STAT 100 Final Examination

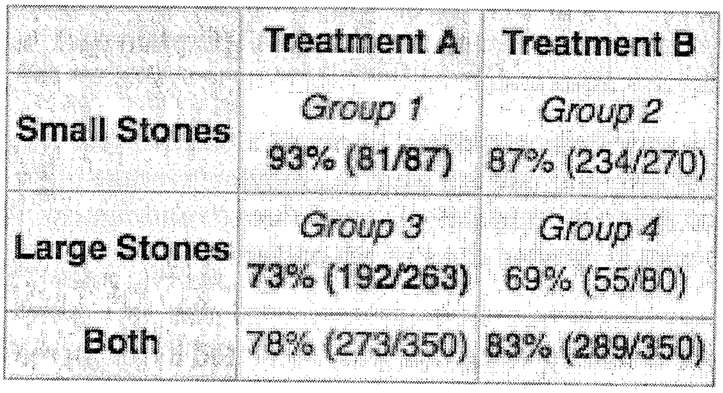
Instructions: Answer all questions. Budget your time so that you have a chance to answer all the questions. You have **three hours** for the exam. The exam is **Open Book**: Texts, Notes, and Calculators are allowed. Be careful to answer the questions posed.

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.

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

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.

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?

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.

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?

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?

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 a poor idea from a purely financial perspective?

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

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.

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

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?

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.)

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?

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.

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?