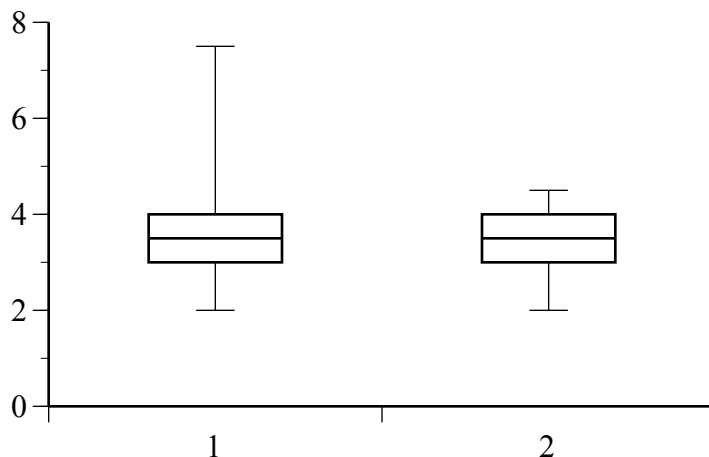


Final Examination (150 minutes)

No calculators allowed. Just set up your answers, for example,  $P = 49/52$ . BE SURE TO EXPLAIN YOUR REASONING. If you want extra time, you can buy time at a price of 1 point a minute; for example, if your test is handed in 10 minutes after the scheduled finish time, 10 points will be subtracted from the test score.

1. A study found that, controlling for the number of cars and miles driven, people driving sports cars were more likely to get ticketed than were people driving minivans. Can you think of a reasonable explanation other than the police like to pick on people driving sports cars?

2. Here are two box plots. Which data set has the higher median? Higher mean? Higher standard deviation?



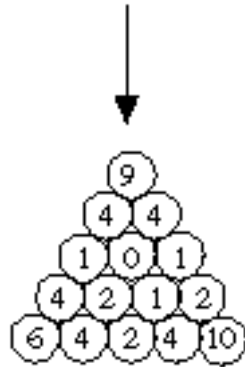
3. You are offered the following game. A fair coin will be flipped until it lands tails. Then the game ends. If it lands tails on the first flip, you win \$1. If it takes two flips, you win \$2. If it takes three flips, you win \$4. And so on....

a. What is the expected value of the payoff?

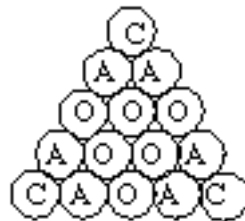
b. How much would you be willing to pay to play this game once?

4. The equation  $Y = \alpha + \beta X + \varepsilon$  was estimated by ordinary least squares. How would a doubling of the number of observations, with the new data exactly replicating the original data, affect
  - a. the estimates of the slope and intercept?
  - b. the standard error of the estimate of the slope?
  - c. the t-value of the estimate of the slope?
  - d.  $R^2$ ?
5. Smith loves to play squash. Andrabi has set up an April 1 match in which Smith will be given a prize if he can win two consecutive games in a three-game match against Andrabi and Ernst alternately, either Andrabi-Ernst-Andrabi or Ernst-Andrabi-Ernst. Assume that Andrabi is a better player than Ernst and that Smith's chances of winning a game against either player are independent of the order in which the games are played and the outcomes of other games. Which sequence should Smith choose and why?
6. A 1981 retrospective study published in the prestigious *New England Journal of Medicine* looked at the use of cigars, pipes, cigarettes, alcohol, tea, and coffee by patients with pancreatic cancer and concluded that there was "a strong association between coffee consumption and pancreatic cancer." This study was immediately criticized on several grounds, including being a fishing expedition in which multiple tests were conducted and the most statistically significant result reported. Subsequent studies failed to confirm an association between coffee drinking and pancreatic cancer. Suppose that six independent tests are conducted, in each case involving a product that is, in fact, unrelated to pancreatic cancer. What is the probability that at least one of these tests will find an association that is statistically significant at the 5 percent level?
7. Explain the error(s) in this reasoning: "The p-value and F-statistic are both very small, which demonstrates a lack of statistical significance and proves the null hypothesis."
8. Use one statistical argument to explain these two aphorisms: "The grass is always greener on the other side of the fence." "Familiarity breeds contempt."

9. Ninety-nine break shots on a full rack of billiard balls resulted in 54 sunk balls, allocated among the 15 balls as follows:



Thus the balls in the three corners of the rack were sunk 9, 6, and 10 times. Dividing the rack into three categories (corners, adjacent, and other),



is there a statistically significant relationship between a ball's position in the rack and its likelihood of being sunk on the break?

10. Karl Pearson flipped a coin 24,000 times and obtained 12,012 heads. Use these data to estimate a 95% confidence interval for the probability of heads with this coin.
11. (continuation) If the probability of a heads with Pearson's coin is 0.5, what is the exact probability that 24,000 flipped would be within 12 of 12,000, that is, between 11,988 and 12,012?

12. A researcher used these data for 60 randomly selected universities:  $Y$  = graduation rate (mean 59.65);  $X_1$  = student body's median math plus reading SAT score, (mean 1,030.5); and  $X_2$  = percent of student body with GPAs among the top 10 percent at their high school (mean = 46.6). He found a statistically significant positive relationship between graduation rate and SAT scores (the t-values are in brackets):

$$Y = -33.39 + 0.090X_1, \quad R^2 = 0.71$$

[2.34]            [6.58]

He also found a statistically significant positive relationship between graduation rate and GPA:

$$Y = 42.20 + 0.375X_2, \quad R^2 = 0.52$$

[8.93]            [4.46]

But when he included both SAT scores and GPAs in a multiple regression equation, the estimated effect of GPA on graduation rates was very small and not statistically significant at the 5 percent level:

$$Y = -26.20 + 0.081X_1 + 0.056X_2, \quad R^2 = 0.71$$

[1.24]            [3.30]            [0.47]

- a. He suspected that there was an error in his multiple regression results. Is it possible for a variable to be statistically significant in a simple regression, but not significant in a multiple regression?
  - b. Do you think that  $X_1$  and  $X_2$  are positively correlated, negatively correlated, or uncorrelated?
  - c. If your reasoning in (b) is correct, how would this explain the fact that the coefficients of  $X_1$  and  $X_2$  are each lower in the multiple regression equation than in the simple regression equations?
13. Data were collected for 25 new cars on the car's weight ( $X$ ) and estimated highway miles per gallon ( $Y$ ). The model  $Y = \alpha + \beta X + \varepsilon$  was estimated by ordinary least squares in three different ways: (a) the data were arranged alphabetically by car name; (b) the data were arranged numerically from the lightest car to the heaviest; and (c) the data were arranged numerically from the lowest miles per gallon to the highest. Which procedure yielded the smallest estimate of the intercept? Which yielded the highest estimate?
14. A researcher estimated a simple regression model using stock market returns for these  $X, Y$  pairs:  $X$  = 1930 return,  $Y$  = average return 1930-1939;  $X$  = 1940 return,  $Y$  = average return 1940-1949, and so on. The estimated correlation was positive, leading him to conclude that annual stock market returns could be used to predict returns over the next 10 years. What statistical flaw do you see in his procedure?

15. A dean claims that humanities professors' salaries  $Y$  are determined solely by teaching experience  $X$ . To see if there is a statistically significant difference in the salaries of male and female humanities professors, the equation  $Y = \alpha + \beta_1 D + \beta_2 X + \beta_3 DX + \epsilon$  was estimated, using the variable  $D = 0$  if male, 1 if female.
- Interpret each of the parameters  $\alpha$ ,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ .
  - What advantage does a regression model have over a simple comparison of average male and female salaries?
  - Describe a specific situation in which a comparison of male and female average salaries shows discrimination against females while a regression equation does not.
  - Describe a specific situation in which a comparison of male and female average salaries shows no discrimination against females while a regression equation indicates discrimination against females.
16. Long ago, the astragali (heel bones) of animals were used as dice. An astragalus of a hooved animal has four sides. Experiments have shown that the probabilities of each of these four sides are 0.39, 0.37, 0.12, and 0.12. In ancient Greece, one game was to roll four astragali simultaneously, with the best outcome being a "Venus," in which each of the four different sides appears. What is the probability of rolling a Venus?

17. A researcher asked 76 college students to identify the food they grew up eating and their favorite food:

Grew Up Eating	Favorite Food				Total
	Asian	American	European	Latin American	
Asian	10	0	1	0	11
American	8	12	7	11	38
European	3	0	7	3	13
Latin American	3	2	2	7	14
Total	24	14	17	21	76

He used the binomial distribution to test these null hypotheses:

$H_0$ : probability[prefer Asian if grew up eating Asian] = 0.25

$H_0$ : probability[prefer American if grew up eating American] = 0.25

$H_0$ : probability[prefer European if grew up eating European] = 0.25

$H_0$ : probability[prefer Latin American if grew up eating Latin American] = 0.25

What would have been a better statistical procedure? You do not need to do the procedure; just identify it.

18. A researcher used least squares to estimate the equation  $Y = \alpha + \beta X + \epsilon$ , where  $Y$  = college GPA and  $X$  = high school GPA. First, he estimated the equation using all his data. Then, he separated his data into male and female students and estimated separate equations for each gender. Identify any apparent errors you see in these reported results.

	Men	Women	Total
sample size	407	482	869
intercept	-0.22 (-0.32)	-0.04 (-0.24)	-0.12 (-0.08)
slope	0.88 (-0.32)	0.80 (-0.24)	0.83 (-0.08)
R-squared	1.01	1.54	1.27

( ): standard error

19. A police department recorded the clothing colors worn by pedestrians who died after being struck by cars at night. They found that four-fifths were wearing dark clothes and concluded that it is safer to wear light clothes. Under what conditions is the probability of being struck by a car if wearing dark clothing equal to the probability of wearing dark clothes if struck by a car?

20. Use a specific example to explain why regression to the mean either is or is not the same as the law of averages.