Fall 2022 Midterm 2

The exam is closed book and closed notes. You are allotted one double sided "cheat sheet" which may contain typed or handwritten notes. You may also use a calculator. Your phone is not allowed as a calculator. Using any resources outside of the aforementioned items is strictly prohibited.

While you take the exam, you are prohibited from discussing the test with anyone. If you are taking the test after your classmates, you are also prohibited from talking to them about the test before you take it. Evidence of cheating may result in a 0 on the exam and be reported to the Student Conduct Board.

Berkeley's code of conduct is here: https://sa.berkeley.edu/code-of-conduct. See Section V and Appendix II for information about how UC Berkeley defines academic misconduct. In particular note the sections on cheating and plagiarism.

UC Berkeley Honor Code

"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others." Please carefully read the statements below, and indicate your understanding and intent to adhere to the UC Berkeley Honor code by typing your name in the space below. I agree not to engage in any of the following behaviors:

- Copying or attempting to copy from others during an exam or on an assignment.
- Communicating answers with another person during an exam.
- Pre-programming a calculator or other personal electronic device to contain answers, or using other unauthorized information for exams.
- Using unauthorized materials, i.e. prepared answers.
- Allowing others to do an assignment or a portion of an assignment for you, including the use of a commercial term-paper service.
- Submitting the same assignment for more than one course without prior approval of all the instructors involved.
- Collaborating on an exam or assignment with any other person without prior approval from an instructor.
- Taking an exam for another person or having someone take an exam for you.
- Altering a previously graded exam or assignment for the purpose of a grade appeal or of gaining points in a re-grading process.
- Submitting an electronic file the student knows to be unreadable or corrupted instead of a completed assignment.

Write your name and SID below.

Enter your name: □

Enter your SID: □

INSTRUCTIONS:

Hand write your responses using a pencil or pen in the space provided. Use only the space provided for your questions and clearly label your final answer. Do not write answers on the back of the exam. Any additional space, including the back, may be used as scratch paper but will not be graded.

Phones should be turned OFF prior to the start of the exam and secured in your backpack or another secure location. Do not leave your phone or other electronic devices out. If you need to leave the room for any reason during the exam please flag a GSI and let them know prior to exiting the room. Time will still accrue when you leave the room.

The length of the midterm is 50 minutes. If you finish early and are satisfied with your work you may leave early. Hand in your exam to a GSI, who will verify that they received it.

Exam Format:

Short Format Question: 1a, 1b, 1c, 1d, 2, 3, 4a, 4b, 5, 6 , 7 [12 points]
Quick Response Grouped: 8a, 8b, 8c, 8d[2 points]
Question on Research Intern Phone Calls: 9a, 9b [6 points]
Question on Medicare Fraud Detection: 10a, 10b [2 points]
Question on Syndrome A and Sampling Distributions: 11a, 11b, [4 points]
Question on Lifestyle Counseling, Hypothesis Testing, and Inference: 12a, 12b, 12c, 12d [5 points]
Optional Feedback Question: 13 [0 points] **Total Points**: 31 points

1a-c) Fill in the correct boxes below the sentences with the correct terms. [3 points]

Probability always begins with defining the ______________[a]________________, which is the set or collection of all possible outcomes. A/An ____________[b]__________________ is one possible outcome or a set of outcomes from the collection of all possible outcomes. We then need to define the ___________[c]____________________ which defines the probability associated with different events.

Possible terms: complement, event, statistical parameter, probability model, random variable, sample space, statistical assumptions

- 1a. ____
- 1b. ____
- 1c. ____

1d) True or False: If A and B are independent, then $P(A|B) = P(B|A)$. [1 point]

□ A. True
□ B. False

```
#!a: sample space
#1b: event
#1c: probability model
#1d: False -- If A and B are independent, then P(A|B) = P(A) and P(B|A) = P(B)
```

2) Which one of these variables is a binomial random variable? [1 point]

□ A. time it takes a randomly selected person to recover from COVID
□ B. number of textbooks a randomly selected student bought this semester
□ C. number of people taller than 70 inches in a random sample of 15 people
□ D. number of cups a randomly selected person owns

```
#Answer - C the number of people taller than 68 inches in a random sample of 15 people
```

3) A normal distribution can be fully described by its _________ and __________. [1 point]

□ A. range, mode
□ B. median, maximum
□ C. mean, median
□ D. standard deviation, mean

```
#Answer = D. standard deviation, mean
```

4) Suppose we draw a random sample of size n from a population with a mean of μ and a standard deviation σ. When we increase n, the mean of the sampling distribution of the mean, $\bar{x}$ is ________(a)____________ μ and $\bar{s}$, the standard derivation of the sampling distribution of the mean gets ________(b)___________ σ. [2 points]

Fill in the blanks with one of the following: *smaller than*, *larger than*, *on average gets closer to*

- 4a. ____
- 4b. ____

```
#Answer - 4a  on avergae gets closer to
#Answer - 4b  smaller than
```

5) A student found that it takes 60 minutes on average to complete a specific programming exercise. The standard deviation for this task is 10 minutes and the times for completing this exercise are normally distributed. What is the probability that on any given day it will take 70 minutes or more to complete the exercise? Select the R code that can calculate this probability. [1 point]

□ A.`qnorm(q = 70, mean = 60, sd = 10, lower.tail = F)`

□ B.`pnorm(q = 70, mean = 60, sd = 10, lower.tail = T)`

□ C.`pnorm(q = 70, mean = 60, sd = 10, lower.tail = F)`

□ D.`qnorm(q = 70, mean = 60, sd = 10, lower.tail = T)`

```
# Solution
# C
```

6) Select all of the following statements that are true for a Poisson distribution: [2 points]

□ A. It describes the count of occurrences of a defined event in fixed, finite intervals of time or space.
□ B. Is used as a model for continious random variables
□ C. The mean is equal to µ and standard deviation is sqrt(µ).
□ D. The occurrences have to be independent, and the probability of an occurrence is the same over all possible intervals.
□ E. There is a fixed upper bound to the distribution

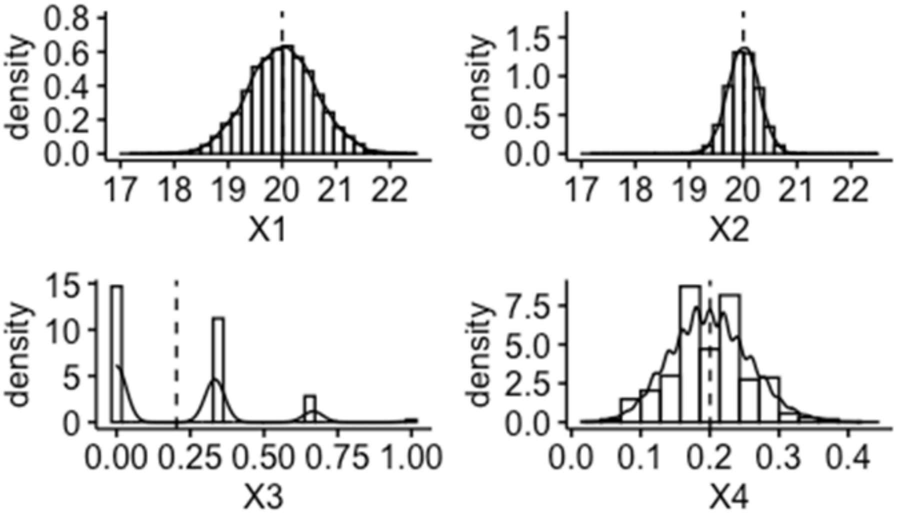
```
# solution: a, c, d
```

7) When our p-value is large is it correct to conclude that our null hypothesis is correct. [1 point]

□ True
□ False

```
#False, we fail to reject the null hypothesis
```

8) Below are plots of the sampling distribution of sample means in four different scenarios. Connect the following plot (using X1, X2, X3, X4) to the description of how the data was generated. [2 points]



_____ a. $n = 50, X \sim N(20, 2)$ An average based on sampling 50 independent observations.

_____ b. $X \sim \text{binom}(n = 50, p = 0.2)$ with the average being the same as the proportion of these 50 draws, or the sampling distribution of $\hat{p} = \frac{X}{n}$

_____ c. $X \sim \text{binom}(n = 3, p = 0.2)$, with the average being the same as the proportion of these 3 draws, or the sampling distribution of $\hat{p} = \frac{X}{n}$

_____ d. $n = 10, X \sim N(20, 2)$ An average based on sampling 10 independent observations.

```
#answers: A = X2 B=X4, C=X3, D=X1
```

Question 9. A mostly dedicated intern working on a clinical trial team is decidcated 90% of days at work and is lazy the other 10% of days. When they are dedicated, the intern calls 20 people and aks them to fill out a survey. When the intern is lazy, they only call 10 people. When they call, each person has a 20 % chance of agreeing to participate. Each person's response is independent of other responses.

9a) Fill out the following values. You can leave these values as mathematical expressions without multiplying them out. Example: .5^10. If you choose to multiply these values out then present your answer as a percent and round to 2 decimal places. [4 points]

A. P(Intern is dedicated that day) = ______________________________

B. P(Intern is lazy at least 1 day in the next 5 days) = __________________________

C. P(Intern is dedicated and only one person takes the survey that day) = __________________________

D. P(Intern is lazy and only one person takes the survey that day) = ________________________

```
#Answers, 1 point each
#P(\text{Dedicated})$ = .9000
#P(Intern is lazy at least 1 day in the next 5 days = 1 - Intern is dedicated all 5 days
# = 1 - .9^5 = 1 - 40.95%
# P(Intern is dedicated that day and only one person takes the survey that day)
# = .9*choose(20, 1) * (1/5)*(4/5)^19 = .0519 = 5.19%
# P(Intern is lazy that day and only one person takes the survey that day)
# = 0.1*choose(10, 1)*(1/5)*(4/5)^9 = 0.0268 = 2.68%
```

9b) One day, exactly 1 person agrees to take the survey. What is the probability the intern was lazy that day? Show your work and present your answer as a percent rounded to 2 decimal places. [2 points]

```
# P(lazy|one person takes the survey) =
# P(one person takes the survey and lazy)/P(one person takes the survey)
# = 0.3405 = 34.05%

#P(lazy | 1 person takes the survey) =
# P(one person takes the survey | lazy) * p(lazy)/p(one person takes the survey)
#                                    = .1 * .2685 / (.00139) = .3405 = 34.05%

#Give points for 7b if calculation for 7a was incorrect but method was correct for 7b.
```

Question 10. The Centers for Medicare and Medicaid Services have developed simple but sophisticated algorithms to detect rates of Medicare fraud among providers. The following tree digram describes detection of Medicare fraud within a particular physician group. The two variables of interest are:

- $P(F = \{\text{no, yes}\})$ which indicates if a billed treatment is fraudulent or not.
- $P(I = \{-, +\})$ where $-$ indicates that the treatment is not flagged as fraudulent by the algorithm and $+$ indicates a treatment is flagged as fraudulent by the algorithm

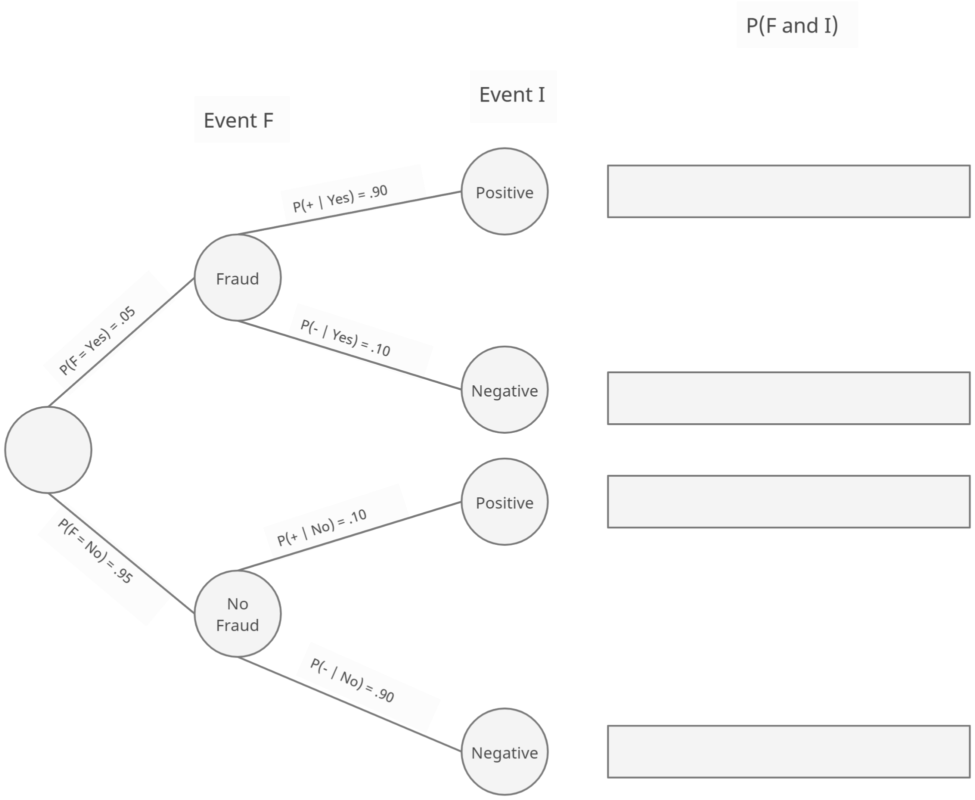


Figure 1: Tree Diagram for Insurance Fraud

Note: The boxes are available for you to fill them in, but they will not be graded.

10a) What is the sensitivity of the algorithm in finding fraud? [1 point]

□ A. 4.50 %
□ B. 10.00 %
□ C. 90.00 %
□ D. 99.41 %

```
# Solution:
# C.
```

10b) What is the marginal probability of the algorithm being negative for fraud? $\mathbb{P}(I = -)$ [1 point]

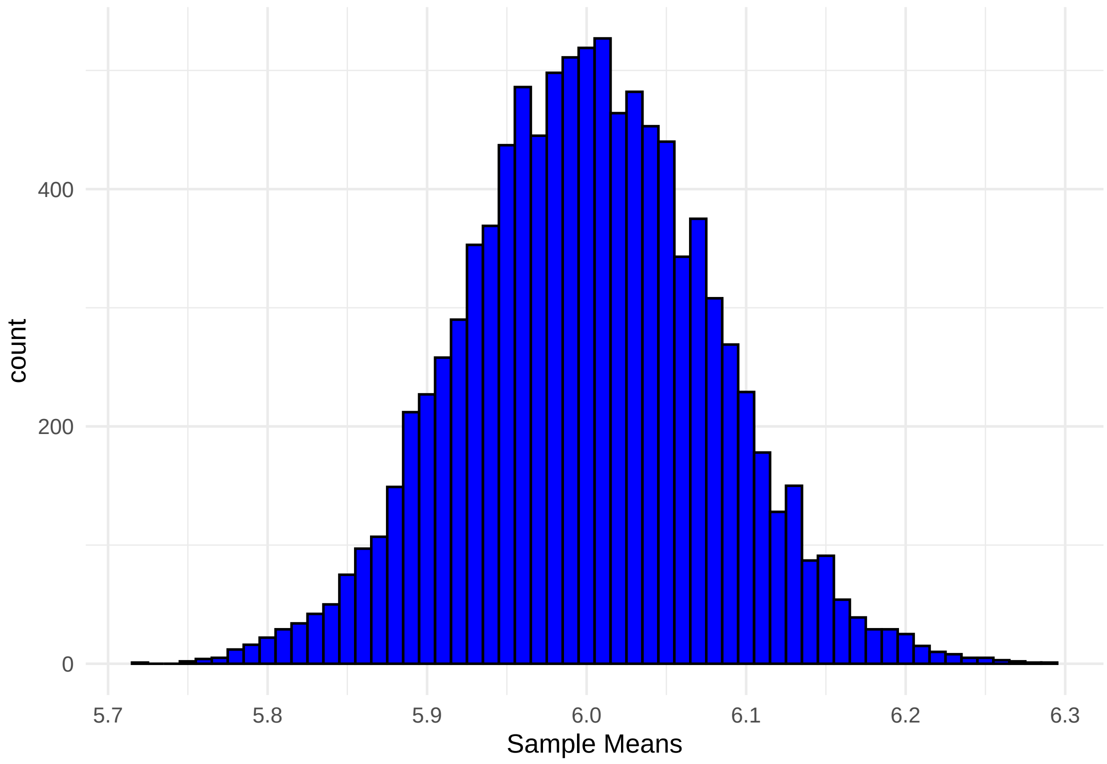
□ A. .860
□ B. .855
□ C. .460
□ D. .140

```
# Solution:
# A.
```

Question 11: A researcher is looking into modeling the occurrence of Syndrome A in the United States. They found that an average occurrence rate of 6 babies per month are reported.

11a) We take a random sample of size n=1,000 from data $X \sim Pois(6)$ and we calculate the mean of this sample; we repeat this step B=10,000 times. What is the approximate distribution of these samples? Specify the theorem that states this? (Just provide the name of the distribution and the name of the theory) [2 point]

```
# Solution:
# Normal
```



11b) Pictured above is a histogram showing the 10,000 sample means. What plots or concepts might better assess if these data follows a normal distribution? Select all that apply. [2 points]

- □ Bar Chart
- □ QQ Plot
- □ the 68-95-99 rule
- □ Statistical Power

```
#QQplot plot and the 68-95-99 rule
```

Question 12: A new set of protocols hope to increase the number of minutes patients received nurse-guided lifestyle counseling among patients at risk for hypertension. 45 patients were enrolled in a trial where clinical staff were trained in this new protocol. Prior information says that the amount of counseling recieved by at-risk patients was 4 minutes per patient.

12a) Using standard notation, rephrase the statement above specifying the null and alternative hypotheses. [1 point]

H_0: ____

H_A: ____

```
#H_0 = H_A = 4 minutes
#H_A > 4 minutes
```

12b) Patients enrolled in the trial received an average of 6.75 minutes of counseling and researchers generated a 95 % confidence interval of (6.5, 7.0). What is the best interpretation of this confidence interval? [1 point]

- □ A. There is a 95% probability that the mean is between (6.5, 7.0)
- □ B. If we repeated this procedure over and over 95 % of confidence intervals would capture the true mean. (6.5, 7.0) is one of those intervals.
- □ C. 95 % of all data values in the population fall between (6.5, 7.0)
- □ D. We are 95% confident that our confidence interval of (6.5, 7.0) is from a large enough sample to ensure that our results are statistically significant.

```
#Answer: B.
```

12c) Suppose that our confidence interval is again (6.5, 7.0) as above. Do we have evidence to reject the null hypothesis given ths information? Provide a 1 sentence explanation. [1 point]

12d) Suppose the researchers had made a type I error. What would this mean in the context of this question? [2 points]

```
#Answer: A type I error occurs when a researcher rejects a null hypothesis that is actually true.
#In this context, the researcher would conclude that there is a difference between
#the new protocol and the old protocol when in reality there was not a difference.
```

13) Please provide feedback on the exam on any issues you might have experienced here.