Name _____ Period _____

AP Statistics Semester 1 Review
Fall 2013

PART 1: FREE RESPONSE PRACTICE QUESTIONS

In the paper "Reproduction in Laboratory colonies of Bank Vole," the authors presented the results of a study of litter size. (A vole is a small rodent with a stout body, blunt nose, and short ears.) As each new litter was born, the number of babies was recorded, and the accompanying results were obtained.

1	4	4	5	5	6	6	7	7	8
2	4	5	5	5	6	6	7	7	8
2	4	5	5	6	6	6	7	7	8
3	4	5	5	6	6	6	7	8	8
3	4	5	5	6	6	7	7	8	9
3	4	5	5	6	6	7	7	8	9
3	4	5	5	6	6	7	7	8	9
3	4	5	5	6	6	7	7	8	10
3	4	5	5	6	6	7	7	8	10
4	4	5	5	6	6	7	7	8	11

The authors also kept track of the color of the first born in each litter. (B = brown, G = gray, W = white, and T = tan)

B	B	T	W	T	G	G	G	B	B
W	B	W	B	T	T	G	B	T	B
B	T	B	B	B	G	W	B	B	G
G	G	G	B	B	T	B	W	T	T
B	T	B	B	T	W	W	B	G	B
B	B	B	G	T	B	B	T	T	G
G	B	B	B	B	G	W	G	T	G
B	B	B	B	G	G	T	T	W	G
G	W	T	G	T	B	B	G	B	B
B	G	T	W	B	G	T	W	G	W

1. Which variable, litter size or color, is categorical?

2. Which variable is quantitative?

3. Make a bar chart of the colors.

4. Make a histogram of the litter sizes.

5. Make a dotplot of the litter sizes.

6. Are there any outliers in the histogram or dotplot?

7. Describe the shape of the histogram (symmetric or skewed).

8. Find the mean of the litter sizes.

9. Is the mean resistant to outliers?

10. Find the median of the litter sizes.
11. Is the median resistant to outliers?

12. Find the range of the litter sizes.

13. Find the 5-number summary of the litter sizes.

14. What is the interquartile range?

15. Make a boxplot of the litter sizes.

16. Find the variance of the litter sizes.

17. Find the standard deviation of the litter sizes.

18. Is standard deviation resistant to outliers?

19. What is the area under a density curve?

20. The (mean or median) of a density curve is the equal-areas point, the point that divides the area under the curve in half.

21. The (mean or median) of a density curve is the balance point, at which the curve would balance if made of solid material.

22. If a density curve is skewed to the right, the (mean or median) will be further to the right than the (mean or median).

23. What is the difference between x-bar and μ?

24. What is the difference between s and σ?

25. How do you find the inflection points on a normal curve?

26. Sketch the graph of N(266, 16), the distribution of pregnancy length from conception to birth for humans.

27. What is the 68-95-99.7 rule?

28. Using the empirical rule (the 68-95-99.7 rule), find the length of the longest 16% of all pregnancies. Sketch and shade a normal curve for this situation.

29. Using the empirical rule, find the length of the middle 99.7% of all pregnancies. Sketch and shade.

30. Using the empirical rule, find the length of the shortest 2.5% of all pregnancies. Sketch and shade.

31. Using the empirical rule, what percentile rank is a pregnancy of 218 days?

32. What percentile rank is a pregnancy of 298 days?

33. What percentile is a pregnancy of 250 days?

34. What is the percentile of a pregnancy of 266 days?

35. What z-score does a pregnancy of 279 days have?

36. What percent of humans have a pregnancy lasting less than 279 days? Sketch and shade a normal curve.

37. What z-score does a pregnancy of 257 days have?

38. What percent of humans have a pregnancy lasting less than 257 days? Sketch and shade.

39. What percent of humans have a pregnancy lasting longer than 280 days? Sketch and shade.

40. What percent of humans have a pregnancy lasting between 260 and 270 days? Sketch and shade.

41. Would you say pregnancy length is a continuous or discrete variable? Justify.

42. How long would a pregnancy have to last to be in the longest 10% of all pregnancies?

43. How short would a pregnancy be to be in the shortest 25% of all pregnancies?

44. How long would a pregnancy be to be in the middle fifth of all pregnancies?

45. Does the vole information from the beginning of this review seem to be normal? Justify by checking actual percentages within 1, 2, and 3 standard deviations of the mean.

46. Make a back-to-back split stemplot of the following data:

Reading Scores

4th Graders		12	15	18	20	20	22	25	26	28	29
31	32	35	35	35	36	37	39	40	42		
7th Graders		1	12	15	18	18	20	23	23	24	25
27	28	30	30	31	33	33	33	35	36		

47. Make a comparison between 4th grade and 7th grade reading scores based on your stemplot.

48. What is the mode of each set of scores?

49. Is the score of "1" for one of the 7th graders an outlier? Test using the 1.5 IQR rule.

50. What is the difference between a modified boxplot and a regular boxplot? Why is a modified boxplot usually considered better?

51. What are the four principles that guide the examination of data? (SOCS)

52. Graph the following hot dog data:

Calories	Sodium (milligrams)
108	149
130	350
132	345
135	360
138	360
140	375
144	380
145	390
150	400
163	415
167	400
172	420
176	450
180	500
184	505
195	500
200	515

53. What is the response variable?

54. What is the explanatory variable?

55. What is the direction of this scatterplot? (positive, negative…)

56. What is the form of this scatterplot? (linear, exponential…)

57. What is the strength of this scatterplot? (strong, weak…)

58. Are their clusters? If so, where?

59. Are their outliers? (Outliers in a scatterplot have large residuals.)

60. If there are outliers, are they influential?

61. Calculate the correlation.

62. Calculate the correlation without the point (108, 149).

63. Using problems 61 and 62, check your answer to problem 60.

64. What two things does correlation tell us about a scatterplot?

65. If I change the units on sodium to grams instead of milligrams, what happens to the correlation?
66. What is the highest correlation possible?

67. What is the lowest correlation possible?

68. Correlation only applies to what type(s) of relationship(s)?

69. Is correlation resistant to outliers?

70. Does a high correlation indicate a strong cause-effect relationship?

71. Sketch a scatterplot with a correlation of about 0.8.

72. Sketch a scatterplot with a correlation of about –0.5.

73. Find the least-squares regression line (LSRL) for the calories-sodium data.

74. What is the slope of this line, and what does it tell you in this context?

75. What is the y-intercept of this line, and what does it tell you in this context?

76. Predict the amount of sodium in a hot dog with 155 calories.

77. Predict the amount of sodium in a hot dog with 345 calories.

78. Why is the prediction in problem 76 acceptable but the prediction in problem 77 not?

79. Find the error in prediction (residual) for a hot dog with 180 calories.

80. Find the residual for 195 calories.

81. The point (x-bar, y-bar) is always on the LSRL. Find this point, and label it on your scatterplot.

82. Find the standard deviation of the calories.

83. Find the standard deviation of the sodium.

84. Using the equations on your formula sheet for b1 and b0, verify the slope and intercept of the LSRL.

85. Find the coefficient of determination for this data.

86. What does r^2 tell you about this data?

87. How can you use a residual plot to tell if a line is a good model for data?

88. If you know a scatterplot has a curved shape, how can you decide whether to use a power model or an exponential model to fit data?

89. Graph the following data:

Time (days)	Mice
0	6
30	19
60	60
90	195
120	597

90. Perform the appropriate logarithmic transformation (power or exponential) on the above data to get an equation.

91. Make a residual plot to support your choice for problem 90.

92. What is extrapolation, and why shouldn't we trust predictions using extrapolation?

93. What is interpolation?

94. What is a lurking variable?

95. Why should we avoid using averaged data for regression and correlation?

96. What is causation? Give an example.

97. What is common response? Give an example.

98. What is confounding? Give an example.

99. Why is a two-way table called a two-way table?

Use this table for questions 100–107:

	Smoking Status		
Education	Never smoked	Smoked, but quit	Smokes
Did not complete high school	82	19	113
Completed high school	97	25	103
1 to 3 years of college	92	49	59
4 or more years of college	86	63	37

100. Fill in the marginal distributions for this table.

101. Display this table on a segmented bar chart.

102. What percent of these people smoke?

103. What percent of never-smokers completed high school?

104. What percent of those with 4 or more years of college have quit smoking?

105. What percent of those with some college smoke?

106. What percent of smokers did not finish high school?

107. What conclusion can be drawn about smoking and education from this table?

108. What is Simpson's Paradox?

109. What is the difference between an observational study and an experiment?

110. What is a voluntary response sample?

111. How are a population and a sample related but different?

112. Why is convenience sampling biased?

113. SRS stands for what kind of sample? Name and define.

114. Discuss how to choose a SRS of 4 towns from this list:

Allendale	Bangor	Chelsea	Detour	Edmonton	Fennville
Gratiot	Hillsdale	Ionia	Joliet	Kentwood	Ludington

115. What is a stratified random sample?

116. What is a multistage sample?

117. What is undercoverage?

118. What is nonresponse?

119. What is response bias?

120. Why is the wording of questions important? Give an example.

121. How are experimental units and subjects similar but different?

122. Explanatory variables in experiments are often called _____.

123. If I test a drug at 100 mg, 200 mg, and 300 mg, I am testing one variable at three _____.

124. What is the placebo effect?

125. What is the purpose of a control group?

126. Give an example of when we may not want to use a placebo/control group.

127. What are the two types of matched pairs used in experiments?

128. What are the three principles of experimental design?

129. What does double-blind mean, and why would we want an experiment to be double-blind?

130. Give an example of how lack of realism in an experiment can give false results.

131. What is block design?

132. I want to test the effects of aerobic exercise on resting heart rate. I want to test two different levels of exercise, 30 minutes 3 times per week and 30 minutes 5 times per week. I have a group of 20 people to test, 10 men and 10 women. I will take heart rates before and after the experiment. Draw a chart for this experimental design.

133. Why is simulation useful?

134. What are the five steps of a simulation?

135. Design and perform a simulation of how many children a couple must have to get two sons. (A simulation involves many trials. For this simulation, perform 10 trials.

136. What is independence?

137. You are going to flip a coin three times. What is the sample space for each flip?

138. You are going to flip a coin three times and note how many heads and tails you get. What is the sample space?

139. You are going to flip a coin three times and note what you get on each flip. What is the sample space?

140. Make a tree diagram for the three flips.

141. There are three ways I can drive from Fremont to Grand Rapids and four ways I can drive from Grand Rapids to my home. How many different ways can I drive from Fremont to my home through Grand Rapids?

142. How many different four-digit numbers can you make (digits: 0, 1, 2, …, 9)?

143. How many different four-digit numbers can you make without repeating digits?

144. What is an event in probability?

145. Any probability is a number between (and including) _____ and _____.

146. All possible outcomes together must have probability of _____.

147. If S is the sample space, P(S) = _____.

148. What are complements? Give an example and draw a Venn diagram.

149. What are disjoint events? Give two examples and draw a Venn diagram.

Use the following chart for questions 150-153:

M & M Color	Brown	Red	Yellow	Green	Orange	Blue
Probability	0.3	0.2	0.2	0.1	0.1	?

150. What is the probability that an M & M is blue?

151. What is the probability that an M & M is red or green?

152. What is the probability that an M & M is yellow and orange?

153. What is the probability that an M & M is not brown or blue?

154. Bre can beat Erica in tennis 9% of the time. Erica can swim faster than Bre 8% of the time. What is the probability that Bre would beat Erica in a tennis match and in a swimming race?

155. What assumption are you making in problem 154? Do you think this assumption is valid?

156. Using two dice, what is the probability that you would roll a sum of seven or eleven?

157. Using two dice, what is the probability that you would roll doubles?

158. Using two dice, what is the probability that you would roll a sum of 7 or 11 on the first roll and doubles on the second roll?

159. What assumption are you making in problem 158? Do you think this assumption is valid?

160. Using two dice, what is the probability that you would roll a sum of 7 or 11 that is also doubles?

161. What is the union of two events?

162. What is an intersection of two events?

163. How can we test independence?

164. Perform an independence test on the smoking/education chart from problem 110 to show that smoking status and education are not independent.

165. Make a Venn diagram for the following situation:

45% of kids like Barney
25% of kids like Blue
55% of kids like Pooh
15% of kids like Blue and Pooh
25% of kids like Barney and Pooh
5% of kids Barney, Blue, and Pooh
5% of kids like Blue but not Barney or Pooh

166. What is the law of large numbers?

167. Where are the mean and median located on a normal distribution.

For problems 168-171 consider the process of a drawing a card from a standard deck and replacing it. Let A be drawing a heart, B be drawing a king, and C be drawing a spade.

168. Are the events A and B disjoint? Explain.

169. Are the events A and B independent? Explain.

170. Are the events A and C disjoint? Explain.

171. Are the events A and C independent? Explain.

172. Give me an example of two events that are disjoint and independent.

173. What does the symbol $\cup$ mean?

174. What does the symbol $\cap$ mean?

PART 2: MULTIPLE CHOICE PRACTICE QUESTIONS

1. A five number summary of a distribution is given a 3.2, 3.9, 4.6, 5.4 and 6.9. The IQR is

   a. 4.60 b. 3.70 c. 1.50 d. 2.25

2. z corresponds to the area under which graph?

   a. normal curve b. density curve c. regression line d. standard normal curve

3. The data below are the number of days on which hail was observed at Evansville, Indiana in each of the years 1960 – 1970: 4, 5, 3, 2, 4, 2, 0, 5, 0, 1, 1. Consider this as a sample of size 11. Find the mean, median, and standard deviation.

   a. 2.45, 2, 1.86 b. 2.7, 2, 1.78 c. 2.5, 4, 1 d. 2, 2.45, 1.78

4. A mill makes castings containing a hole which is specified to have diameter 6 cm. Due to variability in the manufacturing process the actual diameter is a normal random variable with mean 6 cm and standard deviation .25 cm. What proportion of these castings have hold diameter within ½ cm of the specification of 6cm?

   a. 0.0228 b. 0.5498 c. 0.9544 d. 0.9772

5. What can be said about the distribution given in this stemplot?

   0 12258
   1 269999
   2 3347
   3 269
   4 0
   5 8

   a. It is skewed negative.
   b. It is skewed down.
   c. It is skewed positive.
   d. It is skewed up.

6. Scores (in a reference population) on the College Board SAT Verbal exam are normally distributed with mean 500 and standard deviation 100. What scores make up the top 25% of this distribution?

   a. 526 and above b. 575 and above c. 567 and above d. 601 and above

7. The scores on a statistics exam are strongly skewed to the left so it is best to describe the distribution by reporting

   a. the five-number summary b. the mean and standard deviation
   c. the mean, median and mode d. the correlation and its square

8. You record the age, marital status, and earned income of a sample of 1463 women. The number of variables you have recorded is

   a. 1463 b. 4 c. 3 d. 2

9. A standardized test designed to measure math anxiety has a mean of 100 and a standard deviation of 10 in the population of first year college students. Which of the following observations would you suspect is an outlier.

   a. 150 b. 100 c. 90 e. all of these

10. In a normal quantile plot

    a. if the points lie close to a straight line, the plot indicates the data are normal.
    b. outliers appear as points far away from the overall pattern of the plot.
    c. systematic deviations from a straight line indicate a nonnormal distribution.
    d. all of these.

11. A strong correlation implies a cause-and-effect relationship.

    a. true b. false

12. Month-by-month data for production of cottage cheese x (millions of pounds) and ice cream y (millions of gallons) for a recent year are collected and give the following regression line $\hat{y} = -29.5 + 1.4x$. If the cottage cheese production in one month is 78.2 million pounds, what is the predicted value of the ice cream production in millions of gallons?

    a. 77 b. 80 c. 85 d. 94

13. In the analysis of sugar solutions with a spectrophotometer, the intensity of the color produced depends on the amount of phenol solution added to each 2 mL specimen of sugar solution. From data collected for varying amounts of phenol (x in microliters μL) and the corresponding absorbance (y on a 0 to 1 scale), the following regression equation was obtained $\hat{y} = 0.58 - 0.0039x$. For one observation the amount of phenol was 24 μL and the absorbance was 0.47. Find the residual for this observation.

    a. -28.2 b. 0.578 c. -0.016 d. 0.195

14. A LSRL is fitted to the weights (kg) of a group of children between the ages of 12 and 36 months. The equation of the line is $\hat{y} = 6.9 + 0.27t$ where y is the weight and t is the age in months. What is the predicted value of the weight for a child from this group who is 20 months of age?

    a. 6.9 kg b. 12.3 kg c. 27 kg d. 48.5 kg

15. Use the same regression line from #14 to predict the weight of a child who is 50 months of age.

    a. 141.9 kg b. 159.6 kg c. 200 kg d. the prediction is not reliable because it involves extrapolation

16. A scatterplot of the month-by-month data for production of cottage cheese (x) and ice cream (y) for a recent year shows a somewhat linear pattern. Computer output contains the following information:

VARIABLE	MEAN	STD. DEV.
X	76.1917	5.08267
Y	64.3917	9.98621

    CORRELATION X, Y = 0.60737

    The equation of the LSRL is

    a. $\hat{y} = 11.8 - 4.9x$ b. $\hat{y} = 11.8 + .61x$ c. $\hat{y} = 26.53 - .51x$ d. $\hat{y} = -26.53 + 1.2x$

17. A random sample of size 10 was taken from a population. The sample has a variance of zero. Which of the following statements must be true?

I. The population also has a variance of zero.
II. The sample mean is equal to the sample median.
III. The ten data points in the sample are equal in numerical value.

a. I and II b. I and III c. I, II and III d. II and III

18. The following is an example of Simpson's Paradox: At one point during WWII, the percentage of women among US industrial workers was rising in every individual industry, but falling overall. Total employment was rising most rapidly in heavy industry, where few women were employed, so male employment rose faster even though women were gaining in every industry.

a. true b. false

19. A study found correlation r = 0.61 between the gender of a worker and his/her income. You can conclude

a. women earn more than men on average. b. women earn less than men on average.
c. an arithmetic mistake was made. d. r makes no sense here.

20. In a statistics course a linear regression equation was compute to predict the final exam score from the score on the first test. The equation was $\hat{y} = 10 + 0.9x$ where y-hat is the final exam score and x is the score on the first exam. Eve scored 95 on the first test. What is the predicted value of her final exam score?

a. 95 b. 85.5 c. 90 d. 95.5

21. The number of acres of tomato plants infested by aphids on a particular farm during a growing season is given below. Take the natural logarithm of each y and plot the logarithms against the month. Find the LSRL in the form of ln y = a + bx. Plot this line. What can you conclude?

a. The original distribution is normal. b. The original distribution is linear.
c. The original distribution is exponential. d. The original distribution is logarithmic.

22. Which of the following would lease likely have an adverse effect on a sample survey?

a. extrapolation b. nonresponse c. response bias d. undercoverage

23. Identify the statistic and parameter in the following situation: A telemarketing firm in New York uses a device that dials residential telephone numbers in that city at random. Of the first 100 numbers dialed, 21 are unlisted. This is not surprising because 35% of all New York residential phones are unlisted.

a. stat = 100, par = 21 b. stat = 21, par = 35% c. stat = 35%, par = 21 d. 100, 21, 35% are all stats

24. The presence of flu virus is tested by inoculating an egg and seeing if the virus multiplies. Throat swabbings from PA American Legion members who died of a mysteriously after attending their convention were tested this way – no flu virus was found. A virologist pointed out that it is better to also inoculate some eggs with known flu virus to check that it does grow. Given 100 eggs, 50 will be assigned to be inoculated with the known flu virus and the remaining 50 will be in a group that will be inoculated with the throat swabs. Using the random digit table starting at line 109, give the numbers of the first 10 eggs that are assigned to the virus group.

a. 03, 06, 00, 09, 01, 05, 15, 41, 23, 38 b. 36, 00, 91, 93, 65, 15, 41, 23, 96, 38
c. 36, 09, 19, 51, 54, 12, 39, 63, 88, 53 d. 3, 6, 0, 9, 1, 5, 4, 2, 8, 7

25. Students have the option of choosing regular instruction or a self-paced version of STAT301. To compare the effectiveness of self-paced and regular instruction, someone proposes administering the same exam to all STAT 301 students and comparing the average scores of those who took the self-paced option with those in regular sections.

    a. This is a valid experiment.
    b. The experiment is confounded by the fact that an odd number of students (301) will take the exam.
    c. This experiment is worthless because there may be systematic differences between groups due to lack of randomness.
    d. If a double-blind experiment is imposed, the results will be valid.

26. A student wishes to study the opinions of the agriculture faculty at a large university on the advisability of setting up a State Department of Agriculture. There are 380 agriculture faculty members. The population is

    a. all faculty members.
    b. only those faculty chosen for the survey.
    c. the 380 agriculture faculty members.
    d. all people at the university.

27. A new headache remedy is given to a group of 25 patients who suffer severe headaches. Of these, 20 report that the remedy is very helpful in treating their headaches. From this information, you can conclude

    a. the remedy is effective for the treatment of headaches.
    b. nothing, because the sample size is too small.
    c. nothing, because there is no control group for comparison.
    d. the new treatment is better than aspirin.

28. A student organization wants to assess the attitudes of students toward a proposed change in the hours that the library is open. They randomly select 50 freshmen, 50 sophomore, 50 juniors and 50 seniors. This is a

    a. stratified random sample b. SRS c. comparative experiment d. none of these

29. You must choose an SRS of five from a lot of 64 vials of a pharmaceutical product for assay. You label the vials beginning with 01 and use the random digits table beginning at line 133 to choose your sample. Your sample contains the vials labeled

    a. 45, 74, 04, 18, 07 b. 45, 04, 18, 01, 65 c. 45, 40, 41, 55, 61 d. 45, 04, 18, 07, 56

30. Which of the following is not a principle of experimental design?

    a. control b. blocking c. randomization d. replication

31. The following table gives the probability distribution of causes of death in the US. What is the probability that a randomly chosen death was not due to cardiovascular disease?

Cause	Cardiovascular Disease	Malignancies	All other causes
Probability	0.47	0.22	0.31

    a. 0.47 b. 0.22 c. 0.31 d. 0.53

32. Here is the distribution of medical specialties among physicians in the US. What must be the probability that a randomly chosen physician is a psychiatrist (in order for this to be a legitimate distribution)?

Specialty	Probability
General Practice	0.12
Internal Medicine	0.16
Pediatrics	0.06
General Surgery	0.07
OB/Gyn	0.06
Psychiatry	
Anesthesiology	0.04
Ophthalmology	0.03
All others	0.40

a. 0.12
b. 0.06
c. 0.04
d. none of these

33. A study of health care recorded the number of visits X to a physician made by a person in a year. Below is the probability distribution of the random variable X. Is the distribution discrete or continuous? Is it legitimate?

Visits	0	1	2	3	4	5	6	7	8	9	10	11	12
Probability	.10	.08	.06	.08	.12	.13	.12	.10	.09	.06	.03	.02	.01

a. discrete - legitimate b. continuous – legitimate c. discrete - illegitimate d. continuous - illegitimate

34. Statistics is the

a. study of descriptive and inferential operations.
b. mathematics of randomness and variability
c. science of collecting, organizing and interpreting data.
d. the royal road to learning.

35. A randomly selected student is asked to respond yes, no or maybe to the question "Do you intend to vote in the next presidential election?" The sample space is (yes, no, maybe). Which of these following represent a legitimate assignment of probabilities for the sample space?

a. 0.4, 0.4, 0.2 b. 0.4, 0.6, 0.4 c. 0.3, 0.3, 0.3 d. 0.5, 0.3, -0.2

36. Which of the following states that the actually observed mean outcome in independent trials must approach the mean of the distribution outcomes?

a. Rule for Means b. Law of Averages c. Rule for Variances d. Law of Large Numbers

37. Which of the following is not correct?

a. If A and B are independent, then P(A or B) = P(A) + P(B).
b. If B is the complement of A, then P(A) = 1 – P(B).
c. If A and B are independent, then P(A and B) = P(A)P(B).
d. P(A) = (count of outcomes in A)/(count of outcomes in S) where S is the collection of all possible outcomes of which A is an event.