Material covered on test: Chapters 7-9, 12 in text.

WARNING! This sample exam may not cover all topics for which you are responsible on exam 2.

1. A copy machine dealer has data on the number \( x \) of copy machines at each of 89 customer locations and the number \( y \) of service calls in a month at each location. Summary calculations give \( \bar{x} = 8.4 \), \( s_x = 2.1 \), \( \bar{y} = 14.2 \), \( s_y = 3.8 \), and \( r = .86 \). What is the slope of the least squares regression line of number of service calls on number of copiers?

2. In the setting of the previous problem, about what percent of the variation in number of service calls is explained by the linear relation between number of service calls and number of machines?

3. Outdoor temperature influences natural gas consumption for the purpose of heating a house. The usual measure of the need for heating is heating degree days. The number of heating degree days for a particular day is the number of degrees the average temperature for that day is below 65°F, where the average temperature for a day is the mean of the high and low temperatures for that day. An average temperature of 20°F, for example, corresponds to 45 heating degree days. A homeowner interested in switching to solar heating panels collects the following data on her natural gas use for the months October through June, where \( x \) is heating degree days per day for the month and \( y \) is gas consumption per day in hundreds of cubic feet.

<table>
<thead>
<tr>
<th>Month</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>June</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>15.6</td>
<td>26.8</td>
<td>37.8</td>
<td>36.4</td>
<td>35.5</td>
<td>18.6</td>
<td>15.3</td>
<td>7.9</td>
<td>0</td>
</tr>
<tr>
<td>( y )</td>
<td>5.2</td>
<td>6.1</td>
<td>8.7</td>
<td>8.5</td>
<td>8.8</td>
<td>4.9</td>
<td>4.5</td>
<td>2.5</td>
<td>1.1</td>
</tr>
</tbody>
</table>

a. Calculate the correlation coefficient \( r \) and interpret its value; draw a scatterplot of the data.
b. Calculate the least squares regression line \( y = b_0 + b_1 x \) of gas consumption \( y \) on heating degree days \( x \). Draw the regression line on the scatterplot.

4. Each of the following statements contains a blunder. In each case explain what is wrong.
a. “There is a high correlation between the sex of American workers and their income.”
b. “We found a high correlation \( (r = 1.09) \) between students' ratings of faculty teaching and ratings made by other faculty members.”
c. “The correlation between planting rate and yield of corn was found to be \( r = .23 \) bushel.”

5. A study of 1,000 families gave the following results:
   - average height of husband \( = \bar{x} = 68 \) inches, \( s_x = 2.7 \) in.;
   - average height of wife \( = \bar{y} = 63 \) inches, \( s_y = 2.5 \) in.; \( r = .25 \).
Estimate the height of a wife when her husband is 72 inches tall.
a. 63 inches  b. 72 inches  c. 64 inches  d. none of these  e. need more information

The information below is needed for questions 6 and 7.

In finance, million dollar investments are made with the assistance of the Capital Asset Pricing Model (CAPM). The CAPM uses a least squares line to predict the annual rate of return \( (y) \) of a stock based on the rate of return for the overall stock market \( (x) \). The slope of the line is used to evaluate the risk of investing in the stock:

- \( \text{Slope} = 1 \): Average risk (neutral stock)
- \( \text{Slope} > 1 \): High risk (aggressive stock)
- \( \text{Slope} < 1 \): Low risk (conservative stock)
The data in the accompanying table are the annual rates of return for Disney stock (y) and the rate of return for the overall stock market (x) for an 8 year period (where rates of return are measured as a percent). The least squares line for the CAPM model is shown below the table.

<table>
<thead>
<tr>
<th>Year</th>
<th>Disney rate of return (y)</th>
<th>Overall market rate of return (x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>84</td>
<td>-10</td>
<td>-5</td>
</tr>
<tr>
<td>85</td>
<td>18</td>
<td>12</td>
</tr>
<tr>
<td>86</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>87</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>88</td>
<td>-12</td>
<td>0</td>
</tr>
<tr>
<td>89</td>
<td>2</td>
<td>-6</td>
</tr>
<tr>
<td>90</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Results: \( \hat{y} = -0.94 + 1.72x \), sum of squares of residuals = 7.996

6. Give an interpretation of the slope of the least squares line:
   a. For a year with an overall market rate of return of 1%, we estimate Disney's rate of return to be 1.72%.
   b. For every 1.72% increase in overall market rate of return, we estimate Disney's annual rate of return to increase by 1%.
   c. For every 1% increase in overall market rate of return, we estimate Disney's annual rate of return to increase by 1.72%.
   d. For a year with a market rate of return of 0%, we estimate Disney's rate of return to be -.94%.

7. Which of the following is the best interpretation of the sum of the squares of the residuals?
   a. No other line will produce a sum of squares of the residuals greater than 7.996.
   b. The least squares line obtained from these data should yield predictions of Disney's rate of return that are accurate to within \( \pm 2 \sqrt{7.996} \).
   c. The small value of 7.996 for the sum of squares of residuals indicates that the straight-line CAPM model is not useful for predicting Disney's rate of return (y).
   d. No other straight line fit to these data will produce a sum of squares of residuals smaller than 7.996.

Questions 8 and 9 refer to the following:
Data were collected to find the relationship between the labor (y in hours) required to produce lots of custom wood products and the size x of the lot. The following least squares regression equation was calculated from the data:

\[ \hat{y} = 13.7 + 1.7x. \]

8. What is the predicted hours of labor for a lot size of 55?

9. One of the original data points is (20, 50.3). What is the residual when the lot size is 20?

Questions 10, 11, and 12 refer to the following:
A marketing research firm wishes to determine if the adult men in Laramie, Wyoming would be interested in a new upscale men's clothing store. From a list of all residential addresses in Laramie, the firm selects a simple random sample of 100 and mails a brief questionnaire to each.

10. In the above survey, what is the sample?
   a. all adult men in Laramie, Wyoming
   b. all residential addresses in Laramie, Wyoming
   c. the members of the marketing firm that actually conducted the survey
   d. the adult men at the 100 addresses the survey was mailed to
11. In the above survey, what is the population of interest?
   a. all adult men in Laramie, Wyoming
   b. all residential addresses in Laramie, Wyoming
   c. the members of the marketing firm that actually conducted the survey
   d. the 100 addresses the survey was mailed to

12. In the above survey, what is the chance that all 100 homes in a particular neighborhood in Laramie end up being the sample of residential addresses selected?
   a. the same as for any other set of 100 residential addresses
   b. exactly 0 since random samples will spread out the addresses selected
   c. reasonably large due to the “cluster” effect
   d. 100 divided by the size of the population of Laramie

13. Which of the following is not a potential source of bias in a survey?
   a. undercoverage
   b. nonresponse
   c. the use of convenience sampling
   d. the use of simple random sampling

14. The January 2005 Gallup Youth Survey telephoned a random sample of 1,028 U.S. teens aged 13-17 and asked these teens to name their favorite movie from 2004. *Napoleon Dynamite* (Vote for Pedro! He has skills!) had the highest percentage with 8% of teens ranking it as their favorite movie. Which is true?
   I. The population of interest is U.S. teens aged 13-17.
   II. 85 is a statistic and not the actual percentage of all U.S. teens who would rank this movie as their favorite.
   III. This sampling design should provide a reasonably accurate estimate of the actual percentage of all U.S. teens who would rank this movie as their favorite.
   A) I only       B) II only       C) III only       D) I and II       E) I, II, and III

15. Suppose your local school district decides to randomly test high school students for attention deficit disorder (ADD). There are three high schools in the district, each with grades 9-12. The school board pools all of the students together and randomly samples 250 students. Is this sample a simple random sample?
   A) Yes, because the students were chosen at random.
   B) Yes, because each student is equally likely to be chosen.
   C) Yes, because they could have chosen any group of 250 students from throughout the school district.
   D) No, because we cannot guarantee that there are students from each school in the sample.
   E) No, because we cannot guarantee that there are students from each grade in the sample.

16. A chemistry professor who teaches a large lecture class one day surveys the students in attendance about how he can make the class more interesting, hoping he can get more students to attend lectures. This method suffers from
   A) voluntary response bias       B) nonresponse bias       C) response bias
   D) undercoverage       E) none of the above
17. Shown below is a scatterplot with the corresponding least squares line.

Choose the residual plot that corresponds to this scatterplot and least squares line.

a. I        b. II        c. III        d. IV       e. none

18. When every possible sample with the same number of observations is equally likely to be chosen, the selected sample is called a:
   a. simple random sample.
   b. stratified random sample
   c. cluster random sample
   d. systematic random sample

19. When the population is divided into mutually exclusive sets, and then a simple random sample is drawn from each set, this is called:
   a. simple random sampling.
   b. stratified random sampling.
   c. cluster random sampling.
   d. systematic random sampling.
20. A marketing research firm divides the population of a state into geographic areas, and randomly selects some of the areas and takes a simple random sample of each selected area. This is an example of a
   a. cluster random sample
   b. systematic random sample
   c. simple random sample
   d. stratified random sample

21. A simple random sample of 20 undergraduates at Johns Hopkins University found that 60% of those sampled felt that drinking was a problem among college students. A simple random sample of 20 undergraduates at Ohio State University found that 70% felt that drinking was a problem among college students. The number of undergraduates at Johns Hopkins University is approximately 2,000; the number at Ohio State is approximately 50,000. We conclude
   a. the sample from Johns Hopkins is much less representative of its population than the sample from Ohio State is of its population.
   b. the sample from Johns Hopkins is much more representative of its population than the sample from Ohio State is of its population.
   c. the samples from Johns Hopkins and Ohio State are equally representative of their respective populations.
   d. it is impossible to make any statements about which sample is more representative of its corresponding population since the students surveyed attended different schools.

22. A simple random sample of 2% of the undergraduates at Johns Hopkins University found that 60% of those sampled felt that drinking was a problem among college students. A simple random sample of 2% of the undergraduates at Ohio State University found that 70% felt that drinking was a problem among college students. The number of undergraduates at Johns Hopkins University is approximately 2,000; the number at Ohio State is approximately 50,000. We conclude
   a. the sample from Johns Hopkins is much less representative of its population than the sample from Ohio State is of its population.
   b. the sample from Johns Hopkins is much more representative of its population than the sample from Ohio State is of its population.
   c. the samples from Johns Hopkins and Ohio State are equally representative of their respective populations.
   d. it is impossible to make any statements about which sample is more representative of its corresponding population since the students surveyed attended different schools.

23. Consider the following scatterplot.

Which of the following is a plausible value for the correlation coefficient between weight and MPG?
   a. -0.9  b. -1.0  c. +0.2  d. +0.9  e. +0.7

24. If a least squares line was fit to the data shown in the above scatterplot, would the slope of the least squares be positive or negative?
   a. positive  b. negative
Solutions

1. \( b = r\left(\frac{s_y}{s_x}\right) = .86\left(\frac{3.8}{2.7}\right) = 1.56 \)

2. \( r^2 = (.86)^2 = .74 \)

3. a) \( r = .989 \). There is a strong positive linear relationship between heating degree days and gas consumption. b) \( s_x = 13.419; \ s_y = 2.74; \ \text{slope} \ b_1 = .202, \ \text{intercept} \ a = \bar{y} - b\bar{x} = 1.23 \)

4. a. The correlation we are studying measures the linear relationship between 2 quantitative variables; sex is a categorical variable.
   b. \(-1 \leq r \leq 1 \) is violated.
   c. \( r \) has no units.

5. c. method 1: \( b_1 = r\left(\frac{s_y}{s_x}\right) = .25\left(\frac{2.5}{2.7}\right) = .231; \ b_0 = \bar{y} - b_1\bar{x} = 63 - .231 \times 68 = 47.292; \)
   \[ \hat{y} = 47.292 + .231 \times 72 = 63.924 \approx 64. \]
   method 2: The husband is 4 inches, or \( 4/2.7 \) 1.5 standard deviations above the mean husband height. The wife's height is predicted to be above average by \( .25(1.5) = .4 \) standard deviations, or \( .4 \times 2.5 \) inches = 1 inch. (Recall \( b = r(s_y/s_x) \). So the wife's height is \( 63 + 1 = 64. \)

6. c. 7. d. 8. 107.2 9. observed \( y \) – predicted \( y = 2.6 \) hours. 10. d. 11. a. 12. a. 13. d. 14. E. 15. C. 16. D. 17. b. 18. a. 19. b. 20. a. 21. c. 22. a. 23. a. 24. b. negative (from the formula \( b_1 = r\frac{s_y}{s_x} \) for the slope of the least squares line, the correlation and the slope always have the same sign since the standard deviations \( s_y \) and \( s_x \) are positive).