Final Exam Study Guide Questions for Post Exam 3 Material. Key

Question 1

In the same survey students also reported their ACT scores. Here's the multiple regression equation predicting the percent tuition parents pay from ACT scores and # of drinks for the **240** students who responded to the survey:

$$\hat{\text{Tuition}} = -1 + 2.7(\text{ACT}) - 0.5(\text{Drinks})$$

3 parameters n = 240

a) To test the overall regression effect, H_0:R=0 against H_A:R≠0 fill in the missing blanks in the ANOVA table.

Source	SS (Round to nearest whole number)	df	MS (Round to 2 decimal places)	(Round to 2 decimal places)
Model	SSM= 28,393	2 p-1	MSM= 14,196.5 28,393 / 2	F= 15.52 14,196.5 / 914.79
Error	SSE=	237 n-p	MSE = 914.79 216,805 / 237	SD^+_{errors}=
Total	SST= 245,198	n-1	Nothing goes in this box.	R^2= 0.116 SSM / SST = 28,393 / 245,198

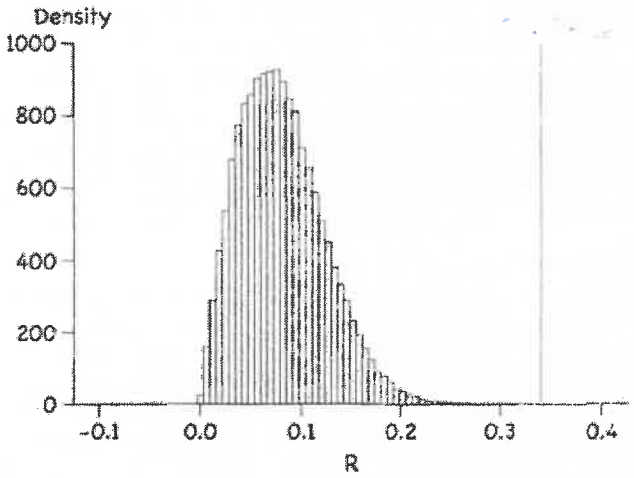
b) When the null is true you'd expect the F stat to be about ___1___.

c) Comparing the F-stat you got in (a) to what you'd expect under the null (or by looking on the F-table) you can estimate the p-value to be **i)** < 1% **ii)** 1% to 5% **iii)** 5% to 10% **iv)** > 10%

d) Another way to compute the p-value is by the re-randomization test. The histogram on the right shows the randomization test results of 50,000 randomizations showing the distribution of R's.

What does the vertical line mark?

- **i.** the specified significance level α
- **ii.** the randomized R's that land at p-value = 0.5 %
- **iii.** the value of our sample R.



e) The p-value given by the randomization test is closest to ….

i) 0
ii) 1%
iii) 5%
iv) not enough info

Final Exam Study Guide Questions for Post Exam 3 Material.

Question 2 Suppose the Final Exam, quiz and lab scores of **33** randomly selected students from a large Physics class of 2000 students yielded the following multiple regression equation with **R= 0.5** $\hat{\text{Final}}$ **= 50 + 0.3(Quiz) + 0.1(Lab)**

a) The above equation describes the best fitting _____ through all the points so as to minimize the squared errors in the _____.

Circle one for the 1st blank: **a)** line **b)** plane **c)** ellipsoid **d)** cube

Circle one for the 2nd blank: **a)** Final Exam Scores **b)** Quiz scores **c)** Lab scores

b) The multiple regression equation for all students with Labs = 70 simplifies to $\hat{F}$ = 57 + 0.3 **(Quiz)**

50 + 0.3 (Quiz) + 0.1(70)

c) The multiple regression equation predicts Doug to have a Final score of 70. If Alex scored 10 points higher than Doug on the quizzes and 20 points higher on the labs what's Alex's predicted Final score? 75

70 + .3(10) + 0.1(20)

d) Do the F test for the overall regression effect for the model: $\hat{\text{Final}}$ **= 50 + 0.3(Quiz) + 0.3(Lab) R= 0. 5** and **n=33**

R^2 = 0.25 **and 1- R^2=** 0.75

F-test

Compute the F statistic. ~~Show work. No work, no credit.~~

$\frac{0.25}{0.75} \cdot \frac{30}{2} = 5$ $\quad \frac{R^2}{1-R^2} \cdot \frac{n-p}{p-1}$

Look at the F table. What is F*(the critical value of F) at $\alpha = 0.01$? F* = 5.3903

Our F test stat < F* so our p-value is > than 1%.

Fill in the two blanks with either <, >, or =

e) Suppose you decided to reject the null at $\alpha = 0.05$, you'd conclude that … Choose one:

i) Both slopes must be significant. ii)The Quiz slope must be significant. iii) The Lab slope must be significant.

iv) The intercept must be significant. v) Either the Quiz or the Lab slope or both must be significant.

f) To see which slope is significant in the multiple regression equation $\hat{\text{Final}}$ **= 50 + 0.3(Quiz) + 0.1(Lab)** the computer ran a Z test and a t-test. Which table shows the t-test? *Circle one:*

i) Table A **ii)** Table B **iii)** Not enough information to determine

Table A

	Slope	SE	t or Z	p
Quiz	0.3	0.1	3	0.185%
Lab	0.1	0.08	1.25	10.565%

bc p is >

Table B

	Slope	SE	t or Z	p
Quiz	0.3	0.1049	2.86	0.365%
Lab	0.1	0.0839	1.192	12.09%

g) How many degrees of freedom for the t-test? 30 n-p

h) Another variable is added to the model. Will R^2 go up or down?

i) R^2 will go up or stay the same, it can't go down.

ii) R^2 will go down or stay the same, it can't go up.

iii) R^2 could go up, down or stay the same depending on the variable.

i) Let's say a 3rd variable that's correlated with the Final is added to the multiple regression model and the Quiz and Lab slopes stay the same. You can conclude the 3rd variable must be

i) correlated with either Quizzes or Labs.

ii) correlated with both Quizzes and Labs.

iii) uncorrelated with both Quizzes and Labs.

iv) negatively correlated with Quizzes and positively correlated with Labs (or vice versa) so their effects cancel out.

j) How was the multiple correlation, **R= 0. 5** calculated?

i. It's the correlation between the Final scores, quiz scores and lab scores after they've been converted to Z scores.

ii) It's the correlation between students' actual Final scores and their predicted ones from the multiple regression equation.

Final Exam Study Guide Questions for Post Exam 3 Material.

Question 3
For each of the following is it appropriate to use logistic regression? **Circle Yes or No.**

a) Predicting income based on years of college. YES NO
b) Predicting ln(income) based on years of college YES NO
c) Predicting graduating college based on family income. YES NO
d) Predicting getting a scholarship based on gender and ethnicity. YES NO
e) Predicting favorite color based on gender YES NO

Y values must be 0-1

Question 4 Circle True or False for each statement below.

a) The logistic regression model only handles X values that can be coded as 1's and 0's. i)True ii)False
b) Transforming non-linear scatter plots into linear ones by converting Y to ln(Y) is called logistic regression. i)True ii)False — Y variable must be ln(odds) not ln Y
c) The assumptions needed to make inferences for linear and logistic regression are the same i)True ii)False — No, we don't assume normal dist. of error.

Question 5
How are the parameters chosen in logistic regression and linear regression?
Fill in the first blank below with "logistic" or "linear" and the second blank with "minimize" or "maximize".

a) In linear regression, the parameters are chosen to minimize the sum of the squared errors

b) In logistic regression, the parameters are chosen to maximize the likelihood of getting our sample data.

Question 6 Are F and t tests ever appropriate to test significance in Logistic regression models?
Choose one:

a) Yes, when the sample size is small the F and t tests give more accurate results.
b) No, because F and t tests can never be done on variables that have undergone log transformations.
c) No, because F and t tests are never done when we are predicting counts (when Y is binary), since the SD can be estimated directly from the count.

Question 7 Part I On our survey, **178** students anonymously answered these 2 questions:
"Would you volunteer to be randomly assigned to either the online or in person section?"(No = 0, Yes =1)
"Which section are you in?" (L1=0, online=1)

To predict the probability of volunteering from section, we fit a logistic regression model. Here's the ln(odds) form of the

regression equation: $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.5261 + -0.7267(\text{Section})$

a) Are online students more or less likely to volunteer? ***Choose one:*** **i)** More **ii)** Less **iii)** Same **iv)** Not enough info

b) What is the probability that an L1 student would volunteer? p= 0.37
$\ln(\text{odds}) = -0.5261 \rightarrow \text{odds} = e^{-0.5261} = 0.59 \Rightarrow p = \frac{0.59}{1.59} = 0.37$

c) What is the probability that an online student would volunteer? p= 0.22
$\ln(\text{odds}) = -0.5261 - 0.7267 = -1.2528 \Rightarrow \text{odds} = e^{-1.2528} = 0.29 \Rightarrow p = 0.29/1.29 = 0.22$

d) The Odds Ratio = _____.
$e^{-0.7267} = 0.48$

e) If we switched the coding for section to online = 0 and L1 =1 what would change? ***Choose one:***
i) Odds **ii)** Probabilities **iii)** Odds Ratio **iv)** All **v)** None

OR would switch from $\frac{\text{ONL odds}}{\text{L1 odds}}$ to $\frac{\text{L1 odds}}{\text{ONL odds}}$

See chart on next page to understand.

Final Exam Study Guide Questions for Post Exam 3 Material.

f) Look at the table showing the 178 responses to the 2 questions.

	No	Yes	Total
L1	44	26	70
Online	84	24	108
Total	128	50	178

Use the table to compute the odds for an L1 and online student volunteering. ***Please leave your answers in fraction form.***

i) Odds for L1 = $\frac{26}{44}$

ii) (Odds for Online = $\frac{24}{84}$

iii) Should you get the same OR as in (d) above? (Assuming you compute the ratio of Online odds to L1 odds.) bc ONL=1 L1=0

a) Yes, within rounding error **b)** No $OR = \frac{24/84}{26/44} \approx e^{-0.7267}$

Question 7 Part II A third question on the same survey was: "How many people have you been in a serious relationship with?" Adding relationships to the model gives us: $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.33 + -1.03(\text{Section}) + 0.64(\text{Relationships})$

a) The χ^2 test for the overall regression effect: H_0: All β's =0 yielded a χ^2 **stat = 26.**
How many **degrees of freedom?** = 3-1 = (2)

b) The p value < 0.1%. This means that the probability that …***Choose only one:***
i) the null is true < 0.1% **ii)** the null is false > 99.9% **iii)** we'd get a χ^2 stat ≥ 26 if the null was true < 0.1%

c) The relationship slope has a SE = 0.14. To test H_0: $\beta_{relationship} = 0$ against H_A: $\beta_{relationship} \neq 0$ compute the Z stat.

$Z = \frac{\text{obs slope} - \text{exp slope}}{SE_{slope}} = \frac{0.64}{0.14}$ Z = 4.57

d) Since p < 5%, a 95% Confidence interval for the Relationship slope does NOT include 0.
Fill in the first blank with > or < , the second with "does" or "does not", and the third blank with a number.

e) The OR for Relationship = $e^{0.64}$ and the OR for Section= $e^{-1.03}$

f) Comparing two people in the same section, the person with 2 more relationships has 3.6 times the odds of volunteering. ***Fill in the blank with a number.***

$e^{0.64} \cdot e^{0.64} = 3.6$

g) Comparing an L1 student with 4 relationships to an online student with 2 relationships, the L1 student has 10.07 times the odds of volunteering. ***Fill in the blank with a number.***

$e^{0.64} \cdot e^{0.64} \cdot e^{1.03} = 10.07$

h) What's the probability that an L1 student with 10 relationships will volunteer? 0.99

$\ln(\text{odds}) = -1.33 + 6.4 = 5.07 \Rightarrow \text{odds} = e^{5.07} \Rightarrow p = \frac{e^{5.07}}{1+e^{5.07}} = 0.99$

i) Would the ln(odds) equation for Part II change if we reversed the coding for Section so that L1=1 and online=0 and kept everything else the same? If so, write the new equation in the blank provided.

a) No, it would not change. **b)** Yes, it would change to $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) =$ -2.36 +1.03(S) +0.64(R)

Don't have to find intercept, except for extra credit.
For L1=0, original equation gives: ln(odds) = -1.33 + 0.64(R)
So to get same ln(odds) when L1=1, intercept = -2.36.

Final Exam Study Guide Questions for Post Exam 3 Material.

Question 8 A predictor of whether esophageal cancer has not metastasized to the lymph nodes is the diameter of the tumor. Below is the log odds regression equation predicting the probability of no metastasis from the diameter of the tumor (measured in cm) from a hypothetical study of 200 patients.

ln (p/(1-p)) = 2 - 0.5 (Diameter)

a) Use the equation to estimate the **odds** and **probability** of no metastasis for a tumor of diameter = 8 cm. ***Show work.***

i) Odds= 0.14 *ii)* Probability = 0.12

$\ln(odds) = 2 - 0.5(8) = -2$ $odds = e^{-2}$ $\frac{e^{-2}}{1+e^{-2}}$

b) How do the estimated *odds* of no metastasis change if the tumor increases in diameter by 1 cm ?
i) odds are multiplied by 0.61 **ii)** the odds decrease by 0.5 **iii)** not enough info

$e^{-0.5} = 0.61$

c) How does the estimated *probability* of no metastasis change if the tumor increases in diameter by 1 cm?
i) the probability is multiplied by 0.61 **ii)** the probability decreases by 0.5 **iii)** not enough info

(Need to know odds, not just OR)

d) How big a tumor would give a 50% probability of metastasis? 4 cm

$p = 0.5 \Rightarrow odds = 1 \Rightarrow \ln(odds) = 0$ $0 = 2 - 0.5D \Rightarrow D = 4$

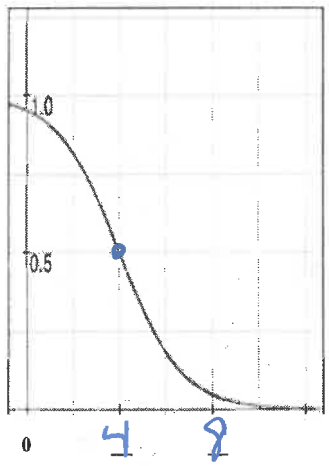
e) How big a tumor would give a 40% probability of no metastasis? 4.81 cm

$p = 0.4 \Rightarrow odds = \frac{0.4}{0.6} \Rightarrow \ln\left(\frac{0.4}{0.6}\right) = -0.405$

$-0.405 = 2 - 0.5D$ $D = 2.405/0.5 = 4.81$

f) Below is a graph of the probability form of the model.

Write its equation: p = $\frac{e^{2-0.5(D)}}{1+e^{2-0.5(D)}}$ and fill in the 2 blanks on the X-axis with the correct diameter values (in cm).



Fill in the 2 blanks above with the correct numbers.

Final Exam Study Guide Questions for Post Exam 3 Material.

Question 9 pertains to the Wilcoxon Mann Whitney test

A randomized double-blind test was done to test the effectiveness of a drug to cure warts. The subjects were 8 people with lots of warts. 4 subjects took the drug and 4 took the placebo. The number of warts that disappeared for each of the 8 subjects is recorded below.

Drug Group: 0, 10, 11, 40 Placebo group: 5, 6, 8, 9

1 6 7 8 2 3 4 5

Part 1

Fill out the chart below. Show work for how you got the observed rank sum for each group.
No partial credit since you should know what the totals should be and you can check your work.

	Observed Rank Sum	Expected Rank Sum	Observed - Expected
Drug Group	22	18	4
Placebo Group	14	18	−4
Total should be....	$\frac{8 \cdot 9}{2} = 36$	36	0

Question 9 Part II

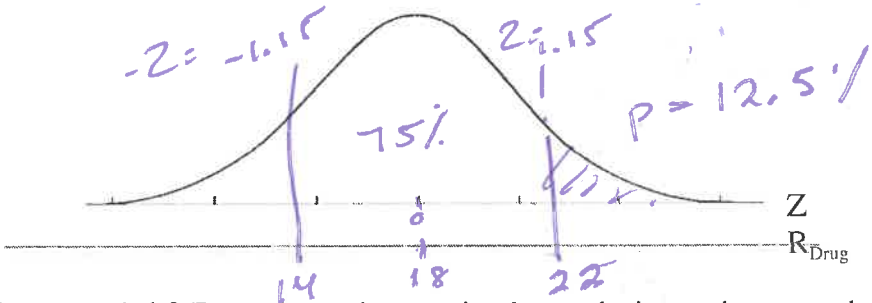
The sample sizes in Part I are too small to use the Normal Approximation but let's just assume for the purpose of this exam that you can use the Normal Approximation anyway.

H_0 : The drug works no better than the placebo in the population
H_A : The drug *does work better* than the placebo in the population for some segments of the population.

a) Compute the Z stat for the drug group. Use $SE_R = \sqrt{\frac{n_1 n_2 (N+1)}{12}} = \sqrt{\frac{4 \cdot 4 \cdot 9}{12}} = \sqrt{12} = 3.464$

$Z = \frac{22 - 18}{3.464} = 1.15$

b) Label the Observed and Expected Value for both the Z and R_{Drug} axes below. Calculate the p-value and shade the appropriate area. (R_{Drug} is the Rank Sum for the Drug group.)



c) What do you conclude? (Remember, we're assuming the sample size was large enough so the normal approximation is valid).

i) Reject the null, we're sure the drug works.
ii) Reject the null, we have strong evidence the drug works.
iii) Cannot reject the null, it's plausible the drug works no better than a placebo.
iv) There's over a 95% chance the drug didn't work.

Final Exam Study Guide Questions for Post Exam 3 Material.

Question 9 cont.
Drug Group: 0, 10, 11, 40 Placebo group: 5, 6, 8, 9

0 4 4 4 1 1 1 1

d) What's the U statistic for the Drug Group? For the Placebo group?

U_{drug} = 12 $U_{placebo}$ = 4

e) The sum of the 2 group U statistics must = 16 for any 2 groups with 4 members each.
(Check that your $U_{drug} + U_{placebo}$ is correct.) 4 × 4 = 16

12 + 4 = 16 ✓

f) Would you get the same Z stat and p-value using U_{drug} as you did using R_{drug} in part (a)?

i) Yes, exactly the same.
ii) Exactly the same values but the Z-scores would be opposite signs.
iii) No, the p-value would be smaller using U.
iv) No, the p-value would be larger using U.

Question 10 pertains to the Kruskal Wallis test *(6 pts)*
There are 3 forms of this Final. Suppose at the grading meeting I randomly select 9 Finals and grade them with these results:
Form A: 80, 81, 82 Form B: 83, 84, 85 Form C: 86,87,89

1, 2, 3 4, 5, 6 7, 8, 9

Null Hypothesis: No difference in difficulty of the exams in the population. We just happen to observe differences in our sample due to chance variation.
Alternative Hypothesis: At least one of the exams is of different difficulty in the population.

a) The Rank Sum for Form A= 6, Form B= 15 and Form C= 24

b) The total Rank Sum for any set of 9 numbers is always= 45. (give a number.) $\frac{N(N+1)}{2} = \frac{9 \cdot 10}{2}$

a) The H-stat = 7.2 Would any other arrangement of 9 numbers into 3 groups of 3 yield a higher H-stat?

i) No ii) Yes iii) Not enough info

b) For large enough samples we can best approximate the distribution of the H stat with
i) Z stat ii) t stat iii) Chi-square stat iv) the F stat

Final Exam Study Guide Questions for Post Exam 3 Material.

Question 11

a) If we decide to do a non-parametric test and use the Spearman correlation coefficient to test the null hypothesis that the population correlation is 0 then the appropriate test-statistic for small samples (<7) is …

i) a t-statistic
ii) Spearman correlation tables that calculate the exact probability distribution
iii) 2 sample t-statistic
iv) an F-test
v) a Chi Square test

b) For large enough samples the appropriate test statistic is

i) Z-test
ii) t-test
iii) either
iv) F-test
v) none of the above

Question 12

Look at the 3 data sets below:

Data Set 1: (1,2), (2, 4) , (3, 6) , (4,8)
Data Set 2: (-1, 5) (-2, 4) (-3, 3)
Data Set 3: (1,1) (8,9) (103,10)

In ranks $r_s = 1$
(1,1) (2,2) (3,3)
(1,1) (2,2) (3,3
(1,1) (2,2) (3,3

For which data set(s) is $r \neq r_s$?

But $r \neq 1$ for Data Set 3
Easiest to see by graphing