e-PG Diploma AI & DS (AUG. '25)

Statistical Foundations of Machine Learning

Quiz 1 Solutions

September 19, 2025

MCQs

Question 1. Which of the following is a categorical variable?

(a) Height of students

(b) Number of siblings

(c) Blood type

(d) Age in years

Answer: (c) Blood type - it represents categories without numerical meaning.

Question 2. The probability of drawing an ace from a standard deck of 52 cards is:

(a) $\frac{1}{13}$

(b) $\frac{1}{4}$

(c) $\frac{1}{26}$

(d) $\frac{4}{52}$

Answer: (a) $\frac{1}{13}$ — there are 4 aces in 52 cards.

Question 3. Which of the following best illustrates ratio data?

(a) Temperature in Celsius

(b) Age in years

(c) Shoe size

(d) IQ scores

Answer: (b) Age in years — it has a true zero and allows ratio comparisons.

Question 4. A population in statistics refers to:

(a) A group of individuals selected for study

(b) The entire set of individuals or items of interest

(c) Only the people living in a country

(d) A small subset of data

Answer: (b) The entire set of individuals or items of interest.

Question 5. A positive correlation between two variables means:

(a) As one increases, the other decreases

(b) As one increases, the other also increases

(c) Both variables are constant

(d) No relationship exists

Answer: (b) As one increases, the other also increases.

Question 6. A bar graph is most appropriate for displaying:

(a) Numerical continuous data

(b) Categorical data

(c) Ratio data

(d) Scatter data

Answer: (b) Categorical data.

Question 7. A histogram differs from a bar graph because:

(a) It uses bars

(b) It displays categorical data

(c) It represents continuous data with no gaps between bars

(d) It is circular in shape

Answer: (c) A histogram represents continuous data with no gaps.

Question 8. Let the total population size be N and we take a sample of size n from it. In simple random sampling without replacement, the probability of selecting any one unit on the first draw is:

(a) $\frac{1}{N}$

(b) $\frac{1}{n}$

(c) $\frac{N}{n}$

(d) $\frac{n}{N}$

Answer: (a) $\frac{1}{N}$. Exactly one of N units is chosen on the first draw, so $P = \frac{1}{N}$.

Question 9. After purchasing an item, a shopkeeper asks you to rate your satisfaction on a 1–5 scale: 1 = Very Dissatisfied, 2 = Dissatisfied, 3 = Neutral, 4 = Satisfied, 5 = Very Satisfied.

This data type is best classified as:

(a) Nominal

(b) Ordinal

(c) Interval

(d) Ratio

Answer: (b) Ordinal.

Question 10. Which chart is best for displaying the frequency distribution of continuous numerical data?

(a) Pie chart

(b) Histogram

(c) Line chart

(d) Box plot

Answer: (b) Histogram. It groups data into adjacent bins with touching bars.

Question 11. In sequential sampling, the sampling process stops when:

(a) A predetermined sample size is reached

(b) The acceptance or rejection boundary is crossed

(c) All units in the lot are inspected

(d) The defect rate exceeds a fixed value

Answer: (b) The acceptance or rejection boundary is crossed.

Question 12. Suppose the correlation coefficient between GDP and Investment is $r = 0.82$. Consider: 1) There is a strong positive linear relationship between GDP and Investment. 2) As GDP increases, Investment also tends to increase. 3) Knowing GDP lets us determine Investment exactly.

Which statements are correct?

(a) 1 and 2 only

(b) 1 only

(c) 1 and 3 only

(d) 1, 2, and 3

Answer: (a) 1 and 2 only. Correlation does not imply exact determination.

Question 13. Consider a dataset of numerical observations $x_1, \ldots, x_n$ with mean $\bar{x}$. Which statements are correct? 1) $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. 2) $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$ (unless all x_i are equal). 3) $\sum_{i=1}^{n}(x_i - \bar{x})$ can be positive or negative depending on the data.

(a) 1 only

(b) 2 only

(c) 1 and 2 only

(d) 1 and 3 only

Answer: (c) 1 and 2 only. Deviations sum to zero; squared deviations are nonnegative and strictly positive unless all x_i are equal.

Question 14. An examiner computes the mean, median, and standard deviation of last semester's scores for a class of 150 students, then uses these to predict future performance. Which statements are correct? 1) Computing mean/median/SD is descriptive statistics. 2) Using them to predict future performance is inferential statistics. 3) Both the calculation and prediction are descriptive statistics. 4) Only the prediction is descriptive; calculation is inferential.

(a) 1 and 2 only

(b) 2 and 3 only

(c) 1 and 3 only

(d) 1, 2, and 4 only

Answer: (a) 1 and 2 only.

Question 15. A standard deck is well shuffled. Two cards are drawn without replacement, one at a time. Let A be "first card is a heart," and B be "second card is red." Find $P(A \mid B)$.

(a) $\frac{25}{204}$

(b) $\frac{25}{102}$

(c) $\frac{25}{51}$

(d) $\frac{1}{2}$

Answer: (b) $\frac{25}{102}$. Note $P(B) = \frac{1}{2}$. Also $P(A \cap B) = \frac{13}{52} \cdot \frac{25}{51} = \frac{25}{204}$. Thus $P(A \mid B) = \frac{25/204}{1/2} = \frac{25}{102}$.

Subjective Questions

Question 16. The following data represent the ages (in years) of a random sample of people who attended a recent soccer match:

23	35	14	37	38	15	45
12	40	27	13	18	19	23
37	20	29	49	40	65	53
18	17	23	27	29	31	42
35	38	22	20	15	17	21

(a) Compute the mean, median, and mode of the ages. [1 marks]

(b) Compute the standard deviation of the ages. [1 marks]

(c) Provide a five-number summary for the data. [0.5 marks]

(a) **Mean, median, mode.**
Step 1 (sort). Sorted data:

12	13	14	15	15	17	17
18	18	19	20	20	21	22
23	23	23	**27**	27	29	29
31	35	35	37	37	38	38
40	40	42	45	49	53	65

Mean: $\bar{x} = \dfrac{\sum x}{n} = \dfrac{1007}{35} \approx 28.7714$.
Median: $n = 35 \Rightarrow$ middle is the 18th value $\Rightarrow$ median $= 27$.
Mode: value with highest frequency is 23 (appears 3 times) $\Rightarrow$ mode $=$ 23.

(b) **Standard deviation (sample).**
Use $s = \sqrt{\dfrac{\sum(x-\bar{x})^2}{n-1}}$. Here $\sum(x-\bar{x})^2 = 5486.171428571429$, so

$$s = \sqrt{\frac{5486.171428571429}{35-1}} = \sqrt{161.35798319} \approx 12.7027.$$

(c) **Five-number summary.**
min $= 12$, $Q_1 =$ median of lower half $= 18$, median $= 27$, $Q_3 =$ median of upper half $= 38$, max $= 65$.

Question 17. A corporation has just received new machinery that must be installed and checked before it becomes operational. The accompanying table shows a manager's probability assessment for the number of days required before the machinery becomes operational.

Number of days	3	4	5	6	7
Probability	0.08	0.24	0.41	0.20	0.07

Let A be the event "it will be more than four days before the machinery becomes operational," and let B be the event "it will be less than six days before the machinery becomes available."

(a) Find $P(A)$. [0.5 marks]

(b) Find $P(B)$. [0.5 marks]

(c) Find $P(A^c)$. [0.5 marks]

(d) Find $P(A \cap B)$. [0.5 marks]

(e) Find $P(A \cup B)$. [0.5 marks]

(a) **Event A:** $X > 4 \Rightarrow \{5, 6, 7\}$.

$$P(A) = 0.41 + 0.20 + 0.07 = 0.68.$$

(b) **Event B:** $X < 6 \Rightarrow \{3, 4, 5\}$.

$$P(B) = 0.08 + 0.24 + 0.41 = 0.73.$$

(c) **Complement of A:** $A^c = \{3, 4\}$.

$$P(A^c) = 0.08 + 0.24 = 0.32.$$

(d) **Intersection $A \cap B$:** Days both > 4 and $< 6 \Rightarrow \{5\}$.

$$P(A \cap B) = 0.41.$$

(e) **Union $A \cup B$:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

$$P(A \cup B) = 0.68 + 0.73 - 0.41 = 1.00.$$

Question 18. Construct a stem-and-leaf display for the hours that 20 students spent studying for a marketing test:

3.5	2.8	4.5	6.2	4.8
2.3	2.6	3.9	4.4	5.5
5.2	6.7	3.0	2.4	5.0
3.6	2.9	1.0	2.8	3.6

[2.5 marks] **Solution:**

1. Sort the data:

1.0	2.3	2.4	2.6	2.8
2.8	2.9	3.0	3.5	3.6
3.6	3.9	4.4	4.5	4.8
5.0	5.2	5.5	6.2	6.7

2. Construct stem-and-leaf (stem = integer part, leaf = first decimal):

Stem	Leaf
1	0
2	3 4 6 8 8 9
3	0 5 6 6 9
4	4 5 8
5	0 2 5
6	2 7

3. Read as: e.g. "2 — 3" = 2.3 hours.

Question 19. A random sample of data has a mean of 75 and a variance of 25. Use Chebyshev's theorem to determine the percent of observations between 65 and 85. [2.5 marks]

Given: $\mu = 75,\ \sigma^2 = 25 \Rightarrow \sigma = 5$.

The interval $[65, 85]$ has distance 10 from the mean, so

$$k = \frac{10}{\sigma} = \frac{10}{5} = 2.$$

By Chebyshev's theorem,

$$P\big(|X - \mu| < k\sigma\big) \geq 1 - \frac{1}{k^2}.$$

Hence

$$P\big(|X - 75| < 2\sigma\big) \geq 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4}.$$

Therefore, at least 75% of the observations lie between 65 and 85.

Question 20. The following data shows the monthly sales (in thousands of rupees) for Gilotti Pizzeria over a 12-month period:

$$6,\ 8,\ 10,\ 12,\ 14,\ 9,\ 11,\ 7,\ 13,\ 11$$

Based on this data, answer the following:

(a) Give the five-number summary (minimum, Q_1, median, Q_3, maximum). [1 marks]

(b) Calculate the mean, mode, and interquartile range (IQR). [1 marks]

(c) Construct a boxplot to represent the data distribution. (Sketch or describe the boxplot showing median, quartiles, whiskers and any potential outliers.) [0.5 marks]

The given data (monthly sales in thousands of rupees) is:

$$6,\ 8,\ 10,\ 12,\ 14,\ 9,\ 11,\ 7,\ 13,\ 11$$

First, sort the data in ascending order:

$$6,\ 7,\ 8,\ 9,\ 10,\ 11,\ 11,\ 12,\ 13,\ 14$$

(a) **Five-number summary:** Minimum $= 6$, $Q_1 = 8$, Median $= 10.5$, $Q_3 = 12$, Maximum $= 14$.

(b) **Mean, Mode, and IQR:**

$$\text{Mean} = \frac{6+7+8+9+10+11+11+12+13+14}{10} = \frac{101}{10} = 10.1$$

$$\text{Mode} = 11 \quad \text{(since it occurs twice)}$$

$$\text{IQR} = Q_3 - Q_1 = 12 - 8 = 4$$

(c) **Boxplot:**

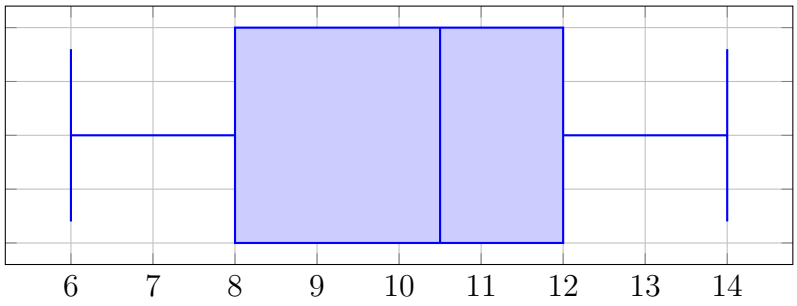
The boxplot has a box from $Q_1 = 8$ to $Q_3 = 12$, with the median at 10.5. Whiskers extend from the minimum value (6) to the maximum value (14).

Since no data point lies outside the range

$$[Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR] = [2, 18],$$

there are no outliers.

Boxplot of Monthly Sales (in thousands of rupees)



Sales (in thousands of rupees)