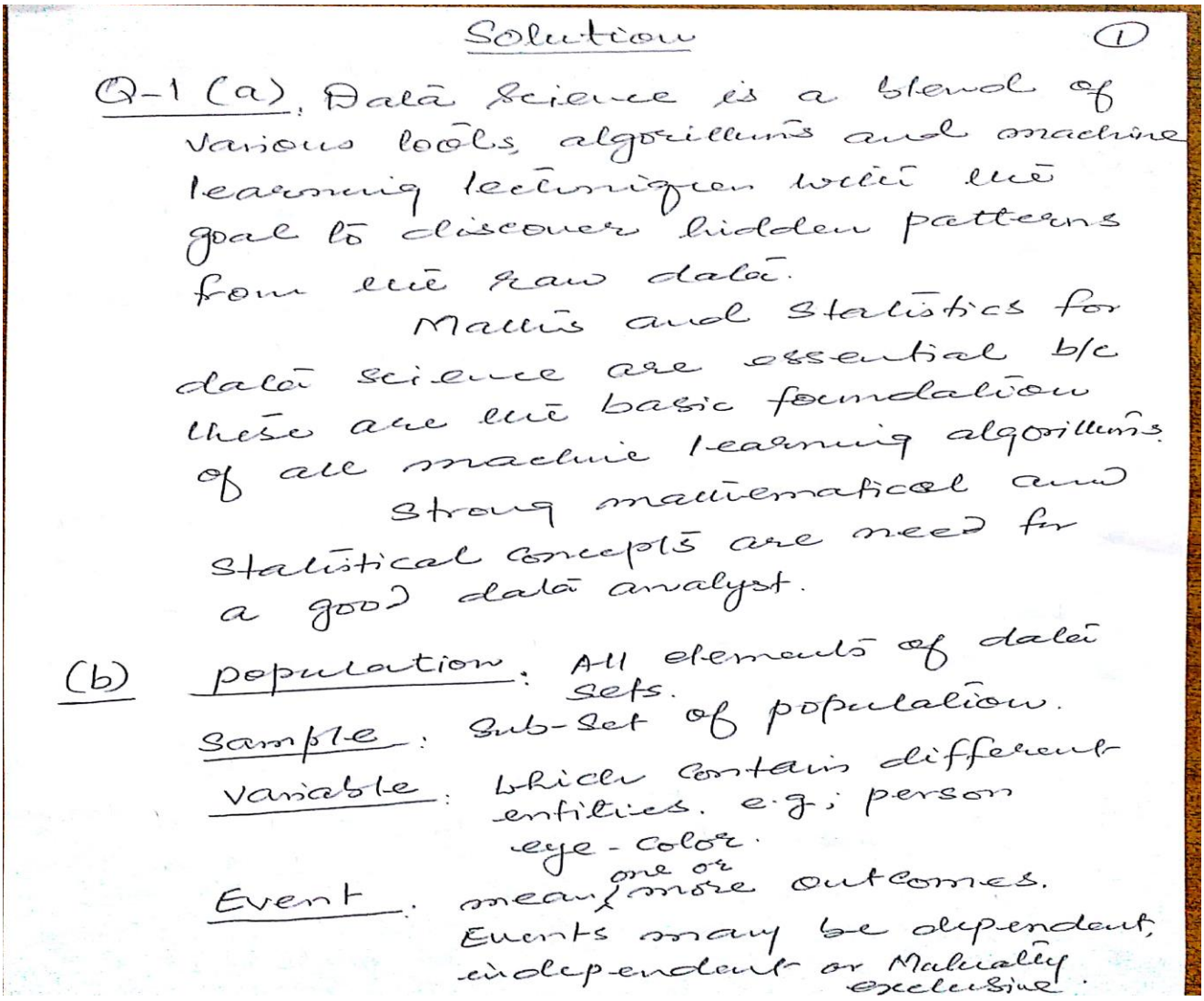
DS501 MIDTERM SOLUTION

Question # 1

a) What is Data science? Why statistics is essential for data science? Differentiate between Descriptive and Inferential Statistics.

b) Define the following terms in reference to Statistics: Population, Sample, Variable, Event and Sample Space.

c) Briefly explain the terms: Box plot and Inter quartile range (IQR) .



Solution ①

Q-1 (a). Data Science is a blend of various tools, algorithms and machine learning techniques with the goal to discover hidden patterns from the raw data.

Maths and Statistics for data science are essential b/c these are the basic foundation of all machine learning algorithms. Strong mathematical and statistical concepts are need for a good data analyst.

(b) Population: All elements of data sets.

Sample: Sub-set of population.

Variable: which contain different entities. e.g.; person eye-color.

Event: mean one or more outcomes. Events may be dependent, independent or Mutually exclusive.

c) A boxplot splits the data set into quartiles. The body of the boxplot consists of a "box" (hence, the name), which goes from the first quartile (Q1) to the third quartile (Q3).
The middle half of a data set falls within the interquartile range. In a boxplot, the interquartile range is represented by the width of the box (Q3 minus Q1).

Question # 2

a) The probability that a student will pass English is $\frac{3}{8}$ and the probability that he will pass Mathematics is $\frac{3}{4}$. If the probability that he will pass either one subject is $\frac{7}{8}$, what is the probability that he:

i. passes both subjects ii. fails English

iii. passes English but fails Mathematics iv. fails both subjects

b) Show that the function $g(x) = \begin{cases} \frac{x}{6} & if\ x = 1,2,3 \\ 0 & otherwise \end{cases}$ is a probability mass function of a random variable x? Show the probability distribution for $\boldsymbol{x}$ and depict it by a graph.

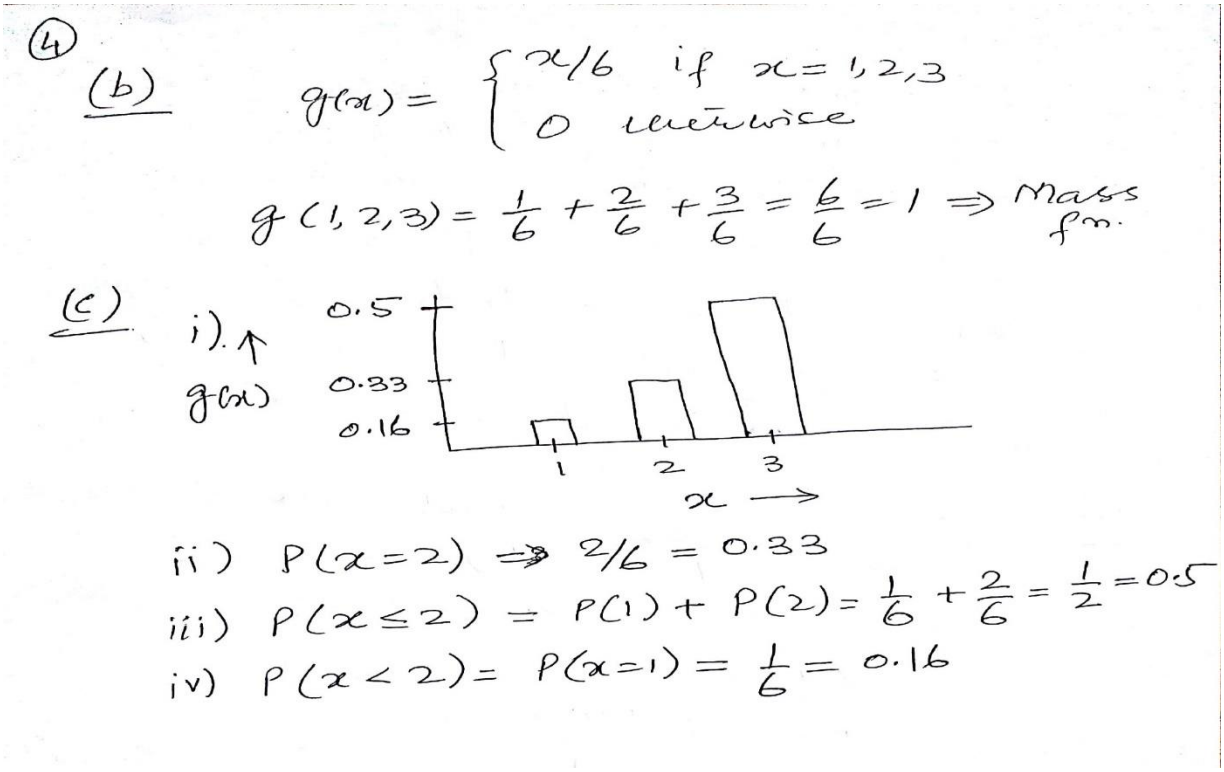
c) For the function given in question "2(b)", solve the following:

i. $P(x = 2)$ ii. $P(x \leq 2)$ iii. $P(x < 2)$ iv. $P\ (x\ \leq 3)$

SOLUTION:

a) Formula "Rule of Addition" : P(AUB) = P(A) + P(B) - P(A ∩ B)

i. (3/8) + (3/4) – (7/8) ii. (1 – (3/8)) iii. (7/8 - 3/ 4) iv. (1 – (7/8))

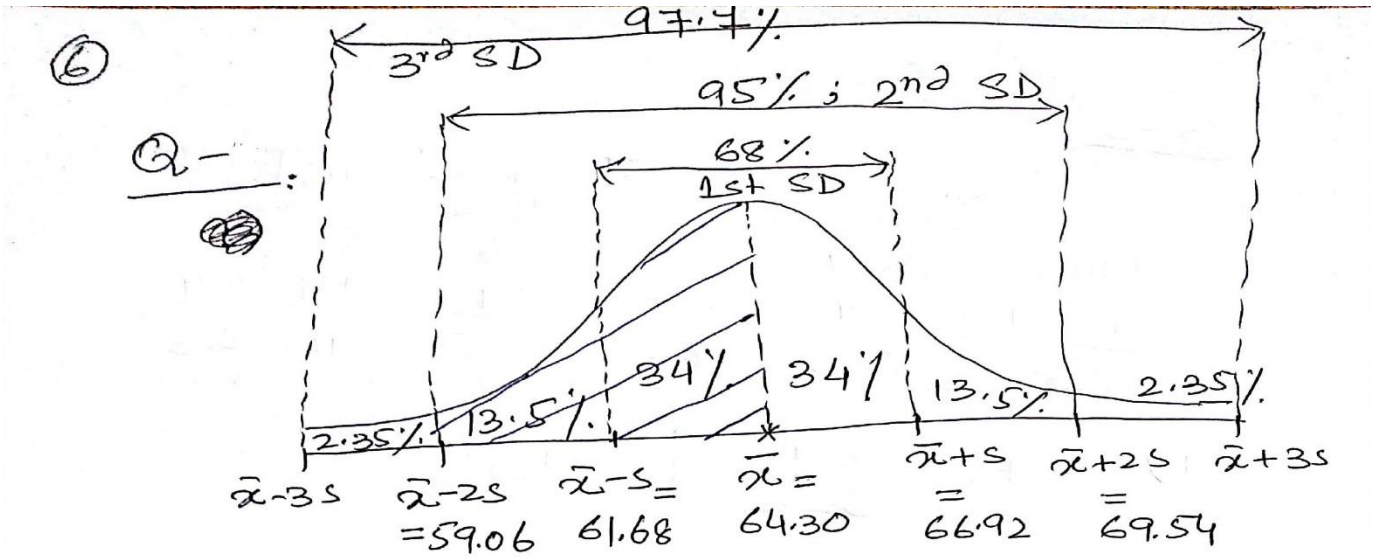


v. P(x<=3) = 1

Question # 3

In a survey, conducted by National Center for Health Statistics (NCHS), the sample mean height of women in the United State (ages 20-29) was 64.30 inches, with a standard deviation of 2.62 inches. Estimate the percent of women whose heights are:

a) Between 59.06 inches and 64.30 inches
b) Between 61.68 inches and 69.54 inches
c) Show the results of part a) and b) with shades in Bell-Shaped Distribution diagram.

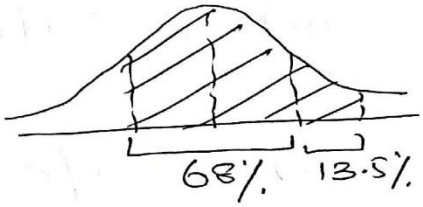


a) % of women b/w 64.30 & 59.06 =

34% + 13.5% = **47.5%** Ans

b) % of women b/w 61.68 and 69.54 =

= 34% + 34% + 13.5% = **81.5%**



c) Shadded.

Question # 4

A study investigated causes of sudden deaths in western region of Paris and France. A sample of 523 such deaths revealed the following:

	Cardiovascular	Cerebral	Respiratory	Others	Total
Males	264	38	36	21	359
Females	89	27	29	19	164
Total	353	65	65	40	523

Suppose one of these cases is randomly selected.

a) What is the probability the person was female? Given the cause was cardiovascular in nature, what is the probability the person was female?
b) Given the person was female, what is the probability the cause was cardiovascular in nature?
c) Given the cause was cerebral or respiratory in nature, what is the probability the person was male?

Q-.

	cardiovas-cular	Cerebral	Respira-tory	others	Σ
Males	264	38	36	21	359
Females	89	27	29	19	164
Σ	353	65	65	40	523

i) $P(F) = \frac{164}{523} \simeq \boxed{0.31}$ Ans

ii) $P(F/CV) = \boxed{\frac{89}{353}}$ Ans

Explanation:

$$P(F/CV) = \frac{P(F \cap CV)}{P(CV)} = \frac{89/523}{353/523}$$

$$\boxed{P(F/CV) = 89/353} \text{ Ans.}$$

iii) $P(CV/F) = \boxed{\frac{89}{164}}$ Ans

iv) $P(M/CUR) = \frac{38+36}{65+65} = \frac{74}{130} = \boxed{0.57}$ Ans

Question # 5

There is a screening test for prostate cancer that looks at the level of PSA in the blood. There are a number of reasons besides cancer that a man can have elevated PSA levels. In addition, many types of cancer develop so slowly that that they are never a problem. Unfortunately, there is currently no test to distinguish the different types and using the test is controversial because it is hard to quantify the accuracy rates and the harm done by false positives.

For this problem we'll call a positive test a true positive if it catches a dangerous type of cancer. We'll assume the following numbers:

Rate of cancer among men over 50 = 0.0005
True positive rate for the test = 0.9
False positive rate for the test = 0.01

Let T be the event a man has a positive test and let D be the event a man has a dangerous type of the disease. Find the probability that a man does have dangerous type of disease given that the tests are positive and negative.

Solution:

We compute all the pieces needed to apply Bayes' rule. We're given

$$P(T|D) = 0.9 \;\Rightarrow\; P(T^c|D) = 0.1, \quad P(T|D^c) = 0.01 \;\Rightarrow\; P(T^c|D^c) = 0.99.$$

$$P(D) = 0.0005 \;\Rightarrow\; P(D^c) = 1 - P(D) = 0.9995.$$

We use the law of total probability to compute $P(T)$:

$$P(T) = P(T|D)\,P(D) + P(T|D^c)\,P(D^c) = 0.9 \cdot 0.0005 + 0.01 \cdot 0.9995 = 0.010445$$

Now we can use Bayes' rule to answer the questions:

$$P(D|T) = \frac{P(T|D)\,P(D)}{P(T)} = \frac{0.9 \times 0.0005}{0.010445} = 0.043$$

$$P(D|T^c) = \frac{P(T^c|D)\,P(D)}{P(T^c)} = \frac{0.1 \times 0.0005}{0.989555} = 5.0 \times 10^{-5}$$