CSCI 3022
Final Exam
Fall 2020

Write **clearly** and **in the box**:

Name:
Student ID:
Section number:

Read the following:

- **RIGHT NOW**! Write your name, student ID and section number on the top of your exam. If you're handwriting your exam, include this information at the top of the first page!
- You may use the textbook, your notes, lecture materials, and Piazza as recourses. Piazza posts should not be about exact exam questions, but you may ask for technical clarifications and ask for help on review/past exam questions that might help you. You may not use external sources from the internet or collaborate with your peers.
- You may use a calculator, Python, or other computational device.
- If you print a copy of the exam, clearly mark answers to multiple choice questions in the provided answer box. If you type or hand-write your exam answers, write each problem on their own line, clearly indicating both the problem number and answer letter.
- Mark only one answer for multiple choice questions. If you think two answers are correct, mark the answer that **best** answers the question. No justification is required for multiple choice questions. For handwriting multiple choice answers, clearly mark both the number of the problem and your answer for each and every problem.
- For free response questions you must clearly justify all conclusions to receive full credit. A correct answer with no supporting work will receive no credit.
- The Exam is due to Gradescope by midnight on Sunday, December 6.
- When submitting your exam to Gradescope, use their submission tool to mark on which pages you answered specific questions. Submitting your exam properly is worth 1/100 points. The other problems sum to 99.

Multiple choice problems: Write your answers in the boxes if using a printed version of the exam.

1. (3 points) It is a well-known fact that 50% of the general population are rascals ($P(R) = 0.5$) and 50% of the general population are not rascals ($P(R^C) = 0.5$). Furthermore, scientists have determined that about 90% of rascals wear a top hat at all times, while only 30% of non-rascals wear top hats. Suppose that you meet a person who is wearing a top hat. Given the information that they are wearing a top hat, what is the probability that they are a rascal?

   A. 0.09

   B. 0.45

   C. 0.6

   D. 0.75

   E. 0.9

2. (3 points) Suppose that the random variable X has mean 2 and standard deviation 4. Let Y be the random variable given by $Y = X^2 + 1$. What is the expected value of Y?

   A. 3

   B. 5

   C. 12

   D. 21

   E. 25

3. (3 points) Suppose you compute a sample mean for a population that is normally distributed with known variance σ^2. Which combination of significance level and sample size n produces the *narrowest* confidence interval for the mean?

   A. $\alpha = 0.12$ and $n = 100$

   B. $\alpha = 0.12$ and $n = 25$

   C. $\alpha = 0.04$ and $n = 100$

   D. $\alpha = 0.04$ and $n = 25$

   E. $\alpha = 0.01$ and $n = 100$

   F. $\alpha = 0.01$ and $n = 25$

4. (3 points) Let X be normally distributed with mean of 1 and variance of 9. Which of the following represents $P(-2 < X < 1)$, using `scipy.stats` syntax?

   A. 0

   B. `stats.norm.cdf(1)` − `stats.norm.cdf(-2)`

   C. `stats.norm.ppf(1)` − `stats.norm.ppf(-2)`

   D. .5 − `stats.norm.cdf(-1)`

   E. `stats.norm.cdf(-1)` − `stats.norm.cdf(0)`

   F. 1

5. (3 points) Which of the following statements is **True**?

   A. You cannot make a Type I error when the null hypothesis is true.

   B. You cannot make a Type II error when the null hypothesis is true.

   C. The test that minimizes Type I error rate α will also be the one that minimizes the Type II error rate β.

   D. The p-value is the probability that the null hypothesis is true.

6. (3 points) Consider performing a multiple linear regression on a data-set with full and reduced models of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ and $y = \beta_0 + \beta_1 x_1 + \beta_4 x_4$, respectively. Suppose that you perform a partial F test and fail to reject the null hypothesis. What is the strongest conclusion you can draw?

   A. Nothing.

   B. $\beta_k \neq 0$ for *some* $k \in \{1, 2, 3, 4\}$.

   C. $\beta_k = 0$ for *all* $k \in \{1, 2, 3, 4\}$.

   D. $\beta_k \neq 0$ for *all* $k \in \{1, 2, 3, 4\}$.

   E. $\beta_1 = \beta_4 = 0$.

   F. $\beta_2 = \beta_3 = 0$.

7. (3 points) Suppose you generate 8,000 confidence intervals for the mean of a population, using fixed significance level α. You discover that 782 of them FAIL to cover the true mean. Which of the following is the most appropriate estimate of the significance level α?

   A. 0.01

   B. 0.025

   C. 0.05

   D. 0.1

   E. 0.2

   F. 0.782

8. (3 points) In the context of a simple linear regression predicting Y based on the single feature X, if the true slope coefficient satisfies $\beta_1 = 0$, then which of the following can you **always** conclude:

   A. We will not reject the null hypothesis of $\beta_1 = 0$ using the F-test for our linear model.

   B. The sum of squared regression (SSR) will exceed the sum of squared errors (SSE).

   C. The sum of squared errors (SSE) will exceed the sum of squared error (SSR).

   D. X an Y are independent.

   E. There is not in reality a linear relationship between X and Y

9. (3 points) You're performing a simple linear regression, and someone spills ink all over your beautiful regression table. As a result, you can only read the following, though you also do recall that the data set had 107 observations:

Coefficient	Estimate	Std. error	t-value	Pval
(Intercept)	**MISSING**	1.12566	4.436	**MISSING**
Slope	0.73111	0.02455	**MISSING**	$<$ 2e-16

What is the correct (exact) **MISSING** value for the *intercept*'s p-value?

A. `stats.t.ppf(4.436, df=107)`

B. 2(1-`stats.t.cdf(4.436, df=105)`)

C. $1.12566 \cdot 4.436$

D. 2(1-`stats.t.cdf(4.436, df=107)`)

E. (1-`stats.norm.cdf(4.436)`)

10. (3 points) From the same table as the prior question, what is the (exact) **MISSING** value for the t-value of the *slope*?

A. 0.73111/0.02455

B. `stats.t.ppf(0.73111)- stats.t.ppf(0.02455)`

C. `stats.t.ppf(0.73111, df=105)- stats.t.ppf(0.02455, df=105)`

D. `stats.norm.ppf(2e-16)`

E. $0.73111 \cdot 0.02455$

F. $0.73111 \cdot 0.02455/(2e-16)$

11. (3 points) Suppose that you are performing a binary classification to assign a class label $y \in \{0, 1\}$ to each data point and you model the probability that the data point x belong to class 1 by the logistic regression model

$$p(y = 1|x) = \text{sigm}(\hat{\beta}_0 + \hat{\beta}_1 x)$$

where $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = 1$. How would your model a data point with $x = -1.5$?

A. inconclusive

B. $\hat{y} = 0$

C. $\hat{y} = .5$

D. $\hat{y} = 1$

E. The limit does not exist.

12. (3 points) For the same logistic regression model given in the previous problem, what happens if x increases by one unit?

A. the odds that $y = 1$ increase by a factor of e

B. the odds that $y = 1$ increase by 1 unit

C. the probability that $y = 1$ increases by 1 unit

D. the probability that $y = 1$ increases by a factor of e

E. the probability that $y = 1$ decreases by a factor of e

Free Response problems: Write your answers in the spaces following each prompt if possible. Make note if your work continues elsewhere!

13. (18 points) Suppose the random variable X represents how delicious a donut is. Under Anna's donut recipe, X has the probability density function f_A and using Jacob's donut recipe, X has the probability density function f_J:

$$f_A(x) = \frac{9}{10}x - \frac{3}{10}x^2, \quad \text{for} \quad x \in [0,2] \quad (0 \quad \text{otherwise})$$

$$f_J(x) = \frac{1}{3}x + \frac{1}{6}, \quad \text{for} \quad x \in [0,2] \quad (0 \quad \text{otherwise})$$

Show all your work and simplify your answers as much as possible. Box your final answers. You must perform these calculations by hand unless otherwise stated.

(a) Which recipe has the higher expected value of donut deliciousness? Use integrals to compute both.

(b) Which recipe has the higher median donut deliciousness? Use integrals to set these calculations up, but you may use software of your choice to solve for the medians if needed. [Hint: The median is the 50th percentile.]

(c) A donut is classified as "officially gross" if it has a deliciousness less than 1/2. Which recipe has the lowest probability of producing a donut that is officially gross? Use integrals and compute by hand.

(d) Which recipe would you want to use and why? *There is no single "right" answer here, as long as you have justification consistent with your responses in a, b, and c.*

14. (15 points) [40 points] A current leading drug provides pain relief within an average of 10 minutes of taking the drug. Another company aims to produce a new drug that will provide pain-relief symptoms in less than ten minutes. It is fair to assume that the time from taking the drug to the time pain relief begins is normally distributed. You are planning the study, and need to give the company information on how their clinical trial will perform. The hypothesis being tested is:
$H_0 : \mu = 10; \quad H_a : \mu < 10.$ where μ is the average length of time to pain relief provided by the new drug. Previous studies have shown the standard deviation of time to pain relief to be 12 minutes, so you decide to use that as a *known* value of σ:

    (a) You want to fix the probability of a Type I error at 15%. What cut-off point or rejection region should you use for rejection of the null hypothesis?

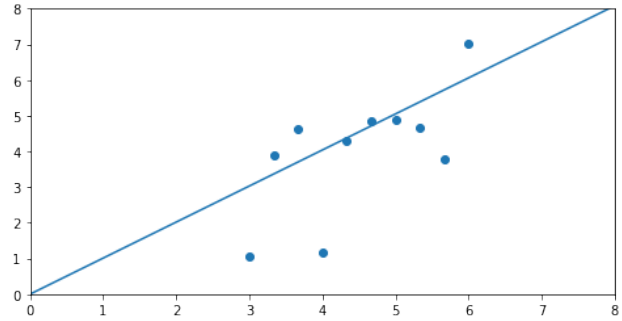
    (b) If the new drug acts 1 minute faster than the old drug, what sample size do you need to have 80% power for your test?

    (c) What will happen to the power of this test if we only increase our probability of a Type I error? (You do not need to give specific calculations here, just explain.)

    (d) What will happen to the power of this test if the new drug actually acts an average of 90 seconds faster than the old drug? (You do not need to give specific calculations here, just explain.)

    (e) How would the 85% confidence interval for μ implied by your sample relate to your results in part (a)?

15. (12 points) It's line-drawing time! The next 4 questions refer to the following plot, with *fitted* least-squares line by model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ shown:



The above plot was created by running SLR=SM.OLS(X,Y).FIT() on a pair of numpy arrays holding the x and y coordinates of the data. For each question, if the statement is always true mark "True"; if it is *possible* for the statement to be false, mark "False." You need to **justify** your answer with (at least) a full sentence:

(a) For the given plot, $\beta_1 > 0$.

(b) For the given plot, $\hat{\beta}_1 > 0$.

(c) For the given plot, if we run SLR.PREDICT(0), Python will return exactly $\hat{\beta}_0$

(d) For the given plot, the simple linear regression estimators satisfy $\hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \bar{Y}$

16. (18 points) A study is performed to investigate the effects of three different studying techniques - or "treatments" - vs. a control on 500 people. The outcome of interest is an improvement score (out of 125 points), reported by the student.

    In the table below, the mean and standard deviation of the improvement score for each treatment category is reported.

Group	Mean	SD
Control	60.93	19.87
Treatment 1: Despair (tx1)	56.66	18.39
Treatment 2: Frequent Gaming Breaks (tx2)	64.51	19.53
Treatment 3: 4am Cramming (tx3)	70.61	17.55

(a) Based on this information, perform three separate hypothesis tests investigating:
$H_0 : \mu_c - \mu_{tx} = 0$; $H_a : \mu_c - \mu_{tx} \neq 0$ for *each* treatment group against the control group. Use $\alpha = 0.10$.

(b) As a more sophisticated data scientist, you instead use a regression model. Your results for the coefficients of the model are:

Group	Estimate	Std. error	t-value	Pval
(Intercept)	60.756	1.678	36.209	<2e-16
tx1	-4.259	2.363	-1.802	0.0722
tx2	4.115	2.354	1.748	0.0811
tx3	9.823	2.363	4.156	3.81e-05

(c) What do you notice about the differences between treatment effects in the regression model when compared to the hypothesis tests? Why are there differences here?

(d) Using the regression model output, calculate 99% confidence intervals for all 4 coefficients: $\beta_0, \beta_1, \beta_2$, and β_3.

(e) Suppose that the first ten subjects of the study are in the following groups: control, tx1, tx1, tx3, tx2, control, control, tx3, tx3, tx2. What does the X (aka the "design matrix") look like for the first ten subjects?