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

Statistical Foundations of Machine Learning

Quiz 2

November 15, 2025

---

## Multiple Choice Questions (MCQs)

**Question 1.** Consider the following statements, where $X, Y, Z$ are discrete random variables. Which of these statements are true?

(A) If $X$ and $Y$ are independent and $Y$ and $Z$ are independent, then $X$ and $Z$ are independent.

(B) If $X$ and $Y$ are independent, then they are conditionally independent given $Z$.

(C) If $X$ and $Y$ are conditionally independent given $Z$, then they are independent.

Choose the correct option:

(a) (A) and (B)

(b) (B) and (C)

(c) Only (A)

(d) **None of the above**

**Question 2.** Four fair dice are rolled. Find the expected total of the rolls.

(a) 10

(b) 12

(c) **14**

(d) 16

**Question 3.** Let $X, Y \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Consider:

(A) $X+Y$ and $X-Y$ are i.i.d. $\mathcal{N}(0,2)$.

(B) $\text{Cov}(X+Y, X-Y) = 0$.

(C) $X+Y$ is independent of $X-Y$.

Which of the above statements are true?

(a) (A) and (B)

(b) Only (A)

(c) (A) and (C)

(d) **(A), (B), and (C)**

**Question 4.** Which of the following statements about the Student-$t$ distribution $t_n$ are true?

(A) If $T \sim t_n$, then $-T \sim t_n$.

(B) As $n \to \infty$, the $t_n$ distribution approaches the standard Normal distribution.

(a) Only (A)

(b) Only (B)

(c) **Both (A) and (B)**

(d) None

**Question 5.** If $X \sim N(25, 16)$, what is the probability that $X$ lies between 13 and 37?

(a) Approximately 68%

(b) Approximately 95%

(c) **Approximately 99.7%**

(d) Cannot be determined

**Question 6.** A **Type II error** occurs when:

(a) We reject the null hypothesis when it is true.

(b) **We fail to reject the null hypothesis when it is false.**

(c) We reject the alternative hypothesis when it is true.

(d) We accept the alternative hypothesis when it is false.

**Question 7.** Which of the following statements are true?

(A) A statistical hypothesis is a statement about the nature of a population.

(B) A test statistic is determined from the sample data.

(C) The critical region is the set of values of the test statistic for which the null hypothesis is accepted.

(a) **(A) and (B) only**

(b) (B) and (C) only

(c) (A) and (C) only

(d) (A), (B), and (C)

**Question 8.** When the population standard deviation $\sigma$ is known, the hypothesis about the population mean is tested by:

(a) $t$-test

(b) $Z$**-test**

(c) $\chi^2$-test

(d) $F$-test

**Question 9.** As the sample size increases, the $t$ distribution becomes more similar to the:

(a) $\chi^2$ distribution

(b) Uniform distribution

(c) **Normal distribution**

(d) $F$ distribution

**Question 10.** The $p$-value represents:

(a) **The smallest significance level at which $H_0$ can be rejected**

(b) The probability of Type II error

(c) The largest significance level at which $H_0$ can be rejected

(d) The probability of Type I error

**Question 11.** A professor sees students during office hours. Time spent follows an exponential distribution with mean 10 minutes. What is $P(X < 20)$?

(a) 0.1353

(b) 0.5

(c) **0.8647**

(d) 0.9817

**Question 12.** The lives of spark plugs are $N(60{,}000, 4{,}000^2)$. A sample of 16 plugs has $\bar{X} = 58{,}500$. What is $P(\bar{X} \leq 58{,}500)$?

(a) **0.0668**

(b) 0.4332

(c) 0.9332

(d) 0.0175

**Question 13.** A café records daily customer counts: $12, 8, 15, 7, 9, 10, 6, 5, 14, 8$. Estimate the proportion of days with $\leq 8$ customers. [Give marks for 0.5 and cut for 0.37]

(a) 0.25

(b) 0.37

(c) **0.50**

(d) 0.63

**Question 14.** A coating's thickness (mm) from 9 samples is given by

$$19.8,\ 21.2,\ 18.6,\ 20.4,\ 21.6,\ 19.8,\ 19.9,\ 20.3,\ 20.8.$$

Find a 90% confidence interval for the population variance (assume normality). [Give marks for everyone]

(a) **(0.406, 2.305)**

(b) (0.637, 1.518)

(c) (0.553, 1.122)

(d) (2.733, 15.507)

**Question 15.** A company claims that its new battery lasts at least 10 hours on average. A consumer group tests this claim with the following hypotheses:

$$H_0 : \mu = 10 \quad vs \quad H_1 : \mu < 10 \tag{1}$$

Which of the following describes a Type I error in this context?

**(a) Concluding that the average battery life is less than 10 hours when it actually is 10 hours or more.**

(b) Concluding that the average battery life is 10 hours or more when it actually is less than 10 hours.

(c) Failing to test enough batteries to detect a difference from 10 hours.

(d) Using the wrong level of significance for the test.

**Question 16.** Which best distinguishes a sample from a population?

(a) A sample includes every member of a group, while a population includes only a few selected members.

**(b) A sample is a subset of the population that is used to draw conclusions about the entire population.**

(c) A population is always smaller than a sample.

(d) A population consists only of data collected from experiments, while a sample comes from surveys.

**Question 17.** Which of the following best describes the Central Limit Theorem?

(a) It states that the mean of a population is always normally distributed, regardless of the population's shape.

**(b) It states that as the sample size increases, the sampling distribution of the sample mean approaches a normal distribution, regardless of the shape of the population.**

(c) It states that large samples always have the same mean as the population mean.

(d) It states that population data become normal when the population size is large enough.

**Question 18.** When attempting to land a drone on a target in two-dimensional space, suppose the horizontal and vertical positioning errors are independent normal random variables each with mean 0 and standard deviation 1.5 meters. Find the probability that the distance between the actual landing point and the target exceeds 2.5 meters.

**(a) 0.249**

(b) 0.135

(c) 0.317

(d) 0.05

**Question 19.** If $X_1, X_2, \ldots, X_n$ are independent exponential random variables with respective rate parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$, which of the following statements is **true** about $Y = \min(X_1, X_2, \ldots, X_n)$?

(a) $Y$ is not exponential for $n > 1$.

(b) $Y$ follows an exponential distribution with parameter $\frac{1}{n}\sum_{i=1}^{n} \lambda_i$.

**(c) $Y$ follows an exponential distribution with parameter $\sum_{i=1}^{n} \lambda_i$.**

(d) $Y$ follows a gamma distribution with shape $n$ and rate $\lambda_i$.

**Question 20.** For a population with $\mu = 100$, $\sigma^2 = 81$, and sample size $n = 25$, find mean and variance of $\bar{X}$:

(a) Mean $= 100$, Variance $= 81$

(b) **Mean = 100, Variance = 3.24**

(c) Mean = 20, Variance = 81

(d) Mean = 100, Variance = 9

**Question 21.** Given a population with a mean of $\mu = 100$ and a variance of $\sigma^2 = 81$, the central limit theorem applies when the sample size is $n > 25$. A random sample of size $n = 25$ is obtained. What are the mean and variance of the sampling distribution for the sample means? [Award marks for everyone.]

(a) $S^2$ always underestimates $\sigma^2$

(b) $S^2$ **is an unbiased estimator of** $\sigma^2$**, i.e.,** $E[S^2] = \sigma^2$

(c) $S^2$ equals $\sigma^2$ only when $n$ is large

(d) $S^2$ estimates the population mean

**Question 22.** Which of the following statements about the sample variance $S^2$ is true?

(a) $S^2$ always underestimates the population variance $\sigma^2$.

**(b)** $\boldsymbol{S^2}$ **is an unbiased estimator of the population variance, i.e.** $\boldsymbol{E[S^2] = \sigma^2}$

(c) $S^2$ equals the population variance only when the sample size is large.

(d) $S^2$ estimates the population mean, not the variance.

**Question 23.** Which of the following best describes the principle of the maximum likelihood estimator (MLE)?

(a) The MLE chooses the parameter values that make the sample variance smallest.

**(b) The MLE chooses the parameter values that make the observed data most probable.**

(c) The MLE always equals the sample mean, regardless of the distribution.

(d) The MLE minimizes the expected value of the squared error between the estimate and the true parameter.

**Question 24.** Let $X$ be a random variable that follows a Gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\lambda > 0$, denoted by

$$X \sim \text{Gamma}(\alpha, \lambda).$$

Which of the following statements is **true** about the Gamma distribution?

(a) The mean and variance of $X$ are $E[X] = \lambda$, $\text{Var}(X) = \alpha$.

**(b) The mean and variance of $X$ are $E[X] = \frac{\alpha}{\lambda}$, $\text{Var}(X) = \frac{\alpha}{\lambda^2}$.**

(c) The Gamma distribution is always symmetric about its mean.

(d) The probability density function of $X$ is $f(x) = \dfrac{1}{\sqrt{2\pi\lambda}} e^{-\frac{(x-\alpha)^2}{2\lambda}}$.

# Subjective Questions

**Question 1**

There are 100 slips of paper in a hat, each of which has one of the numbers $1, 2, \ldots, 100$ written on it, with no number appearing more than once. Five of the slips are drawn, one at a time. First consider random sampling **with replacement** (with equal probabilities).

(a) What is the distribution of how many of the drawn slips have a value of at least 80 written on them? [0.5 marks]

(b) What is the distribution of the value of the $j$th draw (for $1 \leq j \leq 5$)? [0.5 marks]

(c) What is the probability that the number 100 is drawn at least once? [1 mark]

**Solution:**

(a) Suppose that n independent Bernoulli trials are performed, each with the same success probability p. Let X be the number of successes. The distribution of X is called the Binomial distribution with parameters n and p. By the story of the Binomial, the distribution is

$$\text{Bin}(5, 0.21).$$

(b) Let $X_j$ be the value of the $j$th draw. By symmetry,

$$X_j \sim \text{DUnif}(1, 2, \ldots, 100).$$

There aren't certain slips that love being chosen on the $j$th draw and others that avoid being chosen then; all are equally likely.

(c) Taking complements,

$$P(X_j = 100 \text{ for at least one } j) = 1 - P(X_1 \neq 100, \ldots, X_5 \neq 100).$$

By the naive definition of probability, this is

$$1 - \left(\frac{99}{100}\right)^5 \approx 0.049.$$

---

**Question 2**
Let $X$ and $Y$ have joint PDF

$$f_{X,Y}(x, y) = cxy, \quad \text{for } 0 < x < y < 1.$$

1. Find $c$ to make this a valid joint PDF. [1 marks]
2. Are $X$ and $Y$ independent? [1 marks]
3. Find the marginal PDFs of $X$ and $Y$. [1 marks]

**Solution:**
**(1) Find** c:

$$\int_0^1 \int_0^y cxy \, dx \, dy = 1$$

Compute the inner integral:

$$\int_0^y x\,dx = \frac{y^2}{2}$$

So,

$$1 = c\int_0^1 y \cdot \frac{y^2}{2}\,dy = c\int_0^1 \frac{y^3}{2}\,dy = c \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{c}{8}$$

$$\boxed{c = 8}$$

**(2) Check Independence:**
For independence, we require

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for all $x, y$ in the support.

We will compute the marginals below and verify that this condition does not hold.

—

**(3) Marginal PDFs:**
<u>For $X$:</u>

$$f_X(x) = \int_{y=x}^1 f_{X,Y}(x,y)\,dy = \int_{y=x}^1 8xy\,dy = 8x\left[\frac{y^2}{2}\right]_{y=x}^1 = 8x\left(\frac{1-x^2}{2}\right)$$

$$\boxed{f_X(x) = 4x(1-x^2)}, \quad 0 < x < 1$$

<u>For $Y$:</u>

$$f_Y(y) = \int_{x=0}^y f_{X,Y}(x,y)\,dx = \int_{x=0}^y 8xy\,dx = 8y\left[\frac{x^2}{2}\right]_0^y = 8y \cdot \frac{y^2}{2} = 4y^3$$

$$\boxed{f_Y(y) = 4y^3}, \quad 0 < y < 1$$

—

**Independence:**

$$f_{X,Y}(x,y) = 8xy, \quad f_X(x)f_Y(y) = [4x(1-x^2)][4y^3] = 16xy^3(1-x^2)$$

These are not equal for all $x, y$, hence

$$\boxed{X \text{ and } Y \text{ are not independent.}}$$

---

**Question 3**
The life of a particular brand of television picture tube is known to be normally distributed with a population standard deviation of $\sigma = 400$ hours. A random sample of $n = 20$ tubes resulted in a sample mean of $\bar{X} = 9000$ hours. Obtain a 90% confidence interval estimate of the mean lifetime of such a tube.(The $Z$-score that leaves 0.05 in the upper tail is 1.645.) [2 marks]

**Solution:**

$$\text{Given: } \sigma = 400, \ n = 20, \ \bar{X} = 9000, \ Z_{0.05} = 1.645$$

$$\text{Standard Error (SE)} = \frac{\sigma}{\sqrt{n}} = \frac{400}{\sqrt{20}} = 89.44$$

$$\text{Confidence Interval: } \bar{X} \pm Z_{\alpha/2} \times SE = 9000 \pm 1.645(89.44)$$

$$= 9000 \pm 147.1$$

$$\boxed{(8852.9, \ 9147.1)}$$

Hence, the 90% confidence interval for the mean lifetime is:

$$(8852.9, \ 9147.1)$$

---

**Question 4** To test the hypothesis

$$H_0 : \mu = 105 \quad \text{against} \quad H_1 : \mu \neq 105,$$

a sample of size $n = 9$ is drawn. If the sample mean is $\bar{X} = 100$, find the $p$-value when the population standard deviation is known to be 15. [2 marks]

**Solution:**

$$\mu_0 = 105, \quad \bar{X} = 100, \quad \sigma = 15, \quad n = 9$$

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{100 - 105}{15/\sqrt{9}} = \frac{-5}{5} = -1.00$$

$$p = 2P(Z > |z_{\text{obs}}|) = 2P(Z > 1.00)$$

From standard normal tables,

$$P(Z > 1.00) = 0.1587$$

$$\therefore \quad p = 2(0.1587) = 0.3174$$

$$\boxed{p = 0.3174}$$

---

**Question 5** Given a population with mean $\mu = 400$ and variance $\sigma^2 = 1,600$ (Note: $\sigma^2 = 1600$ implied, if $\sigma = 1600$ is intended, the question would be $\sigma = 40$), the central limit theorem applies when the sample size is $n \geq 25$. A random sample of size $n = 35$ is obtained.

(a) What are the mean and variance of the sampling distribution for the sample means? [0.5 marks]

(b) What is the probability that $\bar{x} > 412$? [0.5 marks]

(c) What is the probability that $393 \leq \bar{x} \leq 407$? [1 mark]

(d) What is the probability that $\bar{x} \leq 389$? [0.5 marks]

**Solution:**

$$\text{Given: } \mu = 400, \ \sigma^2 = 1600 \ (\sigma = 40), \ n = 35.$$

$$\text{By the CLT, } \bar{X} \approx N\left(\mu,\ \frac{\sigma^2}{n}\right) = N\left(400,\ \frac{1600}{35}\right).$$

(a) Mean and variance of the sampling distribution:

$$E(\bar{X}) = 400, \qquad \mathrm{Var}(\bar{X}) = \frac{1600}{35} \approx 45.7143,$$

and the standard error is

$$\sigma_{\bar{X}} = \sqrt{\frac{1600}{35}} \approx 6.7612.$$

(b) $P(\bar{X} > 412)$.

$$z = \frac{412 - 400}{\sigma_{\bar{X}}} = \frac{12}{6.7612} \approx 1.7748,$$

$$P(\bar{X} > 412) = 1 - \Phi(1.7748) \approx 0.03796.$$

(c) $P(393 \le \bar{X} \le 407)$.

$$z_1 = \frac{393 - 400}{6.7612} \approx -1.0353, \qquad z_2 = \frac{407 - 400}{6.7612} \approx 1.0353,$$

$$P(393 \le \bar{X} \le 407) = \Phi(1.0353) - \Phi(-1.0353) \approx 0.69948.$$

(d) $P(\bar{X} \le 389)$.

$$z = \frac{389 - 400}{6.7612} \approx -1.6269,$$

$$P(\bar{X} \le 389) = \Phi(-1.6269) \approx 0.05188.$$

$$\boxed{\begin{array}{l} (a)\ E(\bar{X}) = 400,\ \mathrm{Var}(\bar{X}) = \frac{1600}{35} \approx 45.7143, \\ (b)\ P(\bar{X} > 412) \approx 0.03796, \\ (c)\ P(393 \le \bar{X} \le 407) \approx 0.69948, \\ (d)\ P(\bar{X} \le 389) \approx 0.05188. \end{array}}$$

---

**Question 6** A firm employs 189 junior accountants. In a random sample of 50 of these, the mean number of hours over-time billed in a particular week was 9.7, and the sample standard deviation was 6.2 hours.

(a) Find a 95% confidence interval for the mean number of hours overtime billed per junior accountant in this firm that week. [1 mark]

(b) Find a 99% confidence interval for the total number of hours overtime billed by junior accountants in the firm during the week of interest. [1 mark]

**Solution:**

$$\text{Given: } n = 50,\ \bar{x} = 9.7,\ s = 6.2,\ df = n - 1 = 49.$$

$$\text{Standard error: } \mathrm{SE}(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{6.2}{\sqrt{50}} \approx 0.8768124.$$

$$t_{0.025,49} \approx 2.0096, \qquad t_{0.005,49} \approx 2.678.$$

(a) 95% CI for the mean:

$$\bar{x} \pm t_{0.975,49}\,\mathrm{SE}(\bar{X}) = 9.7 \pm 2.0096(0.8768124) = 9.7 \pm 1.7620422,$$

$$\boxed{95\% \text{ CI for } \mu :\ (7.93796,\ 11.46204)}$$

(b) 99% CI for the total hours (first find 99% CI for the mean):

$$9.7 \pm t_{0.995,49}\,\mathrm{SE}(\bar{X}) = 9.7 \pm 2.678(0.8768124) = 9.7 \pm 2.3481036,$$

so the 99% CI for the mean is $(7.35190,\ 12.04810)$.

Multiplying by $N = 189$ gives the CI for the total:

$$\boxed{(1389.51,\ 2277.09)}$$

---