STAT 200 Sample Midterm Exam

(Time: **50 minutes**)

Name _____ Student ID _____

Lab (session number and day/time, e.g., L2A, Mon, 4pm) _____

Closed notes/books, but one "cheat sheet" ($8.5' \times 11'$*, two sided) is allowed.*

Problem 1 (21 pts). Circle the correct answer: true (T) or false (F) (one answer only).

(1) To check if two binary variables are associated, a 2×2 table is usually used. T F

(2) If a dataset is skewed to the right (i.e., longer right tail), the mean is likely to be larger than the median. T F

(3) The sample standard deviation for the data with an outlier is larger than that for the same data but without the outlier. T F

(4) If the sample size $n \geq 50$, the distribution of a sample proportion will be approximately normal. T F

(5) Sampling variability arises because of the difficulty in obtaining a simple random sample (SRS) from the population. T F

(6) A researcher finds that cancer rates are much higher among smokers than non-smokers. The researcher can reasonably conclude that smoking causes cancer. T F

(7) In a regression model, the point $(\bar{x}, \bar{y})$ will always be on the least square line. T F

Problem 2 (27 pts). Circle the answer that is the most appropriate (one answer only).

(1) Which of the following is <u>not</u> correct

(a) A parameter is a characteristic of a population.

(b) A statistic is a characteristic of a sample.

(c) The value of a statistic may change with different samples.

(d) A well-designed sample can eliminate all error in estimating a parameter.

(2) A sample consists of the following numbers: 2, 3, 5, 6, 7, 10. If the last value 10 is omitted, the sample standard deviation would

(a) increase (b) decrease (c) same (d) cannot tell without calculation

(3) If the sample correlation coefficient based on a sample of five observations on variables X and Y is 0.8, then

(a) Variables X and Y must have a strong linear relationship.

(b) If X and Y have any relationship, it must be a linear relationship.

(c) X and Y may have a linear relationship, but we need to draw a picture to confirm it.

(d) None of the above.

(4) The average time for a student to complete homework is 5 hours per week, with a standard deviation of 2 hours. Four students are randomly selected, the mean and standard deviation of the average time for these four students to complete homework each week are respectively

(a) 5, 2 (b) 5, 1 (c) 20, 8 (d) 20, 4

(5) Referring to question (4). For 100 randomly selected students, there is approximately 95% chance that the average time for these 100 students to complete homework each week is
(a) between 4.8 and 5.2 hours (b) between 1 and 9 hours (c) between 4.6 and 5.4 hours (d) between 3 and 7 hours

(6) Suppose that 70% students need at least 6 hours to complete homework each week. For 100 randomly selected students, what are the mean and standard deviation for the proportion of the 100 students who need at least 6 hours to complete homework each week?
(a) 0.7, 0.21 (b) 0.7, $\sqrt{0.21}$ (c) 0.7, $\sqrt{0.0021}$ (d) 0.7, $\sqrt{0.21}/100$

(7) In a large class, 20% students take a math course, 30% take a biology course, and 10% take both. The probability that a student takes neither a math course nor a biology course is
(a) 0.06 (b) 0.56 (c) 0.6 (d) 0.4

(8) Here is a set of two numbers: $\{0, 20\}$. I want to add one number to the set so that the standard deviation for all three numbers is as small as possible. What number should I add?
(a) 0 (b) 2 (c) 10 (d) 18

(9) In a biology class, the time required for students to complete a homework assignment has an average of 5 hours, with standard deviation of 2 hours. A group of 50 students are randomly selected. The distribution of the times these students take to complete a homework assignment will be (approximately)
(a) normally distributed, by the Central Limit Theorem.
(b) normally distributed, by the Law of Large Number.
(c) a binomial distribution.
(d) cannot be determined based on the information given.

This table is for instructor use only

Problem 1	Problem 2	Problem 3	Problem 4	Problem 5	Total Mark

Problem 3 (18 pts). A researcher is developing a new drug for reducing pain. To test the effectiveness of the new drug, the researcher obtained a random sample of 50 patients, and divided these patients randomly into two groups, with 25 people in each group. Each subject in one group took the new drug, while each subject in the other group took a placebo. Neither the subjects nor the researcher knew which group each subject was randomized to. In the new drug group 20 subjects reported improvements, while in the placebo group 15 subjects reported improvements.

(1) (5 pts) This study provides an example of (*circle all apply, and penalty for wrong choices*):

(a) replication (b) blocking (c) randomization (d) blinding (e) confounding

(2) (6 pts) The population is _____.

The statistics are _____.

The parameters are _____.

(3) (3 pts) Which of the following modifications made to the study design would turn this study into an observational study? (choose one answer)

(a) The participating subjects choose whether they receive the new drug or not.
(b) The researcher has information about which group each subject was randomized to.
(c) Remove any possible confounders.
(d) Obtain a simple random sample.

(4) (4 pts) (i) Summarize the data using a 2×2 table (show the contingency table in the space below, fill in the cell counts and marginal totals, and clearly label the row and column).

(ii) Is there an association between group membership (drug/placebo) and reduction in pain? _____ (Yes or No)

Problem 4 (14 pts). Consider the IQ's of husband/wife pairs. A sample of 100 husband/wife pairs is obtained. Based on the data, the average IQ of husbands is 120 and the average IQ of wives is 110. The sample standard deviation of the husbands' IQ's is 25 and the sample standard deviation of the wives' IQ's is 20. The correlation between husbands' and wives' IQ's is 0.8.

(1) (5 pts) Find the least square line for predicting a wife's IQ based on her husband's IQ.

(2) (3 pts) Predict the IQ of a wife whose husband has an IQ of 100.

(3) (3 pts) The scatterplot of the wife's' IQ's and her husbands' IQ's will look like (choose one answer)

(a) linear (b) nonlinear (c) cannot tell based on the given information.

(4) (3 pts) The residual plot will look like (choose one answer)

(a) reasonable (b) unreasonable (c) cannot tell based on the given information.

Problem 5 (20 pts). Suppose that midterm scores of a large statistics class approximately follow a normal distribution with mean 70 and standard deviation 10. (You may need the following information: let $Z \sim N(0,1)$, then $P(Z < -1.645) = 0.05, P(Z < 1) = 0.84, P(Z < 1.5) = 0.93, P(Z < 2) = 0.98, P(Z < 2.5) = 0.99$.)

(1) (5 pts) The probability that a randomly selected student has a score at least 80 is _____.

(2) (5 pts) Suppose that we do not know the mean and standard deviation of the normal distribution of the exam scores, but we know that the median is 68 and the 5th percentile is 30. The mean and standard deviation are respectively _____ and _____.

(3) (5 pts) If **two** students are randomly selected, what is the *distribution* for the number of students who score at least 80 (based on $\mu = 70, \sigma = 10$)?

(4) (5 pts) If 100 students are randomly selected, what is the probability that **at most** 40% of these 100 students would get at least 80?