# *Key to Test 1 - Statistics for Data Science*

Test 1 Statistics for Data Science, 4th Semester Course,
RVCE Bangalore Marks 50

1. In an online examination for the course on ML for Sports, the average score was 20 on 50. List out a maximum of 5 possible inferences that you can make about the class performance from this information. **5 Marks**

**Answer:** **a)** The class average is lower than half of the total marks, indicating that the performance is subpar.

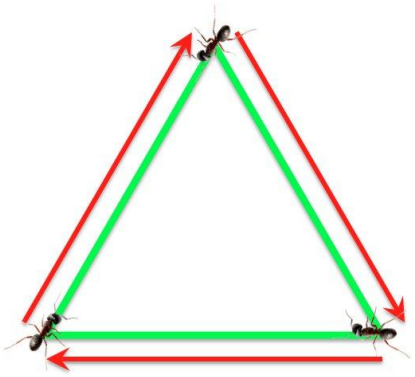
b) Without the knowledge of variance and standard deviation, we will not be able to comment on the distribution of marks.

c) The class average being 20, probably indicates, the students did not indulge in any malpractice.

*The above inferences are only indicative.* Any 5 logical inferences will be fine. If you had written since only an average score is given, we cannot draw any conclusions from here, you will get full marks!

2. There are three ants on a triangle, one at each corner. At a given point of time they all set off for a different corner at random. What is the probability that they don't collide? **5 Marks**

**Answer:**



Each ant has 2 paths to take from the vertex that it is currently located. Therefore, there are 2x2x2 = 8 possibilities for the movement of 3 ants. Out of these 8 possibilities, there are only 2 possibilities which correspond to no collision. Hence

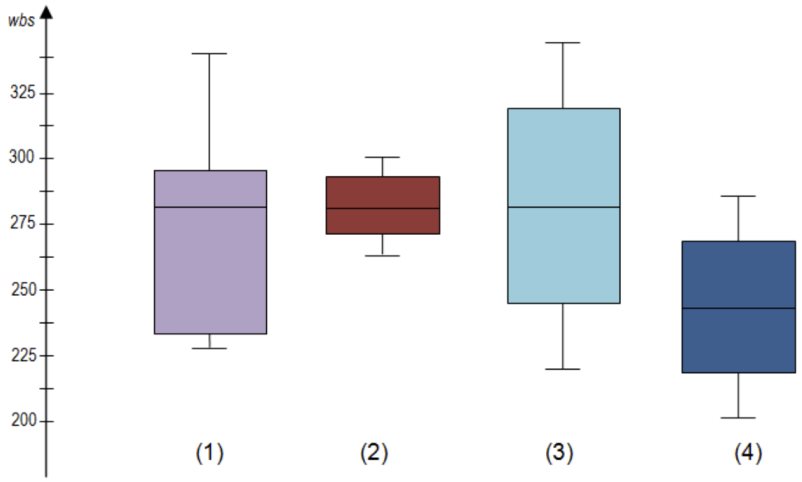
P(no collision) = 2/8 =¼ = 0.25

# *Key to Test 1 - Statistics for Data Science*

3. Given a set of data, we tend to compute the variance and then obtain the standard deviation as the square-root of variance. Why is it that we prefer standard deviation and not variance? **5 Marks**

**Answer:** Standard deviation is on the same scale of measurement as the data and the variance is in squared units. Hence the standard deviation makes sense.

4.




The above boxplot gives the response of students to 4 different questions related to their well-being in college/school. The *y*–axis is the well-being scale.

(a) Which of the 4 questions have a very high level of agreement among students? Justify your answer. **5 Marks**

Question 2 as the boxplot range is very small, depicting high levels of agreement.

(b) Some questions seem to have high levels of disagreement among students, indicating different opinions. Identify the boxplot(s) which indicate this scenario. Justify your answer. **5 Marks**

Questions 1 and Questions 3 have high levels of disagreement as the range of values in both are very large.

# *Key to Test 1 - Statistics for Data Science*

5. Let $X$ be the data as given: $X = 2, 3, 4, 0, 1, 2, 3, 4, 2, 9$.
Recall from our class lecture, the geometrical interpretation of mean as a projection of data onto the vector with all its components equal to 1. Use this interpretation to obtain the deviation vector and subsequently, the length of the deviation vector for the data given above. **10 Marks**

**Answer:** Mean = 3.

$$\mathbf{x} = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 2 \\ 9 \end{pmatrix}; \qquad \bar{\mathbf{x}} = 3 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix};$$

$$Deviation\ vector\ \mathbf{d} = \mathbf{x} - \bar{\mathbf{x}} = \begin{pmatrix} 2-3 \\ 3-3 \\ 4-3 \\ 0-3 \\ 1-3 \\ 2-3 \\ 3-3 \\ 4-3 \\ 2-3 \\ 9-3 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 1 \\ -3 \\ -2 \\ -1 \\ 0 \\ 1 \\ -3 \\ 6 \end{pmatrix}$$

$$\|d\| = \sqrt{(-1)^2 + 0^2 + 1^2 + (-3)^2 + (-2)^2 + (-1)^2 + 0^2 + 1^2 + (-1)^2 + 6^2} = \sqrt{54}$$

Divide by $\sqrt{9}$ or $\sqrt{10}$, both answers will be accepted!

6. An integer is randomly selected between 1 and 50, inclusively.
   (a) Find the probability that the number is not divisible by 8. **3 Marks**
   **Solution:** There are 6 multiples of 8 in numbers from 1 to 50. Therefore the number of numbers not divisible by 8 is 50-6 = 44. Therefore, the probability is 44/50 = 22/25

(b) Find the probability that the number is divisible by 9. **3 Marks**

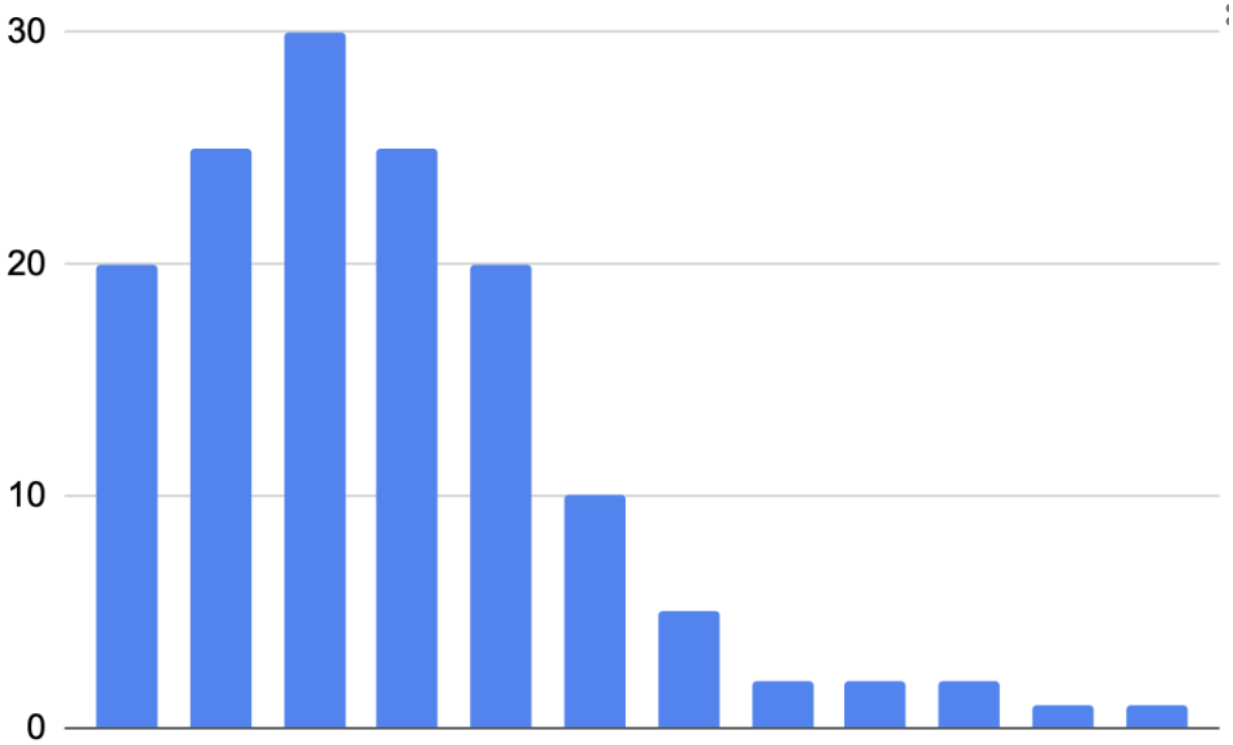
**Solution:** There are 5 multiples of 9 in numbers from 1 to 50. Therefore the number of numbers divisible by 9 is 5. Therefore, the probability is 5/50 = 1/10.

(c) Find the probability that the number is neither divisible by 8 nor by 9. **3 Marks**

**Solution:** There are 39 numbers that are neither divisible by 8 or 9. Therefore the probability is 39/50.

7. Assert or Reject each of the following statements with proper justification.

(a) **Statement 1**: The histogram given below depicts the average salary offered to students in a college through campus placement programme. (*x*-axis depicts the salary band) **3 Marks**




**Solution:** It could very well depict salary distribution among students who have received offers through the campus placement drive. The reason is majority of students get average salary while there are only a few students who get very high package.

# *Key to Test 1 - Statistics for Data Science*

Dear Students, please do not write the figure is not a histogram and is a bar graph etc!

**(b) Statement 2:** Disjoint events are necessarily independent events. **3 Marks**

*They are not independent. If they have to be independent, then the probability of at least one of the events must be 0.*