Fall 2020 Midterm II SOLUTIONS

The exam is open book. This means you can use electronic or hard copies of notes from class. 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. Evidence of cheating may result in a 0 on the exam.

No copies of the exam are to be saved.

I affirm that I 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: [1 point]

Problem 2: [1 point]

Problem 3: [1 point]

Problem 4: [1 point]

Problem 5: [1 point]

Problem 6: [2 points]

Problem 7: [1 point]

Problem 8: [1 point]

Problem 9: [7 points]

Problem 10: [5 points]

Problem 11: [3 points]

Problem 12: [6 points]

Problem 13: [3 points]

Total: 33 points

Question 1 [1 point total]

1. [1 point] For a normal distribution, we can use the `pnorm()` function in R to calculate the probability of x being exactly equal to some value, i.e., $P(X = x)$

a) True
b) False

```
# Solution: b) False.
# `pnorm()` calculates a cumulative probability (lower tail by default)
```

Question 2 [1 point total]

2. [1 point] Power = 1 - P(type II error)

a) True
b) False

```
# Solution: a) True. This is the definition of Power.
```

Question 3 [1 point total]

3. [1 point] Which of the following functions could you use to calculate the $P(X < k)$ where k is the number of occurrences of some rare event in a given time interval with known mean?

a) `dbinom()`
b) `pbinom()`
c) `dpois()`
d) `ppois()`
e) `pnorm()`

```
# Solution: d) `ppois()` since ppois calculated the cumulative probability
# of lower tail at or below some x.
# Here would specify x = k-1 to get <k rather than <=k.
```

Question 4 [1 point total]

4. [1 point] Suppose you have data on *all* gonorrhea cases in the US in 2010 and calculated the proportion that were resistant to a major antibiotic. This proportion was 27%. This number is a _____

a) sampling distribution
b) parameter
c) statistic

```
# Solution: b) parameter
```

Question 5 [1 point total]

5. [1 point] If events A and B cannot occur at the same time and the P(A) $\neq$ 0 and P(B) $\neq$ 0, select all that apply.

a) A and B are Mutually Exclusive
b) $P(A \cap B) = P(A)*P(B)$
c) $P(A \cup B) = P(A) + P(B)$
d) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

```
# Solution: a), c), d)
# A and B are mutually exclusive, disjoint, dependent.
# P(A and B) = 0 != P(A)P(B), P(A or B) = P(A) + P(B) -  P(A $\cap$ B) = P(A) + P(B)
```

Question 6 [2 points total]

6. [2 points] Fill in the blanks to complete the statement about the Central Limit Theorem: As _____ increases, the _____ (2 words) of the mean approaches a _____ distribution with mean μ and standard deviation _____.

A:

B:

C:

D:

```
# Solution:

# A: n, sample size
# B: sampling distribution
# C: Normal
# D: sigma/sqrt(n)
```

Question 7 [1 point total]

7. [1 point] P(reject the null hypothesis | null hypothesis is true). What type of error is this?

a) Type I error
b) Type II error
c) This is not an error
d) None of the above

```
# Solution: a). This is the definition of a type I error, written in probability form.
```

Question 8 [1 point total]

You designed an experiment to test if a new drug is effective for patients with a particular disease, and found that the drug improves the health of 70% of the participants in the experiment.

8.1 [0.5 point] After looking at your data, you decide to use a random variable X which has a binomial distribution with probability of success of 0.7 for further investigation. What does the random variable X represent?

a) Number of participants who took the drug
b) Number of participants for whom the drug improves health
c) Number of participants in the experiment
d) None of the above

```
# Solution: b). X ~ Binomial(n, 0.7) and represents the number of "successes".
# Here a success occurs when the drug improves health,
# such that X is the number of participants for whom the drug is effective.
```

8.2 [0.5 point] Upon careful inspection of your data, you realize that age influences the effectiveness of the drug; Younger participants appear to respond to the drug better than older participants. Which assumption about the binomial distribution is violated?

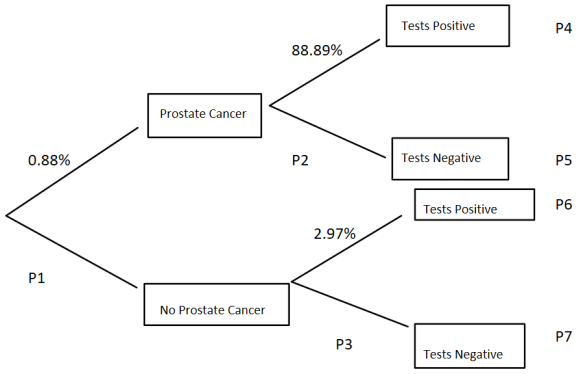
a) There are a fixed number of n observations
b) The n observations are independent.
c) Each observations is either a "success" (1) or a "failure" (0).
d) The probability of success is the same for each observation.

```
# Solution: d), we can find the probability of effectiveness
# of the drug is not the same for individual.

# This is because older individuals will have a different
# probability of success than younger individuals.
```

Question 9 [7 points total]

Suppose that 101,800 men were screened for prostate cancer for the first time. The probability that a screening participant has prostate cancer is 0.88%. Given that a participant has prostate cancer, the probability that they test positive is 88.89%. Given they do not have prostate cancer, the probability that a participant tests positive is 2.97%. A tree diagram with these probabilities is shown below.



9.1 [1 points] Use notation to define prostate cancer and the result of the test as random variables. Write out their respective sample spaces. You do not need to assign or calculate the probabilities to each outcome for this question.

```
# Solution:
# Prostate cancer = P, sample space = {yes, no} , {0,1} , etc.
# Test result = T, sample space = {positive, negative} / {0,1} etc.
```

9.2 [2 points] Write a probability statement to represent specificity and sensitivity using the random variables you defined above. Then, calculate these values. Write your answers as percentages between 0 and 100 rounded to 2 decimal places.

Sensitivity:

Specificity:

```
# Solution:
# Sensitivity: P(T+|C+) = 88.89%
# Specificity: P(T- | C-) = 97.03%
```

9.3 [1 point] Calculate the probability of testing positive on this screening exam. Write your answer as a percentage between 0 and 100 rounded to 2 decimal places.

```
# Solution:
# P(T) = 3.72%
# P(T+) = P(T+|P+)*P(P+) + P(T+|P-)*P(P-)
# 0.8889*(0.0088) + 0.0297*(0.9912)
```

9.4 [3 points] Write a probability statement and calculate the positive predictive value (PPV) of this test. Write your answer as a percentage between 0 and 100. In one sentence or less, give an explanation as to why this value is relatively low (or high) in comparison to the sensitivity.

```
# Solution:
# PPV = P(C+ | T+) = P(P+|T+)/P(T+)
# = [P(T+|P+)*P(P+)]/P(T+)
# = [ 0.8889 * 0.0088 ]/ 0.0372 = 21.02%

# The **prevalence** of prostate cancer in this population is very **low**
# so it makes sense there will be a low probability of testing positive
# when the proportion of false negative may outweigh it.
```

Question 10 [5 points total]

The CDC recommends that adults consume no more than 2300 mg of sodium per day. We know the true population distribution of sodium intake in the US follows a normal distribution with a standard deviation of 200 mg of sodium. You take a simple random sample of 100 adults representative of the population and find that their mean daily intake is 3225 mg. You want to test whether this sample mean is significantly *different* from the recommended daily sodium intake.

10.1 [1 point] What is the null and alternative hypothesis for this test? You may answer in words or notation.

H_0:

H_A:

```
# Solution:
# H0: mu = 2300 mg ; HA: mu does NOT = 2300
# can also say: H0: average sodium intake of sample = recommended intake ;
# HA: average of sodium intake of sample does NOT = recommended intake
```

10.2 [1 point] Calculate the 95% confidence interval and list the the lower and upper bounds below. Round each to one decimal place.

Lower Bound:

Upper Bound:

```
# Solution: CAN USE EITHER z* of 2 or 1.96!
# If use z* = 2 then lower bound = 3185 ; upper bound = 3265

# 3225 - 1.96 * (200/sqrt(100)) = 3185.8
# 3225 + 1.96 * (200/sqrt(100)) = 3264.2
```

10.3 [1 point] What is the margin of error for the 95% confidence interval you calculated?

a) about 20
b) about 39.2
c) about 3.92

```
# Solution: b).
# Margin of error = z* * sigma/sqrt(n) = 1.96 * 200/sqrt(100).
```

10.4 [1 point] Based on the confidence interval you calculated in (ii), say one thing you know about the corresponding p-value for the two-sided hypothesis test corresponding to the hypotheses you listed in 10.1.

```
# Solution:
# p < 0.05 or p < 5% because 2300 is not in the range of values
# (3185.8 to 3264.2)
```

10.5 [1 point] Which of the following would increase the power of this test?

a) increase sample size
b) increase alpha
c) a and b
d) you cannot increase the power of a hypothesis test

```
# Solution: c) Increase sample size and alpha level.
```

Question 11 [3 points total]

Assume that you have recruited some randomly selected, unrelated individuals into a study to test the effectiveness of a new vaccine to prevent tuberculosis. A previous controlled trial established the probability of immunity gained after vaccine administration. Note that none of the persons in your study had gained immunity before the trial. After the vaccine administration, each participant can either be immune to tuberculosis or not be immune. The results of the immunity test are independent. Each participant has the same probability of being immune. The code below displays a probability of immunity given certain parameters.

```
dbinom(x = 20, size = 30, prob = 0.50)
```

```
## [1] 0.0279816
```

11.1 [1 point] What do `size` and `prob` depict in the code above?

```
# Solution:
# a) Size represents the total number of participants recruited into the current study,
# while prob is the probability of immunity gained after vaccine administration
# that was established from a previous study.
```

11.2 [1 point] Interpret the `dbinom()` code and its output in your own words.

```
# Solution:
# The probability that exactly 20 people gain immunity
# after administration of the new vaccine is 0.02798 or 2.798%.
```

11.3 [1 point] What does "The results of the immunity test are independent" mean in your own words?

```
# Solution:
# One person being immune does not affect another person's chance of being immune.
# One person's chance/risk of immunity does not affect another person's chance of being immune.
```

Question 12 [6 points total]

An illness involving E.coli causes a breakdown of red blood cells and intestinal hemorrhages when infected. Approximately 64 people in the United States die as a result of contracting E.coli every year. Assume the number of people in the US who die by E.coli each year follows the Poisson distribution.

12.1 [2 points] What is the mean and standard deviation of this distribution?

Mean:

SD:

```
# Solution:
# mean: 64
# sd: 8
```

12.2 [1 point] What is the probability that exactly 50 people succumbed to the illness this year? Show your work (*not R code*) for full credit. Write your answer as a percentage between 0 and 100 rounded to 2 decimal places.

```
# Solution:
# X ~ Poisson(lambda = 64)
# P(X = 50) = 64^50 * e^(-64) / 50! = 0.0107418 = 1.07%
```

12.3 [1 point] Write R code to calculate the probability that more than 70 people contracted and died from E.coli in the US this year. You do not need to calculate the probability.

```
# Solution:
# ppois(q = 70, lambda = 64, lower.tail = F)
# 1 - ppois(q = 70, lambda = 64)
```

12.4 [1 point] In theory, there can be at most 1000 deaths of the illness observed each year. Why?

a) True

b) False
Explanation:

```
# Solution : b) False. Poisson random variables have no upper bound.
```

12.5 [1 point] It is found that the rate of the illness increases dramatically in recent years, so that 100 people die within one year. With the change in rate, the shape of the distribution will _____, and the range of the distribution is _____.

a) blank 1: become flatter; blank 2: wider
b) blank 1: become sharper; blank 2: narrower
c) blank 1: remain the same; blank 2: remain the same

```
# Solution: a) When the mean of the distribution increases,
# its standard deviation also increases,
# thus making the distribution flatter and have a wider range.
```

Question 13 [3 points total]

Of individuals with cancer, breast cancer is the most common cancer diagnosis and represents 25.9% of all cancer diagnoses among women globally. It is followed by colorectal cancer (9.7% of all diagnoses), lung cancer (8.8% of all diagnoses), and uterine/cervical cancer (6.9% of all diagnoses). *Assume that a woman will only be diagnosed with one type of cancer.*

13.1 [1 point] Write the sample space for types of cancer.

```
# Solution: {breast cancer, colorectal, lung, uterine/cervical, other}
```

13.2 [1 point] What is the probability that a woman with cancer has a diagnosis of breast cancer or uterine/cervical cancer? Write as a percentage between 0 and 100 rounded to 1 decimal place.

```
# Solution: P(B or UC) = 0.259+0.069 = 32.8%
```

13.3 [1 point] Ten women who have cancer know their diagnoses. What is the probability that at least one of them has breast cancer? Write your answer as a percentage between 0 and 100 rounded to 2 decimal places.

```
# Solution: P(at least 1) = 1 - P(none) =1 - (1-0.259)^10 = 1 - 0.0499 = 0.9501 = 95.01%
```

END