November 7, 2005 NAME:_____

Stat 401 Section G - Fall 2005 - Exam 2

Please write your answers in the enclosed spaces. If you need more room, continue on the back of the page. If that is not enough, ask for extra paper. Tables are attached to the exam. The exam consists of 7 problems. The number of points for each question are listed next to the question number.

1. (3pts.) In reading a scientific article you encounter the following table:

Analysis of Variance

Source	SS	df	MS	F
Between samples	722.7	4 → 5	180.68	15.3**
Within samples	473.3	40	11.83	
Total	1196.0	(44) ⇒ n = 45		

Further reading indicates that all sample sizes are equal. Then you can deduct that the experimenter used

a. 4 samples of size 10.
b. 5 samples of size 10.
c. 4 samples of size 9.
(d.) 5 samples of size 9.
e. 4 samples of size 11.
f. None of these.

2. (4pts.) What would you say the value of the correlation coefficient is for the pair of variables: "number of cavities" and "vocabulary size" of elementary school children?

a. Approximately 0.0
(b.) Approximately 0.4
c. Approximately -0.4
d. Approximately -0.9
e. Approximately 0.9

Briefly justify your choice.

3. (3pts.) In a given group, the correlation between height measured in feet and weight measured in pounds is .68. Which of the following would alter the value of r?

(a.) height is divided by weight
b. weight is expressed in kilograms.
c. height is expressed in centimeters.

4. (20pts.) A researcher is evaluating ways to reduce erosion from freeway roadsides by adding material to the soil along the roadside. Plots on a newly build roadside were randomly assigned to one of 5 treatments listed below (7 plots per treatment). Grass was seeded into all plots in the spring. In late summer, the amount of soil erosion (grams of soil eroded) was measured on all plots. The treatments and the sample averages were:

Treatment	average erosion	contrast coefficients
1) control, nothing added	57.2	0
2) topsoil added	30.5	1
3) composted sewage sludge from Cedar Rapids	25.3	$-1/3$
4) composted sewage sludge from Dubuque	27.5	$-1/3$
5) composed sewage sludge from Des Moines	29.1	$-1/3$

These data were analyzed using the one-way analysis of variance.

(a) The p-value from the 4 d.f. F-test that compares the equal means model to the different means model is 0.011. Clearly state an appropriate conclusion from this test (phrased in the context set by the question). not all means are equal $\Rightarrow$ different road locations. lead to diff. soil erosion.

(b) One of the goals of the study is to estimate the effect of compost, compared to just adding topsoil. In the table above, please write the coefficients for the linear contrast to estimate the difference between the topsoil treatment and the average of the three sewage sludge treatments.

$\mu_2 = \frac{\mu_3+\mu_4+\mu_5}{3}$. $\quad \gamma = \mu_2 - \frac{\mu_3}{3} - \frac{\mu_4}{3} - \frac{\mu_5}{3}$.

(c) Please estimate this contrast.

$g = 30.5 - \frac{25.3+27.5+29.1}{3} = 3.2$.

(d) A second goal was to see if there were any differences between the three sewage sludge treatments. This question can be answered by comparing two models. Using words or symbols (μ's with subscripts), please describe the two models:

full model: 5 params: $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$.

reduced model: 3 param: $\mu_1, \mu_2, \mu_3 = \mu_4 = \mu_5$.

(e) List the number of degrees of freedom for each of the above models.

$n = 35$: 30 df FULL

32 df Red.

5. (20 pts.) One general belief held by observers of the business world is that taller men earn more money than shorter men. In a University of Pittsburgh study (reported in the Wall Street Journal, December 30, 1986), 250 MBA graduates, all about 30 years old, were polled and asked to report their height (in inches) and their annual income (to the nearest \$1,000). Assuming that heights of the 250 MBA graduates are recorded to the nearest inch, it turns out that this data consists of 10 groups (given by the different heights).

(a) Complete the ANOVA table for a simple linear regression model of income on height:

Source	DF	Sum of Squares	Mean Square	F Ratio	p-value
Regression	1	~~905.597~~	905.597	13.2025	0.0003
Error	248	17010.967	68.593		
Total	249	17916.564			

(b) Based on the p-value of 0.003 from the table, what would you conclude about the linear regression model of income on height?

The slope of the regr. line fitted is signif. diff. from 0.

(c) Calculate the coefficient of determination for the linear regression model described above. Based on this, what would you conclude about this model?

$$R^2 = \frac{SSR}{SSTO} = \frac{905.597}{17916.564} = 0.05 \text{ or } 5\%.$$

(d) If the residual sum of squares for the one-way analysis of variance is 17000, calculate the lack-of-fit F-statistic.

RSS Full: 17000 df = 240

RSS Red: 17010.967 df = 248

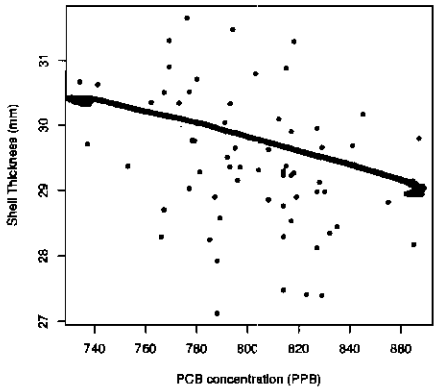
$$F = \frac{10.967/8}{17000/240} = 0.02$$

(e) Regardless of your answer in part (d), assume that the p-value associated with the test statistic computed previously is 0.92. What can you conclude about the regression model in part (a)?

No evidence of lack-of-fit.

6. (35 pts.) The pesticide DDT (dichlorodiphenyltrichloroethane) was banned in part because it led to severe thinning of eggshells of many birds. Since then, the effects of many pollutants on eggshell thickness have been measured. The following analysis came from a study on the effects of PCB, polychlorinated biphenyl, in eggs of brown pelicans on Anacapa Island. This is an area of California with a lot of offshore oil drilling platforms.

The investigators collected 65 eggs from brown pelican nests. Assume they were careful to collect only one egg from each nest. In each egg, they measured the thickness of the egg shell and the concentration of PCB in the yolk. A plot of the data is enclosed below.



The fitted linear regression line, predicting shell thickness in millimeters from the PCB concentration in parts per billion is:

$$\hat{Y} = 39.13 - 0.012\,X$$

Various diagnostic tools indicate no problems with the usual regression assumptions. Some other relevant information:

S.E. (β_0)	3.40
S.E. (β_1)	0.004
S.E.($\bar{Y}_0$ at $X_0 = 400$)	3.40
$\hat{\sigma}$	0.979

(a) Calculate the error degrees of freedom.

65 - 2 = 63.

(b) On the scatter-plot included above, sketch the estimated regression line given above. Which two points did you use to draw your line (give the two pairs of coordinates for each point).

(740, 30.25), (860, 28.81).

(c) Calculate a 95% confidence interval for the slope. The appropriate t statistic is 1.998.

$-0.012 \pm 1.998 \times 0.004 =$
$(-0.0019, -0.004)$.

(d) The p-value from the t-test of slope = 0 is 0.058. The investigators concluded, "These data show that PCB causes thinning of brown pelican eggs." Is this conclusion appropriate?

No → OBSERV. STUDY - CAN'T EST. CAUS.

(e) The investigators were careful to collect only one egg from each nest. Which assumption of linear regression might be a concern if they collected more than one egg from a nest?

IND.

(f) Use the regression line to predict the average thickness of eggs with 400 parts per billion PCB.

$39.13 - 0.012 \times 400 = 34.33$

(g) Calculate a 95% prediction interval for the thickness of eggs with 400 parts per billion PCB. The appropriate t statistic is 1.998. All the other quantities you need can be calculated from your answer to question (6f) and the other relevant information given by the text of the problem.

$34.33 \pm 1.998 \times 0.979 \sqrt{13.0409}$

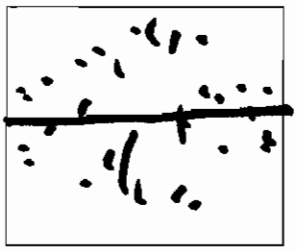
$= (27.26, 41.40)$.

(h) Use the regression line to predict the average thickness of eggs with 1350 parts per billion PCB. Does this prediction require making any additional assumptions? If so, what?

22.93.

(i) **Sketch** the residual plot in the space provided below. You don't need to calculate all the residuals, just make an approximate sketch of what the residual plot looks like here. Don't forget to label the axes (this means writing explicitly what the variables represented by each of the axis and the range of values for each axis).

Resid.



0

$\hat{y}$ (Predicted egg thick).

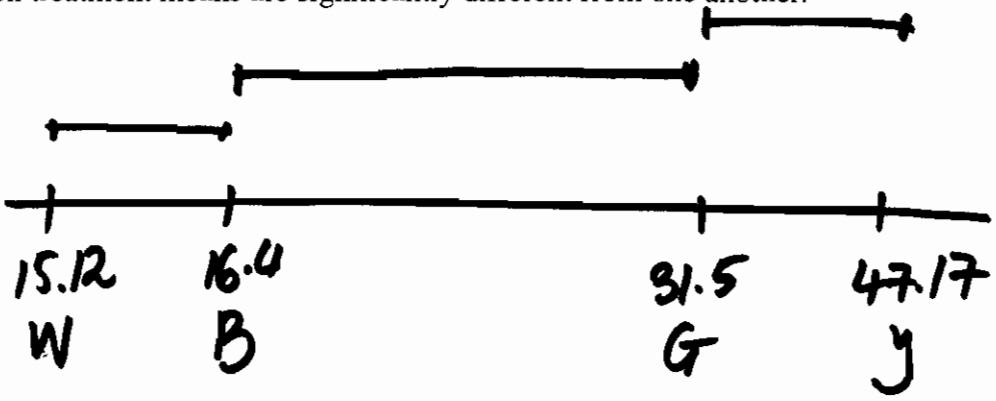
7. (15 pts.) The presence of harmful insects in farm fields is detected by erecting boards covered with a sticky material and then examining the insects trapped on the board. Some colors are more attractive than others. In an experiment aimed at determining the best color for attracting cereal leaf beetles, six boards in each of four colors were placed in a field of oats in July. Data is summarized in the following table (X is the average number of beetles found on a sticky board):

15.12 16.4 31.5 47.17

Color	$\bar{X}$	Var(X)
Lemon yellow	47.17	46.17
White	15.12	11.07
Green	31.5	98.3
Blue	16.4	17.3

1.28 16.38 32.05
15.1 30.77
15.67

(a) Suppose we want to test for differences between all pairs of treatment means while keeping the probability of one or more mistaken rejections of a true null hypothesis below 0.05. The minimum significant difference associated with the Tukey method for all pairwise comparisons among four means is about 16.22 for this data (i.e., HSD=16.22). Use this information to construct a line plot that illustrates which treatment means are significantly different from one another.



(b) Based on the line plot constructed in part (a) state whether each of the following statements is TRUE or FALSE. We can be at least 95% confident that...

i. F The lemon yellow boards catches the highest number of beetles.

ii. F The white boards are the least attractive of cereal beetles.

iii. T The blue boards may attract the lowest number of beetles.

iv. T The green and the lemon yellow boards attract a higher number of beetles than the white boards.

(c) Compute the least significant difference using the Bonferroni correction for question (a). If you don't find the exact confidence level in the t-table, just pick the closest one there. Compare the Bonferroni adjusted LSD to the HSD given in part (a). Which one is larger? Was this to be expected? Why?

6 CI. t_{20}^{G} $1 - \frac{.95}{2.6}$ $LSD = t \times S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

(d) Draw a line plot based on the Bonferroni correction.