Summer 2021 Midterm 2

The exam is open book. This means you can use electronic or hard copies of all class materials and can use datahub or R/Rstudio if you wish. You may not use the internet to search for the answers or to 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 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.

Type your name and SID below :

Name:

Enter your name:

Enter your SID:

Instructions:

1.Use Adobe Reader or Acrobat as a stand-alone application (NOT in a browser) to complete this assignment. (this software can b e accessed for free for UCB studentshttps://software.berkeley.edu/adobe-creative-cloud)

2.Give your responses ONLY in the space provided. Do NOT add any additional textboxes.

3.Please rename the file LASTNAME_FIRSTNAME_Midterm2_Summer2021.pdf

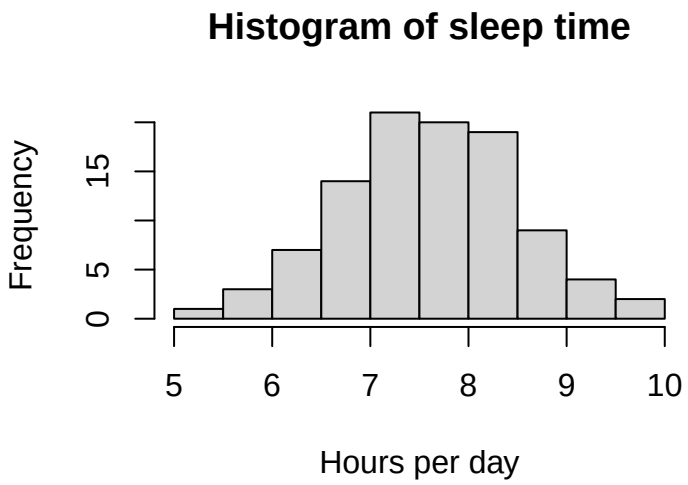
Unless otherwise specified in the question, format your answers according to the following guidelines:

- present your answers rounded to two decimal places
- present proportions as % values (40.50% rather than .405)

** MAKE SURE YOU ARE WORKING WITH THIS DOCUMENT IN ADOBE AND YOU ARE NOT IN A BROWSER WINDOW **

Question 1 [4 pts]

The histogram below is the distribution of student's average sleep time per night in Greendale college. We assign the data to the object `time`. Use these data to answer the following questions



Q1.1 [**2pt**]

Fill in blanks of the code below so that you can calculate the mean, standard deviation of `time` and then convert it into z-scores.

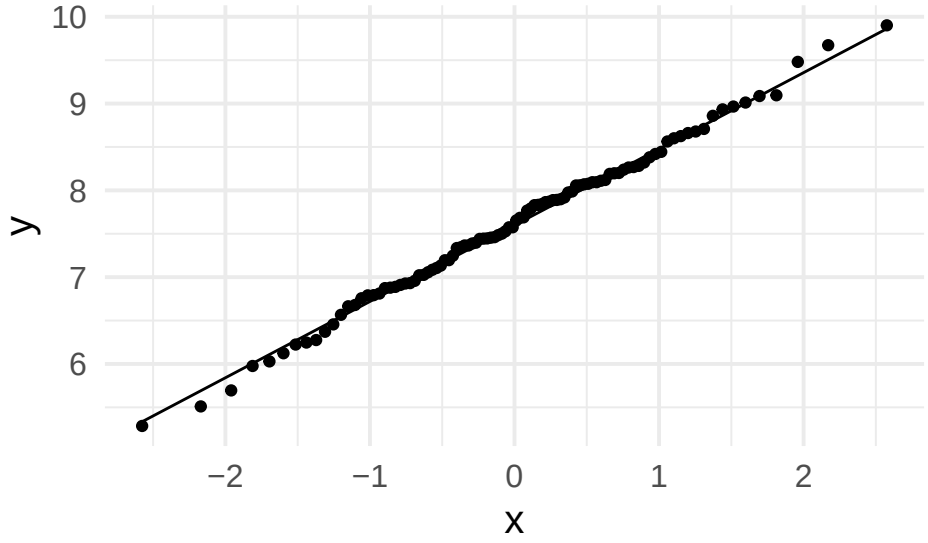
```
mean <-
std <-
zscore <-
print(mean)

print(std)
```

Here is the output generated by this code:

```
## [1] 7.608887
## [1] 0.8981994
```

Q1.2 [1pt]

The below is a QQplot for these data:



In one or two short sentences, explain why we might run a plot like this, and what this plot tells us about our sleep data.

Q1.3 [1pt]

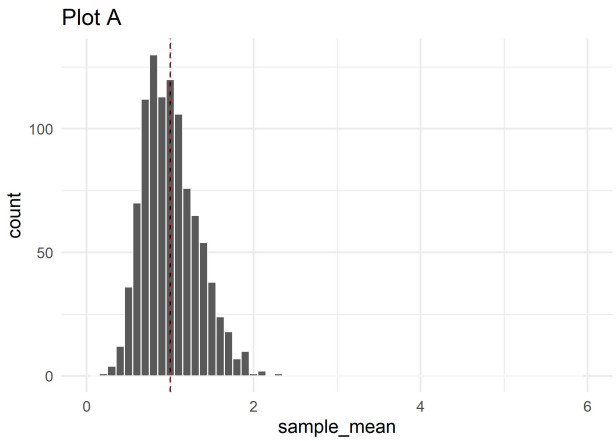
Using the values generated earlier, use a R function to help calculate $P(7 \leq x < 9)$. Show your code.

==

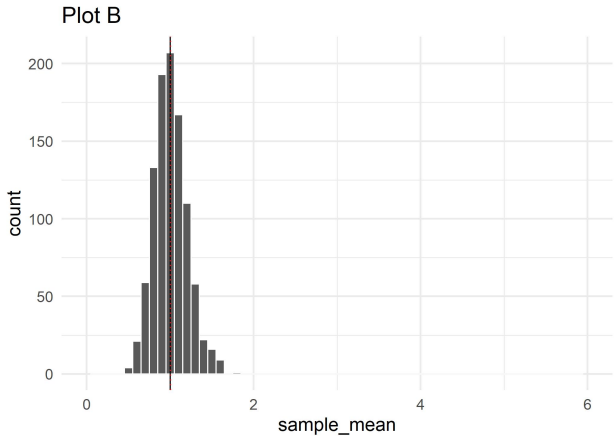
Question 2 [2 points]

Assume each graph shows 1200 samples taken from the same underlying population.

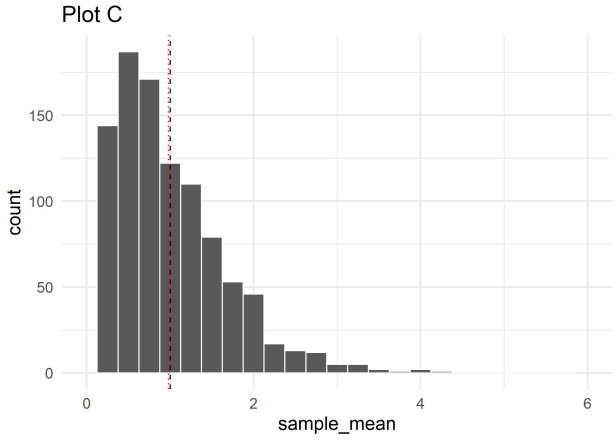
```
knitr::include_graphics("plot_a.jpeg")
```




```
knitr::include_graphics("plot_b.jpeg")
```




```
knitr::include_graphics("plot_c.jpeg")
```



Q2.1 [1 point]

Rank the three sampling distributions in order of decreasing sample size.

Q2.2 [1 point]

Provide reasoning for the ranking in 2-4 sentences:

Question 3 [2pts]

Suppose that we have a dataset for the blood pressure of 1000 students. The mean is 92 and the standard deviation is 12. Please answer the questions below.

Q3.1 [1pt]

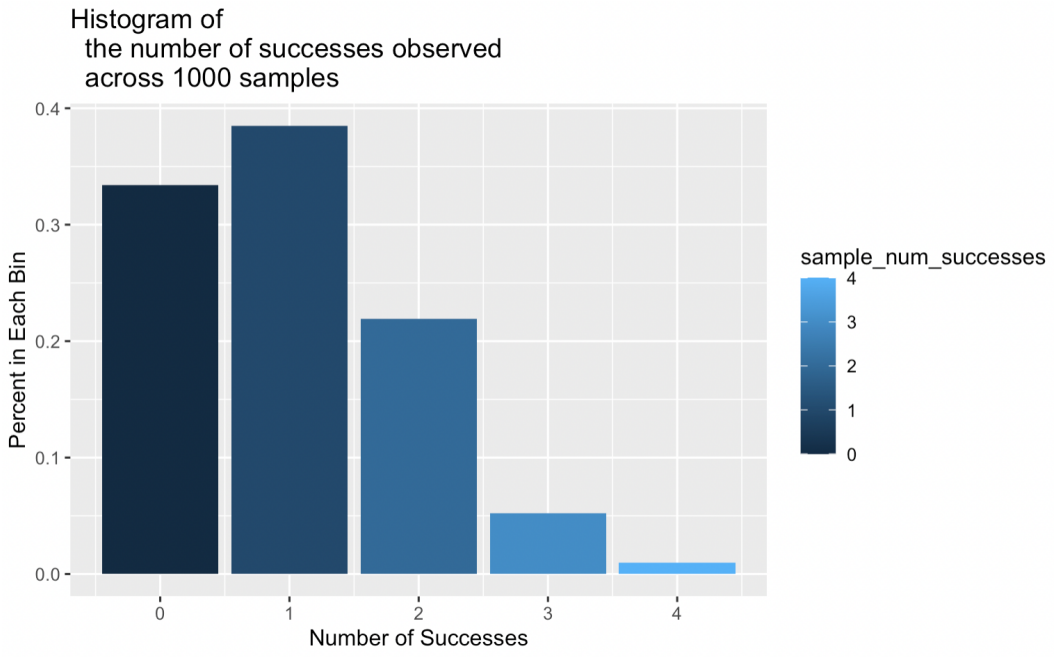
Suppose that this dataset ideally follows the normal distribution with mean 92 and standard deviation 12. Please give an interval so that approximately 99.7% of the data will fall within it.

Q3.2 [1pt]

Suppose that this dataset ideally follows the normal distribution with mean 92 and standard deviation 12. Write the codes to calculate the the median of such normal distribution by using `qnorm`.

Question 4 [5 points]

You are sampling dirty socks out of a very large laundry basket of 10,000 socks, where there are 2 outcomes for your sample: a clean sock (not success) or a dirty sock (success). Every sample is size n = 10 and after you sample, you put the socks back into the laundry basket. You take 1000 samples, record the number of successes across 1000 samples, and generate a histogram of these data.



Q4.1 [1 point]

Based on the histogram, what is the most likely probability of success p?

Q4.2 [2 points]

Based on the information provided in the question, what distribution does this situation most resemble? Justify your answer in one or two sentences.

a) Poisson
b) Binomial
c) Normal

Q4.3 [2 points]

Using your answer in Q4.1, calculate the probability of sampling exactly 1 dirty sock in one sample. Show your setup in the first box and put your final answer in the second box. Please provide your final answer as a percentage and rounded to two decimal points (e.g. 1.42%)

Setup:

Final Answer:

Question 5 [5 Points]

Alice in Wonderland Syndrome

The following excerpt has been adapted from "Alice in Wonderland SYndrome as a Presenting Manifestation of Creutzfeldt-Jakob Disease" by Naarden et al.:

Background Alice in Wonderland syndrome (AIWS) is a rare neurological disorder characterized by distortions of visual perception (metamorphopsias), the body image, and the experience of time.

As noted as early as 1955 by John Todd, these symptoms may be accompanied by derealization and depersonalization. Patients suffering from AIWS may end up consulting a neurologist or a psychiatrist, although in both specialties it is not as well-known as it deserves to be. This is at least partly due to the fact that major classifications such as the Diagnostic and Statistical Manual of Mental Disorders [DSM; (2)] do not list it as a diagnostic category, while others, such as the International Classification of Diseases [ICD; (3)] pay only limited attention to it. Another reason may be the relatively small number of published cases.

A review of the extant literature, published in 2016, indicated that only 169 cases of AIWS had been described since the syndrome's conceptualization in 1955, which boils down to a mean number of 1.1 cases per year.

Q5.1 [1 Point]

What two assumptions about AIWS must be met in order to use the Poisson distribution?

Q5.2 [1 Point]

Assuming that AIWS does indeed follow the Poisson distribution, what is its variance?

Variance:

Q5.3 [1 Points]

Which of the following examples of R code will return the probability of having 3 or more cases of AIWS in a year?

A: `ppois(q = 3, lambda = 1.1, lower.tail = FALSE)`
B: `dpois(x = 0, lambda = 1.1) + dpois(x = 1, lambda = 1.1), + dpois(x = 2, lambda = 1.1)`
C: `1 - (dpois(x = 0, lambda = 1.1) + dpois(x = 1, lambda = 1.1), + dpois(x = 2, lambda = 1.1))`
D: `ppois(q = 3, lambda = 1.1)`

Option:

Q5.4 [2 Points]

What's the probability that there are more than 5 cases of AIWS diagnosed in a year? Provide your answer as a percentage and round to two decimal places. If this event were to actually happen, what conclusions about AIWS could you make?

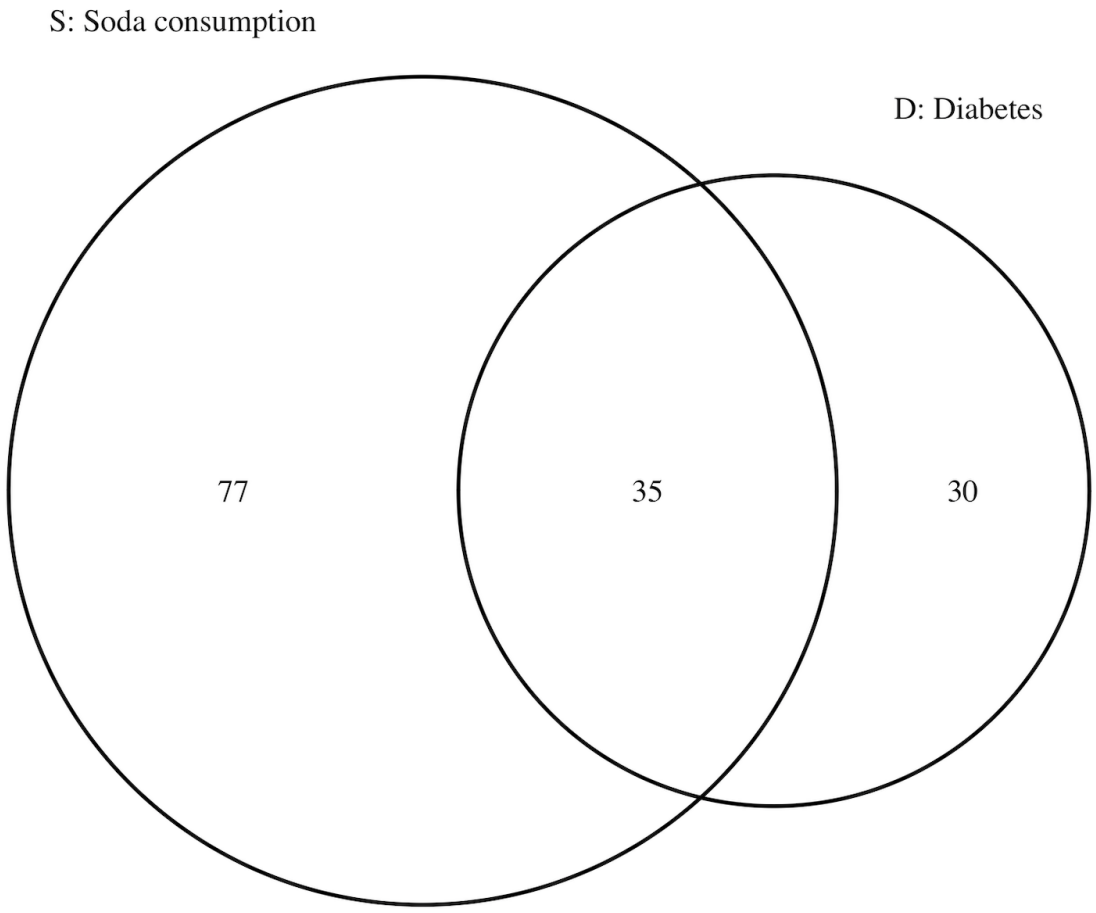
Question 6 [2 points]

Racial concordance is an important concept in medicine and public health and describes the situation in which a patient at a health care facility is receiving care from a provider of the same racial background. Suppose the proportion of patients in California who experience racial concordance with a health care provider is 20%. You take a simple random sample of 5 Californian patients and assess if they experienced racial concordance with their provider. Assume that the chance that a Californian patient experiences racial concordance is equal across all Californian patients.

Based on the information provided in the question, what distribution does this situation most resemble? Justify your answer in one or two sentences.

a) Poisson
b) Binomial
c) Normal

Question 7 [3 points]



Researchers at UC Berkeley randomly surveyed 300 residents in Oakland about their health status. They measured frequent soda consumption (>= 5 times a week) and type 2 diabetes. The diagram below shows the results of the survey. Calculate the following probabilities to the nearest two decimal points (not a percentage).

(Note: The numbers inside the diagram represent the number of Oakland residents for that event alone. For example, 77 represents the number of Oakland residents that indicated soda consumption but doesn't have diabetes.)

Question 7.1 [1 points]

$P(S' \cap D') =$

Question 7.2 [1 points]

$P(S \cap D) =$

Question 7.3 [1 points]

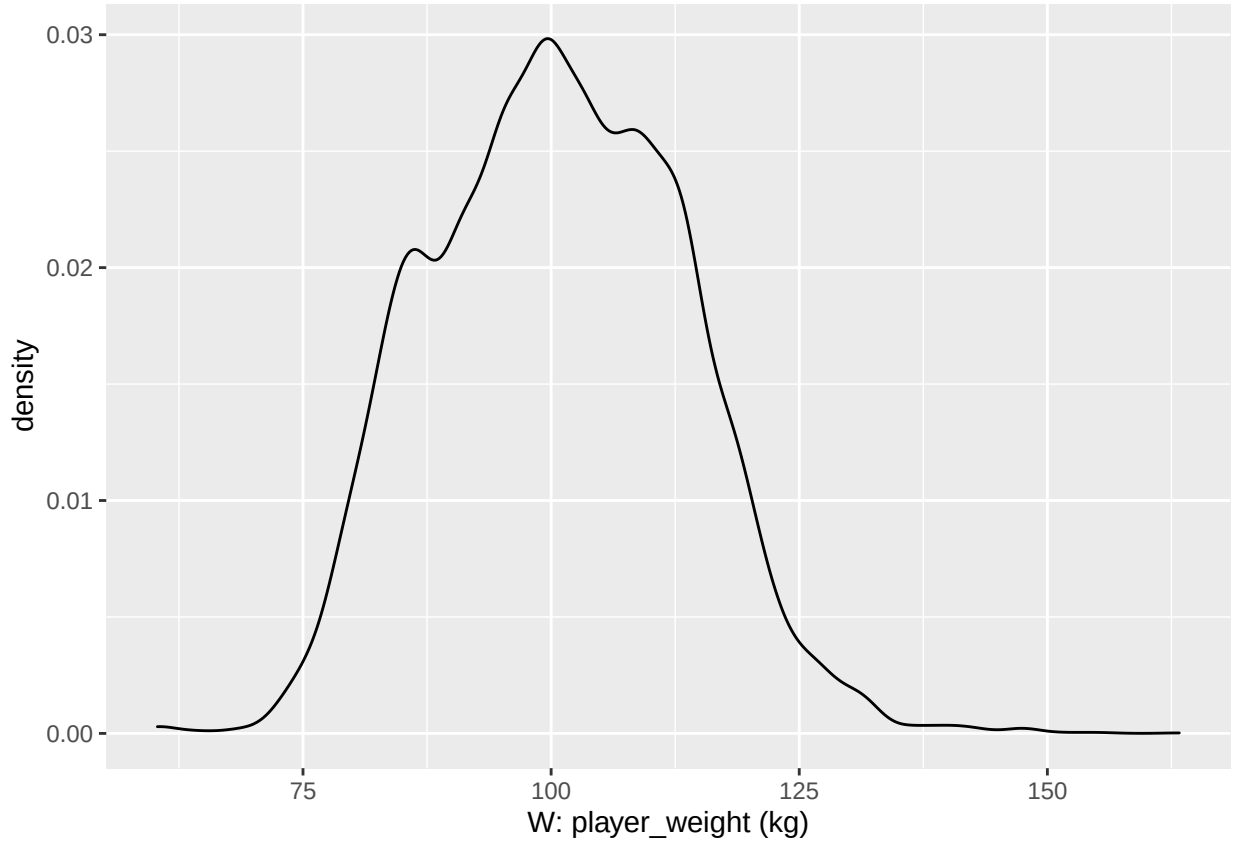
$P(S|D') =$

Question 8 [1 point]

Source: https://www.kaggle.com/justinas/nba-players-data

The following data is based off of all players in the NBA from the 1996 - 2019 seasons.

```
nba %>% ggplot(aes(x = player_weight)) + geom_density()  + labs(x = "W: player_weight (kg)")
```



The graph above shows a density curve of all players' weights. What probability represents the total area underneath the curve?

Question 9 [2 points]

According to the California Department of Public Health, 61.4% of Californians are fully vaccinated from COVID-19. You decided to randomly sample 5 people in California

Question 9.1 [1 point]

What is the probability that all 5 people are not fully vaccinated? Round your probability to two decimal places.

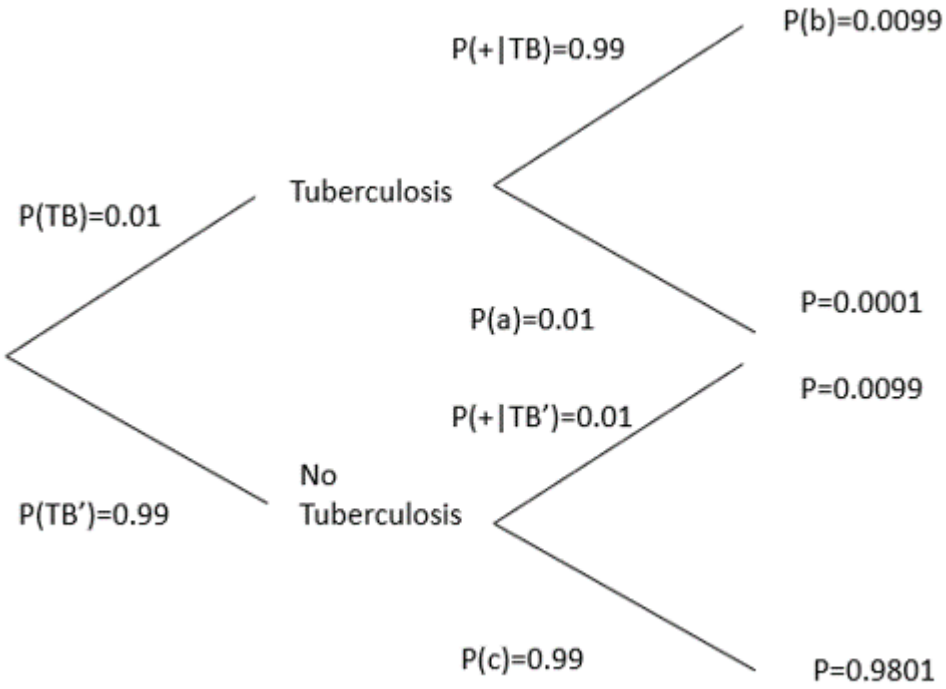
Question 9.2 [1 point]

What is the probability that at least one person is not fully vaccinated?

Question 10 [4 points]

Use the following probability tree about the relationship in patients between having tuberculosis (TB) and the probabilities to have a negative or positive test to answer the next questions.

```
knitr::include_graphics('probabilitytree.jpg.png')
```



Q10.1 [1 point]

What do P(a), P(b), and P(c) represent in the tree (in order of a, b, c)?

a. P(TB'), P(+|TB), PT(-|TB)
b. P(-|TB), P(+|TB), PT(TB|T-)
c. P(-|TB), P(TB ∩ +), P(-|TB')

Option:

Q10.2 [1 point] Calculate the probability that the patient has tuberculosis given that the test is positive. Express the answer as a percentage rounded to one decimal place.

$P(TB|+) =$

Q10.3 [1 point]

What is the quantity in Q10.2 called? What will happen to this quantity of the test is used in a population with a lower prevalence of TB?

Q10.4 [1 point]

If we have a population of 1000 patients, how many will test negative? Round to the nearest integer.

Exam feedback:

If you experienced any issues with your exam please describe them here: