CSCI 3022
Final Exam
Spring 2018

Name: _____

By writing my name I promise to abide by the Honor Code

Read the following:

- **RIGHT NOW**! Write your name on the top of your exam.
- You are allowed two 8 1/2 × 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 **2.5 hours** for this exam.

Problem	Points	Score
1	2	
2	2	
3	2	
4	2	
5	2	
6	2	
7	2	
8	2	
9	2	
10	2	
11	2	
12	2	
13	2	
14	2	
15	2	
16	2	
17	2	
18	2	
19	2	
20	2	

Problem	Points	Score
Mult. Choice Total	40	
21	15	
22	15	
23	15	
24	15	
Total	100	

Potentially Useful Values

Standard Normal Distribution: Here $\Phi(z)$ is the cumulative distribution function for the standard normal distribution evaluated at z. Its equivalent form in Python is $\Phi(z) =$ stats.norm.cdf(z),

$$\begin{array}{lllll}
\Phi(4.75) \approx 1.000 & \Phi(3.00) = 0.999 & \Phi(2.58) = 0.995 & \Phi(2.32) = 0.990 & \Phi(2.00) = 0.977 \\
\Phi(1.96) = 0.975 & \Phi(1.88) = 0.970 & \Phi(1.75) = 0.960 & \Phi(1.64) = 0.950 & \Phi(1.44) = 0.925 \\
\Phi(1.28) = 0.900 & \Phi(1.15) = 0.875 & \Phi(1.04) = 0.850 & \Phi(0.93) = 0.825 & \Phi(0.84) = 0.800 \\
\Phi(0.76) = 0.775 & \Phi(0.67) = 0.750 & \Phi(0.60) = 0.725 & \Phi(0.52) = 0.700 & \Phi(0.45) = 0.675 \\
\Phi(0.39) = 0.650 & \Phi(0.32) = 0.625 & \Phi(0.25) = 0.600 & \Phi(0.19) = 0.575 & \Phi(0.13) = 0.550 \\
\Phi(0.06) = 0.525 & \Phi(0.00) = 0.500 & & &
\end{array}$$

Student's t-Distribution: The following values of the form $t_{\alpha,v}$ are the critical values of the t-distribution with v degrees of freedom, such that the area under the pdf and to the right of $t_{\alpha,v}$ is α. Its equivalent form in Python is $t_{\alpha,v} =$ stats.t.ppf$(1-\alpha, v)$.

$$\begin{aligned}
t_{0.05,38} &= 1.686 \\
t_{0.025,38} &= 2.024 \\
t_{0.05,40} &= 1.684 \\
t_{0.025,40} &= 2.021 \\
t_{0.05,2} &= 2.920 \\
t_{0.025,2} &= 4.303
\end{aligned}$$

F-Distribution: The following values of the form F_{α,v_1,v_2} are the critical values of the F-distribution with v_1 and v_2 degrees of freedom, such that the area under the pdf and to the right of F_{α,v_1,v_2} is α. Its equivalent form in Python is $F_{\alpha,v_1,v_2} =$ stats.f.ppf$(1-\alpha, v_1, v_2)$.

$$\begin{aligned}
F_{0.05,2,7} &= 4.737 \\
F_{0.025,2,7} &= 6.542 \\
F_{0.05,3,5} &= 5.409 \\
F_{0.025,3,5} &= 7.764 \\
F_{0.05,3,10} &= 3.708 \\
F_{0.025,3,10} &= 4.826
\end{aligned}$$

For the following 4 questions, assume the following scenario: two otters have broken into the Rec Center once again and are cruising around the buffalo-shaped swimming pool area. Their names are Valentina and Otto Ferguson. Naturally, quite a crowd has gathered to watch these beloved rascals. At any given time, the probability that Valentina is swimming is 0.20, the probability that Otto is swimming is 0.75, and the probability that both otters are swimming is 0.10.

Look, in the crowd! It is Kyle and Sofie. You happen to know that when Valentina is not swimming, Kyle is stoked only 25% of the time, but when she is swimming, Kyle is stoked 90% of the time. Sofie is stoked 75% of the time if the otters are around *at all*, swimming or not. If there is at least one otter swimming, she is stoked 80% of the time. **Note**: All multiple choice answers have been rounded to three decimal places.

1. (2 points) What is the probability that one or more otters are swimming?

   A. 0.950
   B. 0.850
   C. 0.750
   D. 0.200
   E. 0.150

   $P(V) = 0.2$ $P(S) = 0.75$
   $P(O) = 0.75$ $P(S \mid V \cup O) = 0.8$
   $P(V \cap O) = 0.10$ $P(V \cup O) = P(V) + P(O) - P(V \cap O)$
   $P(K \mid V^c) = 0.25$ $= 0.2 + 0.75 - 0.10$
   $P(K \mid V) = 0.90$

   1. _____

2. (2 points) Given that Valentina is swimming, what is the probability that Otto is swimming?

   A. 0.950
   B. 0.550
   C. 0.500
   D. 0.150
   E. 0.020

   $P(O \mid V) = \frac{P(O \cap V)}{P(V)} = \frac{0.10}{0.2} =$

   2. C

3. (2 points) Suppose you see Sofie in the crowd and observe that she is stoked. What is the probability that no otters are swimming?

   A. 0.093
   B. 0.907
   C. 0.150
   D. 0.750
   E. 0.107
   F. 0.070
   G. Not enough information.

   $P(V^c \cap O^c \mid S) = P(V \cup O)^c \mid S) = 1 - P(V \cup O \mid S)$
   $1 - P(A \cup B) = P(A \cup B)^c = P(A^c \cap B^c)$
   $1 - \frac{P(S \mid V \cup O) P(V \cup S)}{P(S)}$ A

   3. _____

4. (2 points) Suppose you can't see whether either otter is swimming, nor can you see directly if Kyle is stoked. Still, based on what you *do* know, what is the probability that Kyle is stoked?

   A. 0.380
   B. 0.575
   C. 0.770
   D. 0.803
   E. About 50/50.
   F. Not enough information.

   $P(K) = P(K \mid V) P(V) + P(K \mid V^c) P(V^c)$
   $0.90 \times 0.2 + 0.25 \times 0.8$
   $0.18 + :)$

   4. _____

5. (2 points) You have the following 11 samples in a dataset: $2, 1, 0, -1, 21, 13, 8, 30, 5, 1, 3$. Which *one* of the following is true?

A. The mean diverges
B. The range is 30
C. The range is 29
D. The IQR is 9.5
E. The IQR is 12
F. The median is equal to the mean.

$-1, 0, 1, 1, 2, 3, 5, 8, 13, 21, 30$

$Q_1 = 1$

$Q_3 = 10.5$

5. D

6. (2 points) Some engineers use the Rankine temperature scale. To convert from a temperature in Rankine T_R to a temperature in Celsius T_C, the following conversion is used.

$$T_C = \frac{T_R - 491.67}{1.8}$$

Suppose that a temperature is represented by the random variable X, whose units are in Rankine, with standard deviation 6 Rankine. Let Y be the equivalent random variable with units converted to degrees Celsius. What is the variance of Y, rounded to one decimal point?

A. 36.0
B. 1.9
C. The null hypothesis should be rejected.
D. -269.8
E. 269.8
F. 11.1

$X: S_X = 6$

$C = \frac{R - 491}{2}$

$\text{var}(Y) = \text{var}\left(\frac{X - 491}{2}\right)^2$

$= \frac{1}{4}\,\text{var}(X - 491)$

$\frac{1}{4}(\text{var}(X)) = \frac{1}{4}\,36$

6^2

6. _____

7. (2 points) Let $f(x)$ be the pdf of a normal random variable with mean -100 and variance $10,000$. Use this information to compute

$$\int_{-94}^{-40} f(x)\,dx\,.$$

A. 0.05
B. 0.2
C. 1.64
D. 1.96
E. 0.225
F. 0.1

$P(-94 < x < -40)$

$\frac{-100+94}{100} < x < \frac{-100+60}{100}$

$-0.06 < x \leq 0.6$

$0.525 < x < 0.725$

7. B

8. (2 points) The practicum is due in 19 minutes and a student comes to your office hours with the following code, which is supposed to compute a 95% confidence interval for the difference in means between two large datasets, stored in variables *data* and *lore*. Is this code going to do what the student hopes?

   - If there are errors in this code, **clearly** circle those lines numbers *and* write corrected lines of code in the box below the code.
   - If there is no error in the code, simply write "No error" in the box below the code.

```
 0 import numpy as np
 1 from scipy import stats
 2 def get_confidence_interval(x,y,alpha=0.9):
 3     xbar = np.mean(x)
 4     ybar = np.mean(y)
 5     nx  = len(x)
 6     ny = len(y)
 7     z_half_alph= stats.norm.cdf(1-alpha/2)
 8     L = xbar-ybar - z_half_alph * np.sqrt( np.var(x,ddof=1)/nx + np.var(y,ddof=1)/ny )
 9     U = xbar-ybar + z_half_alph * np.sqrt( np.var(x,ddof=1)/nx + np.var(y,ddof=1)/ny )
10     return [L,U]
11
12 CI = get_confidence_interval(data,lore,alpha=0.95)
```

Answer 8 here.

9. (2 points) Consider the following function related to finding an open parking spot in a large parking lot where the probability of an individual spot being open is given by p. What distribution does the return value of the function belong to?

```
import numpy as np
def shoulda_taken_the_bus(p):
    x = 1
    while np.random.choice([0,1], p=[1-p, p]) == 0:
        x += 1
    return x
```

   A. Binomial
   B. Geometric
   C. Poisson
   D. Uniform
   E. Exponential
   F. The function above will produce an error if executed
   G. None of the above

9. B

10. (2 points) In hypothesis testing, the significance level α is the probability that you

A. reject the null hypothesis when the null hypothesis is true

B. reject the null hypothesis when the null hypothesis is false

C. fail to reject the null hypothesis when the null hypothesis is true

D. fail to reject the null hypothesis when the null hypothesis is false

10. A

11. (2 points) Suppose you draw $n = 200$ samples from some distribution and want to test the following hypotheses about the mean μ, $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$. You will reject the null hypothesis if your test statistic is greater than 1.64 in absolute value. What is the significance level α of your test?

A. $\alpha = 0.025$

B. $\alpha = 0.01$

C. $\alpha = 0.05$

D. $\alpha = 0.1$

E. $\alpha = 0.9672$

F. $\alpha = 0.9836$

$p = 1 - \Phi(1.64)$
$= 1 - 0.95 = 0.5$

11. C

12. (2 points) One minus the p-value is:

A. the probability that H_1 is true

B. the probability that H_0 is true

C. the probability, assuming that H_0 is true, that you observe something less extreme than your test statistic

D. the probability, assuming that H_1 is true, that you observe something more extreme than your test statistic

E. the probability of a type-II error

F. the probability of a type-I Tukey-adjusted catastrophe

12. C

13. (2 points) Consider performing a multiple linear regression on a dataset with full and reduced models of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad \text{and} \quad y = \beta_0 + \beta_1 x_1 + \beta_3 x_3,$$

respectively. Suppose that you perform a partial F test and fail to reject the null hypothesis. What can you conclude?

A. $\beta_1 = \beta_3 = 0$

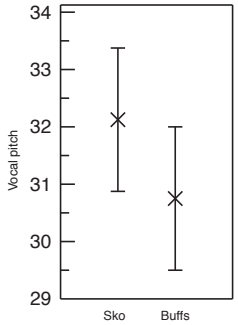
B. $\beta_2 = \beta_4 = 0$

C. $\beta_k = 0$ for $k = 1, 2, 3, 4$

D. $\beta_k \neq 0$ for $k = 1, 2, 3, 4$

13. 

14. (2 points) Before the homecoming football match, a study was conducted of the vocal pitch of undergraduate students as they shouted *sko buffs*. Many, many *sko buffs*es were recorded, but, because *not all of them* were recorded, the analysis team created a 95% confidence interval for the mean vocal pitch of *sko* and another one for *buffs*. What are the researchers able to conclude about the true mean vocal pitches of *sko* and *buffs*?



A. The true mean of *sko* is significantly higher than the true mean of *buffs*.

B. The true mean of *buffs* is 95% lower than the true mean of *sko*.

C. There is not a statistically significant difference between the true means of *sko* and *buffs* at a 95% confidence level.

D. 95% of true pitches will be greater in for *sko* than *buffs*.

E. There is a high confidence that the true means are the same for *sko* and *buffs*, 95% of the time.

F. None of the above.

14. _____

15. (2 points) Suppose a procedure generates confidence intervals with fixed significance level α which **FAIL** to cover the true mean 2 times out of 20 *on average*. What is the significance level α?

A. 0.01

B. 0.05

C. 0.10

D. 0.20

E. 0.90

F. 0.95

$\frac{2}{20} = 0.1$

15. C

16. (2 points) Suppose you compute a sample mean for a population that is normally distributed with known variance σ^2. Which combination of significance level α and sample size n produces the **widest** confidence interval for the mean?

A. $\alpha = 0.05$ and $n = 10$

B. $\alpha = 0.01$ and $n = 10$

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

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

E. Both A and D.

F. Both B and C.

16. A

17. (2 points) You turn the page of your final exam—*whoah how did they know??*—only to find the following problem taken directly from the midterm, which you have now fully mastered: You are stuck in a YouTube clickhole, late on a Thursday night, and cannot stop yourself from clicking Next...Next...Next...on all the videos. Homework is due for CSCI 3022 the next day, but you simply cannot get enough of these amazing clips of river otters playing in the snow. What a bunch of goofballs! A new notification appears on your mobile, and it says:

$$f(x) = kx - x^2 \text{ for } x = 1, 2, 3 \text{ and } f(x) = 0 \text{ for any other integer value of } x$$

What value or values of k make $f(x)$ a valid probability mass function?

A. 3

B. 15/6

C. 6/15

D. $15/6, -15/6$

E. No such value of k exists.

$\int_1^3 kx - x^2 \, dx = k\frac{x^2}{2} - \frac{x^3}{3}\Big|_1^3$

17. _____

18. (2 points) Which of the following is a definition of the *decision boundary* in logistic regression?

A. $P(y \mid x) > 0$

B. The critical value at which we reject $H_0: \beta_1 = \beta_2 = \ldots \beta_p = 0$.

C. Log Odds are equal to one.

D. Log Odds are equal to zero.

E. p-value $< \alpha$

F. p-value $= \alpha$

G. t is more extreme than $t_{1-\alpha/2, n-1}$

18. _____

19. (2 points) Suppose that you perform two multiple linear regressions on a dataset with 57 data points. The first MLR uses 10 features and the second MLR uses a subset of 3 of those 10 features. Suppose you want to answer the questions: Is your 10-feature model a significant improvement on the 3-features model? To answer this question you perform a _____ test where the test statistic follows a _____ distribution.

A. t-test, t_{55}

B. Partial F-test, $F_{7,46}$

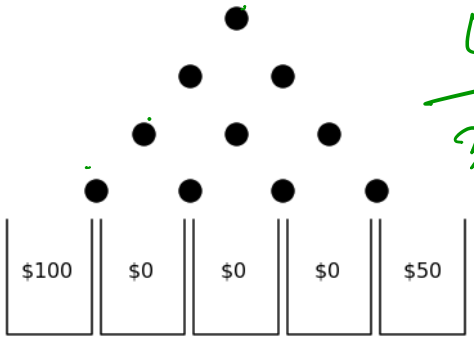
C. Partial F-test, $F_{7,56}$

D. Full F-test, $F_{10,46}$

E. Full F-test, $F_{3,46}$

19. _____

20. (2 points) A game of **Plinko** is to be played on the board shown below. 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 probability that you win a total of exactly \$100 in a game with 2 discs?



A. 30/256
B. 44/256
C. 16/256
D. 28/256
E. 73/256
F. None of the above.

20. _____

The rest of this page may be used for calculations, but will not be graded.

Additional Workspace

21. (15 points) Suppose you run two different email Ad campaigns over many days and record the amount of traffic driven to your website on days that each ad is sent. In particular, suppose that Ad 2 is sent on 50 different days and generates an average of 2 million page views per day with a standard deviation of 1 million views, and Ad 1 is sent on 40 different days and generates an average of 2.5 million page views per day with a standard deviation of a half million views.

    (a) To learn about the difference (if one exists) between Ad 2's average page views per day and Ad 1's average page views per day, compute a 95% confidence interval for the difference between the ads' average page views. Interpret your confidence interval.

    (b) Your boss, Dr. Ana Maria Ferguson, tells you that you can run 10 more days of experiments in order to refine your confidence interval from part (a). Which of the following strategies would most reduce the width of the confidence interval, assuming that your estimates of the means and standard deviations would not change? Explain your choice.

        i. spend 8 days testing Ad 2 and 2 days testing Ad 1
        ii. spend all 10 days testing Ad 2
        iii. spend all 10 days testing Ad 1
        iv. spend 5 days testing each

    (c) Based on the original data (no additional days testing), Dr. Ferguson wants to know if there is enough evidence that Ad 1 outperforms Ad 2 by at least $150,000$ views per day at the $\alpha = 0.01$ level. State the relevant null and alternative hypotheses. Perform a hypothesis test to answer Dr. Ferguson's question.

Additional Workspace

22. (15 points) You are performing logistic regression to classify images based on their features as either otter-related (which in your model corresponds to a 1) or buffalo-related (which in your model corresponds to a 0). In your regression, you have found that $\beta_0 = 2$, $\beta_1 = -1$, and $\beta_2 = 3$.

    (a) For each point below, explain whether you would classify it as otter-related or buffalo-related. As always, show your work and explain why your calculation leads you to classify each point the way that you have.

        i. $(x_1, x_2) = (1, 1)$
        ii. $(x_1, x_2) = (2, 0)$
        iii. $(x_1, x_2) = (3, -1)$

    (b) In words, what are the odds that a picture with features $(x_1, x_2) = (\frac{1}{2}, -\frac{1}{3})$ is otter related? Use the conventions for writing odds discussed in class.

    (c) You add a new feature to your model and find that its coefficient is $\beta_3 = 2$. Supposing that you have already collected the features $(x_1, x_2) = (1, 1)$ and classified it using the reduced model introduced above, which uses only two features, what values of x_3 would cause you to change your classification from otter-related to buffalo-related or vice-versa?

Additional Workspace

23. (15 points) Suppose you use statsmodels OLS to perform a simple linear regression of the form $y = \beta_0 + \beta_1 x$ on data consisting of $n = 40$ observations and obtain the following results:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.133
Model:	OLS	Adj. R-squared:	0.110
Method:	Least Squares	F-statistic:	5.818
Date:	Thu, 03 May 2018	Prob (F-statistic):	0.0208
Time:	18:09:19	Log-Likelihood:	-54.974
No. Observations:	40	AIC:	113.9
Df Residuals:	38	BIC:	117.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.2078	0.304	3.975	0.000	0.593	1.823
x1	1.3155	0.545	?????	?????	?????	?????

Please answer the following questions, and be sure to show your work—for sufficient space, a blank page follows this page.

(a) Give a brief interpretation of the slope parameter in the model in terms of the way that changes in the feature x affect the response y.

(b) Compute the missing 95% confidence interval for the slope parameter.

(c) Based on your CI from part (b), do we have reason to believe that β_1 is different from zero? Explain.

(d) What fraction of the total variation in the response is **NOT** explained by the SLR model?

Additional Workspace

Clauset Classic	Ketelsen Kool	Frongillo Fluff
7	6	5
6	3	6
8	6	5
		8

24. (15 points) Everyone is getting psyched for the algorithms final, looking to earn the coveted Grochow Trophy of Maximum Satisfiability. Of course, everyone includes sea otter Maureen Ferguson, who is trying to decide which fur-care product to use to maximize her saltwater buoyancy. She has three options: Clauset Classic, Ketelson Kool, and Frongillo Fluff. As a data scientist, she conducts 10 independent experiments with the products, and her assistant records her saltwater buoyancy, as shown in the table above. She can, of course, see that the three fur-care products produce different average buoyancies. However, having mastered CSCI 3022, she also remembers that there are ways to test whether the average buoyancies of the three fur-care products are statistically different...

    Unfortunately, she cannot do the calculations herself because she is literally an otter and cannot use Jupyter notebooks or hold a pencil. It is up to you to help her. Be clever like an otter—the Grochow MaxSat Trophy is, for the purposes of this problem, within your reach!

    (a) Name a technique discussed in class that you can use to simultaneously compare the three sets of experimental results and determine whether or not the three fur-care products produce the same average buoyancy, and clearly state your null and alternative hypotheses.

    (b) Use the technique that you named above and perform your analysis at the $\alpha = 0.05$ significance level. Be sure to show your work and state your conclusion in words, as it pertains to Maureen Ferguson's fur-care buoyancy performance.

    (c) What was/were the assumption/s that you made when using your method, and importantly, what is known for this method if the assumption/s is/are violated?

Additional Workspace