Question 01

Researchers scanned past medical charts from a hospital and collected a dataset of patients over the age of 50. The variables in the data set are: patient's age, ejection fraction (a measure of heart health), and whether they have high blood pressure.

PT	age	ejection fraction	high blood pressure
1	75	20	1
2	55	30	1
3	65	25	0
4	50	45	0
5	65	55	1
6	90	30	0

Define an observation in this dataset and classify the three variables as numerical (discrete or continuous) or categorical (ordinal or nominal).

Question 02

Tuberculosis (TB) is an infectious disease caused by the bacterium *Mycobacterium tuberculosis*. The infectious agent typically invades the lungs and spreads from one to another through the air. A study of 137 participants was conducted to study the association between TB bacterial load in patients and CD4 expression. Patient were stratified by latent TB infection (LTBI) vs active TB infection (aTB) and if they were HIV positive or negative. CD4 expression is a measure of our immune system, the higher the CD4 count the more robust a patient's immune system.
A brief snapshot of the data is presented below, presenting the mean CD4 count for LTBI and aTB among HIV+ participants.

	LTBI/HIV-	LTBI/HIV+	aTB/HIV-	aTB/HIV+
N	35	32	33	37
Age (year)	31 (23–38)	34 (31–42)	30 (24–43)	37 (32–45)
Gender (F/M)	19/16	24/8	8/25	11/26
CD4 (cells/mm3)	no data	481 (358–700)	no data	273 (148–435)

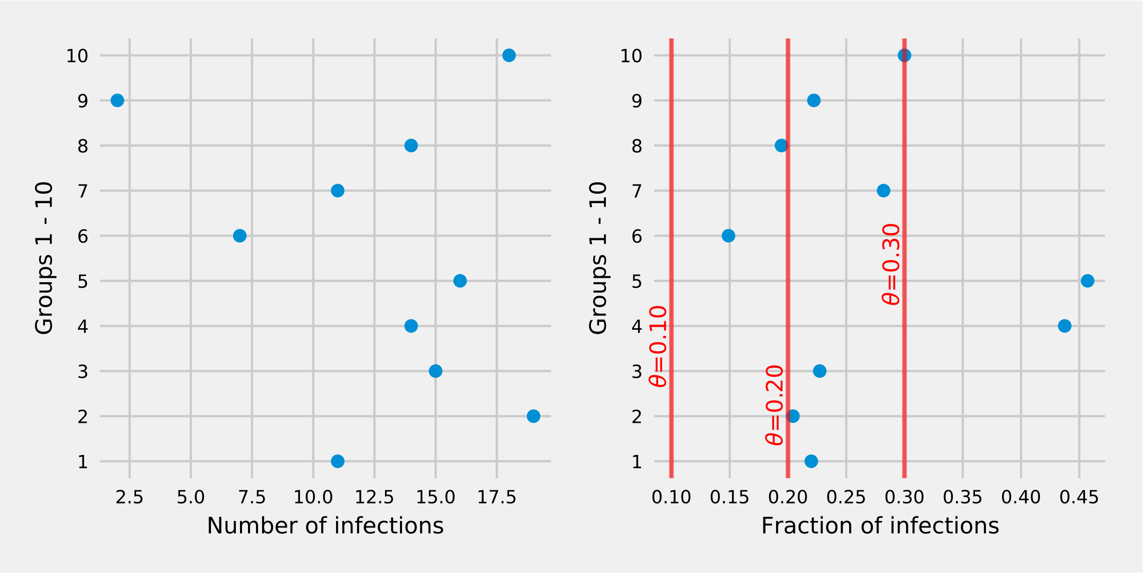
Is the mean CD4 count a parameter or statistic? Why or why not? Given a set of CD4 counts for N patients $c_1, c_2 + \cdots, c_N$, please give a formula using sigma notation to compute the mean.

Question 03

The probability of a set A is 0.45 and the probability of a set B is 0.22 and the probability of the intersection of A and B is 0.10. Please compute $p(A \cup B)$. Do the events A and B contain all outcomes in the sample space (why or why not)?

Question 04

Suppose we collected data on 10 different group of participants and counted the number of those infected with COVID-19 (data below). We can assume the number of infected is a r.v. that follows a Binomial distribution with a differing number of trials and the **same** probability of "success" (θ).

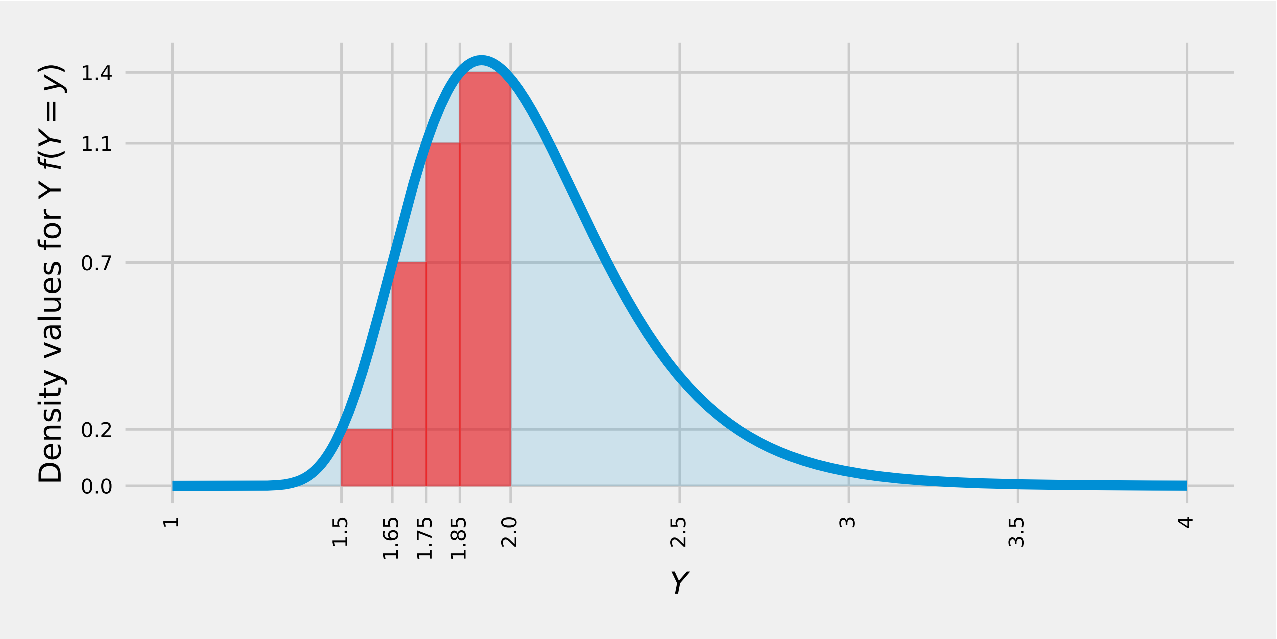


Please compute $p(\theta|\mathcal{D})$ for $\theta = 0.10$,0.20, and 0.30. Which θ value is most likely?

$p(\mathcal{D}|\theta = 0.10) = 0.10$	$p(\theta = 0.10) = 0.30$
$p(\mathcal{D}|\theta = 0.20) = 0.23$	$p(\theta = 0.20) = 0.25$
$p(\mathcal{D}|\theta = 0.30) = 0.50$	$p(\theta = 0.30) = 0.45$

Question 05

Consider a continuous random variable Y that has the following density



Please approximate the CDF at 1.85 minus the CDF at 1.5 (or F(1.85) - F(1.50)) using several rectangles.

Question 06

Let $Z \sim \text{Binom}\,(10, 0.20)$ Please compute $p(Z > 1)$

Question 07

Suppose we define a random variable Y with the following distribution

value	probability
0	0.20
1	0.01
2	0.42
3	0.33
4	0.04

Please compute the expected value and variance of Y