Please put your name on the back of the last page

We grade anonymously. That means we don't want to see your name until we have graded your answers. Put your name only on the back of the last page.

There are 3 questions, each with multiple parts. Each part is worth 3 or 5 points. The total is 100 points. The third question requires a packet of computer output (your choice of JMP, R, or SAS, but only one).

Please write your answers in the enclosed spaces. Continue on the back of the page if necessary. If you do continue on the back, clearly label the appropriate question / part for each extended answer.

1. 36 points. Before the Salk polio vaccine was introduced, a large randomized experiment was conducted to evaluate its effectiveness. The experimental group was 400,000 elementary school students, whose parents consented to their children being in the study. These were randomly assigned to be given either the Salk vaccine or a placebo (control with no active ingredient). Infantile paralysis is the disease caused by the polio virus. A year after injection, the number of students in each group who developed infantile paralysis was counted. Incidence of polio is the probability of developing infantile paralysis. The data for the experimental groups are:

	Infantile Paralysis		
Treatment	Yes	No	Total
Polio vaccine	56	199,944	200,000
Placebo	142	199,858	200,000
Total	198	399,802	400,000

   (a) 3 pts. The data from this study is best described as (circle your choice):

   prospective binomial | retrospective binomial | multinomial | something else

   (b) 3 pts. Briefly explain your choice.

   (c) 5 pts. Consider the group of individuals getting the polio vaccine. Estimate the incidence of polio and its standard error for that group.

(d) 5 pts. Calculate a 95% confidence interval for the incidence of polio **in the group that received the polio vaccine**.

Some quantiles are:

Distribution	0.90	0.95	0.975
T, 3 df	1.638	2.353	3.182
T, 55 df	1.297	1.673	2.004
Z	1.282	1.645	1.960

Show your work.

(e) 5 pts. Consider the null hypothesis that the incidence of polio is the same in the vaccinated and placebo groups. Compute the expected number of cases of infantile paralysis in the vaccinated group. Show your work

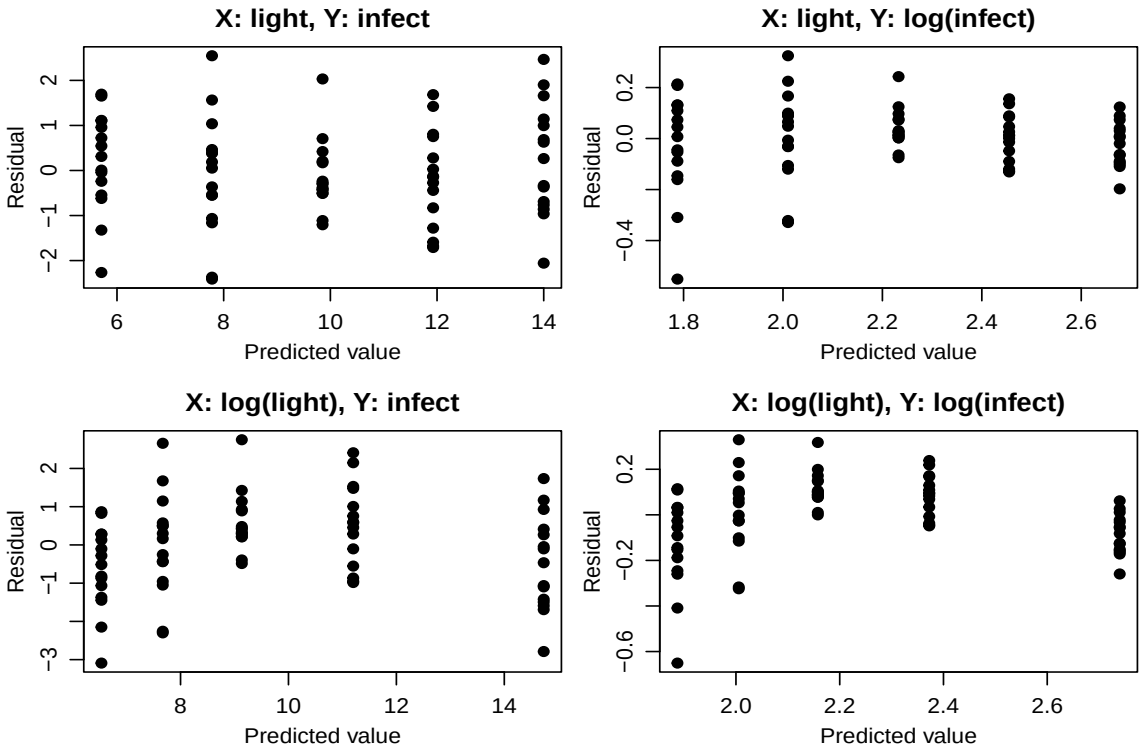
(f) 5 pts. Estimate the odds ratio for the incidence of polio in the vaccinated and placebo treatments. Express the odds ratio as the odds of polio in the placebo group divided by that in the vaccinated group.

(g) 5 pts. Test the null hypothesis that the population value for the odds ratio in part 1f equals 1. Calculate and report the test statistic for this test. Do not use a Chi-square test.
Note: The test statistic could be a T statistic, a Z statistic, an F statistic, or something else.

(h) 5 pts. This study was one of the very first randomized studies of a medical intervention. Many in the medical community felt it was unethical to give a placebo to a patient. Hence, data were collected on a third group of students who were not given anything. These are "natural controls" and were in different schools from the students in the vaccinated or placebo groups. There were 500,000 natural controls. Your friend suggests that the evaluation of the effectiveness of the vaccine should compare the vaccinated group to the natural controls because the

number of natural controls (500,000) is so much larger than the number of students in the placebo group (200,000). Is this suggested comparison a good idea? Explain why or why not.

2. 31 pts. A study of southern corn leaf blight infection looked at the relationship between light levels and the amount of leaf infected. 75 plants were randomly assigned to one of 5 light levels, from 0.4 to 2.0 units of light (15 plants per light level). Each plant was inoculated with the blight, and after an appropriate length of time, the percent of leaf area infected was measured on each plant.



(a) 3 pts. Above are residual vs predicted value plots for four regression models, one for each combination of X = light or X = log(light) and Y = infect or Y = log(infect). The combination is described in the title above each plot. What are the X and Y variables in the most reasonable model? Briefly explain your choice.

No matter how you answered part 2a, all subsequent parts are based on the fitted regression equation: %leaf area infected = 15.2 - 5.25*light. This may or may not be the most appropriate regression model.

Some potentially useful information is:
standard deviation = $\sqrt{\text{MSE}} = s = \hat{\sigma} = 0.91\%$.
Standard error of the predicted mean % leaf area infected at 2.8 units of light = $s_{\hat{Y}_o} = 0.37\%$.

(b) 5 pts. Please provide one sentence each that interprets (or explains) the estimated intercept and slope.
intercept:

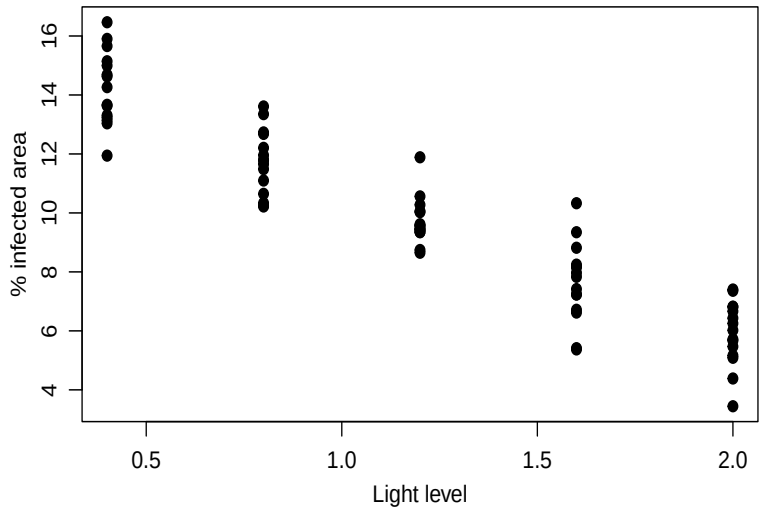
slope:

(c) 5 pts. The list of assumptions for a regression analysis includes "no lack of fit"; the list of assumptions for an ANOVA analysis does not. Briefly explain why a regression needs to assume "no lack of fit" but an ANOVA does not.

(d) 6 pts. You want to test whether a straight line is an appropriate summary of the relationship between light and % infection. Here is information about 3 models fit to these data:
Error SS for intercept only (equal means) model: 85.47
Error SS for regression model: 60.45
Error SS for ANOVA (different means) model: 59.28
In the table below, fill in the appropriate information for the ANOVA lack-of-fit test. You do not need to calculate any quantities not in this table.

Source	d.f.	SS	MS
Difference	_____	_____	_____
Full model	_____	_____	_____
Reduced model	_____	_____	

The information on the top of page 4 and this plot of the data may help answer the last few questions.



(e) 3 pts. The researcher wants to make conclusions about individual plants grown in 2.8 units of light. Predict the % infected leaf area for a plant grown in 2.8 units of light.

(f) 3 pts. Do you have any concerns about using the regression line to make this prediction? Why or why not?

(g) 3 pts. The regression model will be used to predict infection for **individual** plants. The standard error of a predicted % infection for an **individual** plant grown at 2.8 units of light is closest to (circle your choice):

0.28%, 0.37%, 0.80%, 0.90%, 0.98%

(h) 3 pts. Briefly explain your choice in question 2g.

3. 33 pts. Consider a randomized experiment to compare the quality of pepperoni that has been sterilized in different ways. One important quality measure is the hardness of the pepperoni stick. 32 pepperoni sticks (8 per treatment) were prepared from one batch of ingredients. The four treatments (listed below) were chosen to answer specific questions about the effects of radiation and heat on hardness of pepperoni. In the computer packets, the treatments are identified by the Code variable in the table.

Code	Treatment	Radiation dose	Heat level	average hardness
a	control, no radiation	0 kGray	usual	14.1
b	low radiation dose	0.5 kGray	usual	12.5
c	high radiation dose	2.5 kGray	usual	18.3
d	heat, no radiation	0 kGray	high	22.9

The computer packets contain:

	Page number in		
	JMP	R	SAS
first 6 observations	1	1	1
summaries of each group	1	1	1
untransformed values	2	2, 3	2, 3
log transformed values	3	4, 5	4, 5

(a) 1 pt. Which packet are you using?

(b) 5 pts. The packet has two sets of analyses, one for Y = hardness and the 2nd for Y = log(hardness). Which analysis is more appropriate? Briefly explain your choice.

For all subsequent parts, use results from the analysis you chose in question 3b.

(b) 5 pts. Test the null hypothesis that all treatments have the same means. Report the p-value and write a one sentence conclusion.

(c) 6 pts. The investigators are interested in three specific contrasts:

1. the difference between the control and the heat treatment

2. the difference between the heat treatment and the average of the other three treatments

3. the linear trend contrast using the control, low dose, and high dose treatments (codes a, b, and c). The radiation dose, in kGray, for each treatment is given in the table at the start of the problem.

Fill in the table below with the appropriate coefficients for each of these three contrasts.

Treatment	coefficients for contrast:		
	# 1	#2	#3
Control			
Low dose			
High dose			
Heat			

(d) 3 pts. Consider the standard error for contrast # 1 and the standard error for contrast # 2 from part 3c. Will these two contrasts (# 1 or #2) have the same standard error?
If not, which will have the smaller standard error? Briefly explain your answer

(e) 5 pts. The computer packet has results for 3 contrasts. One of them is the comparison between the control and heat treatments. Fill in the estimate and confidence interval for the conclusion appropriate for your analysis. Note: **Complete one of these conclusions, not both.**

Heat increases the mean hardness by _____ units (95% CI: _____, _____).

Heat multiplies the mean hardness by _____ units (95% CI: _____, _____).

(f) 5 pts. The investigators notice that there is a large difference between the means of the heat and low radiation treatments. The investigators decide to test whether this difference = 0 even

though the comparison was not one of the original questions in the study. The investigators ask for your advice. Do you recommend:

1. Don't do the test because it wasn't one of the original questions

2. Do the test using a traditional T test

3. Do the test using a multiple comparisons adjustment. If so, which adjustment is the most appropriate?

Circle your choice and provide a brief explanation.

(g) 3 pts. Rightly or wrongly, the investigators decide to use a Tukey adjustment to assess the difference between the heat and high radiation treatments. The p-value for the traditional T-test is 0.0006. Will the p-value using a Tukey adjustment for 6 tests be:

1. something like 0.0001

2. very similar to 0.0006

3. something like 0.0030?

Circle your choice and provide a brief explanation

That's all! Enjoy the rest of the afternoon.