LAST NAME (Please Print): KEY

FIRST NAME (Please Print): _____

HONOR PLEDGE (Please Sign): _____

Statistics 111

Midterm 4

- This is a closed book exam.
- You may use your calculator and a single page of notes.
- The room is crowded. Please be careful to look only at your own exam. Try to sit one seat apart; the proctors may ask you to randomize your seating a bit.
- Report all numerical answers to at least two correct decimal places or (when appropriate) write them as a fraction.
- All question parts count for 1 point.

I have really enjoyed this semester, and I hope you have too. It has been a lot of work, but the intent was to make it fun as often as possible, and always worthwhile and rewarding.

David Banks

1. What is the purpose of nonparametric regression?

   To predict Y from the explantory variables when the relationship is not of a known form (e.g., linear, quadratic, etc.

2. Explain how a running line smooth with a window containing five observations works. (2 points)

   (1) To find the expected value of Y at X, identify the five X-values closest to X and do simple linear regression on those. (2) Use that line to predict Y.

3. Suppose you are doing cross-validation to assess predictive accuracy in nonparametric regression. But by mistake, some of the (X, Y) pairs that you use to fit the model have been duplicated in the data set. How will this affect your assessment and why? (2 points)

   (1) One will overestimate predictive accuracy. (2) Because the holdout samples will contain copies of the same data used to fit the model.

4. In terms of predictive accuracy, what is the implication of the Curse of Dimensionality? (1 point)

   The accuracy of regression predictions quickly becomes poor as the number of explanatory variables increases.

5. Consider the following design and data.

run	A	B	C	D	E	F	obs
1	-	-	-	+	+	+	3
2	+	-	-	-	-	+	7
3	-	+	-	-	+	-	-2
4	+	+	-	+	-	-	-4
5	-	-	+	+	-	-	6
6	+	-	+	-	+	-	10
7	-	+	+	-	-	+	1
8	+	+	+	+	+	+	13

What are the three generating relations?

I = ABD = ACE = BCF

Look for columns whose sign product equals those of other columns.

How does one use the three generating relations to find all the interactions that are equal to the identity element (i.e., the defining relations)? (1 point)

One multiplies all possible combinations of the generating relations together.

What two-way interactions are confounded with A?

BD, CE

0.25 What is your numerical estimate of the AB interaction?

$\frac{1}{8}(3-7-(-2)-4+6-10-1+13)$ where the signs are the sign products of columns A and B.

In this context, how do we interpret things when a main effect is confounded with a two-way interaction. (2 points)

(1) The effect estimate is the sum of the main effect and the two-way interaction (and anything else that is confounded). (2) We assume that higher order interactions are negligible compared to main effects.

How would you decide whether an estimate of the A effect is significant?

One finds all main effect estimates, and looks for outliers.

2^{6-3}_{III} Give the symbolic name for this design (include the resolution).

There are six factors, but the design is halved three times, giving 2^{6-3}. The resolution is 3 since main effects are confounded with two-way interactions, so I + II = III.

6. An investment firm wants to compare the Bloomberg terminal with the Reuters 3000 Xtra. Five traders are chosen from among their employees, and each uses one of the two terminals for two days, and then switches to the other one. The response of interest is how much profit or loss a trader makes on each of the four days.

   Complete the following ANOVA table. To minimize error ripple, only the blanks with parentheses will be scored.

Source	df	SS	MS	F
Terminal	1	10	10	(2)
Trader	4	5	1.25	(2.99)
Interaction	4	20	5	(11.99)
Error	10	4.17	0.417	
Total	_____	39.17		

   3.48 What is your critical value for testing whether there is an effect due to Trader, at the 0.05 level?

   This is a mixed-effects ANOVA, so the cv is $F_{4,10}$, or 3.478.

   Yes Is there a Trader effect at the 0.05 level?

   The interaction is significant, so everything else is too.

   When do you use Fisher's LSD?

   When the ANOVA test is significant.

7. The lifespan of a mayfly, in days, has density $f(t) = 6t^5$ for $0 \le t \le 1$.

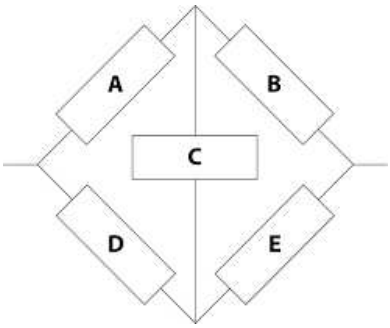
   0.19 What is the value of the hazard function at $t = 1/2$?

   The hazard function is $f(t)/[1 - F(t)] = \frac{6t^5}{1-t^6}$.

   increasing Which, among the four types discussed in class, is this hazard function?

   Graph it. Or look at the sign of the derivative.

8. 0.09 Consider the system shown in the figure. Components A and B fail with probability 0.2, component C fails with probability 0.3, and component D fails with probability 0.4 and E never fails. What is the probability that the system fails?



   Make a table for working and failing:

run	A	B	C	D	System
1	w	w	w	w	w
2	f	w	w	w	w
3	w	f	w	w	w
4	f	f	w	w	w
5	w	w	f	w	w
6	f	w	f	w	w
7	w	f	f	w	w
8	f	f	f	w	w

   and there are eight more rows. Of these, only (f, w, w, f), (f, f, w, f), (f, w, f, f), (w, f, f, f) and (f, f, f, f) correspond to system failure. Multiply the probabilities out and sum to get 0.0928.

9. Five years A $10,000 car depreciates by $1,000 for each year that you own it. An insurance policy that costs $800/year will reimburse the current value of the car if it is totaled. The probability your car gets totaled in a year is 1/7. How old is the car when you stop buying insurance?

   The expected cost without insurance when the car is purchased is $10,000/7 = $1,428.57 > $800, so buy insurance. When the car is one year old, the expected cost without insurance is $9,000/7 = $1,285.71 > $800, so buy the insurance. When the car is five years old, the expected cost is $5,000/7 = $714.28, so don't buy.

10. List all, and only, the true statements. C, E, G, I, J, L

    A. Human beings have increasing hazard rates.

    B. A fractional factorial with resolution II is better than one with resolution III.

    C. The Cox proportional hazards model is good when something wears out at a rate determined by several factors.

    D. People tend to underestimate the risk of being struck by lightning.

    E. The more money someone has, the less they value $10.

    F. People are about evenly split between risk seekers and risk avoiders.

    G. Cost-benefit analysis in a stochastic world requires statistics.

    H. Roger Boisjoly said "All models are wrong, but some are useful."

    I. Competing risks models are appropriate for most first-person shooter games.

    J. In high dimensions, data tend to concentrate on a subspace.

    K. In regression, extrapolation is safer than interpolation.

    L. Statistics can be hard, but it is useful.