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