MA 180/418, Fall 2010 Final Exam, Version A

Student Name (PRINT):....

Student Signature:

The test consists of 20 questions.

Questions 1 through 10 are multiple-choice and worth 2.5 points each. Questions 11 through 20 are computational and worth 7.5 points each.

For questions 1 through 10, circle the correct answer (a, b, c, or d) after each question. Each question is 2.5 points. No partial credit.

Which statement is correct about the mean, median, and mode of a sample? Q1

- (a) The mean is the average of the largest value and the smallest value in the sample, while the mode is the value with the greatest frequency (there may be more than one mode).
- (b) The mean is the arithmetic average of all the sample values, while the median is the middle value when the sample is arranged in order of increasing magnitude.
- (c) The median is the middle value in the original data list, while the mode is the largest value.
- (d) The mean is the average of the median and the mode.

Correct answer (circle one): (a)(b) (c)(d)

Q2 What is the standard normal distribution?

- (a) The standard normal distribution is any bell-shaped probability distribution.
- (b) The standard normal distribution is a bell-shaped probability distribution with $\mu = 1$ and $\sigma = 0$.
- (c) The standard normal distribution is a bell-shaped probability distribution with $\mu = 1$ and $\sigma = 1$.
- (d) The standard normal distribution is a bell-shaped probability distribution with $\mu = 0$ and $\sigma = 1$.

Correct answer (circle one): (a)(b) (c)(d)

Q3 What is the meaning of the term **probability distribution**?

- (a) A probability distribution is a statement of all possible elementary outcomes of a procedure, with a stated probability for each outcome such that the sum of the probabilities is between 0.0 and 1.0.
- (b) A probability distribution is a statement of all possible elementary outcomes of a procedure, with a stated probability for each outcome such that the sum of the probabilities is above 0.0.
- (c) A probability distribution is a statement of all possible elementary outcomes of a procedure, with a stated probability for each outcome such that the sum of the probabilities is equal to 1.0.
- (d) A probability distribution is a statement of all possible elementary outcomes of a procedure, with a stated probability for each outcome such that the sum of the probabilities is at least 1.0.

Correct answer (circle one): (a) (b) (c) (d)

Q4 What is the difference between the null hypothesis and alternative hypothesis?

- (a) The null hypothesis is a statement about a property of a population, while the alternative hypothesis is a statement about a property of a sample.
- (b) The null hypothesis is a statement that the value of a population parameter is known, while the alternative hypothesis is a statement that the value of a population parameter is not known.
- (c) The null hypothesis is a statement that the value of a population parameter is equal to a claimed value, while the alternative hypothesis is a statement that the parameter has a value different from the claimed value.
- (d) The null hypothesis is a statement that the value of a sample statistic is equal to a claimed value, while the alternative hypothesis is a statement that the value of a sample statistic is different from the claimed value.

Correct answer (circle one): (a) (b) (c) (d)

What is a **critical value**?

 O_{5}

- (a) A critical value is the margin of error in a confidence interval.
- (b) A critical value is the degree of confidence when constructing a confidence interval estimate of a population parameter, such as mean or standard deviation.
- (c) A critical value is the length of a confidence interval.
- (d) A critical value is the number on the borderline separating sample statistics that are likely to occur from those that are unlikely to occur.

Correct answer (circle one): (a) (b) (c) (d)

- Q6 What is the meaning of the term **margin of error**?
 - (a) It is the maximum likely distance between the true value of the population parameter and its point estimate.
 - (b) It is a possible error in the construction of a confidence interval estimate of a population parameter, such as mean or standard deviation.
 - (c) It is the probability of a Type I error.
 - (d) It is the length of a confidence interval.

Correct answer (circle one): (a) (b) (c) (d)

Q7 Which definition of **multiple regression equation** is stated correctly?

- (a) A multiple regression equation expresses a linear relationship between a response variable y and two or more predictor variables (x_1, \ldots, x_k) .
- (b) A multiple regression equation expresses a linear relationship between a predictor variable y and two or more response variables (x_1, \ldots, x_k) .
- (c) A multiple regression equation expresses a linear relationship between two or more response variables (y_1, \ldots, y_k) and a predictor variable x.
- (d) A multiple regression equation expresses a nonlinear relationship between a response variable y and a predictor variable x.

Correct answer (circle one): (a) (b) (c) (d)

Q8 Which formula is used to compute the **pooled sample proportion**?

- (a) $\bar{p} = \frac{x_1}{n_1} + \frac{x_2}{n_2}$. (b) $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$. (c) $\hat{p} = \hat{p}_1 + \hat{p}_2$.
- (d) $E = z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}.$

Correct answer (circle one): (a) (b) (c) (d)

Q9 What is the meaning of the term **prediction interval**?

- (a) A prediction interval is an interval estimate of a predicted value of the dependent variable given some value of the independent variable.
- (b) A prediction interval is the interval between the observed value and the predicted value of the response variable y.
- (c) A prediction interval is a confidence interval with a 100% confidence level.
- (d) A prediction interval is the interval between explained and unexplained variation, provided there is a linear correlation between two variables.

Correct answer (circle one): (a) (b) (c) (d)

Q10 What is the meaning of the term **coefficient of determination**?

- (a) The coefficient of determination is the sum of squares of residuals.
- (b) The coefficient of determination is the square of the slope of the regression line, provided there is a linear correlation between two variables.
- (c) The coefficient of determination is the square root of the linear correlation coefficient, provided there is a linear correlation between two variables.
- (d) The coefficient of determination is the amount of the variation in y that is explained by the variation in x.

Correct answer (circle one): (a) (b) (c) (d)

For questions 11 through 20, write your answer in the space provided. Show your work. Each question is worth 7.5 points.

Q11 Scores on the SAT test have a mean of 1590 and a standard deviation of 325. Scores on the ACT test have a mean of 22 and a standard deviation of 4.4. Which is relatively better: a score of 1957 on the SAT test or a score of 27 on the ACT test? Why? (Give both Z-scores and show how you computed them.)

Q12 A ski gondola can carry 16 passengers and has maximal capacity of 3200 lb. That capacity will be exceeded if 16 passengers have weights with a mean greater than 3200/16 = 200 lb. Suppose weights of passengers are normally distributed with a mean of 160 lb and a standard deviation of 45.8 lb.

(a) Find the probability that a randomly selected passenger has weight greater than 200 lb.

(b) Find the probability that 16 randomly selected passengers have a mean weight greater than 200 $\rm lb$

(c) If you used the Central Limit Theorem, then state the corresponding requirements and explain why they are met in this problem

(d) Which question, (a) or (b), is more relevant to the safety of the gondola? What is the chance that the gondola's capacity will be exceeded?

Q13 In a poll, 86% of 4050 adults surveyed said that they use the Internet.

(a) Construct a 98% confidence interval estimate for the proportion of all adults who use the Internet (round off the confidence limits to three decimal places).

(b) Is it correct for a newspaper reporter to write that "9 out of 10 adults use the Internet"? Why or why not?

Q14 Twelve different video games showing substance use were observed and the duration of game play (in seconds) are listed below. Assume that the data comes from a normal population.

 $4015 \ \ 3996 \ \ 3815 \ \ 4022 \ \ 4411 \ \ 4614 \ \ 4586 \ \ 4029 \ \ 5002 \ \ 4855 \ \ 4331 \ \ 4296$

(a) Find the sample mean and the sample standard deviation (round off to one decimal place).

(b) Construct a 90% confidence interval estimate of μ , the mean duration of game play (round off the confidence limits to the nearest whole numbers).

(c) Construct a 99% confidence interval estimate of μ , the mean duration of game play (round off the confidence limits to the nearest whole numbers).

(d) Which interval is wider? Why?

Q15 Use the data from the previous problem (Q14) to test the claim that the standard deviation of the duration of game play is less than 550 sec. Use a 0.10 significance level. (a) State the null and the alternative hypotheses.

(b) Compute the test statistic (round off to three decimal places).

(c) Draw a diagram. Find and mark the critical value(s).

(d) Make an initial conclusion and a final conclusion.

(e) What is the requirement on the population in this test? How strict is it?

[Bonus] For an extra credit, find an interval for the P-value by using a table.

Q16 Sitting heights of 33 adult males are measured and a mean of 90.7 cm is obtained. Assume that adult males have sitting heights with standard deviation of 3.8 cm. Use a 0.20 significance level to test the claim that adult males have sitting heights with mean above 90 cm.

(a) State the null and the alternative hypotheses.

(b) Compute the test statistic (round off to two decimal places).

(c) Draw a diagram. Find and mark the critical value(s).

(d) Make an initial conclusion and a final conclusion.

(e) Find the P-value of the test.

(f) Make an initial conclusion based on the P-value. How do you use the P-value to make a conclusion?

(g) What are the requirements on the population/sample in this test? How strict are they?

Q17 Use a 0.05 significance level to test the claim that the weights of pre-1983 pennies and post-1983 pennies have the same amount of variation. Assume that all populations are normal. Sample results are summarized below:

| Weights of pre-1983 pennies: | n=31 | $\bar{x} = 3.02 \text{ g}$ | s = 0.0187 g |
|-------------------------------|------|----------------------------|---------------|
| Weights of post-1983 pennies: | n=41 | $\bar{x} = 3.01 \text{ g}$ | s = 0.0275 g |

(a) State the null and the alternative hypotheses.

(b) Compute the test statistic (round off to four decimal places).

(c) Draw a diagram. Find and mark the critical value(s).

(d) Make an initial conclusion and a final conclusion.

(e) Just glancing at the sample results, does it appear that the technology of penny production has improved since 1983? Why or why not? If the technology did improve, what changes in the mean or standard deviation would you expect?

Q18 Listed below are concentrations (in parts per million) of CO_2 and temperatures (in degrees) for different years.

| CO_2 | 308 | 312 | 315 | 319 | 322 | 323 | 327 | 330 | 333 | 337 |
|-----------------|------|------|------|------|------|------|------|------|------|------|
| Temperature | 14.0 | 14.1 | 13.9 | 14.0 | 13.8 | 14.0 | 13.7 | 13.6 | 13.8 | 13.9 |

(a) Find the coefficient of determination (round off to three decimal places)

(b) Find the linear correlation coefficient (round off to three decimal places)

(c) Find the standard error of estimate (round off to four decimal places)

(d) Use a 0.05 significance level to test the hypothesis that there is a linear correlation between the CO_2 concentration and the temperature:

Give the critical value and explain how you conduct the test

Make the conclusion

Give the P-value (but if your calculator does not return the P-value, just say that and ignore it)

Q19 Use the data from the previous problem (Q18) to do the following.

(a) Find the equation of the regression line (keep at least three decimal places for the y-intercept and at least five decimal places for the slope)

(b) Find the best predicted temperature in the year when the CO_2 concentration is 340 (round off to two decimal places).

Show which formula you use and explain why you use it.

(c) What proportion of the variation in the temperature is explained by the variation in the CO_2 concentration?

(d) Find the sum of squares of residuals, i.e., find $\sum (y - \hat{y})^2$. Round it off to three decimal places

[Bonus] Find a 95% prediction interval for the estimate of the temperature in the year when the CO_2 concentration is 340.

Q20 The table below shows several multiple regression equations for the selling price of house (y) with three predictor variables: the list price (LP), the living area (LA), and the lot size (Lot).

| Predictor (x) | | | Adjusted | |
|-----------------|---------|----------------|----------|--|
| Variables | P-value | \mathbb{R}^2 | R^2 | Regression Equation |
| | | | | |
| LP,LA,Lot | 0.002 | 0.988 | 0.933 | $\hat{y} = 1200 + 0.97 \text{LP} + 0.3 \text{LA} + 470 \text{Lot}$ |
| LP,LA | 0.002 | 0.966 | 0.913 | $\hat{y} = -80 + 0.98 \mathrm{LP} - 0.4 \mathrm{LA}$ |
| LP,Lot | 0.002 | 0.965 | 0.912 | $\hat{y} = 990 + 0.95 \mathrm{LP} + 450 \mathrm{Lot}$ |
| LA,Lot | 0.005 | 0.873 | 0.815 | $\hat{y} = 110,000 + 97 \mathrm{LA} + 18,000 \mathrm{Lot}$ |
| LP | 0.049 | 0.707 | 0.650 | $\hat{y} = 90 + 0.988 \mathrm{LP}$ |
| LA | 0.055 | 0.707 | 0.650 | $\hat{y} = 140,000 + 102 \mathrm{LA}$ |
| Lot | 0.312 | 0.503 | 0.403 | $\hat{y} = 280,000 + 20,100 \mathrm{Lot}$ |

(a) If only one predictor variable is used, which single variable is best? Why?

(b) If exactly two predictor variables are used, which two variables should be chosen? Why?

(c) Which regression equation is best for predicting the selling price? Why? (*Note: there are two equations for which you can argue that they are better than others. Choosing either one of those two is acceptable provided your explanations are reasonable.*)

(d) A home for sale with a list price of \$350,000 has a living area of 2,500 square feet and is on a 1.0 acre lot. What is the best predicted value of the selling price?