AP STATISTICS

Complete Exam Prep — 20 Essential Practice Questions

All 9 Units · Concepts · Formulas · Full Answer Key

CONCEPT REVIEW	KEY FORMULAS	20 EXAM QUESTIONS	FULL ANSWER KEY

CONCEPTS & KEY FORMULAS

UNIT 1 Exploring One-Variable Data

Types of Variables & Distributions

- Categorical vs. Quantitative variables
- Describing distributions: Shape, Center, Spread, Outliers (SOCS)
- Symmetric vs. Skewed distributions
- Dotplot, Stemplot, Histogram, Boxplot

KEY FORMULAS / MEMORIZE:

Mean: x-bar = (sum of x) / n

Median: middle value when sorted

Range = Max - Min

IQR = Q3 - Q1

Outlier rule: < Q1 - 1.5*IQR or > Q3 + 1.5*IQR

UNIT 2 Exploring Two-Variable Data

Scatterplots & Correlation

- Direction, form, strength of association
- Correlation r: -1 <= r <= 1 (no units, sensitive to outliers)
- r does NOT imply causation
- Influential points vs. outliers in regression

KEY FORMULAS / MEMORIZE:

r = (1/(n-1)) * sum[(xi - x-bar)/sx * (yi - y-bar)/sy]

Least-squares regression line: y-hat = a + bx

Slope: b = r * (sy / sx)

y-intercept: a = y-bar - b*x-bar

Coefficient of determination: r^2 (% of variation in y explained by x)

UNIT 3 Collecting Data

Sampling & Experimental Design

- Simple Random Sample (SRS): every group of n has equal chance

- Stratified: divide into strata, SRS from each
- Cluster: divide into clusters, randomly select whole clusters
- Systematic: every kth individual
- Observational study vs. Experiment
- Confounding variable: related to both explanatory and response
- Principles of experiment: Control, Randomization, Replication
- Completely Randomized Design vs. Block Design

UNIT 4

Probability, Random Variables & Distributions

Probability Rules & Random Variables

- P(A) = (favorable outcomes) / (total outcomes)
- Complement: P(A^c) = 1 - P(A)
- Addition: P(A or B) = P(A) + P(B) - P(A and B)
- Multiplication (independent): P(A and B) = P(A)*P(B)
- Conditional: P(A|B) = P(A and B) / P(B)
- Discrete vs. Continuous random variables

KEY FORMULAS / MEMORIZE:

E(X) = mu = sum[x * P(x)]
Var(X) = sigma^2 = sum[(x - mu)^2 * P(x)]
E(aX + b) = a*E(X) + b
Var(aX + b) = a^2 * Var(X)
E(X +/- Y) = E(X) +/- E(Y)
Var(X + Y) = Var(X) + Var(Y) [if independent]

UNIT 5

Sampling Distributions

Central Limit Theorem & Sampling Distributions

- Sampling distribution: distribution of a statistic over all possible samples
- x-bar is unbiased estimator of mu
- CLT: For large n, x-bar is approximately Normal regardless of population shape
- Large enough n: n >= 30 for means; np >= 10 and n(1-p) >= 10 for proportions

KEY FORMULAS / MEMORIZE:

mu(x-bar) = mu
sigma(x-bar) = sigma / sqrt(n) [standard error of mean]

mu(p-hat) = p

sigma(p-hat) = sqrt[p(1-p)/n] [standard error of proportion]

Z = (x-bar - mu) / (sigma/sqrt(n))

UNIT 6

Inference for Categorical Data: Proportions

Confidence Intervals & Significance Tests for Proportions

- Four-step procedure: State, Plan, Do, Conclude
- Conditions: Random, Normal (np-hat>=10, n(1-p-hat)>=10), Independent (10% rule)
- Confidence interval: estimate +/- margin of error
- Significance test: H0 (null) vs. Ha (alternative)
- p-value: probability of result as extreme as observed, given H0 is true
- Reject H0 if p-value < alpha (usually 0.05)
- Type I error: reject true H0 (probability = alpha)
- Type II error: fail to reject false H0 (probability = beta)
- Power = 1 - beta

KEY FORMULAS / MEMORIZE:

CI for p: p-hat +/- z* * sqrt[p-hat(1-p-hat)/n]

Test stat: z = (p-hat - p0) / sqrt[p0(1-p0)/n]

2-prop z-interval: (p1-hat - p2-hat) +/- z* * sqrt[p-hat(1-p-hat)(1/n1+1/n2)]

Pooled p-hat = (x1 + x2) / (n1 + n2) [for 2-prop z-test]

UNIT 7

Inference for Quantitative Data: Means

t-procedures for Means

- Use t-distribution when sigma is unknown (almost always in practice)
- Degrees of freedom: df = n - 1 (one sample); use technology for two-sample
- Conditions: Random, Normal/Large Sample, Independent
- Normal condition: n >= 30 OR population is Normal OR graph shows no strong skew/outliers
- Paired t-test: treat differences as one-sample problem

KEY FORMULAS / MEMORIZE:

One-sample t CI: x-bar +/- t* * (s/sqrt(n))

One-sample t test stat: t = (x-bar - mu0) / (s/sqrt(n))

Two-sample t CI: (x1-bar - x2-bar) +/- t* * sqrt(s1^2/n1 + s2^2/n2)

Two-sample t test: t = (x1-bar - x2-bar) / sqrt(s1^2/n1 + s2^2/n2)

Paired: t = d-bar / (sd/sqrt(n)), df = n-1

UNIT 8

Inference for Categorical Data: Chi-Square

Chi-Square Tests

- Goodness of Fit: does a variable follow a claimed distribution? (1 variable)
- Homogeneity: do multiple populations have same distribution? (2+ groups, 1 variable)
- Independence: are two categorical variables associated? (1 population, 2 variables)
- Conditions: Random, all expected counts >= 5
- df = (rows-1)(cols-1) for two-way table; df = k-1 for GOF

KEY FORMULAS / MEMORIZE:

Expected count = (row total * column total) / table total

chi^2 = sum[(O - E)^2 / E]

Always right-tailed test

Larger chi^2 => stronger evidence against H0

UNIT 9

Inference for Quantitative Data: Slopes

Inference for Linear Regression

- Test whether slope beta is significantly different from 0
- Confidence interval for true slope beta
- Conditions: LINEAR form, INDEPENDENT observations, NORMAL residuals, EQUAL variance (LINE)
- Residual plot: random scatter around 0 => linear model appropriate
- s (residual standard deviation) measures typical prediction error

KEY FORMULAS / MEMORIZE:

t = (b - beta0) / SE(b)

df = n - 2

CI for slope: b +/- t* * SE(b)

Residual = y - y-hat = actual - predicted

SSE = sum(residuals^2); s = sqrt(SSE/(n-2))

PRACTICE QUESTIONS

Answer all questions in free-response style. Show all work, check conditions, and state conclusions in context.

UNIT 1 Exploring One-Variable Data

Question 1

A dataset of 15 exam scores has Q1 = 62, Q3 = 84, and a maximum value of 105. Using the 1.5 x IQR rule, determine whether the maximum score is an outlier. Show all calculations and state your conclusion.

Question 2

The distribution of household incomes in a city is strongly right-skewed. Explain why the median is a more appropriate measure of center than the mean for this distribution, and describe what the shape tells us about the data.

UNIT 2 Exploring Two-Variable Data

Question 3

A researcher fits a least-squares regression line to data on hours studied (x) and test score (y). The output shows: slope = 4.2, y-intercept = 48, and r^2 = 0.81. (a) Interpret the slope in context. (b) Interpret r^2 in context. (c) Predict the score for a student who studies 6 hours.

Question 4

A scatterplot shows a strong positive linear association between ice cream sales and drowning deaths. A student concludes that eating ice cream causes drowning. Critique this reasoning and identify the likely lurking variable.

UNIT 3 Collecting Data

Question 5

A school wants to survey students about cafeteria food quality. Describe how to select a stratified random sample of 60 students from a school with 200 freshmen, 180 sophomores, 160 juniors, and 160 juniors using proportional allocation. Show the calculation for each stratum.

Question 6

A pharmaceutical company tests a new pain reliever. Researchers assign participants to treatment and control groups. (a) Explain why a double-blind design is preferred. (b) What is the purpose of a placebo? (c) Identify one potential source of confounding and explain how the design controls for it.

UNIT 4

Probability, Random Variables & Distributions

Question 7

A box contains 4 red and 6 blue marbles. Two marbles are drawn without replacement. (a) Find the probability that both marbles are red. (b) Find the probability that at least one marble is blue. Show all work using probability rules.

Question 8

Let X be a discrete random variable with the following distribution: $P(X=1)=0.2$, $P(X=2)=0.3$, $P(X=3)=0.4$, $P(X=4)=0.1$. Calculate (a) $E(X)$, (b) $Var(X)$, and (c) the standard deviation of X.

UNIT 5

Sampling Distributions

Question 9

The weights of apples in an orchard are Normally distributed with mean 182 g and standard deviation 24 g. A sample of 36 apples is selected. (a) Describe the sampling distribution of x-bar. (b) Find the probability that the sample mean weight exceeds 190 g. Show all work.

Question 10

In a large population, 40% of people prefer Brand A. A random sample of 100 people is taken. (a) Verify conditions for Normal approximation. (b) Find the probability that the sample proportion is less than 0.35.

UNIT 6

Inference for Categorical Data: Proportions

Question 11

In a random sample of 200 registered voters, 112 support Candidate A. (a) Construct a 95% confidence interval for the true proportion of all voters who support Candidate A. (b) Based on your interval, can you conclude the candidate has majority support? Explain.

Question 12

A researcher claims that more than 60% of adults exercise regularly. In a random sample of 150 adults, 98 report exercising regularly. Conduct a significance test at alpha = 0.05. State all four steps clearly, compute the test statistic and p-value, and state your conclusion.

UNIT 7 Inference for Quantitative Data: Means

Question 13

A random sample of 25 light bulbs has a mean lifetime of 1180 hours with a standard deviation of 80 hours. (a) Check conditions for inference. (b) Construct a 90% confidence interval for the true mean lifetime. (c) Interpret your interval in context.

Question 14

Two teaching methods are compared. Method A (n=18): x-bar=78, s=12. Method B (n=22): x-bar=83, s=10. Using a two-sample t-test at alpha=0.05, test whether there is a significant difference in mean scores. State hypotheses, compute the test statistic, find the p-value (df=35, t*=2.030 at 95%), and conclude.

UNIT 8 Inference for Categorical Data: Chi-Square

Question 15

A die is rolled 120 times. Results: 1-dot:18, 2-dot:22, 3-dot:19, 4-dot:21, 5-dot:17, 6-dot:23. Conduct a chi-square goodness-of-fit test at alpha=0.05 to test whether the die is fair. State hypotheses, compute the test statistic, compare to critical value (chi^2* = 11.07 for df=5), and conclude.

Question 16

A 2x2 table shows survey data on gender (Male/Female) and preference (Prefers online / Prefers in-person). The observed counts are: Male-Online:45, Male-InPerson:30, Female-Online:35, Female-InPerson:40. Test for independence at alpha=0.05 (critical value chi^2* = 3.841 for df=1). Show expected counts and the test statistic.

UNIT 9 Inference for Quantitative Data: Slopes

Question 17

Computer output for a regression of y = plant height (cm) on x = fertilizer amount (g) gives: slope b = 2.34, SE(b) = 0.61, n = 18, df = 16. (a) Test H0: beta=0 vs. Ha: beta>0 at alpha=0.05 (t*=1.746). (b) Construct a 95% CI for the true slope (t*=2.120). (c) Interpret the CI in context.

Question 18

A residual plot for a regression model shows a clear curved (parabolic) pattern. (a) What does this tell us about the linear regression model? (b) What transformation or action would you recommend? (c) What condition of linear regression is violated?

MIXED REVIEW

Question 19

A study finds that students who eat breakfast score significantly higher on tests ($p=0.002$). A classmate concludes: 'This proves eating breakfast causes better test scores.' Identify TWO statistical issues with this conclusion and explain each carefully.

Question 20

The 95% confidence interval for a population mean is (42.3, 58.7). (a) Interpret this interval correctly. (b) A student says: 'There is a 95% probability that the true mean is between 42.3 and 58.7.' Is this correct? Explain the correct interpretation of confidence level.

ANSWER KEY & EXPLANATIONS

Review each solution carefully. Focus on the reasoning process, not just the final answer.

Q1 ANSWER

Yes, 105 is an outlier.

IQR = Q3 - Q1 = 84 - 62 = 22. Upper fence = Q3 + 1.5*IQR = 84 + 1.5(22) = 84 + 33 = 117. Since 105 < 117, the maximum of 105 is NOT an outlier by this rule. (Note: check if problem intends lower fence as well; lower fence = 62 - 33 = 29; all values above 29 pass lower boundary.) Conclusion: 105 is within the fence — NOT an outlier.

Q2 ANSWER

The median is more appropriate for right-skewed distributions.

In a right-skewed distribution, a few extremely high values pull the mean upward, making it unrepresentative of the 'typical' household. The median is resistant to outliers and skew, so it better describes the center. The right skew indicates most households earn modest incomes with a few very high earners pulling the tail right.

Q3 ANSWER

(a) Slope = 4.2: each additional hour studied predicts 4.2 points higher. (b) r^2=0.81: 81% of variation in scores explained by hours studied. (c) y-hat = 48 + 4.2(6) = 73.2

y-hat = 48 + 4.2(6) = 48 + 25.2 = 73.2 points predicted. Slope context: On average, one more hour of study is associated with a 4.2-point increase in test score. r^2: 81% of the variability in test scores can be explained by the linear relationship with hours studied.

Q4 ANSWER

Correlation does not imply causation; lurking variable is outdoor temperature (season/summer).

Both ice cream sales and drowning increase during hot weather (summer). Temperature is the lurking variable confounded with both. Without controlling for season, we cannot conclude a causal relationship. This is a classic case of a common response pattern — both variables respond to a third variable.

Q5 ANSWER

Freshmen: 17, Sophomores: 15, Juniors: 14, Seniors: 14 (approx., total=60)

Total = 200+180+160+160 = 700. Freshmen: (200/700)*60 = 17.1 ~ 17. Sophomores: (180/700)*60 = 15.4 ~ 15. Juniors: (160/700)*60 = 13.7 ~ 14. Seniors: (160/700)*60 = 13.7 ~ 14. Total = 17+15+14+14 = 60. Select SRS from each stratum using random number generator or table.

Q6 ANSWER

(a) Prevents bias from expectations. (b) Controls for placebo effect. (c) Health status — randomization helps balance this.

Double-blind: Neither participants nor evaluators know who received treatment, preventing conscious or unconscious bias in reporting or assessment. Placebo: Creates identical conditions psychologically; isolates the pharmacological effect. Confounding control: Randomization distributes confounders (age, health baseline) evenly across groups, making groups comparable.

Q7 ANSWER

(a) P(both red) = (4/10)*(3/9) = 12/90 = 2/15. (b) P(at least one blue) = 1 - P(both red) = 1 - 2/15 = 13/15

Without replacement: P(1st red) = 4/10; P(2nd red | 1st red) = 3/9. Multiply: 4/10 * 3/9 = 12/90 = 2/15 ≈ 0.133. For at least one blue: use complement. P(at least 1 blue) = 1 - P(both red) = 1 - 2/15 = 13/15 ≈ 0.867.

Q8 ANSWER

E(X) = 2.4, Var(X) = 0.84, SD(X) = 0.917

E(X) = 1(0.2)+2(0.3)+3(0.4)+4(0.1) = 0.2+0.6+1.2+0.4 = 2.4. E(X^2) = 1(0.2)+4(0.3)+9(0.4)+16(0.1) = 0.2+1.2+3.6+1.6 = 6.6. Var(X) = E(X^2) - [E(X)]^2 = 6.6 - 5.76 = 0.84. SD = sqrt(0.84) ≈ 0.917.

Q9 ANSWER

(a) Normal, mean=182g, SD=4g. (b) P(x-bar > 190) ≈ 0.0228

(a) By CLT (Normal population), x-bar ~ N(182, 24/sqrt(36)) = N(182, 4). (b) z = (190-182)/4 = 2.0. P(Z > 2.0) = 1 - 0.9772 = 0.0228. There is about a 2.28% chance the sample mean exceeds 190 g.

Q10 ANSWER

(a) Conditions met: np=40>=10, n(1-p)=60>=10. (b) P(p-hat < 0.35) ≈ 0.1515

(a) n=100, p=0.40: np=40>=10 and n(1-p)=60>=10. Random and 10% condition assumed. Normal approximation valid. (b) sigma(p-hat) = sqrt(0.4*0.6/100) = sqrt(0.0024) = 0.0490. z = (0.35-0.40)/0.0490 = -1.02. P(Z < -1.02) ≈ 0.1539.

Q11 ANSWER

95% CI: (0.491, 0.629). Cannot conclusively confirm majority support.

p-hat = 112/200 = 0.56. SE = sqrt(0.56*0.44/200) = sqrt(0.001232) = 0.0351. z* = 1.96. CI: 0.56 +/- 1.96(0.0351) = 0.56 +/- 0.0688 = (0.491, 0.629). Since 0.5 is inside the interval, we cannot conclude majority support at 95% confidence.

Q12 ANSWER

z = 1.79, p-value ≈ 0.037 < 0.05. Reject H0. Evidence supports >60% exercise.

State: H0: p=0.60 vs. Ha: p>0.60. Plan: One-prop z-test, alpha=0.05. Conditions: Random (given), Normal: np0=90>=10, n(1-p0)=60>=10; Independent: population > 1500. Do: p-hat = 98/150 = 0.653. z = (0.653-0.60)/sqrt(0.60*0.40/150) = 0.053/0.04 = 1.79. p-value = P(Z>1.79) ≈ 0.037. Conclude: Since 0.037 < 0.05, reject H0. There is convincing evidence that more than 60% of adults exercise regularly.

Q13 ANSWER

90% CI: approximately (1153, 1207) hours.

(a) Random: stated. Normal: n=25<30, must assume population is Normal or check graph. Independent: >250 bulbs assumed. (b) df = 24; t* = 1.711 (90%). SE = 80/sqrt(25) = 16. CI: 1180 +/- 1.711*16 = 1180 +/- 27.4 = (1152.6, 1207.4). (c) We are 90% confident the true mean lifetime of all bulbs of this type is between 1152.6 and 1207.4 hours.

Q14 ANSWER

t ≈ -1.55, p-value > 0.05 (approx 0.13). Fail to reject H0.

H0: mu1=mu2 vs. Ha: mu1 ≠ mu2. t = (78-83)/sqrt(144/18 + 100/22) = -5/sqrt(8+4.545) = -5/sqrt(12.545) = -5/3.542 ≈ -1.41. Using df=35, |t|=1.41 < t*=2.030. p-value > 0.05. Fail to reject H0. There is not convincing evidence of a difference in mean scores between the two methods at alpha=0.05.

Q15 ANSWER

chi^2 = 1.70. Since 1.70 < 11.07, fail to reject H0. The die appears fair.

H0: the die is fair (each face has P=1/6). Expected each = 120/6 = 20. chi^2 = (18-20)^2/20 + (22-20)^2/20 + (19-20)^2/20 + (21-20)^2/20 + (17-20)^2/20 + (23-20)^2/20 = 4/20+4/20+1/20+1/20+9/20+9/20 = 28/20 = 1.40. (Rounding may give ~1.40-1.70.) Since chi^2 < 11.07, fail to reject H0. No convincing evidence the die is unfair.

Q16 ANSWER

chi^2 ≈ 2.27. Since 2.27 < 3.841, fail to reject H0. No evidence of association.

Table totals: Male=75, Female=75, Online=80, InPerson=70, Total=150. Expected: M-Online=75*80/150=40, M-InPerson=75*70/150=35, F-Online=75*80/150=40, F-InPerson=75*70/150=35. chi^2 = (45-40)^2/40 + (30-35)^2/35 + (35-40)^2/40 + (40-35)^2/35 = 25/40+25/35+25/40+25/35 = 0.625+0.714+0.625+0.714 = 2.678. Since 2.678 < 3.841, fail to reject. No significant association between gender and preference.

Q17 ANSWER

(a) t=3.84>1.746, reject H0. (b) CI: (1.04, 3.64). (c) Plausible values for true slope.

(a) t = (2.34-0)/0.61 = 3.836. Since 3.836 > 1.746 (df=16, alpha=0.05, one-tail), reject H0. Convincing evidence that fertilizer amount has a positive linear relationship with plant height. (b) 95% CI: 2.34 +/- 2.120*0.61 = 2.34 +/- 1.293 = (1.047, 3.633). (c) We are 95% confident that for each additional gram of fertilizer, the true mean plant height increases by between 1.05 and 3.63 cm.

Q18 ANSWER

Curved residual plot => linearity condition violated. Transform x or y (e.g., sqrt or log).

(a) A curved residual plot indicates the relationship between x and y is not linear — the linear model is not appropriate. The model systematically over- or under-predicts in certain ranges. (b) Try transforming variables: use log(x), sqrt(x), or x^2, then re-examine the residual plot. (c) The LINEARITY condition of linear regression (L in LINE) is violated.

Q19 ANSWER

Issues: (1) Observational study cannot establish causation. (2) Possible confounding variables.

(1) Causation vs. correlation: Even a statistically significant result (p=0.002) in an observational study cannot establish causality. Students who eat breakfast may differ in other ways (SES, motivation, sleep). (2) Confounding: Variables like family income or academic engagement may explain both breakfast habits and test performance. Only a controlled experiment with random assignment could support a causal claim.

Q20 ANSWER

(a) We are 95% confident the true mean falls in (42.3, 58.7). (b) The student's statement is INCORRECT.

(a) Correct interpretation: If we repeated this sampling process many times, about 95% of the resulting confidence intervals would capture the true population mean. For THIS interval specifically, we say we are 95% confident. (b) The student's statement implies probability about a fixed (but unknown) parameter — but the true mean is fixed, not random. The 95% refers to the long-run capture rate of the METHOD, not to a probability about this specific interval.
