks do not create a saw-tooth pattern for the accumulation. These phenomena are useful life lessons!

7. Require verbalization of why, what, when

Students are used to explaining things in words – not so much with formulas or even with pictures. Until students can describe in words the concepts they are learning in the statistics courses, the concepts will be poorly understood and less useful as a result. Most students are more comfortable with words than with mathematical formulas. Even students who are comfortable with the formulas need to learn to describe the ideas to others, if they hope to make use of the ideas in the real world.

8. Sample tests and exams to impart objectives

It is easy to understand why students registering for their first statistics course would expect it to be formula-based. If it is concept based, students need a warning. A good way to illustrate the kind of conceptual learning that is required is to provide sample midterm tests and exams. These can be assignments with lots of time for the student to learn the concepts required for the answers. The necessity of writing out answers in English (or other language if the course is taught in it) will help the students gain the vocabulary they need, so the midterms and exams will not be a surprise.

9. Application material that is an important part of a general education

If a student learns something about natural science or social science in a statistics class, this is a good thing. We need not assure students that certain application material is outside the core material of the course, and will not be examined. If the application is useful for making a statistical concept clear, then the application itself can be examinable. This approach will help the student see how the core concepts of statistics do impact real life. Moreover, at a university, we should not worry too much that the student is getting an education that reaches outside of his or her chosen field.

10. List of Concepts, Contexts and Techniques as an aid to exam preparation

To make sure the student understands which jargon terms refer to big ideas or *concepts*, which refer to applications, or *contexts,* and which refer to calculation methods, or *techniques*, a list can be provided for the final review at the end of the course. This is also helpful in pulling together the overview structure of the course, which may seem somewhat fragmented from the case study approach. An example is given in Appendix 1.

11. Math as a simplification technique, to summarize technique structure.

In a case-study presentation, with techniques introduced as needed, the helpful structure of the techniques list will not necessarily be obvious to students. For example, the nominal-ordinal-interval-ratio ladder of measurement scales is one such structure, and the one-two-many variable display methods is another, and the description-estimation-testing spectrum is yet another. If these structures are presented after they have appeared in several case studies, the useful summary will help students to see the decision trees that an analysts uses to approach data analysis. This is the step that substitutes for the technique-oriented presentation of traditional courses.

The aim with all these strategies is to provide students with some ideas they will find relevant to their intellectual development and their future careers. These strategies were used in a recent course at Simon Fraser University taught by the author, STAT 100. I will conclude with some details about that course and the student feedback from it.

STAT 100 at SFU in 2010.

The course at Simon Fraser University had about 150 students, about 75% were registered in B.A. programs, and about the same percentage were ESL. It is likely that very few students from this course that would ultimately choose to major in statistics. Students are required to have a certain number of credits of "Quantitative" courses in order to graduate, not matter what their major, and this explains the large B.A. cohort. The course was advertized as "The course will involve little calculation but quite a bit of reading and writing, and thinking!" Some of the ESL students were sorry they did not read this warning, and found the dependence on English quite challenging – these students would have liked more calculations and less (English) verbalization. However, there is general agreement that mastering calculations is less useful in an introductory course than gaining an understanding of the strategies and concepts of statistics. Students with adequate English language skill did find the course useful. The average of 58% on the final exam was a little lower than usual but the exam (Appendix 2) was challenging. One surprising result was the low performance on question #1, even though this topic (sampling distribution of sample means) received more emphasis than any other topic! Average score on this question was 2/10. It may be true that some of the concepts that instructors think have been presented unambiguously and repetitively may be so difficult that only a few students will ever understand them.

The topics covered in the course are summarized in a single page in Appendix 1. The final exam involved almost all of these concepts, contexts and techniques. The exam is included as Appendix 2. The course notes appear at www.stat.sfu.ca/~weldon/stat100-10-1.html

Student Response

The students had varying responses to the ten strategies described above.

1. Avoid text: They were at first confused about the "reader" since it looked like a text but was not full of definitions and formulas. Eventually they realized it was a resource and that the posted notes were the "text" for the course. After the first few assignments and the midterm, they became convinced that the contexts were an examinable part of the course, even though it was not a textbook in the usual role.

2. Experiential presentation of techniques: The class had a large component of "arts"

students and more arts examples might have been a better choice than the ones with medical or scientific flavour. However, the examples in investment, advertising, email, and sport went over fairly well.

3. Use of computing by instructor: The illusions of randomness require students to observe the results of simulations. But some of these require advanced software or advanced programming skills. Several useful phenomena were demonstrated this way, using R programs. It is hard to gauge student acceptance of this method but the phenomena would have been omitted without it.

4. No student computing: only a few students were interested exploring statistical software, and the vast majority were happy to avoid the hassle of learning to use it. Of course it is important that students appreciate the difference between simulation and computation, and some small-scale hands-on experiments without the computer were used to make this distinction.

5. Open book tests: the first test was poorly done – average mark was just over 40%. The provision of a sample midterm as an assignment before the first midterm was not enough make the point about the relative lack-of-advantage to the open book for a test hat was designed with the open book feature in mind. Many students thought they would be able to look up the answer during the test! Of course, the test questions were designed to avoid the sort of question whose answers were directly in the notes, and rather depended on an understanding of the notes. The second midterm and the final exam (averaging just under 60%) were done better.

6. Graphics for explanations: although most students eschewed anything mathematical, they were able to understand the basic Cartesian graph in two dimensions. Thus some concepts were able to be conveyed using only math tools from secondary school.

7. Verbalization: As mentioned earlier, the ESL students had difficulty with the frequent requirement to verbalize concepts. A course in their mother tongue would have been better for them, at least for mastering the concepts. Students that were able to cope with English had a chance to learn some useful statistics strategies well enough to explain them to others.

8. Sample tests: In a course with a history, old tests and exams help to convey to students the kind of learning required in the course. A new course can use sample tests as assignments to achieve this. Learning objectives are another way, but the bottom line for the student is what kind of assessment will test those learning objectives. In this course, the sample tests were set as take-home assignments, and were well-done.

9. Context Knowledge: Students were at first surprised that a “math” course would require learning about the real world: stock markets, accident insurance, spam filters, sports leagues, … However, they eventually realized that to describe the utility of a statistical strategy, this contextual information was essential, and were able to answer the context-based test and exam questions.

10. List of Concepts, Contexts, and Techniques: Some details of the case studies are distracting from the central story. While this is a fact of life, and sorting out the pearls from the pebbles is a useful skill to practice, students found an explicit list of the important target items a useful guide, especially for studying for tests and exams.

11. Math as a simplification technique, to summarize technique structure: This strategy was not used. The poor showing on the square root law question (question 1 in Appendix 2) suggested that it might have been helpful.

Principles for Statistics Education

The strategies described above flow from certain principles that have evolved for me over the 40 years or so that I have been teaching statistics. Here is a brief list of those principles.

1. Authenticity of Experiential Learning
Experiential learning in statistics courses has the advantage of automatically providing the tools and concepts necessary for the statistics practitioner: the strategies chosen are those required for the case studies. These useful strategies are likely to be encountered in other courses and in future careers, and this should help with retention of the strategies.

2. Focus on Concepts
Statistics software provides the details of complex calculations, and drill in these may not be necessary to appreciate the underlying concepts. Moreover, the concepts are what alert students to the potential for using statistics strategies in their work.

3. Motivation from Assessment Procedures
Students are encouraged to pursue high marks in all their courses, and this provides incentive to study in a way that will be reflected in their assessment. If the exams test conceptual understanding, then there is the right motivation for students to learn this aspect of the material.

4. Important Role of Context
Students have to learn that the context needs to be taken seriously in applying statistics strategies. The "standard method" seldom works without adaptation to the local context.

5. Access skills more useful than instant recall
The communication revolution has made information easily accessible. Instruction in all subjects has shifted from knowledge transmission toward knowledge exposure or experience. Once the awareness of detail is accomplished in courses, the renewed access to it in practice is fairly direct.

6. Visual trumps algebraic for impact
Use of graphics for illustrating statistical phenomena is more effective than algebraic formulae, for a majority of students in a first course in statistics. The old tendency to

treat graphics as “descriptive statistics” needs to make way for a generation that treats visual instruments as a normal language.

Conclusion

Nothing has been proved! But the discussion has hopefully raised questions that will focus the discussion on how statistics instruction should evolve. The fact that hundreds of brilliant educators have been working for decades to improve statistics education, while the current status is so similar to the status of 50 years ago, suggests that incremental change has been too slow, and perhaps major changes are now indicated.

Appendices:

1. Content Summary of STAT 100
2. Final Exam of STAT 100 with answers.

Concepts Illusions of Randomness Models (are wrong) Probability Averages (Means) and Proportions Time Series Reliability Interaction Covariate Variability	**Concepts** Sample Means Distr'n Density Curve Distribution (Freq. Dist.) Independence (and Dep.) Observational Studies Experiments	**Concepts** Correlation Population Growth Trial and Error Models Random Sampling Coincidences Survival Uniform Spatial Scatter Clusters and Regularity Quality Control
Contexts Coin Toss Sports Leagues Stock Market Washing Machines Turkey Mail Cell Phone Fraud Zipf's "Law" Fuel Consumption Weather Forecasting	**Contexts** Insurance (Auto) Auto Repair Investment Portfolio (Diversification) Memory Load Expt Simpson's Paradox Randomized Response Surveys	**Contexts** Tiger Prey Africanized Bees SPAM filters Anthrax Outbreak Veterans Fund Raising Political Polls Olympics Medals School Choice Expt HIV Survey Lotteries & Birthdays Auto Accidents Ocotillo Distribution Travelling Salesman Six Sigma Strategy
Techniques Simulation Random Walk Random Assignment Histogram Graphical Display Hypothesis Test & p-Value Smoothing Moving Average Standard Deviation Normal Distribution Regression Residual Plot SD of Means	**Technique**s Square Root Law Central Limit Theorem Dotplot Symmetric Random Walk Forecasting Time Series Bar Chart Repetition Randomization Clinical Trial Double Blind Placebo	**Technique**s Logarithms Gamma Distribution Decision Errors Data Mining Sample Proportions Random Sampling Regression Line Sampling With Replacem't Sampling Without Repl't SD of proportions Cell Counts Poisson Distribution Shortest Path Variability Reduction

STAT 100 Final Examination April 22, 2010

Instructions: Answer all questions. Budget your time so that you have a chance to answer all the questions. You have **three hours** for the exam. The exam is **Open Book**: Texts, Notes, and Calculators are allowed. Be careful to answer the questions posed.

1. (10 marks) In a class of 144 students, each student tosses a fair coin 25 times and reports the proportion of heads among the 25 tosses. The 144 proportions varied as one would expect, and the professor makes a list of these 144 proportions. The students are asked to compute the SD of the proportions directly from the list (using the 144 numbers).
a) Estimate what you think that calculated SD would be. Justify your answer.
b) What interval of values would include about 137 of the students' proportions? Justify.

A1. a) The SD of the population is .5 (square root of .5 * .5), and the SD of the proportion (an average of 0s and 1s) is .5/sqrt(25) =0.5/5= .1
b) 137/144 = .95 so the interval would be about .5 ± 2*(.1) or from .3 to .7.

2. (9 marks) Clinical trials involve the strategies "randomization", "control group", and "placebo". What is the purpose of each of these strategies?

A2. **Randomization** of treatments to subjects produces comparison groups that tend to be balanced with respect to all variables except the treatment itself. The purpose of balanced groups is that the treatment difference can be assumed to be the cause of any response difference.
Control Group is the name for the comparison group that receives no new treatment, since it provides a benchmark with which to compare a new treatment.
Placebo is the name given to the treatment device (often a pill without biochemical effect) that will not have any direct biochemical effect on the patient, and it is used to assess the psychological effect to providing apparent treatment, so that this degree of influence will not be confused with the actual effect of the new treatment.

3. (8 marks) Two of the simulations we did in class were the risky company portfolio and the auto insurance company. The risky company simulation showed that even though one company has a good chance of losing money, the portfolio of those companies had a very small chance of losing money. The auto insurance simulation showed that the chance that the insurance company would lose money depended on the number of policies it held, and with enough policies, a loss was rare. Use the sampling theory of sample means to explain both of these effects, making clear the connection between the two applications.

A3. In both cases the long run average return was positive, and although the return itself was variable, and often negative, the average return was usually positive. The positivity of the average return depended on the average being based on a large enough number of independent values, since the square root law reduces the variability as the sample size increases, and so with a large enough sample size, all but the very bottom end of the return distribution is positive. In the case of the risky companies, the number of companies in the portfolio was the “sample size” used in the square root law., and the average return (from {-1,-.5,0,3}) of .375 was positive. In the insurance case, the average return was positive as long as the policy premium was greater than the average claim, which is usually the case, and the number of policies played the role of the “sample size”. In both cases, the square root law ensured that the average return would usually be positive.

4. (8 marks) The example of Simpson’s Paradox involving two treatments for kidney stones is displayed in the following table:

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

The percentages refer to the success rate of the treatments.
What is the “paradox”, and under what general circumstances can it occur? What is the correct conclusion about the relative success of the two treatments, from the above table?

A4. The paradox is that the combined data suggests Treatment B is better and yet for both subgroups Treatment A is better.
This confusing situation can occur in any observational study. If there is a “lurking” variable that is related to the response (in the above case it is stone size), then if the comparison does not allow for it, its effect will be hidden. The correct conclusion is that Treatment A is better.

5. (7 marks) In class, we studied the clustering phenomenon of plants that grow over a certain area. What strategy allowed us to determine if the clusters were more concentrated or less concentrated than would have occurred due to a uniform spatial distribution.

A5. We used a uniform spatial distribution on the unit square to simulate a sample of N plant locations. Then we used a grid of m^2 squares to count the number of non-empty squares. The proportion of empty squares in this case can be predicted as the 0 probability from a Poisson distribution with mean N/m^2. . If there is a greater proportion of empty squares than the Poisson probability predicts, then this would be evidence of real clustering (more concentrated clusters); if there is a lesser proportion, then this suggests more regularity than a uniform spatial scatter (less concentrated clusters).

6. (6 marks) The article on the Africanized bee invasion included an argument about why wild population sizes will stabilize in time. What is that argument?

A6. The crucial graph for this showed two curves: the birth rate as a function of population size, and the death rate as a function of population size. These curves intersected at the stable population size: if the population was greater than this stable value, the death rate was larger than the birth rate, and so the population would decrease. If the population were less than this stable value, the birth rate would exceed the death rate and the population would increase. In every case, a deviation from the stable population size will cause a move in the direction of the stable population.

7. (6 marks) The Six Sigma article includes methods for the reduction of variability in an industrial setting. How does reduction of variability contribute to increased profitability?

A7. It allows a company to produce more closely the advertised product (or amount of product), and thus saves on excess production required to keep the actual production above the advertised level. This reduces cost and increases profit.

8. (6 marks) a) Public lotteries often have a winner of a large jackpot, even though the chance of winning a large jackpot is an extremely rare event. Explain.
b) Why is the purchase of a large number of tickets in a public lottery is a poor idea from a purely financial perspective?

A8. a) There are a lot of tickets sold and so the chance of a jackpot winner is not rare.
b) The prizes are funded from the ticket sales, and roughly half of that income is removed from the prize pool. So the average return to each ticket is about one-half of its cost.

9. (6 marks) Explain the potential illusion of randomness that was described in connection with the league-points tables for sports leagues.

A9. The wide range of league points for each team (e.g. based on several games in which the league points were assigned as 3 for a win, 1 for a tie, 0 for a loss) suggests that the top teams are much better than the bottom teams. However, some variation in league points would occur even if every game was a 50-50 game, and this effect can be simulated to see how much variation is due to randomness rather than a team tendency to win (= team quality). We showed that the league points variability was not much more than was observed in some high profile leagues, suggesting that the league standings did not really reflect a difference in team quality.

10. (6 marks) In the assessment of the accident-free duration of students that was estimated based on information from the class, two elements of information were collected: the date of obtaining the first driver's license, and whether or not the student had been involved in an accident since having that driver's license. Explain how this information provided an estimate of the risk for students in the class.

A10. Each student's exposure time was calculated using the driver's license date and the current date. Then student exposures were lumped into one-year time intervals, and in each one-year group, the proportion of students who had experienced an accident was computed and plotted against exposure in months. This relationship was smoothed with a straight line, and the slope of the straight line gave the monthly increase in the probability of having had an accident, which is the risk estimate.

11. (6 marks) The sample correlation coefficient between two variables is the average product of the coordinates of the sample points once the coordinates are expressed in standard units. Explain why this method would generate a negative value for the correlation between the following two variables for Vancouver weather data:
i) number of millimeters of rain on a day in April
ii) number of hours of sunshine on a day in April

A11. Days high on i) would be low on ii) and low on i) would be high on ii). So in the graph of the two variables expressed in standard units, the upper right and lower left quadrants would be nearly empty, and the upper left and lower right quadrants would have most of the points. But this latter pair of quadrants both produce negative products, making the average product negative, so the correlation is also negative.

12. (5 marks) It has been established that the real-world stock market index is well-modeled by a symmetric random walk, in the short term. What do our simulations with random walks tell us about patterns over time in the real-world stock market index?

A12. Apparent patterns – increasing trends, or decreasing trends, or even oscillating trends, are useless for prediction of future prices in the short term, since there is no reason to expect these trends to persist.

13. (5 marks) What is the relationship between histograms and dotplots? (Explain each method, and how they are the same, and how they are different.)

A13. Both are methods of displaying the frequency distribution of values in a data set, but, in a dotplot, the interval size for lumping values together is roughly 1/100 of the width of the display, whereas in a histogram it is typically 1/5 to 1/20 of the width of the display. The histogram uses rectangles to represent frequency, whereas the dotplot just uses a vertical pile of dots.

14. (4 marks) The fuel consumption time series was smoothed to reveal a pattern that was not clear from the raw data. However, we got a similar pattern from the moving average of independent N(0,1) data, suggesting that the pattern in this case was due to randomness. What characteristic of the fuel consumption series suggested the smooth pattern was not due to randomness?

A14. The cycles were in phase with the calendar over five annual cycles, and this would not likely happen if the pattern were generated from random data.

15. (4 marks) In the Gilbert murder case, a small p-value that was calculated led to increased suspicion that Ms. Gilbert was guilty of murder. Explain the logic of this inference.

A15. The p-value was the probability that the number of deaths involving Gilbert's shifts could have been as large or larger than was observed (40 on 257 shifts) if all shifts by all nurses had the same chance of a death. Since this probability was very small, the assumption "all shifts by all nurses had the same chance of a death" is likely false. Since Gilbert's shifts had more deaths than other nurses, a possible explanation would be that Gilbert caused some of the deaths.

16. (4 marks) The regression method had an important role in the articles "Reducing Junk Mail", "Monitoring Tiger Prey Abundance in the Russian Far East", "Advertising as an Engineering Science" and "Predicting Quality and Prices of Wines". What single role did regression play in all these articles?

A16. Prediction of one variable from one or more other variables.

KLW 2010/04/13