

**Midterm**

Please send your answers to both of us (jjacobsen, mkaparakis) as an email attachment no later than 5:30 pm on Friday June 25.

(1) (10 points) According to the Center for Disease Control, college graduates feel better emotionally and physically than do high school dropouts. This result came from telephone interviews conducted from in which 431,996 people nationwide were asked how healthy they felt physically and emotionally over the previous 30-day period. Overall it was found that college graduates felt healthy an average of 26 days a month, compared with 23.8 days for dropouts.

- a. How reliable do you think people's memories are on these figures? (Think about how many days in the last 30 you felt well). Comment on how memory reliability might affect the findings.
- b. The Center for Disease Control said the report may be biased because not everyone has a phone or is willing to confess their habits. How does each of these factors bias the estimates? Is the estimate of the difference between the two groups biased as well?

(2) (10 points) The women in a certain college engage in various sports in the following proportions:

Basketball, 10% of all women	Basketball and soccer, 5% of all women
Soccer, 20% of all women	Basketball and track/field, 5% of all women
Track/field, 20% of all women	Soccer and track/field, 10% of all women
	All three sports, 1% of all women

If a woman is chosen randomly at the college, what is the probability that she will be:

- a. A basketball player or a soccer player?
  - b. An athlete (someone who plays at least one sport)?
  - c. If an woman athlete is chosen randomly at the college, what is the probability that she will be a soccer player only?
- (3) (10 points) Mammograms have about a 20% false negative rate (failing to detect cancer when it is present) and a 10% false positive rate (signalling the presence of cancer when it is not present).
- a. It has been estimated that 10 in every 1000 women between the ages of 40 and 50 will develop breast cancer in a given year. If a woman in this age range tests positive for breast cancer on her mammogram, what is the probability that she has breast cancer?
  - b. Assume that if a woman receives a positive mammogram, she will be retested with a second mammogram. Then if the second mammogram is also positive, she will be diagnosed as having cancer. Using this two-stage procedure, what is the probability that a woman without cancer will be diagnosed as having cancer? What is the probability that a woman with cancer will be diagnosed as having cancer?

- (4) (10 points) At Dunkin Dough Huts, the only two items for sale are donuts and coffee. The counter help notes the following distribution of donut and coffee sizes (both in ounces):

	coffee size	
donut size	16	24
4	.20	.30
8	.40	.10

- Calculate the mean and standard deviation for donut size.
  - Calculate all eight conditional probabilities that can be calculated from this joint distribution (in two tables, if you like).
  - Without calculating the covariance, explain how donut size and coffee size are correlated, and how you knew to give your answer.
- (5) (10 points)
- Our textbook argues (e.g., the answer to problem 7-14) that it is more important for an estimator to be accurate than to be precise. Give an example that illustrates this. Can you think of a case that goes against what the book says, i.e., a case where it is more important to be precise than to be accurate?
  - Read problem 9-8 on p. 304. This issue is discussed with regard to scientific journals. Can you think of another realm of life where this tendency to report positive results can be problematic? Make sure your example is clearly stated and that the problems that can arise are made clear as well.
- (6) (20 points) Write a short essay (1-2 pp) in answer to each of these questions. Your grade on these answers will depend on the coherence and completeness of the answer and on your demonstrating knowledge of the material covered in this course.
- At various points in the course, examples have been given of how ordinary humans (i.e., people who haven't had QAC 201) might make systematic errors in interpreting statistical evidence. Discuss some of these examples. What conclusion do you, as a budding scientific researcher, draw from these examples?
  - Suppose you are doing a study of whether or not eating healthy food (i.e., fruits, vegetables, and high-fiber low-fat foods in general) affects life expectancy. You are being funded by the American Junk Food Association, which is hoping that your study will show that eating healthy food has no relationship to life expectancy (or, even better, reduces life expectancy). Discuss various ways in which you might be able to conduct the study or present your statistical results so as to support the conclusion that your sponsor prefers.

(7) (20 points) The mayor of your town can really use your help. A group of town residents is trying to close down a small farm because the group claims it pollutes the nearby river. Another group of town residents sides with the farm and denies that there is any pollution due to the activities of the farm. The mayor has learned of your statistical skills and asks you to help resolve the problem. You agree, and decide to collect water samples upstream and downstream from the farm and explore whether or not there are significant differences in the amount of dissolved oxygen in the water. The data you collected are in datafile P:\QAC\SUMMER10\ASGN\sewage.csv (oxy= Dissolved oxygen mg/L; loc =location)

- a. Is there any statistical evidence to support the claim of the group trying to close the farm? (Clearly state your hypotheses and report appropriate test statistics, p-value, etc.)
- b. A friend of your who is a biology majors hears of your project and asks to look at your data. He emails you: "Tell the mayor there is a bigger problem—the water quality is rather poor—the amount of dissolved oxygen should be greater than 4.75 milligrams per liter for the water quality to be considered good." Test your friend's claim (again clearly state hypotheses, etc.). Would you agree with your friend if your selected level of statistical significance is one percent?
- c. Include the stata code (contents of the do file) that you used to do your work.

(8) (10 points) Select two variables from the dataset you are using for your course project that you believe are related.

- a. Produce appropriate univariate statistics to summarize the information content of the two variables. (As part of this set, please make sure to report minimum and maximum values for each variable).
- b. Use appropriate graphics to help us visualize the (hypothesized) relationship between the two variables.