Version 2 - Mid Semester Exam - Statistical Foundations for ML

September, 2025

Information

- Total marks: 30
- There are two sections in the exam.
  - Section A contains Multiple Choice Questions and carries 10 marks. Each question carries 1 mark.
  - Section B contains Numerical questions and carries 20 marks. Each question carries 2 or 3 marks.

Section A: Multiple Choice Questions (10 marks)

Each question carries 1 mark. Select the best option. Correct answers are indicated.

Q1. Which is a measure of dispersion?

(a) Mean
(b) Median
(c) Variance **(Correct)**
(d) Mode

Q2. For $X \sim \text{Bin}(n, p)$, the mean and variance are:

(a) $(np,\ np(1-p))$ **(Correct)**
(b) $(p,\ 1-p)$
(c) $(n,\ p(1-p))$
(d) $(np^2,\ np)$

Q3. The Poisson approximation to the Binomial is appropriate when:

(a) n large, p small, $\lambda = np$ fixed **(Correct)**
(b) n small
(c) $p \approx 0.5$
(d) Variance equals mean

Q4. Central Limit Theorem (sample mean): as n increases,

(a) $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Rightarrow N(0, 1)$ **(Correct)**

(b) $\dfrac{\bar{X} - \mu}{\sigma} \Rightarrow N(0, 1)$

(c) $\dfrac{S_n}{n} \Rightarrow \text{Exp}(1)$

(d) $\dfrac{S_n - n\mu}{\sigma} \Rightarrow N(0, 1)$

Q5. What is relative frequency?

(a) class frequency / total frequency **(Correct)**

(b) class frequency / class width

(c) class frequency / (class width × total frequency)

(d) class frequency

Q6. A dataset has mean 50 and standard deviation 5. Chebyshev's inequality guarantees that at least what proportion of observations lie between 40 and 60?

(A) 50%

(B) 68%

(C) 75% **(Correct)**

(D) 95%

Q7. In a study of hours studied vs. test score, the sample correlation r will most plausibly be:

(A) Negative and strong

(B) Near zero

(C) Positive and strong **(Correct)**

(D) Undefined

Q8. If Y is replaced by $Y^\star = 3Y + 10$, then the sample correlation between X and $Y^\star$ equals:

(A) $3r$ **(Correct)**

(B) $\dfrac{r}{3}$

(C) r

(D) $-r$

Q9. Which of the following represents a measure of central tendency?

(a) Range

(b) Skewness

(c) Standard deviation

(d) Mode **(Correct)**

Q10. We have some continuous data. We calculated Q_1, Q_2, Q_3 and the inter-quartile range (IQR). One value was less than $Q_1 - 1.5 \times \text{IQR}$. If we plot a box-plot, where should this value lie?

(a) In the box region i.e. $[Q_1, Q_3]$

(b) In the upper whisker

(c) In the lower whisker

(d) None of the above **(Correct)**

Section B: Subjective Questions

Each question carries 2 or 3 marks as indicated. Provide complete workings.

Q1. (3 marks) Let $f(x,y) = 2e^{-x}e^{-2y}$ for $x > 0$, $y > 0$. Compute $P(X > 1,\ Y < 1)$.

Q2. (2 marks) Let $X \sim \text{Bin}(10, 0.4)$. Compute $P(X \leq 6)$. Let $Y \sim \text{Bin}(20, 0.4)$. Compute $P(Y \geq 13) = 1 - P(Y \leq 12)$. State your approach clearly.

Q3. (3 marks) Poisson approximation: For $X \sim \text{Bin}(n,p)$ with n large, p small, and $\lambda = np$ fixed, give approximations for $P(X = 0)$ and $P(X \geq 1)$.

Q4. (3 marks) If $X \sim N(\mu, \sigma^2)$, find the distribution of $Y = \alpha X + \beta$. If $X_1, \dots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, find the distributions of $S_n = \sum_{i=1}^n X_i$ and $\bar{X}$.

Q5. (3 marks) Find the sample variance of the first 10 natural numbers $\{1, 2, 3, \dots, 10\}$. Also find the sample variance for $\{5, 6, 7, \dots, 14\}$. Comment on the results.

Q6. (3 marks) A paired/bi-variate data was given to students for analysis. A student reported covariance 10, with variances 16 and 4 for the 1st and 2nd variables respectively. Comment and justify whether the calculation is correct.

Q7. (3 marks) The joint pmf of (X, Y) is

$X\backslash Y$	0	1	2
0	0.1	0.1	0.1
1	0.2	0.1	0.1
2	0.1	0.1	0.1

(a) Verify it is a valid joint distribution. (b) Find marginals of X and Y. (c) Compute $P(X = 1,\ Y \leq 1)$. (d) Find $E[X]$ and $E[Y]$. (e) Are X and Y independent?