CSCI 3022
Final Exam
Fall 2017

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.
- If you need more space for free-response questions, there are blank pages at the end of the exam. If you choose to use the extra pages, make sure to **clearly** indicate which problem you are continuing.

Page	Points	Score
3	8	
4	6	
5	6	
6	6	
7	6	
8	8	
9	15	
11	15	
13	15	
15	15	
For Luck	1	1
Total	100	

Table of 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),

$\Phi(4.95) \approx 1.000$	$\Phi(3.50) \approx 0.999$	$\Phi(3.00) = 0.998$	$\Phi(2.88) = 0.997$	$\Phi(2.70) = 0.996$
$\Phi(2.53) = 0.994$	$\Phi(2.32) = 0.990$	$\Phi(2.28) = 0.980$	$\Phi(2.12) = 0.979$	$\Phi(1.96) = 0.975$
$\Phi(1.80) = 0.960$	$\Phi(1.72) = 0.952$	$\Phi(1.64) = 0.950$	$\Phi(1.44) = 0.925$	$\Phi(1.15) = 0.868$
$\Phi(1.13) = 0.870$	$\Phi(0.90) = 0.830$	$\Phi(0.82) = 0.764$	$\Phi(0.40) = 0.645$	$\Phi(0.38) = 0.638$
$\Phi(0.35) = 0.630$	$\Phi(0.30) = 0.615$	$\Phi(0.19) = 0.575$	$\Phi(0.18) = 0.568$	$\Phi(0.15) = 0.555$
$\Phi(0.13) = 0.525$	$\Phi(0.00) = 0.500$			

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$(1-\alpha, v)$.

$$\begin{aligned} t_{0.05,48} &= 1.677 \\ t_{0.025,48} &= 2.011 \\ t_{0.05,2} &= 2.920 \\ t_{0.025,2} &= 4.303 \end{aligned}$$

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$(1-\alpha, v_1, v_2)$.

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

For the following three questions, assume the following scenario: A weather system is moving into Boulder. Meteorologists have determined that the probability that it will snow today is 0.20, the probability that it will snow tomorrow is 0.75, and the probability that it will snow today **and** tomorrow is 0.10. You are studying for finals in the windowless bowels of CSEL, with only your Ralphie poster to bring you comfort. You know from experience that on any given day, Chris is happy 90% of the time, but on days that it snows he's happy 25% of the time. You also know from experience that on days that it snows, Dan is happy 80% of the time, and on days that it doesn't snow, Dan is happy 70% of the time. **Note**: All multiple choice answers have been rounded to three decimal places.

1. (2 points) Given that it snows today, what is the probability that it will snow tomorrow?

   A. 0.020
   B. 0.100
   C. 0.500
   D. 0.750
   E. 0.950

   snow = 0.20
   tomorrow = 0.10
   today ∩ tomr = 0.10
   $P(C|S) = .25$
   $P(C|S^c) = 0.9$
   $P(T|D) = \frac{P(T \cap D)}{P(D)} = \frac{0.10}{0.20}$

   1. **C**

2. (2 points) Suppose that today you see Chris in CSEL and observe that he is happy. What is the probability that it is snowing?

   A. 0.056
   B. 0.180
   C. 0.200
   D. 0.250
   E. 0.900

   $P(S|C) = \frac{P(C|S)\,P(S)}{P(C)}$
   $= \frac{0.25 \times 0.2}{0.9}$

   2. **A**

3. (2 points) You know that you are going to see Dan later today. Without knowing whether it is snowing outside or not, what is the probability that Dan is happy?

   A. 0.160
   B. 0.300
   C. 0.720
   D. 0.750
   E. 0.775

   $P(D) = P(D|S)P(S) + P(D|S^c)P(S^c)$
   $= 0.8 \times 0.2 + 0.7 \times 0.8$
   $= 0.16 + 0.56 = 0.72$

   3. **C**

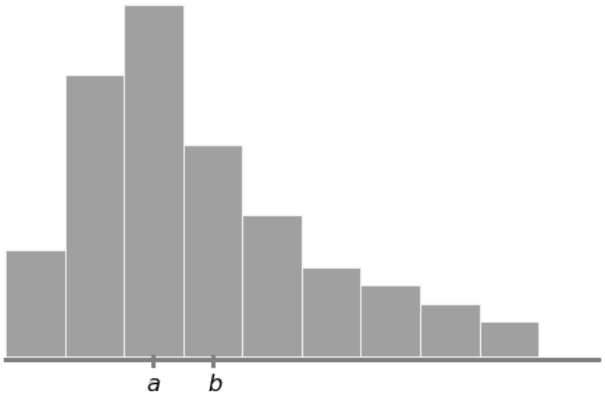
4. (2 points) You have the following 9 samples in a dataset: 2, 1, 21, 13, 8, 1997, 5, 1, 3. Which *one* of the following is true?

   A. The mean diverges
   B. The median is 13
   C. Q1 = 1.5
   D. Q3 = 10.5
   E. The IQR is 11

   1, 1, 2, 3, 5, 8, 13, 21, 1992
   $Q_1 = 2$
   $Q_3 = 13$
   $IQR = 13 - 2 = 11$

   4. **E**

5. (2 points) Which of the following is true about the data represented by the histogram below?



A. **a** is the sample mean and **b** is the sample median

B. b is the sample mean and a is the sample median

C. **a** is the sample variance and **b** is the sample mode

D. **a** and **b** are both technically the sample mean

5. **B**

6. (2 points) Suppose that a temperature is represented by the random variable X, whose units are in Kelvins, with standard deviation 12 Kelvins. Let Y be the equivalent random variable with units converted to degrees Fahrenheit. What is the variance of Y, rounded to the nearest whole number? Recall that the formula for converting a temperature in Kelvin to a temperature in Fahrenheit is $T_F = \frac{9}{5}T_K - 460$

A. 12

B. 467

C. 259

D. -201

E. 7

$s = 12$ $\quad c = \frac{X - 456}{12}$

6. **B**

7. (2 points) Let X be normally distributed with mean 3 and variance 4. What is $P(3 < X < 9)$?

A. 6.0

B. 0.498

C. 0.002

D. 0.4

E. $0.08\bar{3}$

$\frac{3-3}{2} < X < \frac{9-3}{2}$

$0 < x < \frac{6}{2} = 3$

7. **B**

8. (2 points) Consider the following function related to boats entering a harbor. What distribution does the return value of the function belong to?

```
def bbboats(lam):
    x = 0
    t = np.random.exponential(1/lam)
    while t <= 1:
        x += 1
        t += np.random.exponential(1/lam)
    return x
```

A. Binomial

B. Geometric

C. Poisson

D. Uniform

E. Exponential

F. None of the Above

8. **C**

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?

```
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. None of the Above

9. **B**

10. (2 points) Let n be the number of samples that you draw from a population. Which of the following variables of interest is **NOT** well-modeled by a normal distribution.

A. the sample mean from a normal population when $n = 10$ and σ is known

B. the sample mean from a normal population when $n = 1000$ and σ is unknown

C. the sample mean from an exponential population when $n = 1000$

D. the sample variance from a normal population when $n = 1000$

E. the sample proportion when $n = 1000$

10. **D**

11. (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

11. **A**

12. (2 points) Suppose you draw $n = 500$ samples from some distribution and want to test the following hypotheses about the mean μ, $H_0: \mu = \mu_0$ vs $H_1: \mu > \mu_0$. You will reject the null hypothesis if your test statistic is greater than 1.64. What is the significance level α of your test?

A. $\alpha = 0.005$

B. $\alpha = 0.025$

C. $\alpha = 0.01$

D. $\alpha = 0.05$

E. $\alpha = 0.1$

p-val = 1 - 0.95 = 0.05

= 1 - Φ(1.64) = 0.05

12. **D**

13. (2 points) One minus the p-value is: 1 - pvalue

A. the probability that H_1 is true

B. the probability that H_0 is true

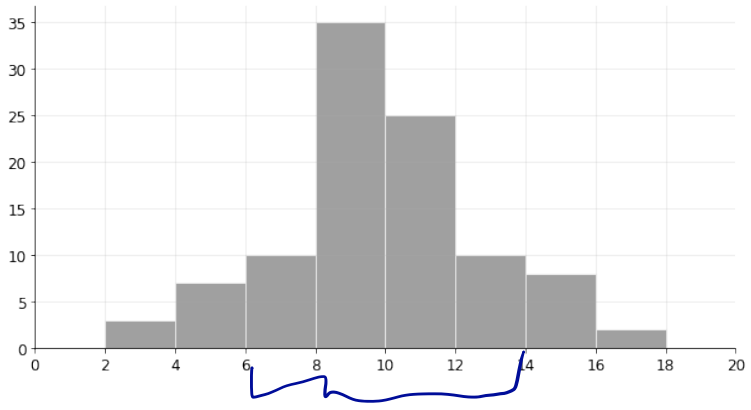
C. the probability, assuming that H_0 is true, that you observe something less extreme than your test statistic

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

13. **C**

14. (2 points) Consider the following histogram of bootstrapped sample means. Which of the following is an 80% bootstrapped confidence interval for the mean?



A. [2, 18]

B. [4, 16]

C. [6, 14]

D. [6, 16]

E. [8, 12]

14. **C**

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. 0.1
D. 0.20

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

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 **narrowest** confidence interval for the mean?

A. $\alpha = 0.05$ and $n = 10$
B. $\alpha = 0.01$ and $n = 10$
C. $\alpha = 0.05$ and $n = 64$
D. $\alpha = 0.01$ and $n = 64$

16. **C**

17. (2 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 1 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 2 is sent on 40 different days and generates an average of 2.25 million page views per day with a standard deviation of a half million views. To learn about the difference (if one exists) between Ad 1's average page views per day and Ad 2's average page views per day, you compute a 95% confidence interval for the difference between the ads' average page views.

Then, your boss, Dr. Anthony Eugene Wong tells you that you can run 10 more days of experiments in order to refine your confidence interval. 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?

A. spend all 10 days testing Ad 1
B. spend all 10 days testing Ad 2
C. spend 5 days testing each
D. spend 8 days testing Ad 2 and 2 days testing Ad 1

$2 \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$

17. **A**

18. (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 $k = 1, 2, 3, 4$

18. **B**

19. (2 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 data point x belongs to Class 1 by

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

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

A. inconclusive
B. $\hat{y} = 0$
C. $\hat{y} = 0.5$
D. $\hat{y} = 1$
E. the limit does not exist

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$= 2 + 1(-1.5) = 0.5$

19. **D**

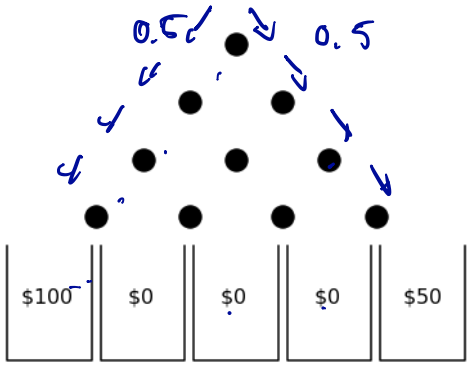
20. (2 points) For the same logistic regression model given in the previous problem, what happens if x increases by one unit?

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

20. **A**

$\frac{1}{1 - e^{-0.5}}$

21. (15 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. Answer the following questions about this Plinko game. Be sure to show your work—for sufficient space, a blank page follows this page.



(a) Let Y be the random variable that describes the winnings when a single disc is dropped. Write down the probability mass function for Y.

(b) What is the probability that you win a total of exactly \$100 in a game with 2 discs?

(c) What is the probability that you win a total of exactly \$200 in a game with 5 discs?

(d) What is the **expected** winnings if you play a game with 5 discs?

a) Probability for bin from left to right

$(1-p)^4$, $4p(1-p)^3$, $4p^2(1-p)^2$, $4p^3(1-p)$, p^4

$P(0) = 4 \cdot \frac{1}{2}\left(\frac{1}{2}\right)^3 + 4\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 + 4\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)$

$= \frac{2}{8} + \frac{4}{16} + \frac{4}{16} = 0.75$

$P(100) = (0.5)^4 = 0.0625$

$P(50) = (0.5)^4 = 0.0625$

Additional Workspace

22. (15 points) Suppose that you have collected 100 samples from a population with known variance $\sigma^2 = 25$ and a sample mean of $\bar{x} = 21$. Please answer the following questions, and be sure to show your work—for sufficient space, a blank page follows this page.

(a) You want to see if there is sufficient evidence to conclude that the true population mean is *different from* $\mu = 20$. State the relevant null and alternative hypotheses.

(b) Construct a confidence interval for the mean at the 95% level. Use your confidence interval to evaluate the hypotheses you stated in part (a) and interpret your result in terms of those hypotheses.

(c) You also want to see if there is sufficient evidence to conclude that the true population mean is *greater than* $\mu = 20$. State the relevant null and alternative hypotheses.

(d) Perform a hypothesis test that involves a p-value at the 99% level, based on the hypotheses you stated in part (c). Use your test to evaluate the hypotheses you stated in part (c) and interpret your result in terms of those hypotheses.

$n = 100$, $\sigma^2 = 25 \Rightarrow \sigma = 5$, $\bar{x} = 21$

(a) $H_0: \mu = 20$

$H_1: \mu \neq 20$

(b) 95% of CI : $\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$, $z_{\alpha/2} = z_{0.05/2} = z_{0.0025} = \Phi(1.96)$

$21 \pm 1.96 \frac{5}{\sqrt{100}} \Rightarrow CI = [20.02, 21.98]$

we fail to reject null hypothesis, we can't conclude that mean is differe from 1.

(c) $H_0: \mu = 20$

$H_1: \mu > 20$

(d) $z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{21 - 20}{5/\sqrt{100}} = 2$

$P = 1 - \Phi(2) = 1 - 0.977 = 0.023 < 0.01$, reject null hypothesys and conclude that there is sufficient evident that true population mean greater than 20

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 $n = 50$ observations and obtain the following results:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.900
Model:	OLS	Adj. R-squared:	0.898
Method:	Least Squares	F-statistic:	377.4
Date:	Sat, 16 Dec 2017	Prob (F-statistic):	2.20e-24
Time:	12:25:46	Log-Likelihood:	-43.114
No. Observations:	50	AIC:	90.23
Df Residuals:	48	BIC:	94.05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6573	0.390	4.249	0.000	0.873	2.441
x	2.0667	0.106	-----	-----	-----	-----

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?

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

(a) For given a unit change in feature x, response y change by 2.0667.

(b) CI: $\beta_1 \pm t_{\alpha/2, n-1} \cdot SS(\beta_1) = 2.0667 \pm t_{0.05, 49} \cdot 0.106$

$= 2.0667 \pm 1.677 \times 0.106$

$CI = [2.049, 2.84]$

(c) Yes, because it does not contain zero.

(d)

Additional Workspace

Shampoo A	Shampoo B	Shampoo C
8	5	7
5	5	5
5	2	4
		4

$\frac{15}{3} = 6$ $\quad \frac{12}{3} = 4$ $\quad \frac{20}{4} = 5$

24. (15 points) Everyone is trying to look their best for Homecoming Weekend in hopes that the Buffs win the match. Of course, everyone includes Ralphie, who is trying to decide which shampoo to use to maximize the silkiness of her coat. She has three options: Shampoo A, Shampoo B, and Shampoo C. As a data scientist, she conducts 10 independent experiments with the shampoos, and her handlers record the silkiness of her coat, as shown in the table above. She can, of course, see that the three shampoos produce different average silkinesses. However, having taken CSCI 3022, she also remembers that there are ways to test whether the average silkinesses of the three shampoos are statistically different.

Unfortunately, she cannot do the calculations herself because she is literally a buffalo and cannot use Jupyter notebooks or hold a pencil. It is up to you to help her. Be quick like a buffalo—the homecoming weekend is, for the purposes of this problem, rapidly approaching.

(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 shampoos produce the same average silkiness, 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 Ralphie's shampoo performance.

(a) we use Anova to determine whether or not the shampoos produce the same average

$H_0: \mu_A = \mu_B = \mu_C$

$H_0: \mu_i \neq \mu_j$ for $i \neq j$

$F = \frac{SSB/df_{SSB}}{SSW/df_{SSW}} = \frac{6/2}{18/7} = \frac{3}{2.57} = 1.167$

$df_{SSB} = I - 1 = 2$, $df_{SSW} = N - I = 10 - 3 = 7$

$\bar{\bar{y}} = \frac{1}{10}(8+5+5+5+5+2+7+5+4+4) = 5$

$SSB = 3(6-5)^2 + 3(4-5)^2 + 4(5-5)^2 = 6$

$SSW = [(8-6)^2 + (5-6)^2 + (5-6)^2] + [(5-4)^2 + (5-4)^2 + (2-6)^2$
$+ [(7-5)^2 + (5-5)^2 + (4-5)^2 + (4-5)^2$
$= 4 + 1 + 1 + 2 + 4 + 4 + 1 + 1 = 18$

Additional Workspace