Final Exam, Fall 2016

1) A dummy variable can be assigned up to three values.
    a) True
    b) False

2) Transformations may be used when nonlinear relationships exist between the response and explanatory variable when performing regression.
    a) True
    b) False

3) The value of the coefficient of determination can never decrease when more variables are added to the model.
    a) True
    b) False

4) For statistical tests of significance about the regression coefficients, the null hypothesis is that the slope is 1.
    a. True
    b. False

5) If the assumptions of regression have been met, residuals plotted against the independent variable(s) will typically show patterns.
    a) True
    b) False

6) The noise in a regression model is assumed to have zero variance.
    a. True
    b. False

7) The SSR value indicates how much of the total variability in the dependent variable is explained by the regression model.
    a) True
    b) False

8) A uniform distribution has a smaller standard deviation than a mound shaped distribution, since the uniform distribution is flat while the mound shaped distribution is not flat.
    a. True
    b. False

9) If a 95% confidence interval for the mean was computed as (25,50), then if several more samples were taken with the same sample size, then 95% of them would have a sample mean between 25 and 50.
    a. True
    b. False

10) 15 cards are selected out of a 52 card deck such that after each card is selected, it is placed back into the deck and the deck is reshuffled. Then the total number of hearts selected follows a binomial distribution.
    a. True
    b. False

11) If the equation of the least squares regression line was computed to be y=45.7+3.1x, then the correlation cannot be less than 0.
    a. True
    b. False

12) If the equation of the regression line that relates percent blood alcohol (x) to reaction time in milliseconds (y) is y=36 - 1.3x, then the slope tells us that for every percent increase in blood alcohol, we can expect reaction time to go down by 1.3 milliseconds
    a. True
    b. False

13) A researcher found the correlation between age of death and number of cigarettes smoked per day to be -0.95. Based just on this information, the researcher can justly conclude that smoking causes early death.
    a. True
    b. False

14) When computing confidence intervals for paired differences, only the difference values (x1 – x2) are used and not the original values of x1 and x2 to compute the mean and standard deviation.
    a. True
    b. False

15) With a large sample size such as n = 100, we do not need to assume the distribution is normal in order to compute a 95% confidence interval for the mean.
    a. True
    b. False

16) If a 95% confidence interval for the mean number of hours that students study per week is (18,23), then there is a 95% chance that a randomly selected student will study between 18 to 23 hours per week.
    a. True
    b. False

17) Consider a large number of countries around the world. There is a positive correlation between the number of Nintendo games per person x and the average life expectancy y. Does this mean that we could increase the life expectancy in Rwanda by shipping Nintendo games to that country?

a) Yes. The correlation says that as the number of Nintendo games per person goes up, so does life expectancy.
b) No. If the correlation were negative, we could accept that conclusion, but this correlation is positive.
c) Yes. Positive correlation means that, if we increase x, then y will also increase.
d) No. The positive correlation just shows that richer countries have both more Nintendo games per person and higher life expectancies.
e) It makes no sense to calculate correlation between these variables.

18) A least-squares regression line is not just any line drawn through the points of a scatterplot. What is special about a least-squares regression line?

a) It passes through all the points.
b) It minimizes the squared values of the data.
c) It has slope equal to the correlation between the two variables.
d) It minimizes the sum of the squared vertical distances of the data points from the line.

19) If the least-squares regression line for predicting y from x is $y = 500 – 20x$, what is the predicted value of y when $x = 10$?

a) 300
b) 500
c) 4800
d) 700
e) 20

20) Suppose that the least-squares regression line for predicting y from x is $y = 100 + 1.3x$. Which of the following is a possible value for the correlation between x and y?

a) 1.3
b) –1.3
c) 0
d) –0.5
e) 0.5

21) According to a college survey, 22% of all students work full time. Find the standard deviation for the number of students who work full time in samples of size 16.

a) 1.94
b) 2.75
c) 1.66
d) 2.63
e) 2

22) The mean annual salary of employees at a company is $40,000 with a standard deviation of $3500. At the end of the year, each employee receives a $2000 bonus and a 4% raise (based on salary). What is the standard deviation of the new salaries?

a) 3360
b) 3872
c) 3546
d) 3640
e) 3905

23) If a card is chosen from a standard deck of cards, what is the probability of getting a five or a seven?

a) 4/52
b) 1/26
c) 8/52
d) 1/169
e) None of the above

24) Let X and Y be independent random variables with $X \sim N(0, 1)$ and $Y \sim N(0, 2)$. The value of $P(X > Y)$ is

a) 0
b) 0.05
c) 0.50
d) 0.95
e) 0.45

25) Which of the following is NOT an assumption of the Binomial distribution?

a) All trials must be identical.
b) All trials must be independent.
c) Each trial must be classified as a success or a failure
d) The number of successes in the trials is counted.
e) The probability of success is equal to .5 in all trials.

26) An industrial designer wants to determine the average amount of time it takes an adult to assemble an "easy to assemble" toy. A sample of 36 times yielded an average time of 19.92 minutes, with a sample standard deviation of 8.6 minutes. Calculate a 95% confidence interval for the mean assembly time.

a) (17.16,22.68)
b) (17.40,22.43)
c) (17.51,22.33)
d) (17.11,22.73)
e) (17.22,22.62)

27) What is the smallest sample size required to provide a 95% confidence interval for a mean, if it important that the interval be no longer than 1cm? You may assume that the population is normal with variance 9.

a) 1245
b) 34
c) 95
d) 139
e) 216

28) A random sample of 100 preschool children in Newton revealed that only 60 had been vaccinated. Provide an approximate 95% confidence interval for the proportion vaccinated in that suburb.

a) (.512,.687)
b) (.503,.682)
c) (.501,.698)
d) (.516,.683)
e) (.504,.696)

29) Suppose X~N(5,32). What is the value of P(X≤2) ?

a) 0.8413
b) 0.1587
c) 0.7258
d) 0.2742
e) 0.2979

30) Suppose *X* is normally distributed with mean 5. If $P(X \le 3) = 0.2$ what is the standard deviation of *X*?

a) 2.52
b) 0.38
c) 0.42
d) 2.38
e) 1.16

31) Suppose that X~N(2,1) and Y~N(3,2). Assuming X and Y are independent what is the distribution of X+Y?

a) N(3,5)
b) N(5,3)
c) N(3,3)
d) N(5,5)
e) None of the above

32) Let X be a random variable with the following probability mass function; $P(X=-1)=0.2, P(X=0)=0.4, P(X=1)=0.4$ Compute $P(X=0 \mid X \le 1).$

a) 0.2
b) 0.4
c) 0.5
d) 0.3
e) None of the above

33) Let X and Y be two random variables with the following joint probability distribution; $P(X=-1 \text{ and } Y=-1)=1/3$, $P(X=0 \text{ and } Y=0)=1/3$ and $P(X=1 \text{ and } Y=1)=1/3$ Compute $P(XY=1).$

a) 0
b) 1/9
c) 1/3
d) 2/3
e) 2/9

34) The weight of a gum drop (piece of candy) in ounces is normally distributed with mean 2 and standard deviation 0.25. A bag contains 10 independent gum drops. The probability that the total weight of the gum drops in the bag exceeds 20 ounces is

a) 0.25
b) 0.5
c) 0.33
d) 0.75
e) 0.35

35) Google and Microsoft are computer stocks that frequently move together. During a 25 day period, the value of Google stock went up 17 days. On 10 of these 17 days, the value of Microsoft stock also went up. On the 8 days when Google stock did not go up, Microsoft stock went up on 2 of these days. If Microsoft does not go up, what is the probability Google will go up?

a) 0.71
b) 0.11
c) 0.54
d) 0.43
e) 0.67

36) The purpose of hypothesis testing is to help the researcher reach a conclusion about _____ by examining the data contained in _____.

a) a population, a sample
b) an experiment, a computer printout
c) a population, an event
d) a sample, a population

37) If the coefficient of determination (R^2) is 0.80, then which of the following is true regarding the slope of the regression line?

a) All we can tell is that it must be positive.
b) It must be 0.80
c) It must be 0.89.
d) Cannot tell the sign or the value.
e) The slope must be significant.

38) Suppose X is normally distributed with mean 5 and standard deviation of 0.4. Suppose $P(X \le x_0) = P(Z \le 1.3)$. What is the value of x_0?

a) 6.94
b) 4.48
c) 2.02
d) 5.52
e) 1.43

39) A multiple regression model with two independent variables exhibits a highly significant F-ratio, but each variable's individual t-statistic is insignificant. The most likely cause of such a situation is

a) Heteroskedasticity
b) Homoskedasticity
c) Multicollinearity
d) Non-normality of residuals

40) Suppose you run a regression with response variable y and three explanatory variables. When the null hypothesis, Ho: $\beta 1 = \beta 2 = \beta 3 = 0$, is rejected, the interpretation should be:

a) there is no linear relationship between y and any of the three independent variables
b) there is a regression relationship between y and at least one of the three independent variables
c) all three independent variables have a slope of zero
d) all three independent variables have equal slopes
e) there is a regression relationship between y and all three independent variables

41) What is the meaning of the term "heteroscedasticity"?

a) The variance of the errors is not constant
b) The variance of the dependent variable is not constant
c) The errors are not linearly independent of one another
d) The errors have non-zero mean

42) Suppose you have estimated wage = 5 + 3education + 2gender – edu*gender, where gender is one for male and zero for female. Suppose instead that gender had been one for female and zero for male. Under this coding what would be the the sum of the coefficients for the gender and interaction variables? (that is we want $b_{gender} + b_{edu*gender}$)

a) -3
b) -1
c) 0
d) 1
e) 2

43) Which of the following can NOT be answered from a regression equation?

a) Predict the value of y at a particular value of x.
b) Estimate the slope between y and x.
c) Estimate whether the linear association is positive or negative.
d) Estimate whether the association is linear or non-linear

44) A regression equation for left palm length (y variable) and right palm length (x variable) for 60 college students gave an error sum of squares (SSE) of 50.7 and $s_y = 1.26$. Find the coefficient of determination

a) 12.6%
b) 84.5%
c) 45.9%
d) 54.1%
e) 76.5%

45) Which of the following is not necessarily true for independent events A and B?

a) P(A and B) = P(A)P(B)
b) P(A|B)=P(A)
c) P(B|A)=P(B)
d) P(A or B) = P(A)+P(B)

46) You plan to determine an approximate 95% confidence interval for a population proportion; you plan to use a conservative margin of error of 2%. How large a sample size do you need?

a) 100
b) 406
c) 2401
d) 1205
e) 200

47) A 95% confidence interval for the proportion of young adults who skip breakfast is .20 to .27. Which of the following is a correct interpretation of the 95% confidence level?

a) In about 95% of all studies for which this procedure is used, the confidence interval will cover the true population proportion, but there is no way to know if this interval covers the true proportion or not.
b) There is a 95% probability that the proportion of young adults who skip breakfast is between .20 and .27.
c) If this study were to be repeated with a sample of the same size, there is a 95% probability that the sample proportion would be between .20 and .27.
d) The proportion of young adults who skip breakfast 95% of the time is between .20 and .27.

48) The smaller the p-value, the

a) stronger the evidence against the alternative hypothesis
b) stronger the evidence for the null hypothesis
c) stronger the evidence against the null hypothesis
d) none of the above

49) Which choice lists two statistics that give information only about the "spread" of a dataset (and not the location)?

a) IQR and standard deviation
b) Mean and standard deviation
c) Median and range
d) Mean and median

For each statement below, determine if the statement is a typical null hypothesis (Ho) or alternative hypothesis (Ha).

50) There is no difference between the proportion of overweight men and overweight women in America.

a) Null Hypothesis
b) Alternative Hypothesis

51) The proportion of overweight men is greater than the proportion of overweight women in America.

a) Null Hypothesis
b) Alternative Hypothesis

52) The average price of a particular statistics textbook over the internet is the same as the average price of the textbook sold at all bookstores in a college town.

a) Null Hypothesis
b) Alternative Hypothesis

53) The average time to graduate for an undergraduate English major is less than the average time to graduate for a history major.

a) Null Hypothesis
b) Alternative Hypothesis

54) A list of 7 pulse rates of undergraduate students is : A list of 7 pulse rates is: 70, 64, 80, 80, 74, 86, 93. What is the median for this list?

a) 81
b) 78
c) 80
d) 91
e) 79

55) A major credit card company has determined that customers charge between $100 and $1100 per month. Given that the average monthly amount charged is uniformly distributed, what percent of monthly charges are between $600 and $889?

a) 15.4%
b) 57.8%
c) 28.9%
d) 38.2%
e) 78.3%

56) A continuous random variable X is uniformly distributed over the interval 0.9 to 2.9. Find the standard deviation of X.

a) 0.29
b) 0.72
c) 0.58
d) 0.52
e) 0.77

57) From past records it is known that the average life of a battery used in a digital clock is 305 days. The battery life is normally distributed. The battery was recently modified to last longer. A sample of 30 of the modified batteries was tested. It was discovered that the mean life was 311 days and the sample standard deviation was 9 days. We want to test at the 0.05 level of significance whether the modification increases the life of the battery. What is our decision ?

a) Fail to reject the null hypothesis
b) Reject the null hypothesis
c) Not enough information

58) A survey at Boston University showed that 870 of 1100 students sampled supported a fee increase to fund improvements to the student recreation center. Using the 95% level of confidence, what is the confidence interval?

a) (0.767, 0.815)
b) (0.759, 0.822)
c) (0.771, 0.811)
d) (0.714, 0.866)
e) (0.753,0.858)

59) [Recall the previous question]. If university officials say that at least 70% of the voting student population supporting the fee increase, what conclusion can be drawn based on a 95% level of confidence?

a) 70% is not in the interval, need to take another sample.
b) 70% is not in the interval, so assume it will not be supported.
c) 70% is below the interval, so assume it will be supported.
d) Since this was not based on population, cannot make conclusion.

60) It has been hypothesized that overall academic success for college freshmen as measured by grade point average (GPA) is a function of IQ scores (X_1) , hours spent studying each week (X_2), and one's high school average (X_3).Suppose the regression equation is: $\hat{Y} = 6.9 + 0.055X_1 + 0.107X_2 + 0.0853X_3$. We also know that $s_e = 6.313$ and $R^2 = 0.826.$ What is the predicted GPA for a student with an IQ of 108, 32 hours spent studying per week and a high school average of 82?

a) 21.7
b) 23.2
c) 22.86
d) 20.55
e) 21.11

61) Suppose we obtain the following regression model for baseball bat sales (Y) when regressed against seasonal indicator variables; $\hat{y} = 100 - 40Spring + 20Wtr - 15Fall$. If we decide to make the baseline season Fall, what would then be the resulting coefficient for Winter (Wtr)?

a) 25
b) -40
c) 30
d) 15
e) None of the above

62) Woof Chow Dog Food Company believes that it has a market share of 15%. They survey *n* = 100 dog owners and ask whether or not Woof Chow is their regular brand of dog food, and 18 people say yes. Based upon this information, what is the value of the test statistic for a two sided hypothesis test?

a) -0.46
b) 0.95
c) 0.46
d) 0.84
e) -0.67

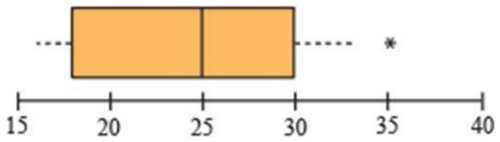
63) The Central Limit Theorem plays an important role in statistics because

a) for any size sample, it says the sampling distribution of the sample mean is approximately normal
b) for a large n, it says the sampling distribution of the sample mean is approximately normal, regardless of the population
c) for any population, it says the sampling distribution of the sample mean is approximately normal, regardless of the sample size
d) for a large n, it says the population is approximately normal

64) The National Association of Realtors reported that 26% of home buyers in the state of Florida are foreigners in 2012. When testing the validity of this report, a Type II error would occur if it was concluded that the proportion of foreign buyers was

A) more than 26% when, in reality, the proportion was 26% or less.
B) not equal to 26% when, in reality, the proportion was equal to 26%.
C) 26% or less when, in reality, the proportion was more than 26%.
D) equal to 26% when, in reality, the proportion was not equal to 26%.

65) Season's Pizza delivers food items to homes in their local area. The following box-and-whisker plot describes the distribution for delivery times in minutes.



Based on this plot, which one of the following statements is correct?

A) The average delivery time is 25 minutes.
B) There are no outliers in this data set.
C) The 75th percentile in this data set is 30 minutes.
D) The second quartile is approximately 18 minutes.
E) None of the above

66) YouTube would like to test the hypothesis that the average length of an online video watched by a user is more than 6 minutes. A random sample of 40 people watched online videos that averaged 6.6 minutes in length with a standard deviation of 1.7 minutes. YouTube would like to set $\alpha = 0.05$. Which one of the following statements is true?

```
One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      40         6.6    .2687936         1.7    6.056314    7.143686
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   2.2322
Ho: mean = 6                                     degrees of freedom =       39

    Ha: mean < 6                 Ha: mean != 6                 Ha: mean > 6
 Pr(T < t) = 0.9843         Pr(|T| > |t|) = 0.0314          Pr(T > t) = 0.0157
```

A) Because the *p*-value is greater than α, we reject the null hypothesis and conclude that the average length of an online video is more than 6 minutes.
B) Because the *p*-value is greater than α, we fail to reject the null hypothesis and conclude that the average length of an online video is less than 6 minutes.
C) Because the *p*-value is less than α, we fail to reject the null hypothesis and conclude that the average length of an online video is less than 6 minutes.
D) Because the *p*-value is less than α, we reject the null hypothesis and conclude that the average length of an online video is more than 6 minutes.

67) Suppose there is a study regarding years of education and hourly wages (in dollars) in the US and it fits a simple linear regression model for hourly wages on years of education. The fitted regression line is (hourly wages) = −4.96 + 1.47 · (years of education). Suppose a 95% confidence interval for the slope (β1) of the regression line is ($ 1.27, $ 1.67). Suppose now that we want to perform hypotheses testing with Ho : β1 = 0 vs Ha : β1 ≠ 0. What should be our conclusion from this test?

a) At significance level α = 0.05, we reject Ho and conclude that there is a significant linear relationship between years of education and hourly wages
b) At significance level α = 0.05, we fail to reject Ho and conclude that there is not a significant linear relationship between years of education and hourly wages
c) At significance level α = 0.05, we reject Ho and conclude that there is not a significant linear relationship between years of education and hourly wages
d) At significance level α = 0.05, we fail to reject Ho and conclude that there is a significant positive linear relationship between years of education and hourly wages

68) As part of a study on student loan debt, a national agency that underwrites student loans is examining the differences in student loan debt for undergraduate students. One question the agency would like to address specifically is whether the mean undergraduate debt of Hispanic students graduating in 2009 is less than the mean undergraduate debt of Asian- American students graduating in 2009. To conduct the study, a random sample of 92 Hispanic students and a random sample 110 Asian- American students who completed an undergraduate degree in 2009 were taken. The undergraduate debt incurred for financing college for each sampled student was collected. Let μ_H denote the population average student loan debt for Hispanic students, and μ_A the population average student loan debt for Asian-American students. Using the Stata output below, test the hypothesis $H_o : \mu_A = \mu_H \quad H_a : \mu_A > \mu_H$. Clearly interpret your results.

```
. ttest hispanic == asianamerican, unpaired unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
hispanic |      92    18659.18    490.0128    4700.037    17685.83    19632.53
asiana~n |     110    20002.54     553.725    5807.517    18905.07       21100
---------+--------------------------------------------------------------------
combined |     202    19390.71    377.2018    5361.045    18646.93    20134.49
---------+--------------------------------------------------------------------
    diff |           -1343.357    739.4078               -2801.401    114.6875
------------------------------------------------------------------------------
    diff = mean(hispanic) - mean(asianamerican)                   t =  -1.8168
Ho: diff = 0                     Satterthwaite's degrees of freedom =  199.798

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0354         Pr(|T| > |t|) = 0.0707          Pr(T > t) = 0.9646
```

a) Reject the null hypothesis
b) Fail to reject the null hypothesis
c) Accept the null hypothesis
d) None of the above

69) Consider the following multiple regression output.

```
. regress y x1 x2 x3 x4 x5 x6 x7 x8 x9

      Source |       SS           df       MS      Number of obs   =       554
-------------+----------------------------------   F(9, 544)       =      7.73
       Model |  [redacted]         9  789.855923   Prob > F        =    0.0000
    Residual |  55572.5479       544  102.155419   R-squared       =  [redacted]
-------------+----------------------------------   Adj R-squared   =    0.0987
       Total |  62681.2512       553  113.347651   Root MSE        =  [redacted]

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1323038   .0723416     1.83   0.068    -.0097993    .2744068
          x2 |   .0924671   .5528632     0.17   0.867    -.9935411    1.178475
          x3 |   .1931725   .1295204     1.49   0.136  [redacted]
          x4 |  -9.86e-06   .0000205    -0.48   0.632    -.0000502    .0000305
          x5 |   1.397048   .4355554     3.21   0.001     .5414719    2.252625
          x6 |    .341846   .5025679     0.68   0.497    -.6453654    1.329057
          x7 |   2.550475   1.009268     2.53   0.012     .5679346    4.533015
          x8 |  -.0615556   .0892273    -0.69   0.491    -.2368278    .1137166
          x9 |   5.328812     1.2751     4.18   0.000      2.82409    7.833533
       _cons |  -8.853647   6.666171    -1.33   0.185    -21.94824    4.240941
------------------------------------------------------------------------------
```

Calculate the value of s_e.

a) 3.31
b) 10.11
c) 6.76
d) 10.41
e) 8.93

70) Consider the following multiple regression output.

```
. regress y x1 x2 x3 x4 x5 x6 x7 x8 x9

      Source |       SS           df       MS      Number of obs   =       554
-------------+----------------------------------   F(9, 544)       =      7.73
       Model |                     9  789.855923   Prob > F        =    0.0000
    Residual |  55572.5479        544  102.155419   R-squared       =
-------------+----------------------------------   Adj R-squared   =    0.0987
       Total |  62681.2512        553  113.347651   Root MSE        =

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1323038   .0723416     1.83   0.068    -.0097993    .2744068
          x2 |   .0924671   .5528632     0.17   0.867    -.9935411    1.178475
          x3 |   .1931725   .1295204     1.49   0.136
          x4 |  -9.86e-06   .0000205    -0.48   0.632    -.0000502    .0000305
          x5 |   1.397048   .4355554     3.21   0.001     .5414719    2.252625
          x6 |    .341846   .5025679     0.68   0.497    -.6453654    1.329057
          x7 |   2.550475   1.009268     2.53   0.012     .5679346    4.533015
          x8 |  -.0615556   .0892273    -0.69   0.491    -.2368278    .1137166
          x9 |   5.328812     1.2751     4.18   0.000      2.82409    7.833533
       _cons |  -8.853647   6.666171    -1.33   0.185    -21.94824    4.240941
------------------------------------------------------------------------------
```

Calculate the value of R^2.

a) 9.41%
b) 36.71%
c) 11.34%
d) 68.65%
e) 15.34%

71) Consider the following information. Calculate the quantity $\bar{y} / s_y$.

$\bar{x}$	s_x	$\bar{y}$	s_y	r	$\hat{y} = b_0 + b_1 x$
16	8	?	?	−0.6	$\hat{y} = 190 - 2x$

a) 7.48
b) 1.65
c) -1.32
d) 5.92
e) -0.59

72) Consider the following multiple regression output.

```
. regress y x1 x2 x3 x4 x5 x6 x7 x8 x9

      Source |       SS           df       MS      Number of obs   =       554
-------------+----------------------------------   F(9, 544)       =      7.73
       Model |                     9  789.855923   Prob > F        =    0.0000
    Residual |  55572.5479        544  102.155419   R-squared       =
-------------+----------------------------------   Adj R-squared   =    0.0987
       Total |  62681.2512        553  113.347651   Root MSE        =

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1323038   .0723416     1.83   0.068    -.0097993    .2744068
          x2 |   .0924671   .5528632     0.17   0.867    -.9935411    1.178475
          x3 |   .1931725   .1295204     1.49   0.136
          x4 |  -9.86e-06   .0000205    -0.48   0.632    -.0000502    .0000305
          x5 |   1.397048   .4355554     3.21   0.001     .5414719    2.252625
          x6 |    .341846   .5025679     0.68   0.497    -.6453654    1.329057
          x7 |   2.550475   1.009268     2.53   0.012     .5679346    4.533015
          x8 |  -.0615556   .0892273    -0.69   0.491    -.2368278    .1137166
          x9 |   5.328812     1.2751     4.18   0.000      2.82409    7.833533
       _cons |  -8.853647   6.666171    -1.33   0.185    -21.94824    4.240941
------------------------------------------------------------------------------
```

Which variable would be removed first when doing backward stepwise regression?

a) Variable 2
b) Variable 3
c) Variable 4
d) Variable 6
e) Variable 8

73) The following table is obtained from a random sample of 30 absences at a local Target store.

Day	Mon	Tue	Wed	Thur	Fri
Number Absent	2	9	6	7	6

You wish to test the claim that the absences occur on the five days with equal frequency. What is the value of the resulting Chi-square goodness of fit statistic?

a) 3.25
b) 4.33
c) 2.6
d) 6.5
e) 3.84

Answers:

1) B
2) A
3) A
4) B
5) B
6) B
7) B
8) B
9) B
10) A
11) A
12) A
13) B
14) A
15) A
16) B
17) D
18) D
19) A
20) E
21) C
22) D
23) B
24) A
25) E
26) D
27) D
28) E
29) E
30) D
31) D
32) B
33) D
34) B
35) C
36) A
37) D
38) D
39) C
40) B
41) A
42) B

43) D
44) C
45) D
46) C
47) A
48) C
49) A
50) A
51) B
52) A
53) B
54) C
55) C
56) C
57) B
58) A
59) C
60) B
61) E
62) D
63) B
64) D
65) C
66) D
67) A
68) A
69) B
70) C
71) D
72) A
73) B

Final Exam, Spring 2017

1) CoStar Realty would like to test if the vacancy rate for warehouse stores is more than 8%. The correct hypothesis statement to test for this vacancy rate would be Ho: p = 0.08; Ha: $p \neq 0.08$.
    c) True
    d) False

2) The coefficient of determination takes on values between -1 and + 1.
    a) True
    b) False

3) The regression model assumes the error terms are dependent.
    a) True
    b) False

4) The errors in a regression model are assumed to have zero variance.
    a) True
    b) False

5) If the assumptions of regression have been met, residuals plotted against the independent variable will typically show patterns.
    a) True
    b) False

6) A high correlation always implies that one variable is causing a change in the other variable.
    a) True
    b) False

7) If the hypothesis, H_o: $\beta_1 = 0$ is rejected, then the sample regression coefficient b_1, indicates the change in the predicted value for a unit change in X_1 when all other X_i variables are held constant.
    a) True
    b) False

8) Stepwise regression analysis is a method that assists in selecting the most significant variables for a multiple regression equation.
    a) True
    b) False

9) A 95% confidence interval for a proportion is (.482, .542). The sample proportion is .512
    a) True
    b) False

10) The Central Limit Theorem says that for large sample sizes the sample mean has an approximately normal distribution.
    a) True
    b) False

11) From the empirical rule we can deduce that, for any distribution, 95% of the observations fall between the mean plus or minus two standard deviations.
    a) True
    b) False

12) A small sample 95% confidence interval for the mean inappropriately using 1.96 instead of the appropriate t statistic will give a wider interval.
    a) True
    b) False

13) If A and B are any two events, then P(A and B) = P(A) + P(B).
    a) True
    b) False

14) The paired t-test tests the null hypothesis that the two population means are equal.
    a) True
    b) False

15) A study was conducted to investigate the relationship between sheep live weight (kg) and its chest girth (cm). A random sample of 66 sheep was weighed and simultaneously had their chest girth measured. Analysis of the data from this study could be performed using a two-sample t-test.
    a) True
    b) False

16) As the number of degrees of freedom increase, the t distribution gets closer and closer to the normal distribution
    a) True
    b) False

17) Customers leaving a subway station can exit through any one of three gates. Assuming that any particular customer is equally likely to select any one of the three gates, find the probability that among a sample of 4 customers, they all exit through the same gate.

a) 0.148
b) 0.037
c) 0.065
d) 0.83
e) 0.56

18) According to government data, 30% of married Americans marry after 30. A study of married people chooses a random sample of 400 married Americans and asks each person in the sample the age at marriage. What is the variance of the number of people in the sample that have married after 30?

a) 84
b) 4000
c) 130
d) 310
e) 120

19) From 1976 to 2002, a mechanical golfer, Iron Byron, whose swing was modeled after that of Byron Nelson (a leading golfer in the 1940s), was used to determine whether golf balls met the Overall Distance Standard. Specifically, Iron Byron would be used to hit the golf balls. If the average distance of 24 golf balls tested exceeded 296.8 yards, then that brand would be considered nonconforming. Under these rules, suppose a manufacturer produces a new golf ball that travels an expected distance of 297.5 yards with a standard deviation of 10 yards. What is the probability that the ball will be determined to be non-conforming when tested if we assume the distance travelled is normally distributed?

a) 0
b) 0.034
c) 0.79
d) 0.5
e) 0.63

20) The health of the bear population in Yellowstone National Park is monitored by periodic measurements taken from anesthetized bears. In a sample of 100 bears, the mean weight was found to be 185 lbs with a standard deviation of 125 lbs. We want to test the claim that the population mean weight of bears is equal to 210 lbs. What is the value of the test statistic?

a) 2.0
b) 0.03
c) -2.0
d) 0.05
e) 12.5

21) For the hypothesis test in the previous question state the final conclusion in simple non-technical terms.

a) There is not sufficient sample evidence to support the claim that the population mean weight of bears is equal to 210 lbs.
b) There is sufficient sample evidence to warrant rejection of the claim that the population mean weight of bears is equal to 210 lbs.
c) There is not sufficient evidence to warrant rejection of the claim that the population mean weight of bears is equal to 210 lbs.
d) The sample data support the claim that more than 80% of adults believe that the population mean weight of bears is equal to 210 lbs.

22) A data set has mean 250 and median 120. An analyst notices that a data point was incorrectly entered as 1250, when the correct value was 125. Once the correct value is entered, which one of the following is true?

a) The mean and median are unchanged.
b) The mean is unchanged but the median decreases.
c) The median is unchanged but the mean decreases.
d) Both the mean and median decrease.
e) None of the above.

23) Each person in a random sample of patrons of a local mall was surveyed regarding a public smoking area outside one of the mall entrances. Each person was asked if they approved of the idea of a public smoking area in the mall. The resulting data is summarized in the table below. The mall management would like to know if there is a relationship between gender and approval of the smoking area. What would be an appropriate set of hypotheses?

Public Smokers

	Approve	Do Not Approve
Males	28	57
Females	39	31

a) Ho: Gender and approval are not independent; Ha: Gender and approval are independent.
b) Ho: Gender and approval are independent; Ha: Gender and approval are not independent.
c) Ho: Knowing a person does not approve of a public smoking area indicates their gender; Ha: Knowing a person does not approve of a public smoking area does not indicate their gender.
d) Ho: There is no difference in gender distribution based on approval; Ha: There is a difference in gender distribution based on approval.
e) Ho: There is an association between gender and approval; Ha: There is no association between gender and approval

24) All but one of the following statements contains an error. Which statement could be correct?

a) There is a correlation of 0.54 between the position a football player plays and his weight.
b) We found a correlation of r = –0.63 between gender and political party preference.
c) The correlation between the gas mileage of a car and its weight is r = 0.71 mpg.
d) We found a high correlation (r = 1.09) between the height and age of children.
e) The correlation between planting rate and yield of tomatoes was found to be r = 0.23.

25) Let X and Y be two random variables that are not independent. Var(3X-4Y) is equal to:

a) 3Var(X) + 4Var(Y) - 8Cov(3X,-4Y)
b) 9Var(X) + 16Var(Y) - 24Cov(3X;,-4Y)
c) 9Var(X) + 16Var(Y) - 24Cov(X, Y)
d) 9Var(X) + 16Var(Y) + 24Cov(3X + 4Y)
e) 3Var(X) + 4Var(Y) - 24Cov(X; Y)

26) Daily sales records for a car dealership show that it will sell 0, 1, 2, or 3 cars, with probabilities as listed in the table below. Calculate $E(X^3)$.

Number of cars (X)	0	1	2	3
Probability (P(X))	0.5	0.3	0.15	0.05

a) 0.75
b) 5
c) 3
d) 2.85
e) Not enough information

27) Wages for workers in a particular industry average $11.90 per hour with a standard deviation of $0.40. If the wages are assumed to be normally distributed what percentage of workers receive less than $11?

a) 0.012
b) 0.123
c) 0.023
d) 0.019
e) 0.044

28) A large retailer wants to open new outlets in Malaysia and Thailand. The probability that the Thailand outlet is profitable is 0.8. However, if the Malaysian outlet is profitable then the probability that the Thailand outlet is profitable increases to 0.9. The probability that the Malaysian outlet is profitable is 0.75. Given the Malaysian outlet is not profitable what is the probability that the Thailand outlet is profitable?

a) 0.12
b) 0.44
c) 0.68
d) 0.32
e) None of the above

29) Suppose the joint probability table of random variables X and Y is given below. Find P(X-Y<0).

		Y	
		2	1
	1	1/8	2/8
X	2	2/8	1/8
	3	1/8	0
	4	0	1/8

a) 2/8
b) 1/8
c) 3/8
d) 4/8
e) None of the above

30) A data set x1, x2, . . . , xn has mean 4, median 10 and standard deviation equal to 7. A new data set y1, y2, . . . , yn is obtained by $y_i = 10 - x_i$. Which one of the following is true of the new data set?

a) The mean is less than the median.
b) The median is less than the mean.
c) The mean and median are equal.
d) The data is left skewed.
e) None of the above.

31) In many farming areas throughout the Midwest, it can be lucrative for landowners to have giant wind turbines for the production of electricity constructed on their land. The turbine will be profitable if the average wind speed at the proposed turbine site is more than 8 mph. For a site north of West Lafayette, the wind speed was monitored every 8 hours for a one year period. A total of 1095 readings had an average speed of 8.319 mph with a standard deviation of 3.909. An appropriate 1-sided hypothesis test was conducted to test whether the average wind speed at the site is more than 8 mph. Which of the below would be a Type I error for this test?

a) We believe the turbine will be profitable, but it will not be.
b) We believe the turbine will not be profitable, but it will be.

32) Which of the following is a correct interpretation of a 90% confidence interval?

a) 90% of the random samples you could select would result in intervals that contain the true population value.
b) 90% of the population values should be close to our sample results.
c) Once a specific sample has been selected, the probability that its resulting confidence interval contains the true population value is 90%.
d) All of the above statements are true.

33) A manufacturer of nails is interested in determining if their nails are meeting specifications. A sample of 100 nails is taken. Which technique would you use to see if the average length of nails is longer than 3 inches?

a) Test of one mean
b) Test of one proportion
c) Matched pairs test of means
d) Test of two independent means
e) Test of two independent proportions

34) The migration of African buffalo herds might be affected by the weight of the transponder used to track them. Last year, the scientists tried their standard transponder on eight herds of buffalo and recorded how far each group traveled. This year they will swap out the heavy transponders with more expensive, lighter ones and see if the same buffalo herds travel farther than last year. Which technique would you use to analyze the data?

a) Test of one mean
b) Test of one proportion
c) Matched pairs test of means
d) Test of two independent means
e) Test of two independent proportions

35) A political action group wonders if college graduates are more likely to support increased penalties for repeat offenders than are those without college degrees.

a) Test of one mean
b) Test of one proportion
c) Matched pairs test of means
d) Test of two independent means
e) Test of two independent proportions

36) Does eating breakfast improve productivity? In a Guess jeans factory, the workers of one sewing crew of 20 are fed a hearty breakfast for a month and then the same sewing crew of 20 is asked to go without eating until lunch time for another month. The teams' productivity is compared for the two months.

a) Test of one mean
b) Test of one proportion
c) Matched pairs test of means
d) Test of two independent means
e) Test of two independent proportions

37) The historical mean final grade in EC 10 is usually assumed to be 78. A random sample of 100 current EC 10 students had an average final grade of 82 with a standard deviation of 16. Based upon the score of the group of students, is there reason to believe that the historical mean is greater than 78? What are the null hypothesis and alternative hypothesis for this test?

a) $Ho: \mu = 81 \quad Ha: \mu \neq 81$
b) $Ho: \mu = 78 \quad Ha: \mu > 78$
c) $Ho: \mu = 78 \quad Ha: \mu \neq 78$
d) $Ho: \mu = 78 \quad Ha: \mu < 78$
e) $Ho: \mu = 81 \quad Ha: \mu > 81$

38) What is the calculated value of the test statistic for the previous question's hypothesis test?

a) 1.64
b) 1.19
c) 2.5
d) 0.13
e) 1.62

39) Based on the information above, what is the conclusion?

a) Fail to reject the null hypothesis, and we can conclude that there is insufficient evidence that the current group of students is performing better than historical levels
b) Fail to reject the null hypothesis, and we can conclude that the current group of students is performing better than historical levels.
c) Reject the null hypothesis, and we can conclude that the current group of students is performing better than historical levels.
d) Reject the null hypothesis, and we can conclude that there is insufficient evidence that the current group of students is performing better than historical levels.
e) The information given is insufficient to perform the hypothesis test.

40) Suppose we have a sample of the heights of Harvard students and want to use the sample mean to get a confidence interval for the mean height in the population. Which of the following would increase the width of this confidence interval?

a) Switching from a 95% confidence interval to a 90% confidence interval.
b) Increasing the sample size used to calculate the sample mean.
c) Switching from a 95% confidence interval to a 99% confidence interval.
d) All of the above.

41) Suppose we regress SAT score on years of mother's education and parent's income. If we run the regression again but also include the student's GPA as an additional explanatory variable:

a) The R^2 for the regression will either stay the same or increase.
b) The adjusted R^2 for the regression will either stay the same or increase.
c) Both (a) and (b) are true.
d) Neither (a) nor (b) is true.

42) At summer camp, one of Carla's counselors told her that you can determine air temperature from the number of cricket chirps. To determine a formula, Carla collected data on temperature and number of chirps per minute on 12 occasions. She entered the data into her calculator and did 2-Var Stats. Here are some results:
$\bar{x} = 166.8, s_x = 31, \bar{y} = 78.83, s_y = 9.11, r = 0.461$
Find the slope of the equation of the least-squares regression line.

a) 0.81
b) 0.26
c) 0.10
d) 0.14
e) 0.74

43) When regressing annual work hours on income, a researcher finds that the variance of the residuals increases as work hours increases. This will affect:

a) The expected value of the slope coefficient for income.
b) The magnitude of the standard error for the slope coefficient for income.
c) Both (a) and (b).
d) Neither (a) nor (b).

44) The American Red Cross wanted to know if people who reside in the northern states are less likely to donate to an emergency relief fund than people who reside in the south. A survey was conducted. One thousand people from the northern states and one thousand people from the southern states were asked if they made a donation to the Hurricane Katrina emergency relief fund. The number of people who reside in the northern states who donated to the Hurricane Katrina fund was 720. While the number of people from the southern states who donated to the Hurricane Katrina fund was 762. Can we conclude that the proportion of people from the north (p1) who donated to the Hurricane Katrina fund is less than the proportion of people from the south (p2)? We have the following Stata output:

```
prtesti 1000 720 1000 762,count

Two-sample test of proportions                     x: Number of obs =     1000
                                                   y: Number of obs =     1000
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |        .72   .0141986                      .6921713    .7478287
           y |       .762   .0134668                      .7356055    .7883945
-------------+----------------------------------------------------------------
        diff |      -.042   .0195693                     -.080355    -.003645
             |  under Ho:   .0195918    -2.14   0.032
------------------------------------------------------------------------------
        diff = prop(x) - prop(y)                                      z =  -2.1438
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0160         Pr(|Z| > |z|) = 0.0321          Pr(Z > z) = 0.9840
```

From the results of the Stata output, what conclusion can the Red Cross make?

a) We fail to reject the null hypothesis and conclude that there is insufficient evidence to suggest that people who reside in the northern states are less likely to donate than people who reside in the southern states.
b) We reject the null hypothesis and conclude that there is sufficient evidence to show that people who reside in the northern states are less likely to donate than those who reside in the southern states.
c) We fail to reject the null hypothesis; therefore, we can conclude that people who reside in the northern states are just as likely as or more likely to donate than people who reside in the southern states.
d) We can reject the null hypothesis because the p-value is larger than the significance level; therefore, there is no evidence that people of northern states are less likely to donate than people of southern states.
e) The results are inconclusive.

45) A statistics student designs an experiment to see whether today's high school students are becoming too calculator dependent. She prepares two quizzes, both of which contain 40 questions that are best done using paper-and-pencil methods. A random sample of 30 students participates in the experiment. Each student takes both quizzes—one with a calculator and one without. To analyze the data, the student constructs a scatterplot that displays the number of correct answers with and without a calculator for each of the 30 students. A least-squares regression yields the equation

$$\text{Calculator} = -1.2 + 0.865(\text{Pencil}) \qquad r = 0.79$$

Which of the following statements is/are true?

i. If the student had used Calculator as the explanatory variable, the correlation coefficient would remain the same.
ii. If the student had used Calculator as the explanatory variable, the slope of the least-squares line would remain the same.
iii. The standard deviation of the number of correct answers on the paper-and-pencil quizzes was larger than the standard deviation on the calculator quizzes.

a) I only
b) II only
c) III only
d) I and II only
e) I and III only

46) When running a simple linear regression, which of the following is not possible?

a) The error sum of squares is larger than the total sum of squares.
b) The error sum of squares is equal to the total sum of squares.
c) The error sum of squares is zero.
d) The error sum of squares is positive.

47) Suppose we run a regression with GPA as the dependent variable and SAT score as the independent variable. Which of the following statements is definitely true?

a) The sign of the estimated slope coefficient will be the same as the sign of the correlation between GPA and SAT score.
b) The sign of the estimated slope coefficient could be different than the sign of the correlation between GPA and SAT score if there are omitted variables.
c) The magnitude of the slope coefficient will be equal to the magnitude of the correlation between GPA and SAT score.
d) The slope coefficient will be statistically significant.

48) Suppose the R^2 for a bivariate regression is equal to 1. This tells us that:

a) The correlation between the dependent and independent variables is equal to 1.
b) The slope coefficient is equal to 1 or -1.
c) The error sum of squares is equal to the total sum of squares.
d) The dependent and independent variables are perfectly correlated.

49) Suppose your data produce the regression result $\hat{y} = 10 + 3x$. If both y and x are multiplied by 2.0, the new intercept and slope estimates will be

a) 20 and 1.5
b) 10 and 3
c) 20 and 6
d) 20 and 3

50) In a multiple regression model, if all the explanatory variables are not significantly individually but are significant as a group, this is most likely due to

a) Heteroskedasticity
b) The presence of dummy variables.
c) The absence of dummy variables.
d) Multicollinearity

51) A hypothesis test is done in which the alternative hypothesis is that more than 10% of a population is left-handed. The p-value for the test is calculated to be 0.25. Which statement is correct?

a) We can conclude that more than 10% of the population is left-handed.
b) We can conclude that more than 25% of the population is left-handed.
c) We can conclude that exactly 25% of the population is left-handed.
d) We cannot conclude that more than 10% of the population is left-handed.

52) Heights of college women have a distribution that can be approximated by a normal curve with a mean of 65 inches and a standard deviation equal to 3 inches. About what proportion of college women are between 66 and 67 inches tall?

a) 0.75
b) 0.50
c) 0.25
d) 0.17
e) None of the above

53) The probability is p = 0.50 that a patient with a certain disease will be successfully treated with a new medical treatment. Suppose that the treatment is used on 40 patients. What is the "expected value" of the number of patients who are successfully treated?

a) 40
b) 20
c) 8
d) 32

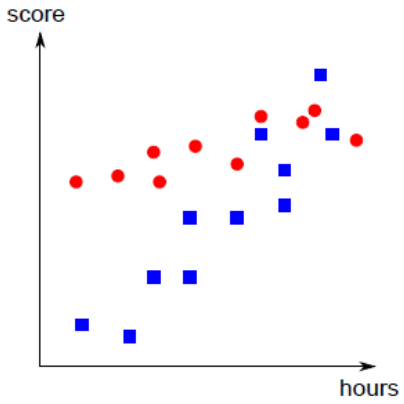
54) An economist is interested in studying the incomes of consumers in a particular region. The population standard deviation is known to be $1,000. What sample size would the economist need to use for a 95% confidence interval if the width of the interval should not be more than $100?

a) 20
b) 40
c) 385
d) 1537

55) The width of a confidence interval estimate for a proportion will be:

a) narrower for 99% confidence than for 95% confidence.
b) wider for a sample size of 100 than for a sample size of 50.
c) narrower for 90% confidence than for 95% confidence.
d) narrower when the sample proportion is 0.50 than when the sample proportion is 0.20.

The figure below is a scatter plot with hours of study on the horizontal axis and final exam score on the vertical axis for 21 students in an ECON 10 class.



The round data points correspond to economics majors. The square data points correspond to non-majors. Suppose we use this data to estimate the following model:

$$SCORE = \beta_0 + \beta_1 Major + \beta_2 Hours + \beta_3 Major * Hours + \varepsilon$$

Where SCORE is a student's final exam score, Major is a dummy variable equal to one if the student is an economics major and zero otherwise, Hours is the number of hours the student studies for the final and ε is a random error term that satisfies all of our assumptions.

56) Based on the scatterplot, we would expect our estimated value of β_1 to be:

a) Positive.
b) Negative.
c) Larger for economics majors than non-majors.
d) Larger for non-majors than economics majors.

57) Based on the scatterplot, we would expect our estimated value of β_3 to be:

a) Positive.
b) Negative.
c) Larger for economics majors than non-majors.
d) Larger for non-majors than economics majors.

58) The predicted score for an economics major who studies ten hours will be:

a) Greater than the predicted score for a non-major who studies ten hours.
b) Less than the predicted score for a non-major who studies ten hours.
c) Equal to the predicted score for a non-major who studies ten hours.
d) Not enough information.

59) The predicted increase in score for an economics major associated with one extra hour of studying will be (for your convenience here is the model again: $SCORE = \beta_0 + \beta_1 Major + \beta_2 Hours + \beta_3 Major * Hours + \varepsilon$).

a) Equal to b_2
b) Equal to $b_2 + b_3$
c) Equal to $b_2 + b_3 * Hours$
d) Equal to b_3

60) A sample size of 200 light bulbs was tested and found that 11 were defective. What is the 95% confidence interval around this sample proportion?

a) 0.055 plus or minus 0.032
b) 0.055 plus or minus 0.009
c) 0.055 plus or minus 0.044
d) 0.055 plus or minus 0.018

61) A university dean is interested in determining the proportion of students who receive some sort of financial aid. The dean randomly selects 200 students and finds that 118 of them are receiving financial aid. The 95% confidence interval for p is 0.59 ± 0.07. Interpret this interval.

a) We are 95% confident that the true proportion of all students receiving financial aid is between 0.52 and 0.66.
b) We are 95% confident that 59% of the students are on some sort of financial aid.
c) We are 95% confident that between 52% and 66% of the sampled students receive some sort of financial aid.
d) 95% of the students get between 52% and 66% of their tuition paid for by financial aid.

62) Suppose you have estimated wage = 5 + 3education + 2gender, where gender is one for female and zero for male. If gender had been one for male and zero for female, this result would have been

a) Unchanged
b) wage = 5 + 3education - 2gender
c) wage = 7 + 3education + 2gender
d) wage = 7 + 3education - 2gender
e) None of the above

63) In a multiple regression model, the error term ε is assumed to be a random variable with a mean of

a) 1
b) 0
c) -1
d) Any value

64) A corporation randomly selects 150 salespeople and finds that 99 salespeople would like to take a self-improvement course. The firm did a similar study 10 years ago in which 96 salespeople out of a random sample of 160 salespeople wanted a self-improvement course. The groups are assumed to be independent random samples. Let $p1$ and $p2$ represent the true proportion of workers who would like to attend a self-improvement course in the recent study and the past study, respectively. If the firm wanted to test if this proportion has changed from the previous study, which of the following is most correct?

```
. prtesti 150 99 160 96,count

Two-sample test of proportion                     x: Number of obs =      150
                                                  y: Number of obs =      160
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |         .66   .0386782                      .5841922    .7358078
           y |          .6   .0387298                      .5240909    .6759091
-------------+----------------------------------------------------------------
        diff |         .06   .0547357                     -.0472801    .1672801
             |  under Ho:    .0549009     1.09   0.274
------------------------------------------------------------------------------
       diff = prop(x) - prop(y)                                      z =  1.0929
   Ho: diff = 0

   Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
Pr(Z < z) = 0.8628         Pr(|Z| < |z|) = 0.2744          Pr(Z > z) = 0.1372
```

a) Reject the null hypothesis and conclude that the proportion of employees who are interested in a self-improvement course has increased over the intervening 10 years.
b) Fail to reject the null hypothesis, there is not enough evidence to conclude that the proportion of employees who are interested in a self-improvement course has changed over the intervening 10 years.
c) Reject the null hypothesis and conclude that the proportion of employees who are interested in a self-improvement course has changed over the intervening 10 years.
d) Fail to reject the null hypothesis, there is not enough evidence to conclude that the proportion of employees who are interested in a self-improvement course has not changed over the intervening 10 years.
e) None of the above

65) Consider three individual tests to determine whether the variablces X2, X3 and X4 are needed in the model. The conclusions from these tests may be summarized as:

a) All three are needed in the model.
b) X2 and X3 are needed, X4 is not needed
c) X2 and X4 are needed, X3 is not needed
d) X3 and X4 are needed, X2 is not needed
e) Either none or one of the three is needed.

66) The economic structure of Major League Baseball allows some teams to make substantially more money than others, which in turn allows some teams to spend much more on player salaries. These teams might therefore be expected to have better players and win more games on the field as a result. Suppose that after collecting data on team payroll (in millions of dollars) and season win total for 2015, we find a regression equation of Wins = 71.87 + 0.101Payroll - 0.060League where League is an indicator variable that equals 0 if the team plays in the National League or 1 if the team plays in the American League. One American League team in the data set had a payroll of $108 million and won 88 games. Calculate the residual for this observation.

a) –1.26
b) 5.28
c) 9.65
d) 11.70
e) 22.61

67) As an intern at a Delaware real estate company over the summer, you are asked to analyze house prices. You collect data on 45 house sales in one week in Delaware, Worthington, and Dublin, and obtain the following output

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	35773.52	18177.88	1.967969	0.056031
Bedrooms	10723.41	4085.702	2.624619	0.012225
House Size	56.26	8.098312	6.947359	2.22E-08
Worthington	-2407.31	7824.012	-0.30768	0.759921
Dublin	26714.37	9358.994	2.854406	0.0068

Which of the following statements **CAN** be inferred from this output?

a) a house in Dublin would costs on average $26,714.37 less than a house in Worthington, holding everything else constant
b) there is no significant difference in price between similar houses in Delaware and Worthington
c) having a swimming pool affects the sale price of a house in the Worthington area
d) holding everything else constant, you pay on average $10,723.41 more for a four-bedroom than a one-bedroom house
e) on average, houses are larger in Delaware than in Worthington

68) A 95% confidence interval for the difference between two population proportions is found to be (0.07, 0.19). Which of the following statements is (are) true?

I. It is unlikely that the two populations have the same proportions.
II. We are 95% confident that the true difference between the population proportions is between 0.07 and 0.19.
III. The probability is 0.95 that the true difference between the population proportions is between 0.07 and 0.19.

a) I only
b) II only
c) I and II only
d) I and II only
e) II and III only

69) In a study on teenage pregnancies, the researchers attempted to determine the relationship between y=weight of baby at birth (in pounds), x1=age of the mother and x2=1 if mother had prenatal care and 0 otherwise. The fitted regression equation is
wt= -1.84 + 0.53x1 + 1.79x2 - .003 x1x2 .

To predict weight of babies born to teenagers who receive prenatal care we use the equation:

a) wt= -0.05 + .527x1
b) wt= -1.84 + 0.53x1
c) wt= 1.79 + .003x1
d) wt= 1.79 + 0.53x1
e) None of the above

70) What would happen if instead of using an ANOVA to compare 10 groups, you performed multiple t-tests?

a) Nothing, there is no difference between using an ANOVA and using a t-test.
b) Nothing serious, except that making multiple comparisons with a t-test requires more computation than doing a single ANOVA.
c) Sir Ronald Fischer would be upset; he put all that work into developing ANOVA, and you use multiple t-tests
d) Making multiple comparisons with a t-test increases the probability of making a Type I error

71) One of the most common questions of prospective house buyers pertains to the cost of heating in dollars (*Y*). To provide its customers with information on that matter, a large real estate firm used the following 4 variables to predict heating costs: the daily minimum outside temperature in degrees of Fahrenheit (X_1) the amount of insulation in inches (X_2), the number of windows in the house (X_3), and the age of the furnace in years (X_4). Given below is the output of a regression model.

Regression Statistics						
R Square	0.8080					
Adjusted R Square	0.7568					
Observations	20					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	4	169503.4241	42375.86	15.7874	0.0000	
Residual	15	40262.3259	2684.155			
Total	19	209765.75				
	Coefficients	*Standard Error*	*t Stat*	*P*		*r 90.0%*
Intercept	421.4277	77.8614	5.4125			57.9227
X_1 (Temperature)	-4.5098	0.8129	-5.5476			-3.0847
X_2 (Insulation)	-14.9029	5.0508	-2.9505			-6.0485
X_3 (Windows)	0.2151	4.8675	0.0442			8.7484
X_4 (Furnace Age)	6.3780	4.1026	1.5546			13.5702

The estimated value of the parameter β_1 means that

a) holding the effect of the other independent variables constant, an estimated expected $1 increase in heating costs is associated with a decrease in the daily minimum outside temperature by 4.51 degrees.
b) holding the effect of the other independent variables constant, a 1 degree increase in the daily minimum outside temperature results in a decrease in heating costs by $4.51.
c) holding the effect of the other independent variables constant, a 1 degree increase in the daily minimum outside temperature results in an estimated decrease in mean heating costs by $4.51.
d) holding the effect of the other independent variables constant, a 1% increase in the daily minimum outside temperature results in an estimated decrease in mean heating costs by 4.51%.

72) Analysis of variance is a statistical method of comparing the _____ of several populations.

a) standard deviations
b) variances
c) means
d) proportions
e) none of the above

73) Consider the following multiple regression output.

```
. regress y x1 x2 x3 x4 x5 x6 x7 x8

      Source |       SS           df       MS      Number of obs   =     1,757
-------------+----------------------------------   F(8, 1748)      =     85.06
       Model |  28382.2069         8  3547.77586   Prob > F        =    0.0000
    Residual |                 1,748  41.7081053   R-squared       =
-------------+----------------------------------   Adj R-squared   =    0.2769
       Total |  101287.975     1,756  57.6810791   Root MSE        =

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.0620625   .0587997    -1.06   0.291    -.1773877    .0532627
          x2 |   .0862966   .0217482     3.97   0.000     .0436415    .1289518
          x3 |   .5305659   .4560234     1.16   0.245
          x4 |   .7515198   .3365928     2.23   0.026      .091353    1.411687
          x5 |   .9864597   .0864999    11.40   0.000     .8168055    1.156114
          x6 |   .4936952   .0639827     7.72   0.000     .3682046    .6191859
          x7 |   .2865146   .0413485     6.93   0.000     .2054169    .3676123
          x8 |   .0027432   .0212261                0.897    -.0388881    .0443745
       _cons |  -27.89086   3.153943    -8.84   0.000    -34.07676   -21.70496
------------------------------------------------------------------------------
```

Calculate the value of s_e.

f) 3.31
g) 10.11
h) 6.46
i) 10.41
j) 8.93

74) Using the output above, Calculate the value of R^2.

f) 9.41%
g) 36.71%
h) 11.34%
i) 68.65%
j) 28.01%

75) Consider the following multiple regression output.

```
. regress y x1 x2 x3 x4 x5 x6 x7 x8

      Source |       SS           df       MS      Number of obs   =     1,757
-------------+----------------------------------   F(8, 1748)      =     85.06
       Model |  28382.2069         8  3547.77586   Prob > F        =    0.0000
    Residual |              1,748  41.7081053   R-squared       =
-------------+----------------------------------   Adj R-squared   =    0.2769
       Total |  101287.975     1,756  57.6810791   Root MSE        =

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.0620625   .0587997    -1.06   0.291    -.1773877    .0532627
          x2 |   .0862966   .0217482     3.97   0.000     .0436415    .1289518
          x3 |   .5305659   .4560234     1.16   0.245
          x4 |   .7515198   .3365928     2.23   0.026      .091353    1.411687
          x5 |   .9864597   .0864999    11.40   0.000     .8168055    1.156114
          x6 |   .4936952   .0639827     7.72   0.000     .3682046    .6191859
          x7 |   .2865146   .0413485     6.93   0.000     .2054169    .3676123
          x8 |   .0027432   .0212261             0.897    -.0388881    .0443745
       _cons |  -27.89086   3.153943    -8.84   0.000    -34.07676   -21.70496
------------------------------------------------------------------------------
```

Which variable would be removed first when doing backward stepwise regression?

f) Variable 2
g) Variable 3
h) Variable 4
i) Variable 6
j) Variable 8

76) Using the output above, what is the value of the test statistic for testing the hypothesis $H_0 : \beta_8 = 0 \quad H_a : \beta_8 \neq 0$?

a) 6.93
b) 0.13
c) 0.05
d) 0.24
e) 1.16

Do certain car colors attract the attention of police more than others, so that they are more likely to get speeding tickets? A few years ago a curious newspaper columnist tabulated the car color on a random sample of 120 speeding citations at the local courthouse. Here are his results.

Color	Red	White/Silver	Gray/Black	Other
Number of speeding tickets	16	33	39	32

He then went to the state motor vehicle registry and obtained data on the distribution of car colors for all cars registered in his state:

Color	Red	White/Silver	Gray/Black	Other
Percentage of cars on highway	14%	35%	23%	28%

77) To answer the question posed above about car color and speeding tickets, the appropriate null hypothesis is:

a) The observed counts are all equal to 30.
b) The observed number of speeding tickets is the same for all four color groups.
c) At least one of the four car color percentages is different from the other three.
d) The distribution of car colors for the speeding citations is the same as the distribution of colors for cars on the highway.
e) The observed counts are equal to the expected counts.

Answers:

1) b
2) a
3) b
4) b
5) b
6) b
7) a
8) a
9) a
10) a
11) b
12) b
13) b
14) b
15) b
16) a
17) b
18) a
19) e
20) c
21) b
22) e
23) b
24) e
25) c
26) d
27) a
28) e
29) b
30) b
31) a
32) a
33) a
34) c
35) e
36) c
37) b
38) c
39) c
40) c
41) a
42) d
43) b
44) b
45) a
46) a
47) a
48) d
49) d

50) a
51) d
52) e
53) b
54) d
55) c
56) a
57) b
58) d
59) b
60) a
61) a
62) d
63) b
64) b
65) a
66) b
67) b
68) c
69) a
70) d
71) c
72) c
73) c
74) e
75) e
76) b
77) d