CSCI 3022
Final Exam
Spring 2021

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

Name: Jaryd Meek

Student ID:

Section number: 001

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 resources. 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 or Python terminal to check numerical results.
- 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, May 2.
- 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) Which of the following tests are appropriate for the goal of establishing uncertainty or confidence intervals on the interquartile range of a probability process?

   A. Because of the central limit theorem, we can use normals for any problem like this.

   B. For small samples, we can create a confidence interval on teh interquartile range via t-distributions.

   [✓ C.] We could simulate the probability process and bootstrap to find confidence intervals for an interquartile range.

   D. None of the above will work: we'd have to use exact sums or integrals of the probability process to get theoretically consistent estimates.

   C

2. (3 points) You are sampling the weights of various puppies from a population with a known mean of 15 pounds and variance of 16 pounds^2. You obtain a measurement from an adorable Beagle of $X = 19$ pounds. What is the corresponding value of the standardized normal random variable, Z?

   A. 0.25

   B. 0.5

   [✓ C. 1]

   D. $\frac{19}{16}$

   E. 2

   F. $\frac{19}{4}$

   G. 15

   $\sqrt{16} = 4$

   $19 - 15 = 4$

   $4 = 1$ std

   C

3. (3 points) You are in awe of your desk plant Fernoulli Jr.'s grandeur. It's growing so successfully that you're considering renaming it Fernomial! a botanist tells you that the distribution of colors of leaves (via spectral analysis) is independently distributed and comes from a normal distribution with a mean wavelength of 515nm. You decide to verify this claim, and carefully measure the spectral intensity of 40 of Fernoulli Jr.'s leaves. You then compute the statistic $\left(\bar{X} - 515\right)/s_x$, using your sample mean and variance $\bar{X}, s_x$. What tests can you apply this statistic to?

   A. A t- distribution is appropriate here, a standard normal is not.

   B. A standard normal distribution is appropriate here, a t is not.

   C. Neither a t not a standard normal is appropriate.

   [✓ D.] Either a t or a standard normal is suitable for this problem.

   D

4. (3 points) A random variable U has a standard deviation equal to s_U, and a random variable V has a standard deviation equal to s_V. U and V are independent. Let $W = U + V$. What is the standard deviation of W?

   A. $s_U + s_V$.

   B. $s_U^2 + s_V^2$.

   [✓ C.] $\sqrt{s_U^2 + s_V^2}$.

   D. $\sqrt{s_U^2 + s_V^2}/\sqrt{2}$

   E. stats.chi2.ppf(s_U/s_V, len(U)-1, len(V)-1)

   F. None of the above.

   C

Use the following information for Problems 5 – 8. You're performing a simple linear regression, and someone spills ink all over your beautiful regression table. Now you can only read the following, though you also do recall that the data set had 147 observations, and most of their x-values were close to $x = 1$:

Coefficient	Estimate	Std. error	t-value	Pval
(Intercept)	3.4	**MISSING**	3.1	**MISSING**
Slope	5.56	0.19	**MISSING**	< 2e-16

5. (3 points) What is the correct (exact) **MISSING** value for the *intercept*'s p-value?

   A. `stats.t.ppf(3.1, df=147)`

   [✓ B.] 2(1-`stats.t.cdf(3.1, df=145)`)

   C. $3.4 \cdot 3.1$

   D. 2(1-`stats.t.cdf(3.1, df=147)`)

   E. (1-`stats.norm.cdf(3.1)`)

   B

6. (3 points) From the same table as the prior question, what is the (exact) **MISSING** value for the t-value of the *slope*?

   [✓ A.] 5.56/0.19

   B. `stats.t.ppf(5.56)- stats.t.ppf(0.19)`

   C. `stats.t.ppf(5.56, df=145)- stats.t.ppf(0.19, df=145)`

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

   E. $5.56 \cdot 0.19$

   A

7. (3 points) Suppose we were to take this linear regression and add in an additional predictor of x^2, so the model became $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$. Which of the following best describes **all** of the results of this added parameter?

   ✓ A. The R^2 of the model will increase.

   P B. The standard error associated with β_1 will increase.

   ✓ C. Because the coefficient for β_1 was so significant, it's unlikely that adding x^2 will help the model.

   D. Only (A) and (B) are true.

   ✓ E. Only (A) and (C) are true.

   F. Only (B) and (C) are true.

   G. All of (A), (B), and (C) are true.

   E

8. (3 points) Suppose we found ourselves a new data value at the (x, y) location of $(12, 50)$. What would be the effects of this data point on the resulting line of best fit?

   ✗ A. The estimate for β_1 will increase.

   ✓ B. The standard error associated with β_1 will increase.

   ✗ C. The total SSE of the model will decrease.

   D. Only (A) and (B) are true.

   E. Only (A) and (C) are true.

   F. Only (B) and (C) are true.

   G. All of (A), (B), and (C) are true.

   B

9. (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. $\alpha = 0.2$ and $n = 50$

   B. $\alpha = 0.2$ and $n = 12$

   C. $\alpha = 0.01$ and $n = 50$

   D. $\alpha = 0.01$ and $n = 12$

   E. $\alpha = 0.04$ and $n = 50$

   F. $\alpha = 0.04$ and $n = 12$

   A

10. (3 points) Data Scientists are often involved in study planning. You are in charge of a study that examines the mean lifetime (in years) of different cars. You know that the standard deviation of the lifetime of cars is $\sigma = 1.2$ years. What value of n do you need for the maximum 95% confidence interval width to be at most 0.5 years?

   ✗ A. 9

   ✗ B. 10

   ✗ C. 22

   ✗ D. 23

   ✗ E. 30

   ✗ F. 88

   [✓ G. 89]

   H. 100

   I. 101

   J. 1000

   G

11. (3 points) Which of the following statements is **True**?

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

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

   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. ✗

   E. The larger the p-value, the more we doubt the null hypothesis. ✗

   A

12. (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 reject the null hypothesis. What is the strongest conclusion you can draw?

A. Nothing.

[✓ B.] $\beta_k \neq 0$ for *some* $k \in \{2, 3\}$.

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

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

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

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

G. The model with $\beta_1 = \beta_4 \neq 0$ is not significantly better at capturing variance than the model with $\beta_1 = \beta_4 = 0$

B

13. (3 points) Suppose you generate 5,000 confidence intervals for the mean of a population, using fixed significance level α. You discover that 491 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. It's 50-50.

G. 4509/5000

D

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

14. (30 points) Answer the following short answer prompts after each question.

A. (6 points) What is the difference in how we interpret the pdf $f(x)$ of a continuous random variable and the pmf $f(x)$ of a discrete random variable? Do they have the same units? What does each measure?

Pdf ⇒ Probability Density Function, used for Continuous cases. Interpreted by integrating to determine probability.

Pmf ⇒ Probability Mass function, used for discrete cases. Gives exact answer to interpret (as probability)

Units? ⇒ Different units since one is a rate and one is a probability.

Measure? ⇒ $\text{cdf} \Rightarrow P(a < x < b)$, $\text{pmf} \rightarrow P(X = x)$

B. (8 points) Suppose that a sample $X_1, X_2, \ldots X_6$5 comes from a population with an unknown distribution. The population has a mean of 42 and a standard deviation of 12. Find the probability that the sample mean is between 40 and 47. Write your answer **three ways**: in critical value notation (using e.g. $t_{\alpha,\nu}$, z_{α_2}) **and** with exactly how you would find those values using python code (using scipy.stats syntax:.ppf, .cdf, .pdf, etc.) **and** the exact interval. Is this answer *exact* or an *approximation*? Why or why not?

if n=65 (can't tell if 6 or 65)

Critical Value → ?

Python → stats.norm.cdf(47, loc= 42, Scale =12) - stats.norm.cdf(40, loc= 42, Scale =12)

exact interval → $P(40 < x < 47) = \int_{40}^{47} \frac{1}{12\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-42}{12}\right)^2} dx = 0.2277$

C. (8 points) Suppose we're constructing a linear model to test the reaction of COVID vaccines, and suspect their may be a difference based on the sex of the recipient. We decide on the model:

$$y_i = \beta_0 + \beta_1 \cdot W_i + \beta_2 \cdot_i + \varepsilon_i$$

where y_i is the strength of response of patient i, W_i is that patient's weight in pounds, and M_i is an indicator or dummy variable that is true when the patient is a Male. We gather some data, and the first 6 patients are $\{(165, M), (125, W), (220, M), (145, M), (150, W), (185, W)\}$. What are the first 6 rows of the corresponding *design* matrix?

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix} = \begin{bmatrix} 1 & 165 & 1 \\ 1 & 125 & 0 \\ 1 & 220 & 1 \\ 1 & 145 & 1 \\ 1 & 150 & 0 \\ 1 & 185 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}$$

D. (8 points) You have 12 observations drawn from a normal distribution with unknown parameters, and want to test the hypothesis Ho : $\mu = \mu_0$ vs. Ha : $\mu > \mu_0$. You will reject the null hypothesis if your test statistic is greater than 1.84. What is the probability of a Type I error for your test? Write your answer both in critical value notation (using e.g. $t_{\alpha,\nu}$, z_{α_2}) **and** with exactly how you would find those values using python code (using scipy.stats syntax:.ppf, .cdf, .pdf, etc.) **and** the exact probability.

Critical value→ ?

Python→ (1- Stats.t.cdf(1.84,11))

exact→ The formula is too complicated to calculate by hand use a calculator or t-score table

15. (14 points) An e-commerce client claims that more than 20 percent of visitors to their site eventually become buyers (tracked by a cookie). Test this claim if a random check of the web server log indicates that 54 of 200 visitors made a purchase from the web site. In testing this claim, follow these four steps:

1. (5 points) Construct and interpret the 95% confidence interval for the true proportion of customers who become buyers.
2. (2 points) Write down the two hypotheses being tested, and define any parameter used in the hypotheses.
3. (5 points) Calculate the appropriate test statistic and its corresponding p-value.
4. (2 points) Using the p-value, decide whether or not to reject the null hypothesis at the 1% significance level and interpret your conclusion in terms of the original problem.

1 → [0.20847, 0.33153]

We can be 95% confident that the true proportion of customers who become buyers is captured by the interval [0.20847, 0.33153]

2 → $H_0 \rightarrow x \leq 0.20$ $\qquad$ x = true proportion of visitors to the site that become buyers

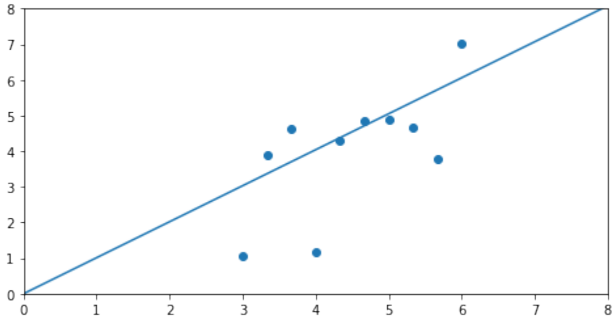
$H_A \rightarrow x > 0.20$

3 → 1 proportion z-test

P= 0.0067 $\qquad$ 0.67%

4 → With a p-value of 0.0067 and a significance level of $\alpha = 0.01$, we have sufficient evidence to reject the null hypothesis, in this case, we have sufficient evidence to say that more than 20% of people who visit the site are converted to buyers.

16. (16 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:



Note that the fitted line goes through exactly (0,0), here. 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 questions (a), (b), and (c), 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, $\beta_0 = 0$.

(b) For the given plot, $\hat{\beta}_0 = 0$.

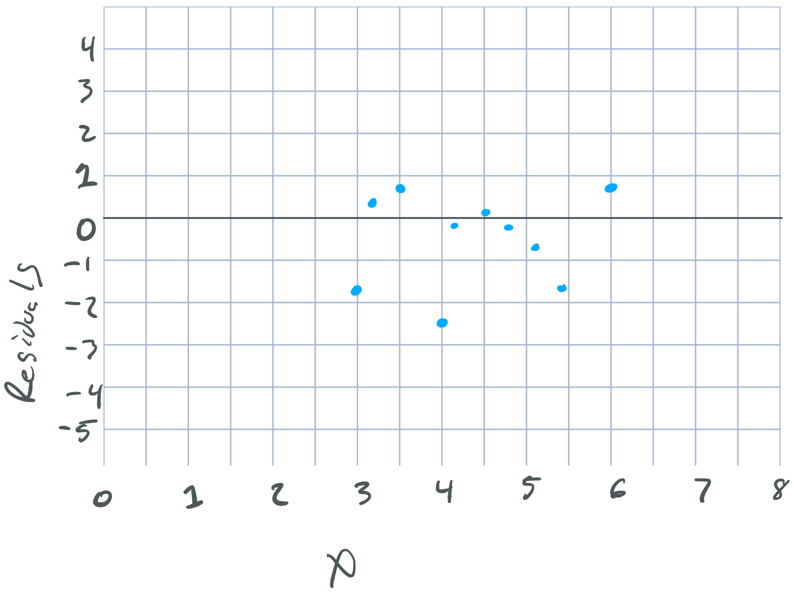
(c) For the given plot, the simple linear regression estimators satisfy $\hat{\beta}_1 = \bar{Y}/\bar{X}$

(d) Sketch a plot of the residuals of the linear model.

A ⇒ False, Our estimate goes through 0, but we don't know if the true value would be 0.

B ⇒ True, our line goes through 0,0, so the y intercept is 0.

C ⇒ True, if the y intercept is 0, the slope will be $\bar{Y}/\bar{x}$. our estimate has y intercept = 0, so the slope(B_1) = $\bar{Y}/\bar{x}$


Residuals
x