📘 Statistics Midterm Comprehensive Practice Question Bank

This question bank covers all major units — from probability to regression diagnostics — based on the study guide and key lecture topics.
The questions mix **conceptual understanding, computation, and interpretation**.

🧮 Unit 1: Probability and Random Variables

Multiple Choice

1. Which of the following statements about probabilities is **true**?
   - (A) The probability of the sample space is 0.
   - (B) Probabilities can be negative.
   - (C) The sum of probabilities of all possible outcomes equals 1.
   - (D) The probability of the complement of an event is 0.

2. If A and B are **mutually exclusive**, then:
   - (A) $P(A \cap B) = P(A) + P(B)$
   - (B) $P(A \cap B) = 0$
   - (C) $P(A \mid B) = 1$
   - (D) A and B are independent

3. Events A and B are **independent** if:
   - (A) $P(A \cap B) = P(A) + P(B)$
   - (B) $P(A \cap B) = P(A)P(B)$
   - (C) $P(A \mid B) = P(A \cap B)$
   - (D) $P(A \mid B) = 1 - P(A)$

4. In a cancer detection test, we want to **maximize sensitivity**. This means:
   - (A) Minimize false positives
   - (B) Minimize false negatives
   - (C) Maximize specificity
   - (D) Maximize total error rate

5. If $X \sim N(100, 10^2)$, what is $P(X < 110)$?
   - (A) 0.16
   - (B) 0.50
   - (C) 0.84
   - (D) 0.95

True / False

6. The complement rule states that $P(A^c) = 1 - P(A)$.
7. Sensitivity measures how often a test correctly identifies **non-diseased** individuals.
8. Independence means knowing one event affects the probability of another.
9. For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
10. If $P(A \cap B) = 0$, then A and B are necessarily independent.

Short Answer

11. Explain the difference between **mutually exclusive** and **independent** events.
12. A test has sensitivity 0.9 and specificity 0.8. Interpret both terms in context.
13. If $P(A) = 0.3$, $P(B) = 0.5$, and $P(A \cap B) = 0.15$, are A and B independent? Show why or why not.

📊 Unit 2: Sampling and Estimation

Multiple Choice

1. The **Central Limit Theorem (CLT)** states that:
   - (A) The population distribution becomes normal as n increases.
   - (B) The sampling distribution of the mean becomes approximately normal for large n.
   - (C) The sample mean equals the population mean for large n.
   - (D) The variance of the population decreases as sample size increases.
2. The **standard error** measures:
   - (A) Variability of raw data
   - (B) Variability of the sampling distribution
   - (C) The bias in the estimator
   - (D) The residual variance from regression
3. A **95% confidence interval** means:
   - (A) 95% of population values fall in the interval
   - (B) 95% of repeated samples produce intervals that capture the true mean
   - (C) There's a 95% chance the true mean lies in this specific interval
   - (D) The true mean changes in 95% of samples
4. Increasing confidence level (e.g., 90% → 99%) will:
   - (A) Narrow the interval
   - (B) Widen the interval
   - (C) Leave interval width unchanged
   - (D) Decrease standard error

True / False

5. The width of a confidence interval increases as confidence level increases.
6. The population parameter is random, not fixed.

7. Larger samples lead to smaller standard errors.
8. Bootstrap confidence intervals rely on the normality assumption.
9. Analytical CIs require fewer assumptions than bootstrap CIs.

Short Answer

10. Explain conceptually how a **bootstrap CI** is obtained.
11. Contrast **analytical** and **bootstrap** confidence intervals.
12. Define **bias** and **consistency** in the context of an estimator.

📈 Unit 3: Hypothesis Testing

Multiple Choice

1. The **null hypothesis (H_0)** is typically:
   - (A) The research hypothesis
   - (B) The hypothesis we try to prove
   - (C) The default assumption of no effect or no difference
   - (D) The hypothesis with a small p-value

2. The **p-value** represents:
   - (A) Probability that H_0 is true
   - (B) Probability of observing data as extreme as this if H_0 were true
   - (C) Probability of Type I error
   - (D) Power of the test

3. A **two-sided test** doubles the p-value because:
   - (A) It tests both tails of the sampling distribution
   - (B) It has twice the variance
   - (C) It uses a larger critical value
   - (D) It assumes a non-normal distribution

4. If $p < \alpha$ (e.g., 0.05), we:
   - (A) Fail to reject H_0
   - (B) Reject H_0
   - (C) Accept H_0
   - (D) Cannot conclude anything

True / False

5. We can “accept” the null hypothesis when $p > 0.05$.
6. Simulation-based inference relies on resampling under the null hypothesis.
7. A smaller α increases the chance of rejecting a true null hypothesis.
8. The power of a test is 1 – P(Type II error).

9. Parametric tests depend on known sampling distributions.

Short Answer

10. Explain the difference between **Type I** and **Type II** errors.
11. Define the **significance level (α)** in hypothesis testing.
12. When should you use a **simulation-based** test instead of a **parametric** one?

📉 Unit 4: Comparing Groups — t-Tests, z-Tests, and Two-Sample Tests

Multiple Choice

1. The **t-test** is preferred over the **z-test** when:
   - (A) The population standard deviation is known
   - (B) The sample size is small and σ is unknown
   - (C) The population is not normally distributed
   - (D) The mean difference is 0
2. A **paired t-test** is used when:
   - (A) Two groups are independent
   - (B) Same subjects are measured twice
   - (C) Samples have unequal variances
   - (D) Observations are from different populations
3. The **z-test** relies on:
   - (A) Student's t-distribution
   - (B) Known population variance
   - (C) Simulation
   - (D) Bootstrapping
4. If two samples are independent but have unequal variances, we should use:
   - (A) Pooled t-test
   - (B) Welch's t-test
   - (C) Paired t-test
   - (D) ANOVA

True / False

5. The z-test assumes the population variance is unknown.
6. A t-distribution has heavier tails than the Normal distribution.
7. The t-test is appropriate for both small and large samples.
8. Paired t-tests analyze within-subject changes.
9. As n → ∞, the t-distribution approaches the Normal distribution.

Short Answer

10. Explain the difference between **paired** and **independent** two-sample tests.
11. When should you use a **z-test** instead of a **t-test**?
12. Why does adding more predictors always increase R^2 in regression?

📉 Unit 5: Regression Concepts

Multiple Choice

1. In simple linear regression, the slope coefficient β_1 represents:
   - (A) The change in X for one unit increase in Y
   - (B) The expected change in Y for one unit increase in X
   - (C) The intercept value
   - (D) The residual variance
2. In multiple linear regression, holding all other predictors constant refers to:
   - (A) Partial effect of one predictor
   - (B) Total variance explained
   - (C) Heteroscedasticity
   - (D) Interaction effect
3. The **Gauss–Markov theorem** states:
   - (A) OLS estimators are unbiased and have smallest variance among all unbiased estimators
   - (B) OLS estimators are consistent only under normality
   - (C) OLS provides maximum likelihood estimates
   - (D) OLS minimizes mean absolute error
4. Which assumption is *not* required for OLS to be unbiased?
   - (A) Linearity in parameters
   - (B) Independence of errors
   - (C) Normality of errors
   - (D) Zero mean of errors

True / False

5. The OLS estimator minimizes the sum of absolute residuals.
6. Including irrelevant predictors can inflate variance without changing bias.
7. R^2 always increases with more predictors.
8. Adjusted R^2 penalizes model complexity.
9. A high R^2 guarantees a good model fit.

Short Answer

10. Explain the meaning of β_1 in the model $Y = \beta_0 + \beta_1 X + \varepsilon$.
11. What does the **F-test** evaluate in multiple regression?
12. What is the difference between the **F-test** and **t-test** in regression?

📈 Unit 6: Model Assessment and Diagnostics

Multiple Choice

1. The average leverage in a regression with n = 100 and p = 5 predictors is:
   - (A) 0.05
   - (B) 0.10
   - (C) 0.50
   - (D) 1.00
2. Leverage values:
   - (A) Can exceed 1
   - (B) Are always between 0 and 1 if model includes intercept
   - (C) Sum to n
   - (D) Are equal for all observations
3. Cook's Distance measures:
   - (A) Nonlinearity of predictors
   - (B) Variance inflation
   - (C) Influence of a single observation
   - (D) Correlation among predictors
4. High leverage points:
   - (A) Always have large residuals
   - (B) Are always influential
   - (C) Have extreme X values
   - (D) Always reduce R^2

True / False

5. Leverage depends only on X-values, not Y-values.
6. A high Cook's D indicates the point strongly affects fitted values.
7. Outliers are always influential points.
8. Homoscedasticity means residuals have constant variance.
9. Nonlinearity can often be diagnosed from a residuals vs fitted plot.

Short Answer

10. Define **leverage** and explain how it differs from **influence**.
11. How would you detect **multicollinearity** in a regression model?

12. What does a **funnel shape** in a residuals vs fitted plot indicate?

📊 Unit 7: Categorical Variables and Interaction Terms

Multiple Choice

1. A dummy variable representing a categorical predictor with K levels requires:
   - (A) K dummy variables
   - (B) K – 1 dummy variables
   - (C) 1 dummy variable
   - (D) No dummy variables if numeric codes are used
2. The **reference category** is:
   - (A) The level excluded to avoid multicollinearity
   - (B) The level with highest mean response
   - (C) The category assigned coefficient 1
   - (D) The least frequent group
3. In an interaction model $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3(X_1 \times X_2) + \varepsilon$, the coefficient β_3 represents:
   - (A) The main effect of X_1
   - (B) The interaction effect between X_1 and X_2
   - (C) The slope of X_1 when $X_2 = 0$
   - (D) The intercept shift for X_2

True / False

4. The presence of a significant interaction term means main effects cannot be directly interpreted alone.
5. Including all K dummy variables for a categorical variable causes perfect multicollinearity.
6. Interaction terms should always be included, even if not theoretically justified.
7. The effect of X_1 depends on X_2 when an interaction term is significant.

Short Answer

8. Explain why we drop one dummy variable when encoding a categorical feature.
9. Interpret the interaction coefficient β_3 in plain language.
10. How would you test whether a categorical variable as a whole improves model fit?

🧩 Unit 8: Multicollinearity

Multiple Choice

1. Multicollinearity occurs when:

- (A) The residuals are correlated
- (B) Predictors are highly linearly related
- (C) Errors are heteroscedastic
- (D) Variance is constant across observations

2. The **Variance Inflation Factor (VIF)** measures:

   - (A) Bias of an estimator
   - (B) The increase in variance of a coefficient due to correlation among predictors
   - (C) The effect of outliers on residuals
   - (D) Multicollinearity between Y and X

3. A VIF value above 10 generally indicates:

   - (A) Severe heteroscedasticity
   - (B) Severe multicollinearity
   - (C) Strong independence
   - (D) Normality violation

True / False

4. Multicollinearity affects coefficient interpretability more than model predictions.
5. Centering or scaling variables can remove multicollinearity completely.
6. High R^2 among predictors indicates potential collinearity.
7. VIF = 1 means no multicollinearity.

🧾 Short Answer Practice Bank — Conceptual & Applied Problems

🔷 Unit 1: Probability, Sampling, and Model Foundations

1. Suppose we are studying the relationship between **average commute time (minutes)** and **stress score** on a 1–10 scale.

   - a. Define the population, parameter, and sample.
   - b. Write a generic simple linear regression model relating stress to commute time.
   - c. State the assumptions needed for the model to yield unbiased OLS estimates.

2. A researcher collects data on **household energy use (kWh)** and fits the model $Y = \beta_0 + \beta_1(\text{Income}) + \varepsilon$.

   - a. Interpret β_1 in context.
   - b. What is the expected energy use when income = 0? Discuss whether this value is meaningful.

3. You conduct a survey on weekly coffee consumption among students. The population mean is unknown.
   - a. Write out the formula for the sampling distribution of the sample mean and its standard error.
   - b. Explain how the **Central Limit Theorem** justifies inference using this statistic even if consumption is skewed.
4. Define **bias**, **variance**, and **mean squared error (MSE)** of an estimator. How do these relate to model accuracy?
5. A 95% confidence interval for the mean commute time is (32 min, 38 min).
   - a. Interpret this interval in proper statistical language.
   - b. What happens to the width if we: (i) double n, (ii) raise the confidence level to 99%?
6. Explain the conceptual difference between the **population regression function** and the **sample regression line**.
7. List and describe the four OLS assumptions (Gauss–Markov conditions). Which assumption ensures unbiasedness?
8. In your dataset, one observation of annual salary is $1,200,000 while most others are near $60,000.
   - a. What type of influence might this point exert?
   - b. How could you detect its influence numerically or graphically?
9. Describe what it means for two random variables to be **independent** vs. **uncorrelated**.
   Can two variables be uncorrelated but not independent? Provide an example.
10. Suppose we estimate `TestScore = β₀ + β₁(ClassSize) + ε`.
    - a. Sketch or describe the expected sign of β_1 based on education-economics reasoning.
    - b. If $\beta_1 = -2.5$, interpret the coefficient.

🔷 Unit 2: Inference, Hypothesis Testing, and Estimation

11. You run a one-sample t-test for whether the average number of hours students study per week differs from 10.
    - a. Write the null and alternative hypotheses.
    - b. Interpret $p = 0.07$ at $\alpha = 0.05$.
12. For the model `Y = β₀ + β₁X + ε`, describe the difference between **estimating β_1** via OLS and **testing whether $\beta_1 = 0$**.
13. Explain why we use a **t-distribution** instead of the normal distribution when conducting inference on small samples.
14. A researcher computes a 90% confidence interval for the slope β_1 as (0.15, 0.65).
    - a. Write the null hypothesis that this CI addresses.
    - b. What would the corresponding two-sided p-value roughly indicate?

15. Define and contrast **Type I** and **Type II** errors in the context of a two-sample test comparing average rents in two cities.

16. When conducting a two-sample t-test:

    - a. Explain how to determine whether the samples are **independent** or **paired**.
    - b. Provide one real-world example for each case.

17. Suppose we perform a permutation test comparing the mean exam scores between two groups.

    - a. Describe how to generate the null distribution.
    - b. What does the empirical p-value represent in this context?

18. Write a short paragraph comparing **parametric inference** (e.g., t-tests) with **simulation-based inference** (e.g., bootstrap or permutation).
    Include at least one advantage and one disadvantage of each.

19. A dataset of 100 households has `Y = household electricity cost`, `X₁ = square footage`, and `X₂ = number of occupants`.

    - a. Write the full theoretical model including both predictors and define each term.
    - b. Suppose $\beta_2 = 8.5$; interpret it in words.

20. You add an interaction term between square footage and occupants.

    - a. Write the new theoretical model.
    - b. Write the implied models when `occupants = 0` and when `occupants = 4`.
    - c. What does the coefficient on the interaction represent conceptually?

21. Explain how a **bootstrap confidence interval** is obtained and how its interpretation differs from a traditional analytical CI.

22. For each situation, identify the correct test and justify your choice:

    - a. Compare the mean blood pressure between two unrelated groups.
    - b. Compare pre-test and post-test scores for the same patients.
    - c. Assess if the correlation between age and cholesterol differs from zero.

23. In a regression with 50 observations and 3 predictors (including the intercept), what is the expected average leverage value?
    How would you flag points with unusually high leverage?

24. A regression yields $R^2 = 0.80$ and Adjusted $R^2 = 0.65$.

    - a. Interpret both values.
    - b. What might cause Adjusted R^2 to drop when adding predictors?

25. Given partial output below:

Predictor	Estimate	Std. Error	t value	p value
Intercept	52.3	4.1	12.76	< 0.001

Predictor	Estimate	Std. Error	t value	p value
Income	0.32	0.11	2.91	0.005
Age	−0.41	0.09	−4.56	< 0.001

- a. Write the fitted model.
- b. Interpret the coefficient on Age.
- c. State the null hypothesis tested for each coefficient.
- d. Suppose another variable "Education" is added and the p-value for Age rises to 0.25. What might explain this change?

26. Describe the steps you would take to check whether the assumptions of the regression model are satisfied (list the key diagnostic plots and what to look for).

27. In your fitted model, Cook's D = 1.2 for one observation.

    - a. What does this suggest about the influence of that point?
    - b. How would you confirm whether it materially changes model conclusions?

28. Explain the difference between **confidence intervals** for β-coefficients and **prediction intervals** for new observations.

29. Write the theoretical model and corresponding fitted equation for predicting `LifeExpectancy` from `InfantDeaths`, `Measles`, `HepatitisB`, and `Status` (developed = 0, developing = 1).
    Then describe in plain language what the coefficient on `Status` means.

30. Extend the previous model to include an interaction between `InfantDeaths` and `Status`.

    - a. Write the full theoretical model.
    - b. Write out the separate models implied for developed and developing countries.
    - c. Identify the design-matrix dimension (n × p) if there are 134 observations.

31. Provide two examples of when **Adjusted R^2** is a more informative comparison metric than **R^2**, and explain why.

32. Define **leverage**, **residual**, and **influence** in regression. How are they related?

33. In your own words, explain the **Gauss–Markov theorem** and under what assumptions OLS is the BLUE estimator.

34. Suppose you regress `Y = Sales` on `X = Advertising`. The slope is positive and significant.

    - a. Does this prove that advertising causes sales to rise? Why or why not?
    - b. Suggest an alternative variable that could confound this relationship.
