CSCI 3022
Final Exam
Summer 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 **one** 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 **75 minutes** for this exam.

Page	Points	Score
3	12	
4	9	
5	9	
6	9	
7	20	
9	20	
11	20	
For Luck!	1	1
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),

$$\begin{array}{lllll}
\Phi(4.75) \approx 1.000 & \Phi(3.00) = 0.999 & \Phi(2.58) = 0.995 & \Phi(2.32) = 0.990 & \Phi(2.00) = 0.977 \\
\Phi(1.96) = 0.975 & \Phi(1.88) = 0.970 & \Phi(1.80) = 0.964 & \Phi(1.75) = 0.960 & \Phi(1.64) = 0.950 \\
\Phi(1.44) = 0.925 & \Phi(1.28) = 0.900 & \Phi(1.15) = 0.875 & \Phi(1.04) = 0.850 & \Phi(1.00) = 0.841 \\
\Phi(0.93) = 0.825 & \Phi(0.84) = 0.800 & \Phi(0.76) = 0.775 & \Phi(0.67) = 0.750 & \Phi(0.60) = 0.725 \\
\Phi(0.52) = 0.700 & \Phi(0.45) = 0.675 & \Phi(0.39) = 0.650 & \Phi(0.32) = 0.625 & \Phi(0.25) = 0.600 \\
\Phi(0.19) = 0.575 & \Phi(0.13) = 0.550 & \Phi(0.06) = 0.525 & \Phi(0.00) = 0.500 &
\end{array}$$

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$(1-\alpha, v)$.

$$\begin{array}{rcll}
t_{0.05,48} & = & 1.677 & t_{0.025,48} = 2.011 \\
t_{0.05,44} & = & 1.680 & t_{0.025,44} = 2.015 \\
t_{0.05,9} & = & 1.833 & t_{0.025,9} = 2.262 \\
t_{0.05,2} & = & 2.920 & t_{0.025,2} = 4.303
\end{array}$$

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

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

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

   1. **D**

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

   2. **D**

3. (3 points) Let X be normally distributed with mean 1 and variance 9. What is $P(-2 < X < 1)$?

   A. 6.0
   B. 0.341
   C. 0.136
   D. 0.4
   E. 0.819

   3. **B**

4. (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 α and sample size n produces the **widest** confidence interval for the mean?

   A. $\alpha = 0.01$ **and** $n = 16$
   B. $\alpha = 0.01$ and $n = 25$
   C. $\alpha = 0.05$ and $n = 16$
   D. $\alpha = 0.05$ and $n = 25$
   E. $\alpha = 0.1$ and $n = 16$
   F. $\alpha = 0.1$ and $n = 25$

   4. **A**

5. (3 points) Consider the following poorly-named function. The function output constitutes a sample from which one of the following distributions?

```
def worst_function_name_ever(n_sample, mu, sigma):
    X = []
    for k in range(n_sample):
        X.append(9*np.var(stats.norm.rvs(mu, sigma, size=10), ddof=1)/(sigma**2))
    return X
```

A. T
B. Uniform
C. Exponential
D. Normal
E. Chi-squared

5. **E**

6. (3 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 parking_problems(p):
    x = 1
    while np.random.choice([0,1], p=[1-p, p]) == 0:
        x += 1
    return x
```

A. Binomial
B. Poisson
C. Geometric
D. Uniform
E. Exponential

6. **C**

7. (3 points) A manufacturer produces trinkets targeted to weigh 52 g. The weight of the trinkets is known to be normally distributed, but an engineer is concerned that the variation in the produced trinkets is larger than acceptable. In an attempt to estimate the variance, she selects $n = 10$ trinkets at random and weighs them. The sample yields a sample variance of 4 g^2. Which of the following gives a 90% CI for the variance?

A. $\frac{9 \cdot 4}{\chi^2_{0.05,9}} \leq \sigma^2 \leq \frac{9 \cdot 4}{\chi^2_{0.95,9}}$

B. $\frac{9 \cdot 4}{-\chi^2_{0.05,9}} \leq \sigma^2 \leq \frac{9 \cdot 4}{\chi^2_{0.05,9}}$

C. $4 - \chi^2_{0.05,9} \cdot \sqrt{\frac{4}{10}} \leq \sigma^2 \leq 4 + \chi^2_{0.95,9} \cdot \sqrt{\frac{4}{10}}$

D. $4 - t_{0.05,9} \cdot \sqrt{\frac{4}{10}} \leq \sigma^2 \leq 4 + t_{0.05,9} \cdot \sqrt{\frac{4}{10}}$

E. $52 - t_{0.05,9} \cdot \sqrt{\frac{4}{10}} \leq \sigma^2 \leq 52 + t_{0.05,9} \cdot \sqrt{\frac{4}{10}}$

7. **A**

8. (3 points) You would like to test the null hypothesis $H_0 : \mu = 2$ against the alternative hypothesis $H_1 : \mu > 2$. You decide to use the $\alpha = 0.05$ significance level. Suppose you know the population standard deviation is some σ_0. You draw a sample from the population distribution, and calculate a sample mean of $\bar{x} = 2.2$, with a corresponding p-value of $p = 0.07$. Assume all calculations have been done correctly.

   If you repeat this experiment 100 times, in about how many of these samples do you expect to find the sample mean $\bar{x} \geq 2.2$, assuming that the null hypothesis is true?

   A. 5

   B. 7

   C. 50

   D. 93

   E. 95

   8. **B**

9. (3 points) Suppose you generate 10,000 confidence intervals for the mean of a population, using fixed significance level α. You discover that 248 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. 0.20

   9. **B**

10. (3 points) Consider performing a multiple linear regression on a data-set with full and reduced models of the form

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

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

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

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

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

    D. $\beta_1 = \beta_3 = 0$

    E. $\beta_2 = \beta_4 = 0$

    10. **E**

11. (3 points) Suppose that you are performing a binary logistic regression 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 $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = -3$. If $p(y = 1 \mid x) \geq 0.5$, you classify as Class 1, otherwise, you classify as Class 0. How would your model classify a data point with $x = 1$?

$$\text{Recall:} \quad \text{sigm}(z) = \frac{1}{1 + e^{-z}}$$

A. inconclusive

B. Class 0

C. $\hat{y} = 0.5$

D. Class 1

E. the limit does not exist

11. **B**

12. (3 points) For the same logistic regression model given in the previous problem, what is the decision boundary?

A. $x = -1/2$

B. $x = 1/2$

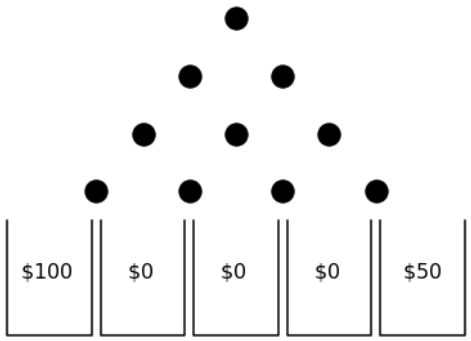
C. $x = 0$

D. $x = -1/3$

E. $x = 1/3$

12. **E**

13. (3 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 expected value of your winnings with a single disc?



A. 0.0625 (= 16/256)

B. 6

C. 8.125 (= 130/16)

D. 9.375 (= 150/16**)**

E. 30

F. None of the above.

13. **D**

14. (20 points) Suppose that you have collected n samples from a population known to be normal, but with unknown variance. The sample mean is $\bar{x} = 2$ and the sample standard deviation is $s = 5$.

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

(a) You want to see if there is sufficient evidence to conclude that the true population mean is *different from* $\mu = 1$. State the relevant null and alternative hypotheses.

Solution:

$$H_0 : \mu = 1$$
$$H_1 : \mu \neq 1$$

(b) Suppose $n = 10$. Construct a confidence interval for the mean at the 95% level ($\alpha = 0.05$). Use your confidence interval to evaluate the hypotheses you stated in part (a) and interpret your result in terms of those hypotheses.

Solution:
(NB: Since $n = 10$ and we know the population we are sampling from is normal, the T test is appropriate with $n - 1 = 9$ degrees of freedom.)
A 95% confidence interval is given by $\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$. We have:

$$\bar{x} = 2, \qquad t_{\alpha/2,n-1} = t_{0.025,9} = 2.262, \qquad s = 5, \qquad \text{and} \quad \sqrt{n} = \sqrt{10}$$

That gives:

$$CI = 2 \pm 2.262 \cdot \frac{5}{\sqrt{10}} = 2 \pm 3.577 = \boxed{[-1.577, \quad 5.577]}$$

Since our confidence interval contains $\mu = 1$, we **fail to reject** the null hypothesis, and cannot conclude that the mean is different from 1.

(c) Now suppose $n = 81$. Construct a confidence interval for the mean at the 95% level ($\alpha = 0.05$). Use your confidence interval to evaluate the hypotheses you stated in part (a) and interpret your result in terms of those hypotheses.

Solution:
(NB: Since $n = 81$ and we know the population we are sampling from is normal, the Central Limit Theorem applies.)
A 95% confidence interval is given by $\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$. We have:

$$\bar{x} = 2, \qquad z_{\alpha/2} = z_{0.025} = 1.96, \qquad s = 5, \qquad \text{and} \quad \sqrt{n} = \sqrt{81} = 9$$

That gives:

$$CI = 2 \pm 1.96 \cdot \frac{5}{9} = 2 \pm 1.089 = \boxed{[0.911, \quad 3.089]}$$

Since our confidence interval contains $\mu = 1$, we **fail to reject** the null hypothesis, and cannot conclude that the mean is different from 1.

(d) You also want to see if there is sufficient evidence to conclude that the true population mean is *greater than* $\mu = 1$. State the relevant null and alternative hypotheses.

Solution:

$$H_0 : \mu = 1$$
$$H_1 : \mu > 1$$

(e) Suppose $n = 81$. Perform a hypothesis test that involves a p-value at the 90% level ($\alpha = 0.1$), based on the hypotheses you stated in part (d). Use your test to evaluate the hypotheses you stated in part (d) and interpret your result in terms of those hypotheses.

Solution:

(NB: Since $n = 81$ and we know the population we are sampling from is normal, the Central Limit Theorem applies.)

Our test statistic is:

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{2-1}{5/\sqrt{81}} = \frac{1}{5/9} = 9/5 = 1.8$$

Since we are testing the alternative hypothesis $\mu > 1$, our p-value is the total probability mass to the **right** of our test statistic. This gives:

$$p = 1 - \Phi(1.8) = 1 - 0.964 = 0.036$$

Since our p-value $< \alpha = 0.1$, we **reject** the null hypothesis, and conclude that the mean is greater than 1 at the 10% significance level.

15. (20 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 $n = 45$ observations and obtain the following results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.105
Model:                            OLS   Adj. R-squared:                  0.084
Method:                 Least Squares   F-statistic:                     5.030
Date:                Sat, 36 Dec 2023   Prob (F-statistic):             0.0301
Time:                        12:25:46   Log-Likelihood:                -150.35
No. Observations:                  45   AIC:                             304.7
Df Residuals:                      43   BIC:                             308.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.05       0.95]
------------------------------------------------------------------------------
const          74.8106      2.180      2.207      0.033      71.148      78.473
x              0.8608       0.384      ?????      ?????       ?????       ?????
==============================================================================
```

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

(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.

Solution:
For a unit change in the feature, x, the response y changes by about 0.8608, on average.

(b) Compute the missing 90% confidence interval for the slope parameter.

Solution:
The confidence interval is a 90% t CI, given by: $\hat{\beta}_1 \pm t_{\alpha/2,n-1} \cdot SE(\hat{\beta}_1)$

The table gives $\hat{\beta}_1 = 0.8608$ and $SE(\hat{\beta}_1) = 0.384$, there are $n = 45$ data points, and we have $\alpha = 0.1$ for a 90% CI.

This yields:

$$\begin{aligned} CI &= 0.8608 \pm t_{0.05,44} \cdot 0.384 \\ &= 0.8608 \pm 1.680 \cdot 0.384 \\ &= \boxed{[0.216, \quad 1.506]} \end{aligned}$$

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

Solution:
Yes, because the CI does not contain 0. (at the 90% confidence level)

(d) What fraction of the total variation in the response is explained by the SLR model? Fully justify your response.

Solution:
The coefficient of determination is defined to be the fraction of variation that is explained by the model. This is $\boxed{R^2 = 0.105}$

(e) Suppose this linear regression model relates the feature X = "coffee consumption by a student" (in cups) and response Y = "exam scores". Use your previous answers and the output above to evaluate the **strength** and **significance** of the relationship between coffee consumption and exam scores. Fully justify your response.

Solution:
Parts (b) and (c) suggest that there **is a significant relationship** between coffee and exam scores.

Buuuuuuut Part (d) indicates that **this relationship is quite weak** because the SLR model using coffee consumption as a feature does not explain much of the variation in exam scores.

You can also note that the slope coefficient $\hat{\beta}_1$ is quite small, but an argument based on R^2 is much better.

Dr. A	Dr. B	Dr. C
3	5	6
4	5	8
5	7	10
	7	

16. (20 points) Fall is rapidly approaching, and you need to decide which of three professors to take a particular data science course with. Your options are Dr. Anthony (A), Dr. Brian (B) and Dr. Chris (C). You take a brief survey of 10 of your friends who have taken classes previously with these three professors, and have them quantify how much fun they had on a scale of 0 (no fun) to 10 (lots of fun). The survey results are above.

    You may assume that your friends' responses are all independent of one another. Your fond memories of CSCI 3022 suggest that a **one-way ANOVA** is an appropriate way to test whether the mean fun for the three professors' classes are statistically different. Be sure to **show all work** for any problems involving calculations.

    (a) Clearly state the null and alternative hypotheses for the one-way ANOVA test to compare the three sets of experimental results and determine whether or not the three professors produce the same mean fun.

    Solution:
    Let μ_A, μ_B and μ_C correspond to the mean fun from Professors A, B and C, respectively.

    $$H_0 : \mu_A = \mu_B = \mu_C$$
    $$H_1 : \mu_i \neq \mu_j \text{ for some } i \neq j$$

    (b) Compute the relevant **test statistic** for the one-way ANOVA to test the hypotheses from part (a). Put a $\boxed{\textbf{box}}$ around your answer for the test statistic.

    Solution:
    For this ANOVA we have $I = 3$ groups and $N = 10$ total observations.
    The test statistic for the one-way ANOVA is

    $$F = \frac{SSB/df_{SSB}}{SSW/df_{SSW}}$$

    where we have $df_{SSB} = I - 1 = 2$ and $df_{SSW} = N - I = 10 - 3 = 7$.

    The group means are: $\bar{y}_A = 4, \quad \bar{y}_B = 6, \quad \bar{y}_C = 8$
    And the grand mean is: $\bar{\bar{y}} = \frac{1}{10}(3 + 4 + 5 + 5 + 5 + 7 + 7 + 6 + 8 + 10) = \frac{60}{10} = 6$

    We calculate the sums of squares:

    $$\begin{aligned} SSB &= \sum_{i=1}^{3} n_i(\bar{y}_i - \bar{\bar{y}})^2 \\ &= 3(4-6)^2 + 4(6-6)^2 + 3(8-6)^2 \\ &= 3 \cdot 4 + 4 \cdot 0 + 3 \cdot 4 \\ &= 24 \end{aligned}$$

$$
\begin{aligned}
SSW &= \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 \\
&= [(3-4)^2+(4-4)^2+(5-4)^2]+[(5-6)^2+(5-6)^2+(7-6)^2+(7-6)^2]+ \\
&\qquad [(6-8)^2+(8-8)^2+(10-8)^2] \\
&= [1+0+1]+[1+1+1+1]+[4+0+4] \\
&= 14
\end{aligned}
$$

So the test statistic is:

$$F = \frac{24/2}{14/7} = \frac{12}{2} = \boxed{6}$$

(c) What distribution does this test statistic follow, including any relevant degrees of freedom?

Solution:
The test statistic F follows an F distribution with numerator degrees of freedom $df_{SSB} = 2$ and denominator degrees of freedom $df_{SSW} = 7$. So:

$$F \sim F(2,7)$$

(d) Using the $\alpha = 0.05$ significance level, what is/are the critical value(s) that bound the rejection region?

Solution:
The **one** critical value is $F_{0.05,2,7} = 4.737$

(e) State your conclusion for this hypothesis test in words, as it pertains to your decision for which professor's class to take. Fully justify your response.

Solution:
Our test statistic is $F = 6 > F_{0.05,2,7} = 4.737$, so we **reject** the null hypothesis and conclude that there is *some* significant (at the 5% level) difference between the mean amounts of fun produced by the three professors.