

Stat 401 B/xm - Fall 2001 - Final Exam

18 Dec 2001

Please put your name on the back of the last page

There are 8 problems on the exam. Most have multiple parts and are worth different numbers of points. Generally, each separate part is worth 5 points, although there are a few exceptions. The total number of points for each question is indicated.

Please write your answers in the enclosed spaces. If you need more room, continue on the back of the page.

A formula sheet is included on the last page. If you want a formula that you don't see, please ask. If you want any tabulated statistic (e.g. a t statistic), just ask.

1. 25 pts. Here are two box plots of observations (control and active treatment). Each group has 50 observations. These are from a randomized experiment. Each experimental unit is randomly assigned to one of the two treatments. For parts a) - c), please consider the distribution in the active group. Part d concerns both groups.

a) Is the distribution of the observations in the active treatment group symmetrical? Why or why not?

b) Approximately, what is the median of the observations in the active treatment group?

c) Will the average of the observations in the active treatment group be larger than the median, the same as the median, or smaller than the median?

d) Here are six possible methods to test the hypothesis of no treatment effect. Circle the two methods that are most appropriate for these data.

- 1 t-test on raw data
- 2 paired t-test on raw data
- 3 t-test on log transformed data
- 4 paired t-test on log transformed data
- 5 F-test on raw data
- 6 Wilcoxon rank sum test on raw data

2. 45 pts. The lab I used to work at had a large research program on radiation effects in natural systems. They did a lot of work in the Chernobyl area, where a nuclear reactor accident over 10 years ago created widespread contamination, especially by Cesium (Cs). One study investigated the relationship between the Cesium concentration in fish muscle and the number of mutations in a specific gene sequence. Radiation is one of many factors that can increase the number of observed mutations. 25 fish were collected from a particular lake near Chernobyl. The variables of concern are:

- Mass log transformed weight of the fish (in kg), this is closely associated with fish age
- Dose the Cesium (Cs) concentration measured in fish muscle
- # Mutations the number of mutations (for the geneticists, the # of restriction fragment length polymorphisms)

Plots of data from the 25 fish are on the next page. There are also two plots of diagnostics that are used in part f).

The fitted regression equation is:

$$\# \text{ mutations} = -1.52 + 5.12 * \text{mass} + 12.0 * \text{dose}$$

The residual s.d. is 2.61.

The standard errors of the coefficients are: intercept 1.08, mass 1.52, dose 6.91

- a) How many degrees of freedom are there for the residual SS?
- b) Use the fitted regression equation to fill in the table:

Case	Weight (kg)	Dose	predicted # mutations
A	1.0	0.10	
B	1.0	0.50	

c) Carefully explain the interpretation of the coefficient for dose in the fitted regression equation.

d) One biological question is whether the predicted # mutations for young uncontaminated fish (mass

= 0, dose = 0) is 0 or something other than 0. Note that mass=0 is biologically meaningful because mass is log transformed weight. Is it possible to construct a 95% confidence interval for the predicted # mutations for young uncontaminated fish (mass = 0, dose = 0) from the information given in the beginning of the problem and an appropriate t-statistic? If so, please construct that confidence interval. The appropriate t-statistic is 2.069. If not, please explain what extra information you need.

e) A residual plot and a plot of Cook's Distance are included on the next page. Does either suggest a concern with the multiple linear regression? Why or why not?

f) If you do a simple linear regression of mutations on dose, the fitted regression equation is:

$$\# \text{ mutations} = -0.8173 + 33.4 * \text{dose}$$

The residual s.d. is 3.139.

The standard errors of the coefficients are intercept: 1.27 and dose: 3.30. The investigator can not understand why the coefficient for dose from this model is so different from the coefficient for dose from the model in part a. Please explain why. (or explain why the coefficients should be the same, so the difference is just due to sampling variation).

3. 20 pts. Nitrates, a component of fertilizer, are good in farm fields and generally bad when they get into rivers and streams. A few years ago, a research group tried to predict nitrate export in rivers. They measured seven characteristics that might 'explain' nitrate export: DIScharge of the river (m^3/sec), annual watershed RUNoff ($1/(\text{sec km}^2)$), Precipitation (cm/yr), watershed AREA (km^2), POPulation density ($\text{people} / \text{km}^2$), nitrate concentration in PRECipitation ($\mu\text{M}/(\text{sec km}^2)$), and annual watershed nitrate DEPosition (precipitation * nitrate conc. in precipitation). A river is one observation; the 42 rivers in the data set are scattered throughout the world.

The investigators looked at a variety of possible regression models. The variables in the model are identified by the capitalized letters in the descriptions of the variables, e.g. P is Precipitation, DIS is DIScharge and PREC is the nitrate concentration in the PRECipitation. Results from some of the models are given below:

Model	Variables in model	Residual SS	MSE	R ²	Cp	BIC	AIC
1	POP, DEP	218.56	5.60	0.774	3.58	80.49	75.27
2	RUN, DIS, DEP	194.80	5.13	0.799	1.28	79.39	72.44
3	RUN, DIS, DEP, PREC	188.46	5.09	0.805	2.13	81.74	73.05
4	RUN, AREA, POP, DEP	192.64	5.21	0.801	2.89	82.66	73.97

a) Of these four models, which is the most appropriate one to predict export for a new river? Why?

b) A friend argues that a model with 7 variables (not included in the above table) should be used because it has the smallest MSE (4.27) and highest R^2 (0.85). Do you agree? Why or why not?

4. 35 pts These data are part of a British experiment comparing 4 ways of growing green peppers in a glasshouse (= a greenhouse on this side of the Atlantic). The different ways were: the usual way (a control), with a 0.5% increase in CO₂ concentration, with a 1% increase in CO₂ concentrations, and with supplemental heating. Because 5 separate buildings (labelled A, B, C, D, E), each with 4 glasshouses were available, the experiment was designed as a randomized complete block experiment. Each building was a block and the treatments were randomly assigned to each glasshouse. Six (6) pepper plants were grown in each glasshouse. The response was the average yield of peppers in each glasshouse. The observations are plotted below

The data were analyzed in two different ways, corresponding to the following SAS or JMP models:

Analysis	SAS PROC GLM;	JMP: Block and Trt are nominal Construct Model box includes:
1	CLASS BLOCK TRT; MODEL YIELD = BLOCK TRT;	BLOCK and TRT
2	CLASS TRT; MODEL YIELD = TRT;	TRT

The ANOVA tables and Least Squares Means (LSMEANS) for each analysis are:

1)	Source	d.f.	Sum of sq	Mean Sq.	F	p value
	Block	4	108.72	27.18	4.70	0.016
	Trt	3	90.89	30.30	5.23	0.015
	Error	12	69.45	5.79		
	treatment	lsmean	s.e.			
	control	13.4	1.1			
	0.5% CO ₂	14.2	1.1			
	1% CO ₂	16.4	1.1			
	heat	10.4	1.1			
2)	Source	d.f.	Sum of sq	Mean Sq.	F	p value
	Trt	3	90.89	30.30	2.72	0.079
	Residuals	16	178.17	11.14		
	treatment	lsmean	s.e.			
	control	13.4	1.1			
	0.5% CO ₂	14.2	1.1			
	1% CO ₂	16.4	1.1			
	heat	10.4	1.1			

a) Identify the experimental unit and observational unit in this experiment.

b) Identify the factors and their levels used in this experiment.

c) Which analysis corresponds to a 1 way ANOVA?

d) Which analysis and ANOVA table are the most appropriate to answer questions about the effects of the CO₂ and heat treatments on pepper yield? Why?

Provide answers for questions e, f, g, and h in the following table.

e) Give the coefficients for the linear contrast to estimate the effect of heating the glasshouse.

f) Estimate the value of this contrast.

g) Give the coefficients for the linear contrast to estimate the linear trend of CO₂ concentration.

h) Estimate the value of this contrast.

Do not estimate the precision of either contrast.

Question	Contrast coefficients for:				Estimated value
	control	0.5% CO ₂	1% CO ₁	heat	
e) and f)					
g) and h)					

5. 25 pts. During my Ph.D thesis I compared seedling survival rates in two different habitats. I created two experimental patches, one in each habitat and planted 25 seedlings in each patch. At the end of the growing season, I counted the number alive and number dead in each area. The data looked like:

	Alive	Dead	Total
Wet meadow	8	17	25
Dry ridge	3	22	25
Total	11	39	50

- a) Estimate the difference (between the two habitats) of the probability of surviving.
- b) Estimate the precision (s.e.) of that difference.
- c) For these data, I used a Chi-square test. How many d.f. are associated with the Chi-square statistic?
- d) The value of the Chi-square statistic was 2.91, with a p-value of 0.088
State the hypothesis being tested in this test.
- e) Is the Chi-square test appropriate for these data and hypothesis? Why or why not?
6. 25 pts. A paired t-test was used to compare two means. The t statistic was large, and the p-value was highly significant. Consider the p-value for that test. For each item in the enclosed list, circle whether the p-value is likely to increase (be less significant), not change, or decrease (become more significant).

the p-value will:

If the sample size is increased	increase	no change	decrease
If the population variance increases	increase	no change	decrease
If the difference between the means increases	increase	no change	decrease
If you ignore the pairing and analyze the data as 2 groups	increase	no change	decrease
If you treat each pair of observations as a block and use an F test for treatments in an RCB	increase	no change	decrease

7. 15 pts. For each of the following, indicate whether it is or is not a randomized experiment.

a) IS IS NOT. An experiment evaluating surgical treatments for lameness in horses. Horses are randomly chosen from a herd. The left foreleg receives the control treatment; the right foreleg receives the active treatment.

b) IS IS NOT. An experiment comparing two methods of measuring soil phosphorus concentration. Soil samples are collected from five arbitrary locations in a field. Each soil sample is divided in two. One half is randomly chosen to be measured by the 'standard' method; the other half is analyzed by the 'new' method.

c) IS IS NOT. An experiment evaluating changes in mushroom abundance after a forest fire. Five plots are randomly located in an area that was burnt two years ago. Five plots are randomly located in an adjacent unburnt area.

8. 10 pts. One approach to summarizing population trends in endangered animals is to calculate the linear regression of $Y = \log(\text{estimated population size})$ on $X = \text{year}$. You probably remember that the standard error of a slope in simple linear regression depends on the number of data points, the variance in the X values and the standard deviation of the residuals. Manatee populations in Florida have been intensively studied. For four populations, the estimated slopes, the standard errors of those slopes and the p-values for a test of slope = 0:

Population	N	est. slope	s.e.	p-value
A	5	-0.1	0.2	0.45
B	15	-0.05	0.05	0.30
C	21	-0.21	0.07	0.001
D	16	-0.30	0.14	0.04

a) In which population is the slope (the rate of decline) known the most precisely? Why?

b) Which population provides the most convincing evidence that the population size is actually declining? (i.e. most convincing evidence that the true slope is not zero)? Why?

Congratulations! You're done. Enjoy the holiday.

Formulae

$$\begin{aligned}
 \text{s.e. } \bar{X} &= \frac{s_x}{\sqrt{N}} \\
 \text{pooled s.d. } s_p &= \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}} \\
 \text{pooled s.e. of } \bar{X}_1 - \bar{X}_2 &= s_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}} \\
 \text{T ratio: } T &= \frac{\text{estimate} - \text{parameter}}{\text{s.e. of estimate}} \\
 \text{estimate of a linear combination: } g &= \sum c_i \bar{X}_i \\
 \text{s.e. of a linear combination: } s_g &= s_p \sqrt{\sum c_i^2 / n_i} \\
 \text{s.d. of obs around a regression line: } s_e = \hat{\sigma} &= \sqrt{\frac{1}{N - k} \sum (Y_i - \hat{Y}_i)^2} \\
 \text{s.e. of slope: } s_{\beta_1} &= s_e \sqrt{\frac{1}{(n - 1)s_x^2}} \\
 \text{s.e. of intercept: } s_{\beta_0} &= s_e \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n - 1)s_x^2}} \\
 \text{s.e. of predicted mean at } X_o: s_{\hat{Y}} &= s_e \sqrt{\frac{1}{n} + \frac{(X_o - \bar{X})^2}{(n - 1)s_x^2}} \\
 \text{s.e. of predicted obs at } X_o: &= s_e \sqrt{1 + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{(n - 1)s_x^2}} \\
 &= \sqrt{s_e^2 + s_{\hat{Y}}^2} \\
 \text{Error sum of squares: SSE} &= \sum (Y_i - \hat{Y}_i)^2 \\
 p &= \# \text{ events} / \# \text{ tries} \\
 \text{se}(p) &= \sqrt{p(1 - p)/n} \\
 p_c &= \text{total events} / \text{total tries} \\
 \text{se of } p_1 - p_2, \text{ assuming no difference} &= \sqrt{\frac{p_c(1 - p_c)}{n_1} + \frac{p_c(1 - p_c)}{n_2}} \\
 \text{se of } p_1 - p_2, \text{ for a c.i.} &= \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} \\
 \chi^2 &= \sum_{i,j} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right] \\
 E_{ij} &= N_i * p_c
 \end{aligned}$$