Applied Comprehensive Examination

August 17, 2001

Instructions:

1. You have four hours to answer questions in this examination.
2. Answer as many of the questions as you can during this time.
3. There are 185 total points on this exam.
4. Write only on one side of the paper, and start each question on a new page.

1. It is conjectured that students who attend class lectures and also study the textbook perform better than students who study the textbook but occasionally miss class. On a recent test, 35 students who swear they never miss class and practice good study habits scored $\overline{Y}_1 = 92.5$ with $S_1^2 = 6.5$. Thirty students who admit to missing class "now-and-then" but study the text, scored $\overline{Y}_2 = 90.2$ with $S_2^2 = 10.2$.

(a) Construct an appropriate confidence bound that can be used to check the validity of the conjecture at the 1% significance level. **(10 pts)**

(b) Use the calculated bound to draw a conclusion about the conjecture. **(5 pts)**

2. A chemical company claims that a new insecticide kills at least 90% of the treated insects within 24 hours. In a random sample of 80 insects, 75 were dead within 24 hours. Should we believe the claim by the chemical company? Explain. **(10 pts)**

3. In an experiment to compare four methods of determining serum amylase values, serum specimens were collected from six patients. Each specimen was divided into four parts and assigned to the four methods at random. The measured serum amylase values were as given below.

	method				
patient	1	2	3	4	total
1	360	435	391	502	1688
2	1035	1152	1002	1230	4419
3	632	750	591	804	2777
4	581	703	583	790	2657
5	463	520	471	502	1956
6	1131	1340	1144	1300	4915
	4202	4900	4182	5128	18412

(a) Write a model for this experiment defining all terms. **(5 pts)**

(b) Given $SS[Methods] = 116,979.3$ and $SS[E] = 24,593.7$, perform an hypothesis test to see if there is a difference between the true mean amylase values determined by the four methods. Be sure to specify the null and research hypotheses that are being tested and the conclusion that can be drawn. **(10 pts)**

(c) The investigators wish to perform a pairwise multiple comparison of the four method means with experimentwise error rate not exceeding 0.05.

1. Explain what is meant by experimentwise error rate. **(5 pts)**
2. Suggest two methods that can be used for this purpose. **(5 pts)**
3. Which method would you prefer? Why? **(5 pts)**

4. In an investigation of the effect of smoking and physical activity on oxygen uptake, 16 individuals were classified into two groups: non-smokers and heavy smokers. Four subjects in each group were randomly assigned to each of the two stress tests: bicycle ergometer and step test. The following data are the times (in minutes) until maximum oxygen uptake for the study subjects in their assigned tests.

	test		
	bicycle	step	total
non-smokers	12.8, 13.5, 11.2, 11.6	22.6, 19.3, 18.9, 19.6	129.5
heavy smokers	8.7, 9.2, 7.5, 8.6	16.2, 16.1, 17.3, 17.8	101.4
total	83.1	147.8	230.9

(a) Write an appropriate ANOVA model for the data. Explain all terms in your model. **(5 pts)**

(b) Use the following data to perform appropriate statistical tests and draw conclusions about the effects of smoking and physical activity on oxygen uptake. **(10 pts)**

source	sum of squares
smoke	49.3506
test	261.6306
smoke×test	0.2756
total	326.8543

5. A drug antibiotic manufacturer randomly sampled 12 different locations in the fermentation vat to try and estimate the mean potency for the batch of antibiotic being prepared. Readings were as follows:

8.9	9.0	9.1	8.9	9.1	9.0	$\overline{Y} = 8.983$
9.0	8.8	9.1	8.9	8.8	9.2	$S = 0.127$

(a) Set up a 98% confidence interval for the mean potency for the batch and interpret the interval. **(5 pts)**

(b) A laboratory technician sampled 10 different locations in the same fermentation vat as the manufacturer above and her readings and 98% confidence interval were as follows:

9.0	9.0	9.1	8.9	9.0	$\overline{Y} = 9.0$	98% CI = (8.897, 9.103)
8.8	8.9	9.1	9.2	9.0	$S = 0.115$	

The manufacturer and the technician are arguing about whose CI is better. They come to you for advice as to what you think they should do to come up with the best interval. What do you recommend they do? Explain your reasoning. **(10 pts)**

6. The listed fill weight of a box of cereal is 20 oz. If the standard deviation among box weights is 1.5 oz., how many boxes must be sampled to estimate the true mean fill weight to within ± 0.5 oz. with 99% confidence? **(10 pts)**

7. The data below were collected from twenty lakes in the Adirondack region of New York State. For each of the twenty lakes, the table gives (i) the number of fish species that live in the lake, (ii) the pH of the water, and (iii) the area of the lake (in hectares).

Lake	Species	pH	Area	Lake	Species	pH	Area
1	9	7.03	27.1	11	5	7.06	2.1
2	6	6.94	10.6	12	8	6.77	36.5
3	0	5.09	18.9	13	6	8.10	1.7
4	4	6.00	0.8	14	0	4.69	0.8
5	6	7.30	8.5	15	1	5.15	4.4
6	3	5.33	22.1	16	3	6.20	2.7
7	4	5.28	6.5	17	8	6.95	3.3
8	12	7.42	152.6	18	1	5.35	47.4
9	6	5.04	7.8	19	9	6.98	14.9
10	5	5.33	54.6	20	6	7.40	40.9

Consider the following SAS program.

```
data fish;
input obs species ph area @@;
cards;
1 9 7.03 27.1 2 5 7.06 2.1 3 6 6.94 10.6 4 8 6.77 36.5 5 0 5.09 18.9
6 6 8.10 1.7 7 4 6.00 0.8 8 0 4.69 0.8 9 6 7.30 8.5 10 1 5.15 4.4
11 3 5.33 22.1 12 3 6.20 2.7 13 4 5.28 6.5 14 8 6.95 3.3 15 12 7.42 152.6
16 1 5.35 47.4 17 6 5.04 7.8 18 9 6.98 14.9 19 5 5.33 54.6 20 6 7.40 40.9
;
run;
data newfish;
  set fish;
  interact=ph*area;
run;
proc reg data=newfish;
   model species=ph area interact/ss1 ss2;
run;
```

(a) Write down the multiple regression model that this SAS program fits. **(5 pts)**

Partial Output from SAS Program

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	xxx	xxx	a	xxx
Error	b	xxx	c	
C Total	xxx	xxx		

Root MSE	xxx	R-square	xxx
Dep Mean	5.10000	Adj R-sq	0.6034
C.V.	39.64045		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0
INTERCEP	1	-7.089727	3.61852906	-1.959
PH	1	d	0.55732832	e
AREA	1	-0.056113	0.12724308	-0.441
INTERACT	1	0.012470	0.01783200	0.699

Variable	DF	Type I SS	Type II SS
INTERCEP	1	520.200000	15.689616
PH	1	105.366184	44.985355
AREA	1	23.041074	0.794826
INTERACT	1	f	1.998812

(b) Fourteen numbers have been removed from the output and replaced with either a single letter or the symbol "xxx". Find the values of a, b, c, d, e, and f. You may use the fact that the first diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ is 3.2. You do not have to find the values of the xxx's. For each letter, explain how you arrived at your answer. The correct answer without an explanation is worth nothing. **(15 pts)**

(c) Compute a 95% confidence interval for the mean number of fish species in a lake with an area of 10 hectares and a pH of 7. You may use the fact that

$$\begin{pmatrix} 1 & 7 & 10 & 70 \end{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} 1 \\ 7 \\ 10 \\ 70 \end{pmatrix} = 0.09125$$

(If you need a number from the Partial Output that you can't compute, just use the corresponding letter in its place.) **(5 pts)**

8. You have fit a linear model using $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ involves r independent variables. Now assume that the true model involves an *additional* s independent variables contained in $\mathbf{Z}$. That is, the true model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\gamma}$ are the regression coefficients for the independent variables contained in $\mathbf{Z}$.

(a) Find $E(\hat{\boldsymbol{\beta}})$ (under the true model). **(4 pts)**

(b) Under what conditions will $\hat{\boldsymbol{\beta}}$ remain an unbiased estimator? **(4 pts)**

9. In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X_1) and sweetness (X_2) of the product, the following results were obtained from the experiment based on a completely randomized design (data are coded):

Y_i=Liking	X_{i1}=Moist	X_{i2}=Sweet
64	4	2
73	4	4
61	4	2
76	4	4
72	6	2
80	6	4
71	6	2
83	6	4
83	8	2
89	8	4
86	8	2
93	8	4
88	10	2
95	10	4
94	10	2
100	10	4

(a) Consider the simple linear regression model whose equation is given by $E(X_{i2}) = \beta_0 + \beta_1 X_{i1}$. Given that $\sum_{i=1}^{16} X_{i1}X_{i2} = 336$, $\sum_{i=1}^{16} X_{i1} = 112$, and $\sum_{i=1}^{16} X_{i2} = 48$, what is the least squares estimate of β_1? **(2 pts)**

(b) Consider the following three models

$$E(Y_i) = \beta_0 + \beta_1(\text{Moist}) + \beta_2(\text{Sweet})$$

$$E(Y_i) = \beta_0 + \beta_1(\text{Moist})$$

$$E(Y_i) = \beta_0 + \beta_2(\text{Sweet})$$

In terms of fitting these three models, what are the implications of your answer to (a)? **(5 pts)**

10. Consider the simple linear regression through the origin model with equation

$$Y_i = \beta x_i + \varepsilon_i \ ,$$

$i = 1, \ldots, n$, and assumptions: (i) the ε_i are iid N$(0, \sigma^2)$, (ii) β and σ^2 are unknown parameters, and (iii) the x_i are known constants.

(a) Show that the least squares estimator of β is

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \ .$$

(5 pts)

(b) Find the distribution of $\hat{\beta}$. (You may use the fact that a linear combination of independent normal random variables has a normal distribution.) **(5 pts)**

(c) Define $\hat{Y}_i = x_i\hat{\beta}$. Find the expected value of $\sum_{i=1}^n (\hat{Y}_i - Y_i)^2$ and use the result to construct an unbiased estimator of σ^2, call it $\hat{\sigma}^2$. **(5 pts)**

(d) Consider testing $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$ for some fixed β_0. Write down an appropriate test statistic. What is the distribution of this statistic. Why? (You don't need to do any calculations, just explain.) **(5 pts)**

(e) Construct a second unbiased estimator of β that depends on the Y_i's only through $\sum_{i=1}^n Y_i$. Call this new estimator $\tilde{\beta}$. Show that the mean square error of $\hat{\beta}$ is smaller than that of $\tilde{\beta}$. **(5 pts)**

11. A researcher wishes to study the effect on the dependent variable Y of a single factor (qualitative variable) with 3 levels, and a quantitative predictor X. He assumes that all observations are independent with (approximately) equal variances.

(a) (Model 1) Initially the researcher wishes to fit the model which assumes that Y depends linearly on X, but with possibly different slopes and intercepts at each level of the factor. Define appropriate indicator variables and write down a representation of this model. **(8 pts)**

(b) (Model 2) Now write down the model corresponding to the hypothesis of parallel regression lines, i.e., the model that assumes that the slope of the regression line is the same for each level of the factor. **(3 pts)**

(c) (Model 3) Finally write down the model which assumes that Y does not depend on the factor (after accounting for X). **(3 pts)**

(d) Suppose that the researcher has 36 observations, and that the residual sums of squares from fitting each of the above models are 120, 144, and 162, respectively. Calculate the F-statistic for each of the following model comparisons, and give its numerator and denominator degrees of freedom. **(11 pts)**

(i) Model 2 vs Model 1.

(ii) Model 3 vs Model 2.