MA206X: Probability & Statistics

Comprehensive TEE Review — Answer Key
AY26-2

How to use this review. This review covers everything on the TEE: probability rules, named distributions, one- and two-sample inference, regression, and ANOVA. Work each problem using only your reference card, calculator, and R-lite (`pnorm/qnorm`, `pt/qt`, `pbinom`, `ppois`, `pexp`, `pf/qf`). Solutions are shown in blue.

1 Multiple Choice

1. In hypothesis testing, the p-value is the probability of:
   - (a) The null hypothesis being true given the data
   - (b) Observing data at least as extreme as what was observed, assuming H_0 is true
   - (c) Making a Type I error
   - (d) The alternative hypothesis being true

2. Two events A and B are mutually exclusive. Which statement must be true?
   - (a) A and B are independent
   - (b) $P(A \cap B) = P(A) \cdot P(B)$
   - (c) $P(A \cup B) = P(A) + P(B)$
   - (d) $P(A|B) = P(A)$

3. For a random variable X that follows a Poisson distribution with mean μ, the variance equals:
   - (a) μ^2
   - (b) $\sqrt{\mu}$
   - (c) μ
   - (d) $1/\mu$

4. The Central Limit Theorem says that the sampling distribution of the sample mean $\bar{X}$ is approximately normal when:
   - (a) The population is normal regardless of n
   - (b) The sample size n is sufficiently large
   - (c) Either (a) or (b)
   - (d) The population variance is known

5. A 95% confidence interval for μ is reported as $(42, 58)$. Which interpretation is correct?
   - (a) There is a 95% probability that the true mean lies in $(42, 58)$.

(b) 95% of all sample means fall in $(42, 58)$.

(c) If we repeated the sampling procedure many times, about 95% of the resulting intervals would contain the true mean.

(d) 95% of the population values fall in $(42, 58)$.

6. In a multiple linear regression, the coefficient on x_1 is interpreted as:

   (a) The total change in y for a one-unit increase in x_1

   (b) The average change in y per one-unit increase in x_1, holding all other predictors constant

   (c) The correlation between x_1 and y

   (d) The proportion of variability in y explained by x_1

7. Failing to reject H_0 at significance level $\alpha = 0.05$ means:

   (a) H_0 is true.

   (b) There is insufficient evidence to reject H_0 at the 0.05 level.

   (c) The probability that H_0 is true exceeds 0.95.

   (d) The Type I error rate is below 5%.

8. Which distribution would be most appropriate for modeling the time until the next email arrives in your inbox, assuming emails arrive at a constant average rate?

   (a) Binomial

   (b) Poisson

   (c) Exponential

   (d) Normal

9. The standard error of the sample mean is:

   (a) σ

   (b) σ^2/n

   (c) $\sigma/\sqrt{n}$

   (d) s

10. In a one-way ANOVA with k groups and N total observations, the test statistic F follows an F distribution with degrees of freedom:

    (a) $(N-1, k-1)$

    (b) $(k-1, N-k)$

    (c) (k, N)

    (d) $(N-k, k-1)$

11. Which of the following will increase the power of a hypothesis test? *Select all that apply*, then choose the best answer.
    (a) Increasing the sample size n
    (b) Increasing the significance level α
    (c) Increasing the effect size
    (d) All of the above

12. For a continuous random variable X, $P(X = c)$ for any specific value c equals:
    (a) Approximately 0 for large n
    (b) $f(c)$, the density at c
    (c) Exactly 0
    (d) It depends on the distribution

2 Event Probability (Intersections, Unions, Complements)

Problem 2.1. A standard urn contains 8 red, 5 blue, and 7 green marbles. One marble is drawn at random.

(a) What is the probability of drawing a red OR blue marble?

(b) What is the probability the marble is NOT green?

(c) What is the probability of drawing a marble that is both red AND blue?

Problem 2.2. A retailer surveys 500 customers about their shopping channels. The survey finds:

- 320 customers use the mobile app
- 280 customers use the website
- 180 customers use BOTH the mobile app AND the website

(a) What is the probability a randomly selected customer uses at least one of the two channels?

(b) What is the probability a customer uses neither channel?

(c) What is the probability a customer uses ONLY the mobile app (not the website)?

3 Conditional Probability (Bayes and Non-Bayes)

Problem 3.1. A medical lab has developed a new screening test for a relatively rare disease. Background information:

- Approximately 2% of the general population has the disease.
- Among those who have the disease, the test correctly returns a positive result 96% of the time (sensitivity).
- Among those who do NOT have the disease, the test correctly returns a negative result 95% of the time (specificity).

(a) What is the probability that a randomly selected person tests positive?

(b) Given that a person has tested positive, what is the probability they actually have the disease?

(c) Briefly explain why your answer to (b) is so much smaller than the 96% sensitivity figure might suggest.

Problem 3.2. Two cards are drawn at random from a standard 52-card deck *without replacement*.

(a) Given that the first card drawn is a King, what is the probability the second card is also a King?

(b) What is the probability that BOTH cards are Kings?

(c) What is the probability that BOTH cards are red (hearts or diamonds)?

4 Binomial Distribution

Problem 4.1. A factory produces electronic widgets, and historical data show that 5% of widgets are defective. A quality-control inspector randomly selects $n = 20$ widgets from the production line. Let X be the number of defective widgets in the sample.

(a) Identify the distribution of X and explain why the binomial assumptions are reasonable here.

(b) What is the probability that NONE of the 20 widgets are defective?

(c) What is the probability that at most 2 widgets are defective?

(d) What is the probability that 3 or more widgets are defective?

Problem 4.2. A college basketball player is an 80% free-throw shooter. In an upcoming game she is expected to attempt 10 free throws. Let X be the number of free throws she makes.

(a) Identify the distribution of X, including parameter values.

(b) Compute $E[X]$ and $V[X]$.

(c) What is the probability she makes exactly 8 free throws?

(d) What is the probability she makes at least 9?

5 Poisson Distribution

Problem 5.1. Customers arrive at a coffee shop during the morning rush at an average rate of 12 per hour, and arrivals can be modeled as a Poisson process.

(a) What is the probability that exactly 10 customers arrive during a randomly chosen 1-hour interval?

(b) What is the probability that exactly 20 customers arrive during a randomly chosen 2-hour interval?

(c) What is the probability that 15 or more customers arrive in a 1-hour interval?

Problem 5.2. A copy editor finds that her typewritten manuscripts contain typos at an average rate of 1.5 typos per page. The number of typos on any given page can be modeled as Poisson.

(a) What is the probability that a single randomly chosen page contains AT LEAST one typo?

(b) For a 5-page chapter, what is the probability the chapter contains exactly 8 typos?

(c) For a 5-page chapter, what is the probability the chapter contains 5 or fewer typos?

6 Exponential Distribution

Problem 6.1. At a customer-service call center, the time between consecutive incoming calls is exponentially distributed with mean 4 minutes. Let T be the time (in minutes) until the next call.

(a) Identify the distribution of T and state λ.

(b) What is the probability the next call arrives within the next 3 minutes?

(c) What is the probability that more than 5 minutes pass before the next call?

(d) What is the probability the next call arrives between 2 and 6 minutes from now?

Problem 6.2. A manufacturer claims that the lifetime of a particular model of LED bulb is exponentially distributed with mean 8,000 hours. Let T denote the lifetime of a randomly chosen bulb.

(a) What is the probability that a randomly chosen bulb lasts more than 10,000 hours?

(b) Find the median lifetime of these bulbs.

(c) For warranty purposes, the manufacturer wants to identify the time t^* such that only 10% of bulbs fail before t^*. Find t^*.

7 Normal Distribution

Problem 7.1. The heights of adult women in the United States are approximately normally distributed with a mean of 64 inches and a standard deviation of 2.5 inches. Let X denote the height of a randomly chosen adult woman.

(a) What is the probability that a randomly chosen woman is less than 60 inches tall?

(b) What is the probability that her height is between 60 and 68 inches?

(c) What is the probability that she is taller than 70 inches?

Problem 7.2. SAT total scores are approximately normally distributed with mean 1050 and standard deviation 200.

(a) What proportion of test takers score above 1300?

(b) What is the 90th percentile of SAT scores — i.e., the score below which 90% of students fall?

(c) What range of scores corresponds to the middle 50% of all test takers?

8 One-Sample t-Test

Problem 8.1. A cereal manufacturer claims that the mean weight of cereal in its 16-ounce boxes is 16 oz. A consumer-protection agency suspects boxes are systematically underfilled or overfilled. The agency randomly samples $n = 25$ boxes and measures their contents. The sample mean weight is $\bar{x} = 15.7$ oz with standard deviation $s = 0.5$ oz.

(a) State the hypotheses for testing the agency's suspicion.

(b) Conduct the test at $\alpha = 0.05$. Compute the test statistic, p-value, and state your conclusion in context.

(c) Construct and interpret a 95% confidence interval for the true mean weight.

Problem 8.2. A laptop manufacturer advertises that its new battery lasts at least 50 hours under normal usage. A reviewer samples $n = 15$ laptops and finds a sample mean lifetime of $\bar{x} = 47.5$ hours with $s = 4.2$ hours. Assume battery lifetimes are approximately normal.

(a) State the appropriate hypotheses for testing whether the manufacturer's claim is overstated.

(b) Conduct the test at $\alpha = 0.05$. State your conclusion in context.

9 One-Proportion z-Test

Problem 9.1. A company is considering replacing its current logo. Before committing, it runs a survey of 400 randomly chosen customers, asking whether they prefer the new logo over the current one. Of the 400 surveyed, 240 prefer the new logo. The marketing team will adopt the new logo only if more than half of customers prefer it.

(a) State the appropriate hypotheses.

(b) Conduct the test at $\alpha = 0.05$. Compute the test statistic, p-value, and state your conclusion in context.

(c) Construct a 95% confidence interval for the true proportion who prefer the new logo.

Problem 9.2. A factory's stated defect rate for one of its products is 5%. A QA inspector wants to know whether the actual rate is higher than the stated rate. She randomly samples $n = 500$ units and finds 32 defective.

(a) State the hypotheses.

(b) Conduct the test at $\alpha = 0.05$. Compute the test statistic, p-value, and state your conclusion in context.

10 Two-Sample t-Test (Independent)

Problem 10.1. A horticulturist wants to compare the effect of two fertilizers on tomato yield (kg per plot). She randomly assigns plots to receive either Fertilizer A or Fertilizer B. Results:

Fertilizer	n	$\bar{x}$ (kg)	s (kg)
A	20	45	4
B	22	42	5

(a) Why are the two samples treated as independent (not paired)?

(b) Test at $\alpha = 0.05$ whether the two fertilizers produce different mean yields. State the hypotheses, test statistic (using conservative df), p-value, and conclusion.

(c) Construct a 95% confidence interval for $\mu_A - \mu_B$ and interpret.

Problem 10.2. An education researcher compares two teaching methods for an introductory statistics course. Random samples of students are assigned to each method, and final exam scores are recorded:

Method	n	$\bar{x}$	s
1 (lecture)	15	78	8
2 (flipped)	18	72	10

The researcher hypothesizes that students taught with lecture (Method 1) score higher on average than those taught with the flipped classroom (Method 2).

Test at $\alpha = 0.05$. State the hypotheses, test statistic (with conservative df), p-value, and conclusion.

11 Two-Proportion z-Test

Problem 11.1. A national retailer surveys customers in two regions about which of two product designs they prefer. Of 200 randomly sampled customers in Region 1, 140 preferred Design A. Of 250 customers in Region 2, 145 preferred Design A. Test at $\alpha = 0.05$ whether the proportion preferring Design A differs between regions.

(a) State the hypotheses.

(b) Conduct the test. Report the test statistic, p-value, and conclusion in context.

(c) Construct a 95% CI for $p_1 - p_2$.

Problem 11.2. An online advertising team is comparing two ad campaigns. Campaign 1 was shown to 1,000 users, of whom 35 clicked. Campaign 2 was shown to 1,200 users, of whom 60 clicked. The team wants to test at $\alpha = 0.05$ whether Campaign 2 has a higher click-through rate than Campaign 1.

State the hypotheses, conduct the test, and state your conclusion in context.

12 Simple Linear Regression

Problem 12.1. A statistics professor fits a simple linear regression of final exam score (y, out of 100) on the number of hours per week the student studied (x). Using data from $n = 30$ students, the fitted regression equation is:

$$\hat{y} = 50 + 4.2\,x$$

The model has $R^2 = 0.65$, and the slope is statistically significant at $\alpha = 0.05$.

(a) Interpret the slope coefficient in context.

(b) Interpret the intercept. Is its interpretation meaningful in this context?

(c) Interpret the value of R^2 in context.

(d) Predict the exam score of a student who studies 8 hours per week.

Problem 12.2. A real-estate analyst regresses home sale price (y, in thousands of dollars) on home size (x, in square feet) using $n = 50$ recently sold homes. The fitted equation is:

$$\hat{y} = 80 + 0.15\,x$$

with $R^2 = 0.78$. The data set contains homes ranging from 1,000 to 3,500 square feet.

(a) Interpret the slope coefficient in dollars per square foot.

(b) Predict the sale price of a 2,000-square-foot home.

(c) A developer asks you to use the model to predict the price of a 5,000-square-foot luxury home. Why should you be cautious?

13 Multiple Linear Regression

Problem 13.1. A labor economist fits an MLR model predicting annual salary (y, in \$1,000s) from years of education (x_1), years of work experience (x_2), and average weekly hours (x_3), using a random sample of $n = 100$ workers in a metropolitan area. Selected R output:

$$\hat{y} = 35 + 2.5\,x_1 + 1.8\,x_2 - 0.4\,x_3$$

Predictor	Estimate	Std. Error	t-value	p-value
(Intercept)	35.00	4.20	8.33	< 0.001
Education (x_1)	2.50	0.55	4.55	< 0.001
Experience (x_2)	1.80	0.30	6.00	< 0.001
Hours (x_3)	-0.40	0.35	-1.14	0.257

$R^2 = 0.72$, Adjusted $R^2 = 0.71$.

(a) Identify which predictors are statistically significant at $\alpha = 0.05$.

(b) Interpret the coefficient on Education in context.

(c) Interpret R^2 in context.

(d) Predict the annual salary of a worker with 16 years of education, 10 years of experience, working 45 hours per week.

Problem 13.2. A property-management firm fits an MLR model predicting monthly rent (y, in dollars) from the number of bedrooms (x_1), bathrooms (x_2), and a categorical variable for neighborhood. The neighborhood variable has three levels: Suburbs (reference), Midtown, Downtown. The fitted model is:

$$\hat{y} = 800 + 250\,x_1 + 100\,x_2 + 150\,I_{\text{Midtown}} + 200\,I_{\text{Downtown}}$$

where I_{Midtown} and I_{Downtown} are indicator variables.

(a) Interpret the coefficient on Bedrooms (x_1).

(b) Interpret the coefficient on I_{Downtown}.

(c) Predict the monthly rent of a 2-bedroom, 2-bathroom apartment in Downtown.

(d) Predict the monthly rent of a 2-bedroom, 2-bathroom apartment in the Suburbs.

14 One-Way ANOVA

Problem 14.1. An agricultural researcher is comparing the effect of $k = 4$ fertilizer formulations (A, B, C, D) on tomato yield (kg per plot). She randomly assigns 10 plots to each formulation ($N = 40$ total). After harvest, an R analysis returns:

$$\text{SSTr} = 240 \qquad \text{SSE} = 540$$

(a) State the hypotheses for the ANOVA F-test.

(b) Calculate the F test statistic.

(c) Find the p-value using R-lite, then state your conclusion at $\alpha = 0.05$ in context.

Problem 14.2. A health researcher randomly assigns 36 sedentary adults equally to one of $k = 3$ exercise programs ($n = 12$ per program). After 12 weeks, weight loss (in lb) is recorded. An R analysis returns:

$$\text{SSTr} = 180 \qquad \text{SSE} = 396$$

(a) State the hypotheses.

(b) Compute the F test statistic.

(c) Find the p-value via R-lite and state your conclusion at $\alpha = 0.05$.

(d) State the three conditions required for the ANOVA F-test to be valid, and briefly comment on whether they appear reasonable here.

End of Review.
If you can confidently work each of these 26 free-response problems and the 12 multiple-choice items, you have demonstrated mastery of every TEE topic: probability and conditional probability, the four named distributions (Binomial, Poisson, Exponential, Normal), one- and two-sample inference for means and proportions, simple and multiple linear regression, and one-way ANOVA.

Final reminders:

- State hypotheses about *population parameters* (μ, p), not about sample statistics ($\bar{x}$, $\hat{p}$).
- Always check conditions before applying a test.
- Use the R-lite functions (`p`*`dist`*, `q`*`dist`*) for p-values and critical values — and remember the always-upper-tailed structure of ANOVA p-values.
- Conclusions need to be stated in context, not just "reject H_0."
- Confidence intervals and hypothesis tests are complementary — use both when asked.