Write **as clearly as you can** and **in the box**:

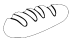
CSCI 3022
Final Exam
Fall 2019

Name:
Student ID:

Read the following:

- **RIGHT NOW**! Write your name on the top of your exam.
- You are allowed **one** 8.5 × 11in sheet of **handwritten** notes (both sides). No magnifying glasses!
- You may use a calculator provided that it cannot access the internet or store large amounts of data.
- You may **NOT** use a smartphone as a calculator.
- Clearly mark answers to multiple choice questions on the provided answer line.
- 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.
- If you do not know the answer to a question, skip it and come back to it later.
- 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.
- You have **150 minutes** for this exam.
- Point allocations: 15 × 2 points multiple choice; 2 × 5 points short answer; 4 × 15 points free response

Potentially Useful Values and Formulas

Bayes' theorem	$p(A \mid B) = \dfrac{p(B \mid A)p(A)}{p(B)}$	Law of total probability	$p(E) = \sum_{i=1}^{N} p(E \mid F_i)p(F_i)$
Union of sets	$p(A \cup B) = p(A) + p(B) - p(A \cap B)$	Conditional probability	$p(A \mid B) = \dfrac{p(A \cap B)}{p(B)}$
Sigmoid function	$\mathrm{sigm}(z) = \dfrac{1}{1+e^{-z}}$	Regression	$\hat{\sigma}^2 = \dfrac{SSE}{n-2}, \quad SE(\hat{\beta}) = \dfrac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}$
	Get that bread!	Binomial coefficients:	$C(n,k) = \binom{n}{k} = \dfrac{n!}{k!\ (n-k)!}$

Three types of F you might use:	$F = \dfrac{(SST - SSE)/p}{SSE/(n-p-1)}$	$F = \dfrac{SSB/df_{SSB}}{SSW/df_{SSW}}$	$F = \dfrac{(SSE_{red} - SSE_{full})/(p-k)}{SSE_{full}/(n-p-1)}$
Some confidence intervals:	$\bar{x} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{\alpha/2,n-1} \dfrac{\sigma}{\sqrt{n}}$	$\left[\dfrac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \ , \ \dfrac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \right]$

Multiple choice problems: Write your answers in the boxes, or they will not be graded!

1. (2 points) The average high temperature for Boulder, CO on October 31 is 58° Fahrenheit with a standard deviation of 11 degrees. If the temperature C in Celcius is calculated from the temperature in Fahrenheit F by $C = \frac{5}{9}(F - 32)$, what is the *variance* of the temperature in Boulder on October 31 in degrees Celcius?

   A. $\frac{5}{9} \cdot (58 - 32)$

   B. $11^2 \cdot \frac{5^2}{9^2}$

   C. $11 \cdot \frac{5}{9}$

   D. $\left(\frac{5}{9}\right)^2 \cdot 26^2$

   E. $11^2 \cdot \frac{5}{9}$

   F. $11 \cdot \left(\frac{5}{9}\right)^2$

2. (2 points) Let $f(x) = kx^3 + x - 1$ for $0 \leq x \leq 1$, and $f(x) = 0$ for $x \notin [0, 1]$, where k is some unknown constant. What value of k will make f a valid probability density function?

   A. 1/4

   B. 1/3

   C. 1

   D. 2

   E. 4

   F. 6

   G. No such value of k exists.

3. (2 points) Suppose you are trying to assess whether of not the mean GPAs of 12 students in an honors seminar course are greater than 3.5. From historical data, you know that students' GPAs typically follow a normal distribution. The 12 seminar students have a mean GPA of 3.65, with a standard deviation of 0.2. Which of the following corresponds to the p-value for this test with a significance level of $\alpha = 0.1$?

   A. stats.norm.cdf($\frac{.15\sqrt{12}}{0.2}$)

   B. stats.t.cdf($\frac{.15\sqrt{12}}{0.2}$, df = 11)

   C. stats.norm.ppf(0.9)

   D. stats.t.ppf(0.9, df = 11)

   E. Not enough information given.

4. (2 points) Suppose we are using a logistic regression model with two features, x_1 and x_2, to classify objects as either Class 0 or Class 1. Which of the following best describes the meaning of a decision boundary in logistic regression?

   A. A decision boundary is the set of all the features we can classify using logistic regression.

   B. A decision boundary is a point, line, plane, or hyperplane in feature space that contains all of the features we would classify as Class 1.

   C. A decision boundary is a point, line, plane, or hyperplane in feature space that contains all of the features we would classify as Class 0.

   D. A decision boundary is a point, line, plane, or hyperplane in feature space that separates sets of features we would classify as Class 0 from those we would classify as Class 1.

   E. A decision boundary is the point when you are making dinner plans with friends where nobody agrees on a restaurant and you all stop being friends anymore.

   □

5. (2 points) Suppose we fit a two-feature logistic regression model for $p(y = 1|x_1, x_2)$ and find $\beta_0 = 1, \beta_1 = -1$, and $\beta_2 = 2$. If we use a threshold of 0.5 for our decision rule, how should we classify a data point with features $(x_1, x_2) = (1, 1)$?

   A. Class 1

   B. Class 0

   C. Unable to classify

   D. I've never heard of this before.

   □

6. (2 points) Suppose you are petting dogs to see which ones will bark at you when you do so. Assume that each dog's probability of barking is independent of the other dogs, and that the dogs will bark when petted with constant probability $p = 0.85$. You pet 10 dogs. Let the random variable X represent the number of dogs that bark when petted.

   What distribution best describes the probability distribution for X?

   A. Geometric

   B. Exponential

   C. Normal

   D. Negative binomial

   E. Uniform

   F. Poisson

   G. Binomial

   H. Bernoulli

   □

7. (2 points) Suppose you roll the same biased die repeatedly. What is the probability of observing your first 3 on the fifth roll? Note, $P(\text{roll} = 3) = .25$.

   A. $\begin{pmatrix} 4 \\ 1 \end{pmatrix} \left(\frac{3}{4}\right)^4 \left(\frac{1}{4}\right)$

   B. $5 \cdot \frac{3^4}{4^5}$

   C. $\frac{3^4}{4^5}$

   D. $\frac{1}{6}$

   E. $\frac{1}{4}$

   □

8. (2 points) You have the ability to sample many exponentially distributed random variables of mean $\frac{1}{2}$ and variance $\frac{1}{4}$, and are interested in estimating the population mean via a 95% confidence interval. How large of a sample do you need to have a confidence interval with width no wider than 0.1?

   A. The smallest integer n larger than $2 \cdot 10 \cdot t_{\alpha,n-1} \cdot \sigma$

   B. The smallest integer n larger than `stats.norm.ppf(.975)*10)**2`

   C. $n = 30$ is always good.

   D. This problem should be done by bootstrapping, so you simulate until the width is less than .1 anyways.

   E. The smallest integer n larger than $(z_{.05} \cdot \frac{0.5}{10})^2$

   F. Cannot be determined.

   □

9. (2 points) Consider the following function. The function output constitutes a sample from which one of the following distributions?

```
def didmyHW(p):
    param=(1+np.sqrt(5))/2
    draw = np.random.random()
    x = -np.log(1-draw)/param
return x
```

   A. Discrete Uniform

   B. Normal

   C. Negative binomial

   D. Continuous Uniform

   E. Poisson

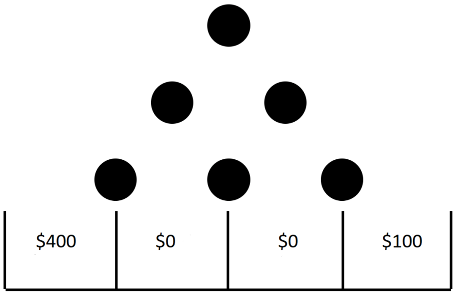
   F. Exponential

   G. Binomial

   H. Log-Normal

   □

The next two questions refer to the following Plinko payouts:



10. (2 points) Let's play Plinko! A game of **Plinko** is to be played on the board shown above. The pegs are unbiased, meaning that the disc has equal probability of moving left or right at each peg. Furthermore, the disc can only be dropped from directly above the top-most peg. What is the expected value of your winnings with a single disc?

    A. $62.5

    B. $75

    C. `stats.binom.cdf(1,3,.5)`

    D. $0

    E. $50

    F. `np.random.choice([400,0,0,100])`

    □

11. (2 points) Bob Barker is back as host for the Price is Right and is a little rusty as host, and we've discovered that we can now control whether the puck is dropped above *any* of the 3 top-most pegs (or the top 2 rows). Where should we drop the puck to maximize our earnings, and what is the variance of dropping the puck from this location?

    A. Drop from row two left side, variance of 36093.75.

    B. Drop at row two right side, variance of 1875.

    C. Drop from center, variance of 17343.75.

    D. Drop from row two left side, variance of 30000.

    E. Drop from row two left side, variance of 1/4.

    □

12. (2 points) Suppose you generate 1,000 confidence intervals for the mean of a population, using fixed significance level α. You discover that 892 of them in fact do contain the true mean. Which of the following is the most appropriate estimate of the significance level α?

    A. $\alpha = 0.05$

    B. $\alpha = 108$

    C. $\alpha = 1.1$

    D. $\alpha = 0.01$

    E. $\alpha = 0.1$

    F. $\alpha = 0.25$

    □

13. (2 points) Suppose you compute a sample mean for a population that is normally distributed with estimated variance s^2. Which combination of significance level α and sample size n produces the widest confidence interval for the mean?

   A. $\alpha = 0.03$ and $n = 30$

   B. $\alpha = 0.03$ and $n = 50$

   C. $\alpha = 0.05$ and $n = 30$

   D. $\alpha = 0.05$ and $n = 50$

   E. $\alpha = 0.10$ and $n = 30$

   F. $\alpha = 0.10$ and $n = 50$

□

14. (2 points) In an attempt to estimate the variance in Zach's legible characters per day, a student selects $n = 20$ days of lecture slides at random and counts the number of legible characters. The sample yields a sample variance of 300 and a sample mean of 3022. Which of the following gives a 95% CI for the variance, assuming the day-to-day distribution is normal?

   A. $\dfrac{19 \cdot 300}{-\chi^2_{0.025,19}} \le \sigma^2 \le \dfrac{19 \cdot 300}{\chi^2_{0.025,19}}$

   B. $300 - \chi^2_{0.025,19} \cdot \sqrt{\dfrac{300}{20}} \le \sigma^2 \le 300 + \chi^2_{0.975,19} \cdot \sqrt{\dfrac{300}{20}}$

   C. $300 - t_{0.025,19} \cdot \sqrt{\dfrac{300}{20}} \le \sigma^2 \le 300 + t_{0.025,19} \cdot \sqrt{\dfrac{300}{20}}$

   D. $3022 - t_{0.025,19} \cdot \sqrt{\dfrac{300}{20}} \le \sigma^2 \le 3022 + t_{0.025,19} \cdot \sqrt{\dfrac{300}{20}}$

   E. $\dfrac{19 \cdot 300}{\chi^2_{0.025,19}} \le \sigma^2 \le \dfrac{19 \cdot 300}{\chi^2_{0.975,19}}$

□

15. (2 points) Building on the previous question: To follow up their prior work, the same student selects $m = 16$ days of lecture slides from Rachel's lectures at random and counts the number of legible characters. The sample yields a sample variance of $100 char^2$. What is the appropriate test statistic to compare whether or not the variances of the two classes are equal?

   A. $F_{stat} = 300/100 \sim F_{15,19}$

   B. $t_{stat} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{300}{20} + \frac{100}{16}}} \sim t_\nu$

   C. Subtract Zach's $\chi^2_Z = \frac{300}{19}$ from Rachel's $\chi^2_R = \frac{100}{15}$ and it's a t statistic

   D. $z_{stat} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{300}{20} + \frac{100}{16}}} \sim Z$

   E. $F_{stat} = \sqrt{100/300} \sim F_{19,15}$

□

Short answer problems: If you answer does not fit in the box provided, make a note of where it is continued!

16. (5 points) What is a p-value? Describe, in words, what a p-value is meant to quantify. Then, provide a definition in terms of a probability (you may use words in your probability terms). What do we use it for?

17. (5 points) You love error bars. It upsets you that when you make box plots the interquartile range looks like an error bar, but doesn't have an error bar of its own. Error bars on your error bars: that's true data science. So you take your sample of a total of $n = 3022$ data points and you commit to finding the 95% confidence interval on the population Q_1. Recall that Q_1 denotes the lowest quartile: for a *sample* it is estimated by the smaller of the two numbers in the interquartile range.

    Describe in full pseudocode (and words where appropriate) a bootstrapping algorithm that will calculate a 95% confidence interval on Q_1.

Free response problems: If your answers do not fit on the page and the one right after it, make a note of where the work is continued!

18. (15 points) Suppose you just bought a new puppy and are trying to feed it. The problem is that the puppy keeps getting distracted and running away right in the middle of its meal. Your friend tells you that the pdf, $f(x)$, represents the probability that any given puppy will get distracted after a certain amount of time. Your friend also tells you that the average amount of time a puppy will eat prior to getting distracted and running away, is 30 seconds (1/2 minute). $f(x)$ is a triangular distribution as defined below:

$$f(x) = \begin{cases} 1 - \frac{1}{2}x & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

x represents the number of minutes after which puppies get distracted from eating. All puppies get distracted once 2 minutes have passed. Please answer the following questions, and be sure to show your work—for sufficient space, a blank page follows this one.

(a) (4 points) You have watched many puppies in your day, and you think that puppies usually get distracted after less than 30 seconds of eating. You want to see if there is sufficient evidence to conclude that the true parameter value is less than $\mu = 1/2$ minutes. State the relevant null and alternative hypotheses.

(b) Devise a test of the form "reject if $X < c$ " where c is how long you wait until the puppy gets distracted from its dinner. Use a significance level of $\alpha = \frac{9}{25}$. How long do you wait before you reject the null hypothesis?

(c) What is the probability of a Type 1 Error?

(d) It turns out that neither the null nor the alternative are correct! The true distribution of puppy distraction was a continuous uniform with pdf:

$$f(x) = \begin{cases} \frac{1}{2} & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Now, what is the probability that you reject the null hypothesis?

Additional Workspace

19. (15 points) We have the somewhat famous 'mtcars' data set, where the miles per gallon of various cars are predicted by numerous features. Unfortunately, we only have two of the features available: whether the car being manual/automatic transmission and the cars' time to drive 1/4 of a mile. We fit the model using `sm.OLS.fit()` and ask for a `.summary()` and get the following:

```
OLS Regression Results
==============================================================================
Dep. Variable:                    mpg
No. Observations:                  32   R-squared:                       0.687
Df Residuals:                      29   Adj. R-squared:                  0.665
Df Model:                           2  F-statistic:                     31.80
Model:                            OLS

==============================================================================
                coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept        (A)      6.597     -2.863      0.008     -32.382      -5.397
am            8.8763        (B)      6.883      0.000       6.239      11.514
qsec          1.9819      0.360        (C)        (D)         (E)       2.718
==============================================================================
```

(a) What are the correct values for the missing numbers in the table above, labeled (A), (B), (C), (D), and (E)? If a critical value or quantile is necessary, denote your final answer using the python code that would generate that value (in scipy.stats syntax).

(b) The first 4 observations in the data set are:

	mpg	qsec	am
Mazda RX4	21.0	16.46	M
Mazda RX4 Wag	21.0	17.02	M
Datsun 710	22.8	18.61	M
Hornet 4 Drive	21.4	19.44	A

What are the first 4 rows of the X matrix (design matrix) used to fit the model above, if A is used as the baseline for the categorical predictor?

(c) You want the sum of squared error for your model, so you astutely ask statsmodels for (.ssr) (Sum of Squared Residual/Error) of your fit. It returns `352.63`. What is the SST (Sum of Squares Total) of the 'mpg' variable?

(d) Based on the summary output above, should any of the 3 terms above be excluded from the model? Why or why not?

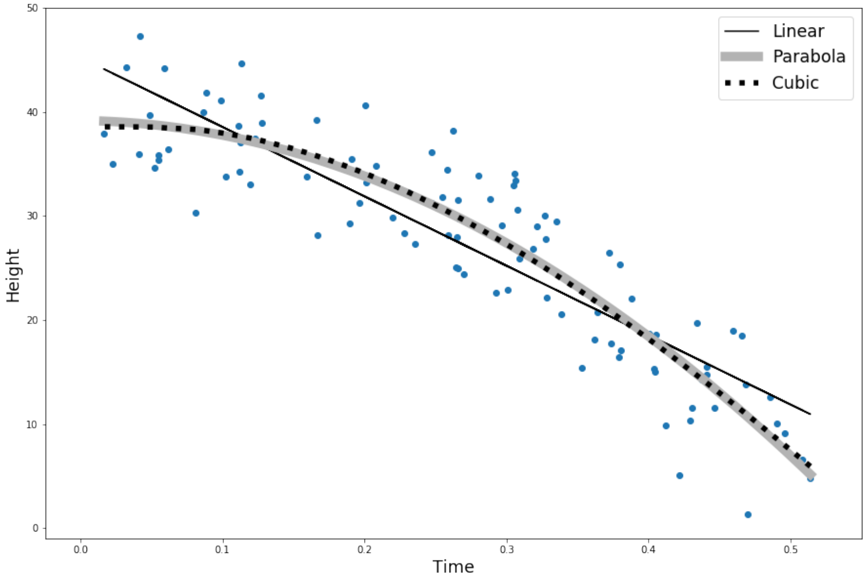
Additional Workspace

20. (15 points) It's almost time for vacation! You giddily head to DIA to prepare for your flight, and settle into the 3 hour long Travel Security Administration (TSA) lines. Under their new protocol, TSA agents invert every passengers' luggage and shake vigorously until all of their belongings fall out. This fascinates you, and you start taking snapshots of the falling objects with your camera. You gather up all 100 of your pictures and create a plot of how high off the ground (in inches) each object is as a function of how long the TSA agent has been shaking their bag (in seconds).

You have plenty of time to run some models, so you decide to evaluate the motion according to the following models:

Model Number	Function	On Graph below
(1)	$h = \beta_0 + \text{error}$	Not shown
(2)	$h = \beta_0 + \beta_1 \cdot t + \text{error}$	thin (straight) black line
(3)	$h = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \text{error}$	thick gray line
(4)	$h = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \beta_3 \cdot t^3 + \text{error}$	medium dotted back line

Note that models 3 and 4 are nearly entirely overlapping on the picture of the resulting best-fit lines, below.



(a) You start with model (4) since you view it as the most complete, and decide to test whether or not it's outperforming model (2), with a significance level of .10.

(i) State the null and alternative hypothesis for your test.

(ii) Write your test statistic and what distribution it comes from, under the null hypothesis.

(b) Of the 4 models, which model do you think has the lowest R^2? Fully **explain** your choice.

(c) Of the 4 models, which model do you think has the lowest adjusted R^2? Fully **explain** your choice.

(d) You take model (2) and decide to make a component-residual plot, because those things are cool. Does what you see tell you anything about the assumptions underlying model (2)? **explain**.

Additional Workspace

21. (15 points) During finals week, you decide to take a study break and go down to the engineering center lobby to pet some dogs. You are interested in analyzing the different amounts of licks that different breeds of dogs give out. You collect a data set of 12 students' experience with petting dogs and being licked. You separate the dogs into three groups: Poodles, Poodle Labrador mixes (Labradoodle), and Corgi Poodle mixes (Corgipoo). The numbers represent the number of times a student was licked in petting a particular dog type.

Poodle (P)	Labradoodle (L)	Corgipoo (C)
3	6	7
4	7	8
6	7	10
7	8	11

(a) Clearly state the null and alternative hypotheses for the one-way ANOVA test to compare the three groups of students and determine whether or not there is evidence that there is some difference in number of dog licks according to dog breed.

(b) Compute the relevant **test statistic** to test the hypotheses from part (a). Put a $\boxed{\textbf{box}}$ around your answer for the test statistic. Show all work!

(c) For a test at the $\alpha = 0.1$ significance level, perform a **rejection region** test for your test statistic from part (b). Be sure to clearly state (i) the distribution you are referencing (including any degrees of freedom), and (ii) the critical value to which you are comparing your test statistic. You may leave your critical value in terms of a Python function.

(d) What does it mean to reject the null hypothesis in the context of this problem? What does it mean to fail to reject the null hypothesis in the context of this problem.

Additional Workspace