Fall 2021 Midterm II

The exam is open book. This means you can use electronic or hard copies of all class materials and datahub if you wish. You may not use the internet to search for the answers or inform your answers. Using the internet is strictly prohibited and any evidence of this may result in a 0 on the exam.

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 may be reported to the Student Conduct Board.

Type your initials to affirm that you have read and agree to the above statements:

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 the sections on cheating and plagiarism).

Problem 1: 5 points

Problem 2: 5 points

Problem 3: 6 points

Problem 4: 6 points

Problem 5: 5 points

Problem 6: 5 points + **1 POINT BONUS**

Total: 32 points + 1 possible point bonus

Question 1 [5 points total]

A research team conducted a study to determine the relationship between an individual's score on a cognitive risk assessment test and developing Alzheimer's disease. Assume that the subjects did not know each other at the time of taking the assessment and that none of the subjects are related to each other. Individual scores and whether or not they developed Alzheimer's is aggregated in the table below.

Score	Develops Alzheimer's (A)	Did NOT Develop Alzheimer's (A')	Total
0-1	0	3	3
1-2	3	14	17
2-3	9	25	34
3-4	14	11	25
4-5	13	5	18
5+	3	0	3
Total	42	58	100

1.a. [2 points] Find the following probabilities and *show your work*. Provide your answer as proportions rounded to two decimal places.

P(A) =

P(A ∩ score 2-3) =

P(score 2-3 | A) =

P(A' ∪ score 0-1) =

```
# SOLUTION:

# P(A) = 42 / 100 = 0.42

# P(A $\cap$ score 2-3) = 9 / 100 = 0.09

# P(score 2-3 | A) = 9 / 42 = 0.21

# P(A' $\cup$ score 0-1) = (58/100) + (3/100) - (3/100) = 0.58

# Rubric: 0.50 points for each probability correct, -0.25 for no work shown (for each part)
```

1.b. [2 points] Is developing Alzheimer's disease independent of an individual's cognitive risk assessment score? Explain why or why not in a sentence and prove your reasoning mathematically.

Your answer:

```
# SOLUTION:
# No; if these two events were independent, then we would expect the conditional probability of
# a randomly chosen person getting Alzheimer's to be the same across all cognitive risk assessment scor
# (the value should just be equal to the probability of choosing a person who develops Alzheimer's, or
# However, as shown below, this is not the case:

# P(A) = 0.42
# P(A | 0-1) = 0
# P(A | 1-2) = 3/17 = 0.176

# Note: Students can show independence in multiple ways -- they can compare the probability of developi
# by comparing one with the probability of getting Alzheimer's.

# 0.5 points for explanation, 0.5 points for mathematical proof
```

From this point onward, assume the researchers found that the data they collected is representative for all 50 year old individuals in the U.S. That is, these data arose from a population-based sample.

1.c. [1 point] If the team randomly picks 15 people from the U.S. population of 50 year old individuals, what is the probability that at least one of them has Alzheimer's? Do not round your answer.

Your answer:

```
# SOLUTION:
# P(at least 1 has Alzheimer's) = P(1 has Alz) + P(2 has Alz) + P(3 has Alz) + ... + P(15 has Alz)
# P(at least 1 has Alz) = 1 - P(0 have Alz)
# P(no Alzheimer's) = 1 - .42 = .58
# P(0 of the 15 having Alzheimer's) = .58 ^ 15
# 1 - .58 ^ 15 = P(at least 1 has Alzheimer's) = 0.9997172

# can also figure this out using pbinom:
# 1 - pbinom(0, size = 15, prob = .42)
```

Question 2 [5 points total]

A cleft lip is a split in the lip tissue connecting the mouth. Suppose you are interested in modeling the number of babies born with a cleft lip in all hospitals in Texas each month. On average, 9 babies are born with cleft lips among all Texas hospitals each month. Assume the babies born with cleft lips are not related to one another.

2.a. [1 point] Which of the following distributions is the best way to model this event?

a) Binomial
b) Normal
c) Poisson
d) Uniform

Your answer:

```
# SOLUTION:
# c) Poisson
```

2.b. [1 point] What are the mean and variance of this distribution?

a) mean = 3; var = 3
b) mean = 3; var = 6
c) mean = 3; var = 9
d) mean = 9; var = 3
e) mean = 9; var = 6
f) mean = 9; var = 9

Your answer:

```
# SOLUTION:
# f) mean = 9; var = 9
```

2.c. [1 point] Calculate (by hand) the probability that exactly 3 babies are born with cleft lips among all Texas hospitals in a given month. Show your work and round your answer to 3 decimal places.

Your answer:

```
# SOLUTION:
# (e^-mu * mu^k) / k!
# (e^-9 * 9^3) / 3! = 0.01499429 = 0.015

# 0.5 points for correct calculation
# 0.5 points for showing work with Poisson equation
```

2.d. [2 points] Write the line of R code to calculate the probability that more than 2 babies are born with cleft lips among all Texas hospitals in a given month. Then use 1 to 2 sentences to describe how you could calculate this value by hand.

Your answer:

```
# SOLUTION:
# 1 - ppois(2, lambda = 9) OR
# ppois(2, lambda = 9, lower.tail = FALSE)

# can calculate by hand by using equation below:
# 1 - P(X < 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]
# There's no upper bound for Poisson distributions, so we would have to calculate
# the individual probabilities of exactly 0, 1, and 2 babies having cleft lips,
# add these together, and subtract from 1.


# 1 point for correct code
# 1 point for correct probability notation
```

Question 3 [6 points total]

You are studying a sample of 500 individuals who underwent heart surgery from 15 different hospitals around California. The probability of a patient having an unsuccessful heart surgery (not surviving) is 0.07. Assume patients are independent of one another.

3.a. [1 point] What is the expected number of patients surviving heart surgery in this sample?

a) 7
b) 35
c) 250
d) 465

Your answer:

```
# SOLUTION:
# 500 * .93
# d) 465
```

3.b. [1 point] Which line of R code calculates the probability that no more than 10 of the 500 patients survive heart surgery?

a) `dbinom(10, 500, 0.93)`
b) `pbinom(10, 500, 0.93)`
c) `dbinom(490, 500, 0.93)`
d) `pbinom(10, 500, 0.93, lower.tail = FALSE)`

Your answer:

```
# SOLUTION:
# b) pbinom(10, 500, 0.93)
```

3.c. [1 point] Which of the following expressions calculates the probability that at least 1 of the 500 patients has a **fatal** heart surgery?

a) $\binom{500}{1} \cdot 0.93 \cdot 0.07^{499}$

b) $\binom{500}{1} \cdot 0.93^{499} \cdot 0.07$

c) $1 - \binom{500}{0} \cdot 0.93^{0} \cdot 0.07^{500}$

d) $1 - \binom{500}{0} \cdot 0.93^{500} \cdot 0.07^{0}$

Your answer:

```
# SOLUTION:

# d) $1 - {500 \choose 0} \cdot 0.93^{500} \cdot 0.07^0$

# P(fatal) = 0.07
# P(at least 1 of 5000 has fatal) = 1 - P(0 have fatal)
```

3.d. [1 point] To calculate the probability that more than 30 out of the 500 patients have **fatal** heart surgeries, you decide to use the Normal approximation. What would the mean and standard deviation of the normal distribution be?

a) $\mu = 35, \sigma = 32.55$
b) $\mu = 35, \sigma = 5.71$
c) $\mu = 30, \sigma = 32.55$
d) $\mu = 30, \sigma = 5.71$

Your answer:

```
# SOLUTION:
# b) $\mu = 35, \sigma = 5.71$

# $\mu$ = np = 500*.07
# $\sigma$ = sqrt(n*p*(1-p)) = sqrt(500*.07*.93) = 5.71
```

3.e. [1 point] Write a line of R code to show how you would use the Normal approximation to calculate the probability of more than 30 patients experiencing fatal heart surgeries. *You do not need to worry about using the continuity correction.*

Your answer:

```
# SOLUTION:
# 1-pnorm(30, mean = 35, sd = 5.71)
```

3.f. [1 point] True or False: the Normal approximation you performed in question 3.e. is extremely accurate because $p = 0.07$ is close to 0.

a) True
b) False

Your answer:

```
# SOLUTION:
# b) False.
# The Normal approximation is not extremely accurate - the probability using pbinom() is
# 0.7821283 [1 - pbinom(30, size = 500, prob - 0.07)]
# The Normal approximation is more accurate when p is close to 0.5,
# and is less accurate when p is close to 0 or 1.
```

Question 4 [6 points total]

The average nightly sleep duration of the population of students across the 9 University of California (UC) schools follows a normal distribution with mean 7.2 hours per night and standard deviation 0.2 hours per night.

4.a. [2 points] Calculate the z-score on the Normal Distribution for a student who receives 7.8 hours of sleep a night. Approximately what percentile does this z-score correspond to?

Your answer:

```
# SOLUTION:
# z-score: (7.8 - 7.2) / 0.2 = 3.0
# percentile: 99th percentile (99.85th percentile)

# 1 point for z-score
# 1 point for percentile
```

4.b. [1 point] Suppose researchers at UC Berkeley are conducting a study about the average nightly sleep duration of students at Berkeley in particular. The researchers want to determine whether their sample of 200 Berkeley students have an average nightly sleep duration that is shorter than the average for all 9 UC schools. The researchers found that the average sleep duration was equal to 7.0 hours in their sample.

State the null and alternative hypotheses in the context of this question. (Hint: reread the description provided under the question 4 header for more information)

H_0:

H_A:

```
# SOLUTION:
# $H_0$: There is no difference in the average nightly sleep duration between
# Berkeley students and UC students.
# Can also say: $\mu$ = 7.2 hours of sleep per night
# (Okay if they just write $\mu = 7.2$ and
# $\mu < 7.2$ for null and alt.

# $H_A$: The average nightly sleep duration of Berkeley students is
# less than that for all of the UC students.
# Can also say: $\mu$ < 7.2 hours of sleep per night

# 0.5 points for correct null
# 0.5 points for correct alternative
```

4.c. [1 point] Calculate the z test statistic for this sample by hand.

Your answer:

```
# SOLUTION:
# (7.0 - 7.2) / (0.2/sqrt(200)) =
# -14.14214
```

4.d. [1 point] Write the line of code that gives you the corresponding p-value associated with the z test statistic you calculated in question 4.c.

Your answer:

```
# SOLUTION:
# pnorm(-14.14214, mean = 0, sd = 1)
# can also say: pnorm(7.0, mean = 7.2, sd = 0.2/sqrt(200))
```

4.e. [1 point] You obtain a p-value of 1.04e-45. At an $\alpha = 0.05$ level, is there evidence that Berkeley students have an average nightly sleep duration shorter than the average for all 9 UCs? Why or why not? (1 to 3 sentences).

Your answer:

```
# SOLUTION:
# Yes, there is evidence that Berkeley students have a shorter average sleep duration
# compared to all 9 UCs.
# The p-value is very small (less than alpha) so assuming the null is true,
# there is a very small chance of seeing our test statistic or more extreme.
# Thus, we reject $H_0$ that Berkeley students sleep the same amount on average
# as students from all 9 UCs.
```

Question 5 [5 points total]

You are interested in modeling the distribution of the mean alcohol intake per week (in grams) among Berkeley seniors. You and 50 of your classmates take random samples of Berkeley seniors until you've each collected data for 100 individuals. Your personal sample of 100 individuals has a mean alcohol intake of 38 grams per week with a standard deviation of 3 grams. The distribution of the means of all 51 samples has a mean of 40 grams with a standard deviation, $\sigma/\sqrt{n}$, of 0.2 grams.

5.a. [1 point] Do we need to know the shape of the distribution of alcohol intake for the true underlying population of Berkeley seniors to determine μ? If yes, explain. If not, what is the approximate value of μ?

Your answer:

```
# SOLUTION:
# No, you don't have to know the shape of the distribution of alcohol intake for
# the underlying population.
# We have 51 samples of size n=100 each so our sampling distribution would be
# approximately normal.
# Thus, the mean of our sampling distribution of the mean would be an
# unbiased estimator of $\mu$.
# $\mu$ is approximately 40 grams.

# 0.5 points for saying no
# 0.5 points for explanation
```

5.b. [1 point] What is the standard deviation of the sampling distribution of the mean alcohol intake?

a) 0.02 grams
b) 0.2 grams
c) 3 grams
d) Cannot determine

Your answer:

```
# SOLUTION:
# b) 0.2 grams
```

5.c. [1 point] What is the standard deviation of the grams of alcohol intake in the population of Berkeley seniors?

a) 0.02 grams

b) 0.2 grams
c) 2 grams
d) 3 grams

Your answer:

```
# SOLUTION:
# c) 2 grams
# $\sigma$ / sqrt(n) = 0.2
# $\sigma$ = .2 * sqrt(100) = 2
```

5.d [2 points] You find that the margin of error for a 92% confidence interval is 0.35. What is the 92% confidence interval based on the sample you collected? Would you reject or fail to reject the null hypothesis that the mean alcohol intake of your personal sample is the same as the mean of the sampling distribution? Explain your choice in 1 sentence.

Lower bound:

Upper bound:

Reject or fail to reject null:

```
# SOLUTION:
# lower: 38 - 0.35 = 37.65
# upper: 38 + 0.35 = 38.35
# Reject the null that the mean alcohol intake of your sample is the same as the
# mean of the sampling distribution.
# $\mu$ = 40 is not included in the 92% confidence interval (37.65, 38.35),
#which confirms that our sample mean would most likely not have been
# 40 92 out of 100 times.

# 0.5 points for each of the bounds
# 0.5 points for rejecting null
# 1 point for explanation
```

Question 6 [5 points total + 1 point BONUS]

A clinic is considering investing in a COVID PCR test that is different from the test they currently use. The clinic recruited you to analyze the accuracy of their current COVID PCR test, which they have been administering since the beginning of the pandemic. The data is shown in the table below.

	Has COVID	Does not have COVID
Tests +	443	75
Tests -	59	879

6.a. [1 point] Calculate the sensitivity of the clinic's current COVID test and provide your answer rounded to two decimal places. Interpret this value in the context of the problem.

Your answer:

```
# SOLUTION:
# sensitivity = P(tests + and Has COVID) / P(Has COVID)
# 443/(443+59) = 0.8824701
# Among those who truly have COVID, 88.25% test positive for COVID

# 0.5 points for correct calculation, 0.5 points for correct interpretation
```

6.b. [2 points] Since COVID is an infectious disease, we want to minimize the number of false negatives to help mitigate the spread. Should the clinic invest in a test that is more sensitive or more specific? Explain why in 1 to 2 sentences.

Your answer:

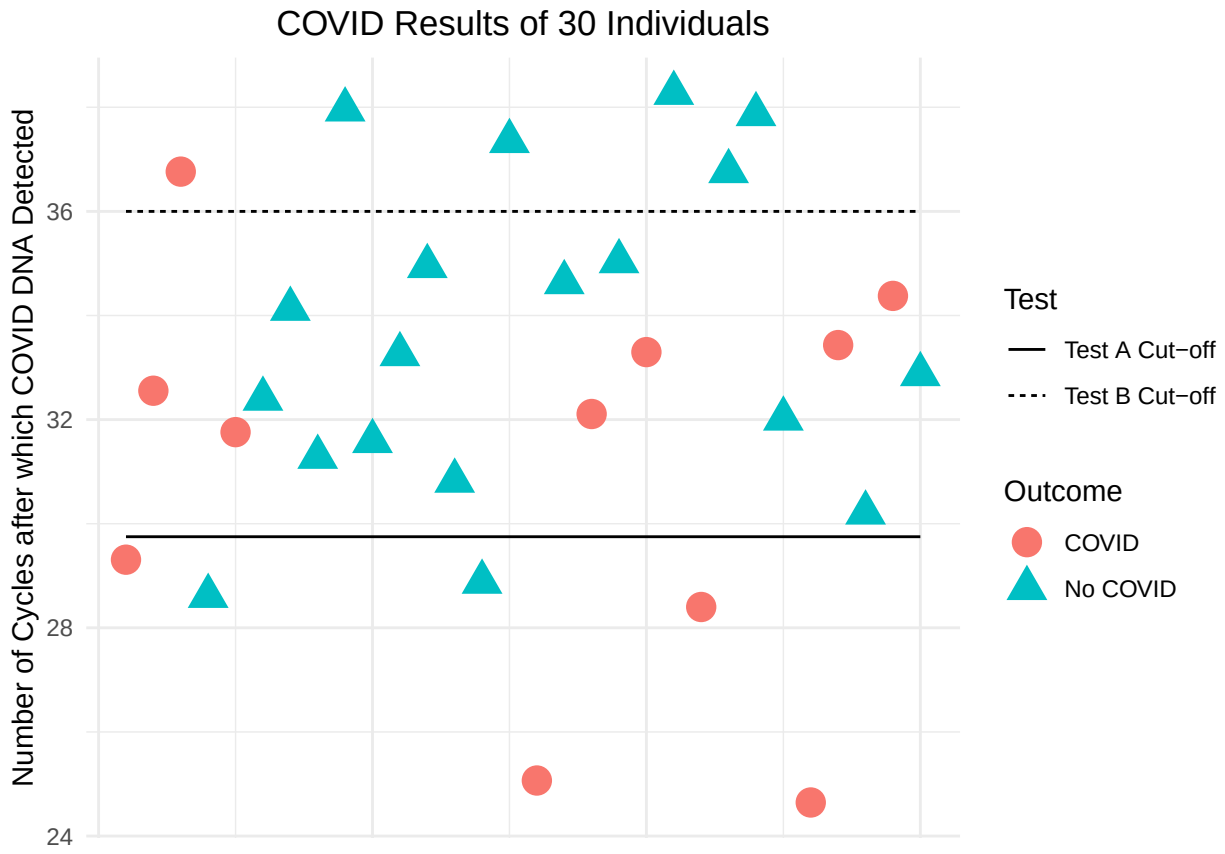
```
# SOLUTION:
# Invest in a more sensitive test - high sensitivity means that we are correctly
# capturing most of the true COVID cases.
# We want to minimize the number of people who test negative but actually have COVID
# (so they don't go about their normal lives and spread the disease),
#which means we want to minimize part of the denominator of the
# sensitivity calculation (increasing sensitivity).

# Alternative explanation: We want to minimize false negatives
# which means minimize P(test negative | truly have COVID),
# this is equivalent to wanting to maximize P(test positive | truly have COVID),
#which implies we want a more sensitive test.

# 1 point for saying more sensitive
# 1 point for correct explanation
```

6.c. [2 points] The COVID PCR test uses a machine that rotates and doubles the amount of DNA each rotation cycle, then tests the amount of COVID DNA present after a specified number of cycles.

The points on the plot below show the COVID status of 30 individuals. The horizontal lines show the threshold cut-off values, chosen by two brands of COVID PCR tests, for detecting a positive COVID case using DNA (points below the threshold line receive a positive test result). Given your answer to question 6.b., would you recommend the clinic invests in test A or test B? Explain why in 1 to 2 sentences.



Your answer:

```
# SOLUTION:
# The clinic should choose test B. Test B has a higher sensitivity because it has
# a higher cutoff value for the cycle threshold -
# this means that the DNA has 36 cycles for which it can replicate and result in
# a positive test.
# sensitivity of test B = 10/11, sensitivity for test A = 4/11
# 1 point for correct test selection
# 1 point for explanation - explanation can be comparing the sensitivities of
# tests A and B
```

BONUS Question 6.d. [1 point] Calculate the positive predictive value of test A.

Your answer:

```
# SOLUTION:
# PPV = P(tests + and Has COVID) / P(tests +)
# test A PPV: 4/6 = 0.6666667
```