AP Statistics First Semester Final Exam Review

1. Which one of the following activities is **not** an example of data gathering?
   A. A telephone survey
   B. Counting the number of flowers produced by a plant nourished by an experimental fertilizer
   C. Having people fill out a questionnaire
   D. Reaching a conclusion about the results of a reading program
   E. Giving school-aged children eye exams as part of a study to see how many children need glasses

2. Which of these are categorical data?
   A. The birth weights of anteaters
   B. The lengths of anteaters
   C. The different types of anteaters
   D. The top speeds of anteaters
   E. The prices of anteaters

<u>Questions 3 – 5</u> refer to the following survey:
You want to know something about your neighbors, so you give them a survey. The survey collects the following data about each family on your block: family size, the kind of pets they have, the grade of the youngest child in the family, the family's average annual income over the last five years, what the dad does for a living, whether the mom works, and their phone number. Each kind of data you collect about a family is a variable.

3. Which of the variables you collect are continuous data?
   A. Average annual income and phone number
   B. Only family size
   C. Grade of youngest child and average annual income
   D. Only average annual income
   E. Average annual income and family size

4. What are the categorical variables in your survey?
   A. Family size, dad's occupation, whether mom works, and kind of pets
   B. Dad's occupation, whether mom works, and kind of pets
   C. Grade of youngest child, dad's occupation, whether mom works, and kind of pets
   D. Everything but average annual income and family size
   E. Everything but average annual income

5. True or False: In the survey of your neighbors described above, the only discrete quantitative data you're collecting about your neighbors is family size.

6. Which of the following would most likely be graphed as a bar chart rather than a histogram?
   A. The number of blue, red, black, and white cars in a random sample of 500 cars
   B. The number of students in a mid-size university who own Macintosh or PC computers
   C. The ethnic distribution for a major city
   D. The number of people in various management positions at a large electronics store
   E. All of the above

7. Consider a complete table of relative frequencies. The sum of the relative frequency column in such a table must be:
   A. 0
   B. .5
   C. 1
   D. 2
   E. It depends on the data.

8. Histograms are most useful in displaying:
   A. Categorical data
   B. Cumulative frequencies
   C. Large numeric data sets
   D. Data trends over time
   E. Mean values

<u>Questions 9 and 10</u> refer to the data below:
Here are some data on the average hours spent studying per day by a sample group of college students.

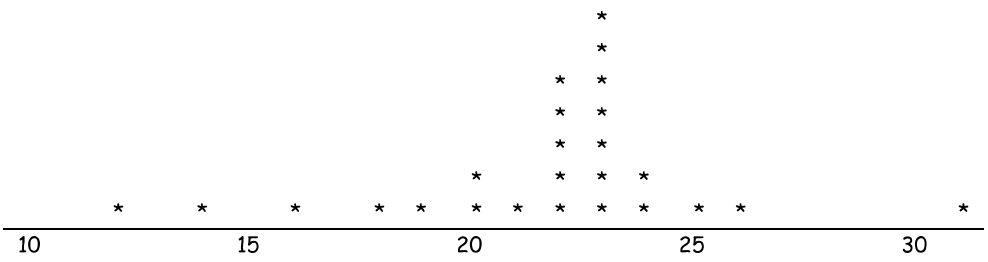
Average Hours of Study	Frequency
0 – 2	45
3 – 4	35
5 – 6	29
7 – 10	21

9. The relative frequency of students who study 5 – 6 hours a day is:
   A. 0.133
   B. 0.223
   C. 0.300
   D. 0.323
   E. 0.600

10. The cumulative relative frequency of students who study 4 hours or fewer a day is:
   A. 0.385
   B. 0.500
   C. 0.615
   D. 0.690
   E. 0.810

11. Estimate the relative frequency of countries with 20 – 22 days of no rainfall in the month of January.

Average Days with No Rainfall
(Sample of 25 European Countries)



A. 0.08
B. 0.25
C. 0.32
D. 0.52
E. 0.75

<u>Questions 12 – 14</u> refer to this plot:

Home Runs per team by American League vs. National League at midseason, 1999

A.L.		N.L.
7	5	
	6	
9	7	17
7554	8	5678
2	9	01259
3321	10	13
83	11	126
	12	
	13	
9	14	

where 5|7 represents 57 homeruns

12. Which league has the team that had hit the most home runs by midseason, and how many home runs had that team hit? Which league had the team with the fewest home runs?
A. American, 149, National
B. National, 116, American
C. National, 149, American
D. American, 118, American
E. American, 149, American

13. True or False: The number of home runs made by most National League teams were in the 90s; the number of home runs made by most teams overall were in the 80s.

14. How many fewer home runs did the team with the fewest home runs in the American League have than the team with the fewest home runs in the National League? How many more home runs did the team with the highest home runs in the American League have than the highest team in the National League?
    A. 2, 798
    B. 4, 25
    C. 14, 33
    D. 4, 33
    E. 14, 35

15. A symmetric distribution can't have which of the following characteristics?
    A. Gaps
    B. Outliers
    C. A long tail on one side
    D. A long tail on both sides
    E. More than one mode

16. True or False: If the population of interest is all day care centers in the United States, a sample of day care centers could be either all day care centers in New York City or a randomly selected group of day care centers throughout the US. Either sample is equally good.

17. The kind of sampling strategy **least** likely to produce statistics that are good estimates of population parameters is a:
    A. random sample
    B. simple random sample
    C. haphazard sample
    D. population sample
    E. representative sample

18. Consider these eight observations: {11,6,2,5,8,4,4,9}. What is the mean?
    A. 8.167
    B. 5.5
    C. 4
    D. 6.875
    E. 6.125

19. Consider these eight observations: {11,6,2,5,8,4,4,9}. What is the median?
    A. 6.5
    B. 6.125
    C. 5.5
    D. 5
    E. 6

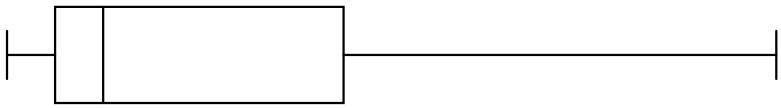
20. Since the distribution of housing prices in a community is usually skewed right, which measure of center should you use for housing prices?
    A. Mean
    B. Median
    C. Mode
    D. Outlier
    E. None of the above

21. Six radio listeners are surveyed. Their favorite FM stations are: 89.1, 89.1 89.1, 94.7, 94.7 and 104.3. Based on these data, you want to name the favorite station of a typical listener. You should name:
    A. The mean, which is 93.5
    B. The median, which is 91.9
    C. The mode, which is 89.1
    D. The mean, the median, or the mode
    E. None of the above

22. Calculate, to the nearest whole number, the sample standard deviation of this data set, which is a sample from a larger population: {71, 75, 65, 73, 69, 77, and 67}.
    A. 2
    B. 4
    C. 6
    D. 8
    E. 10

23. All the following statements about the sample standard deviation are true, **except**:
    A. the standard deviation is negative when there are extreme values in the sample.
    B. the denominator in the formula is (n-1).
    C. It's important to know the degrees of freedom in the computation.
    D. It's based on squared deviations between the observations and the mean.
    E. All the statements are true.

24. The box plot below represents a distribution that is:



    A. right skewed
    B. left skewed
    C. symmetrical
    D. uniformly distributed
    E. None of the above

Questions 25 and 26 refer to the data sets below:
Data sets of the average points scored per game by basketball players on three different teams in the 1998-99 season:

BOSTON		CHICAGO		SEATTLE	
Walker	18.7	Kukoc	18.8	Payton	21.7
Mercer	17.0	Harper	11.2	Schrempf	15.0
Pierce	16.5	Barry	11.1	Baker	13.8
Anderson	12.1	Simpkins	9.1	Hawkins	10.3
Potapenko	10.0	Bryant	9.0	Grant	8.9
Barros	9.3	Brown	8.8	Polynice	7.7
Battie	6.7	David	6.2	Crotty	5.9
McCarty	5.7	LaRue	4.7	Stepania	5.5
Jones	5.2	Carr	4.1	McCoy	5.1
Minor	4.9	Lang	3.8	Williams	4.0

25. The median points per player per game for each team are:
    A. Boston = 9.65; Chicago = 8.9; Seattle = 8.3.
    B. Boston = 10.0; Chicago = 9.0; Seattle = 8.9.
    C. Boston = 9.3; Chicago = 8.8; Seattle = 7.7.
    D. Boston = 14.6; Chicago = 20.1; Seattle = 12.4.
    E. Boston = 10.0; Chicago = 10.0; Seattle = 9.0.

26. Create a modified box-and-whisker plot for each of the three data sets. Then use the plot and the values of the five-number summary to answer the following question: Which team has the greatest variation in the points per game for the middle 50% of their observations?
    A. Boston
    B. Chicago
    C. Seattle
    D. The variation is roughly equivalent for the three teams.
    E. You can't answer this question, given the information in the five-number summary.

27. The following hypothetical data set shows the purchase price (in thousands) for a sample of 3-bedroom, 2-bathroom homes in Essex County, MA over the past year. Compute the five-number summary and create a modified box plot. How many outliers are present in this distribution?
    250, 254, 320, 342, 221, 235, 210, 426, 210, 298, 231, 254, 278, 234, 236, 235, 300, 401, 129, 234, 235, 235, 245

    A. 0
    B. 1
    C. 2
    D. 3
    E. 4

Questions 28 – 32 refer to the data in the table below:

Babies born in U.S. in 1984, in thousands

January	314
February	289
March	291
April	302
May	296
June	297
July	336
August	323
September	329
October	316
November	292
December	311

28. What is the median for the data in the table?
    A. 308
    B. 306.5
    C. 316.5
    D. 302
    E. 294

29. What is the upper quartile for the data in the table?
    A. 322.5
    B. 294
    C. 336
    D. 316
    E. 319.5

30. What is the lower quartile for the data in the table?
    A. 289
    B. 296
    C. 296.5
    D. 294
    E. None of the above

31. What is the interquartile range for the data in the table?
    A. 12.5
    B. 5
    C. 25.5
    D. 47
    E. -25.5

32. Based on the data in the table, what is the largest number of births in a month you could possibly have and still have a lower outlier?
    A. 268.25
    B. 332.25
    C. 268.5
    D. 255
    E. None of the above

33. Why is the IQR considered to be a resistant statistic?
    A. Adding a new extreme observation has little effect on it.
    B. It's equivalent to the distance from Q1 to the median, a resistant measure.
    C. Adding an outlier changes the IQR dramatically.
    D. It uses the mean in its calculation.
    E. None of the above

34. Which measure of central tendency and measure of variation should be used with a normally distributed distribution?
    A. The mean and standard deviation
    B. The mean and interquartile range
    C. The median and interquartile range
    D. The median and standard deviation
    E. The mode and interquartile range

35. Which of the following can stemplots show?

I.	Symmetry	II.	Gaps
III.	Clusters	IV.	Outliers

    A. I, II, and III
    B. I, II, and IV
    C. I, III, and IV
    D. II, III, and IV
    E. I, II, III, and IV

36. Which of the following can outliers affect significantly?
    I. Mean
    II. Median
    III. Standard deviation
    IV. Range
    V. Interquartile range

    A. I, III, and V
    B. II and IV
    C. I and V
    D. I, III, and IV
    E. None of these choices give the correct combination of answers

Questions 37 and 38 – Use the Empirical Rule:
A population of bolts has a mean thickness of 20 millimeters, with a population standard deviation of 0.01 millimeters.

37. Give, in millimeters, a minimum and maximum thickness that includes 68% of the population of bolts.
    A. 20.00 to 20.02 millimeters
    B. 19.00 to 21.00 millimeters
    C. 19.98 to 20.02 millimeters
    D. 19.99 to 20.01 millimeters
    E. 19.97 to 20.03 millimeters

38. Give, in millimeters, a minimum and a maximum thickness that will include 95% of the population of bolts.
    A. 19.98 to 20.02 millimeters
    B. 19.99 to 20.01 millimeters
    C. 19.97 to 20.03 millimeters
    D. 19.8 to 20.2 millimeters
    E. These can't be accurately computed.

39. Assume that normal curve A and normal curve B have identical population means. Assume further that A has a greater population standard deviation than B. Which curve is taller, and why?
    A. Curve A is taller because it has fewer inflection points.
    B. Curve A is taller because smaller standard deviations produce wider curves.
    C. Curve B is taller because its median is greater.
    D. Curve B is taller because smaller standard deviations produce thinner curves.
    E. The curves are the same height

40. Using the empirical rule, you can assume that what percent of the normal distribution is outside two standard deviations of the mean in both directions?
    A. 0.3%
    B. 50%
    C. 5%
    D. 95%
    E. Can't be calculated

Questions 41 – 43:
You measured the weights of members of population W and found the weights to be normally distributed. The distribution has a population mean weight of 160 pounds and a population standard deviation of 25 pounds.

41. For population W, find the z-score associated with a weight of 120 pounds.
    A. $z = -2.6$
    B. $z = -1.0$
    C. $z = -1.6$
    D. $z = 1.0$
    E. $z = 1.6$

42. For population W, what is the percentile for the weight 160 pounds?
   A. 50th percentile
   B. 10th percentile
   C. 30th percentile
   D. 75th percentile
   E. 90th percentile

43. In population W, what is the probability that a randomly selected subject will weigh between 140 and 180 pounds?
   A. 0.202
   B. 0.5
   C. 0.677
   D. 0.950
   E. 0.576

<u>Questions 44 and 45</u>:
Population H is a group of women with normally distributed heights. Population H has a population mean of 66 inches and a population standard deviation of 2.5 inches.

44. In population H, what is the height, to the nearest tenth of an inch, of the 70th percentile?
   A. 67.3 inches
   B. 67.0 inches
   C. 64.7 inches
   D. 66.0 inches
   E. 63.0 inches

45. In population H, what is the z-score, to the nearest tenth, associated with the height 65 inches?
   A. $z = -1.0$
   B. $z = -0.5$
   C. $z = 0.4$
   D. $z = -0.4$
   E. $z = 1.4$

46. To the nearest whole number, what percentile is associated with $z = -0.68$?
   A. 10th percentile
   B. 40th percentile
   C. 50th percentile
   D. 25th percentile
   E. 75th percentile

47. To the nearest whole number, what percentile is associated with $z = 1.2$?
   A. 25th percentile
   B. 50th percentile
   C. 75th percentile
   D. 88th percentile
   E. 12th percentile

48. What area, to the nearest whole percent, of the normal curve is located between z = -0.6 and z = 1.4?
    A. 64%
    B. 91%
    C. 27%
    D. 50%
    E. 95%

<u>Questions 49 – 53</u> refer to the following information:
Applicants to a college psychology department have normally distributed GRE scores with a mean of 544 and a standard deviation of 103. Round all answers to the nearest whole number.

49. What percentage of applicants scored between 500 and 700?
    A. 50%
    B. 60%
    C. 70%
    D. 80%
    E. 90%

50. What percentage of applicants scored above 450 on the GRE?
    A. 82%
    B. –82%
    C. 26%
    D. 11%
    E. 1%

51. What percentage of applicants had a GRE score below 625?
    A. 78%
    B. –24%
    C. 24%
    D. 22%
    E. 5%

52. What is the GRE score at the 77th percentile?
    A. 468
    B. 505
    C. 620
    D. 1
    E. -1

53. Find the GRE score at the upper quartile, Q3.
    A. 1
    B. 475
    C. 470
    D. 0
    E. 613

54. Find the z-score for the lower quartile of any normal curve. Round your answer to the nearest hundredth.
   A. 0.67
   B. 0.1
   C. –0.1
   D. –0.67
   E. 0.33

55. Consider a normal distribution with $\mu = 65$ and $\sigma = 4$. A sample of size 950 is drawn from this population. Approximately how many of the 950 cases would you expect to find between 57 and 73?
   A. 646
   B. 903
   C. 947
   D. All 950 cases
   E. There isn't enough information to tell

56. A normal probability plot:
   A. graphs raw data against percentile ranks.
   B. graphs percentile ranks against raw data
   C. graphs percentile ranks against z-scores
   D. graphs raw scores against z-scores of percentile ranks
   E. graphs standardized scores against raw data

57. You believe that a normal probability plot of 30 data points provides evidence that the original data is normally distributed. If this is true, then the pattern of points on this normal probability plot is:
   A. bell-shaped
   B. u-shaped
   C. a logarithmic curve
   D. randomly scattered in the plot
   E. None of the above

58. All of the following statements are true about the normal distribution except:
   A. The normal curve crosses the x-axis at z-scores above 3.0 and below –3.0.
   B. The normal distribution is a mathematical model of a population distribution.
   C. The total area under the normal curve is always one.
   D. The proportion of observations two standard deviations above the mean is the same as the proportion two standard deviations below the mean.
   E. Values located on the normal curve are not meant as actual observations.

59. The proper notation for a normal distribution with a mean of 250 and standard deviation of 25 is:
   A. N(250, 1)
   B. X(250, 25)
   C. N(250, 25)
   D. N(25, 250)
   E. None of the above.

60. Which of the following is true about the areas described under the normal curve?
    A. The area below the mean is greater than the area above the mean.
    B. The area in the range from z = 0 to z = 1 is the same as the area in the range from z = 1 to z = 2.
    C. Roughly 50% of the observations are located at the mean of the distribution.
    D. Fewer than one percent of the cases are located three standard deviations above or below the mean.
    E. 95% of the observations have values twice that of the mean.

61. In a normal distribution with a mean of 30 and a standard deviation of 5, you'd find the largest proportion of cases between:
    A. x = 20 and x = 30
    B. x = 10 and x = 15
    C. x = 25 and x = 35
    D. x = 35 and x = 50
    E. Answer can't be determined.

62. When you have a normal distribution and you know that the area above a given value of x is 0.35, you also know that:
    A. 35% of the observations have that value of x.
    B. 35% of the observations have values less than or equal to that value of x.
    C. There is a 35% chance of choosing an individual with that value of x.
    D. The relative frequency of that value of x is 0.35.
    E. None of the above.

63. A normal curve table tells you that the probability lying below z = -1 is 0.1587. This can be interpreted as:
    A. 15.87% of the area under the curve lies below z = -1.
    B. 15.87% of the area under the curve lies at or below z = -1.
    C. A random selection from the population has a 15.87% chance of being below z = -1.
    D. A random selection from the population has a 15.87% chance of being at or below z = -1.
    E. All of the above.

64. You have two normally distributed populations:
    Population A:
    Mean = 50 Standard deviation = 12
    Population B:
    Mean = 75 Standard deviation = 15
    The area under the normal curve is greatest in which scenario?
    A. Area below a value of 25 in Population A
    B. Area above 65 in Population A
    C. Area between 75 and 90 in Population B
    D. Area below 60 in Population B
    E. Area above 80 in Population B

Questions 65 – 67:
You have the following regression equation for the effect of streetlights per block (x), on crimes per month (y): $y = 2.4 - 0.2x$.

65. How many crimes a month are predicted when there are 7 streetlights on a block?
    A. 3.8
    B. 1.7
    C. 16.6
    D. –11.6
    E. 1.0

66. Calculate the residual for a block with 10 streetlights and 1 crime a month.
    A. –0.6
    B. 0.6
    C. –0.4
    D. 0.4
    E. –1.2

67. Using the regression equation above, correctly interpret the slope.
    A. The effect of the crime rate on streetlights is given by the coefficient –0.2.
    B. For every additional streetlight per block, the crimes per month go up by 2.4.
    C. For every additional streetlight per block, the crimes per month decrease by 0.2.
    D. For every additional crime per month, the number of streetlights goes down by 0.2 lights.
    E. Every additional streetlight per block is associated with 0.2 more crimes a month, on average.

68. True or False: On the least-squares regression line, the point $\left(\bar{x}, \bar{y}\right)$ always has a residual of 0.

Questions 69 and 70: Use the sample data set {(8,4), (4,3), (7,3), and (5,2)}.

69. Calculate the least-squares regression line.
    A. $y = 0.3 + 1.2x$
    B. $y = -0.02 + 0.9x$
    C. $y = 1.5 + 1.5x$
    D. $y = 4.8 - 3x$
    E. $y = 1.2 + 0.3x$

70. Calculate the correlation coefficient.
    A. 0.3
    B. 0.6708203932
    C. 1.2
    D. 0.45
    E. 0.88

71. True or False: All positive correlations indicate stronger relationships than all negative correlations.

72. Which of the following determines the sign of r?
   A. The pattern of the residuals
   B. The strength of the relationship between variables
   C. Whether the y-intercept is positive or negative
   D. Whether the value of y increases or decreases as the value of x increases
   E. Whether the sum of the squared residuals is positive or negative

73. Suppose that all sample data points are on the same line with a positive slope. What would r be for this sample?
   A. r would be –1.0
   B. r would be +0.99
   C. r would be +100
   D. r would be +1.0
   E. r would be the same as the slope of the regression line.

74. A bivariate scatterplot has an r^2 of 0.85. This means:
   A. 15% of the variation in y is explained by the changes in x.
   B. 15% of the variation in x is explained by the changes in y.
   C. 15% of the variation in x isn't explained by the changes in y.
   D. 85% of the variation in y is explained by the changes in x.
   E. 85% of the variation in x is explained by the changes in y.

75. Consider the data set A: (2,8), (3,6), (4,9), and (5,9). Which of the following is the proper interpretation of the coefficient of determination?
   A. There's a 30% increase in the variation in the data set.
   B. Thirty percent of the y-values can be explained by variation of the x-values.
   C. Thirty percent of the variation in the y-values can be explained by variation of the x-values.
   D. You've reduced the total variation by 70%.
   E. Both B and C above are correct.

76. A residual:
   A. is the amount of variation explained by the least-squares regression line of y on x.
   B. is how much an observed y-value differs from a predicted y-value.
   C. predicts how well x explains y.
   D. is the total variation of the data points.
   E. should be smaller than the mean of y.

77. An outlier:
   A. usually does not have a strong effect on the correlation coefficient and regression line.
   B. can also be an influential point.
   C. may be an error.
   D. usually does not have a strong effect on the regression line, can also be an influential point, and may be an error.
   E. usually has a strong effect on the correlation coefficient and regression line and can also be an influential point.

78. True or False: An r of –1.0 proves a strong cause and effect relationship between x and y.

79. Influential points and outliers:
   A. sometimes have no effect on the regression line.
   B. are useful when they cause a stronger correlation coefficient for a data set.
   C. should be discarded if they cause a weaker correlation coefficient for a data set.
   D. are rarely found in data sets.
   E. should be examined carefully to determine if they're part of the data set.

80. A linear regression line indicates the amount of grams of the chemical $CuSO_4$ (the response variable, y) that dissolve in water at various temperatures in Celsius (the explanatory variable, x). The least-squares regression line is $y = 10.14 + 0.51x$. Give the meaning of the slope of the regression line in the context of the problem.
   A. For each one-degree rise in the temperature, you can dissolve 10.14 more grams of $CuSO_4$.
   B. If the temperature increases by 0.51, you can dissolve one more gram of $CuSO_4$.
   C. When you dissolve one more gram of $CuSO_4$, then the temperature will rise by 0.51.
   D. For each one-degree rise in the temperature, you can dissolve 0.51 more grams of $CuSO_4$.
   E. For each one-degree rise in the temperature, you can dissolve 0.51 fewer grams of $CuSO_4$.

81. When looking at a scatterplot of two variables, the variable along the horizontal axis is typically referred to as the:
   A. response variable.
   B. residuals.
   C. explanatory variable.
   D. dependent variable.
   E. categorical variable.

82. Which of the following would indicate the strongest relationship between two variables?
   A. r = -0.35
   B. r = -0.28
   C. r = 0.21
   D. $r^2 = 0.01$
   E. $r^2 = 0.23$

83. The goal of the least-squares regression is to compute a line that:
   A. connects all the bivariate data points in a scatterplot.
   B. connects all the residuals shown in a scatterplot.
   C. minimizes the sum of the observed values of x and y.
   D. minimizes the sum of the squared residuals.
   E. None of the above.

Questions 84 and 85 refer to the following information.
You've computed the following least-squares regression line using a sample of college students:
$y = 55 + 5x$, where x = hours of study per day and y = test score (ranges from 0 to 100).

84. Using this equation, which is the predicted test score for an individual who studies 8 hours a day?
    A. 68
    B. 63
    C. 85
    D. 95
    E. Can't be determined.

85. Suppose the maximum number of hours of study among the students in your sample is 6. If you used the equation to predict the test score of a student who studied 8 hours a day, your prediction would be considered a (an):
    A. regression constant.
    B. influential point.
    C. interpolation.
    D. extrapolation.
    E. None of the above.

86. If the association between two variables is exponential, which of the following is the general form of the regression equation for the transformed data?
    A. $y = a + b\log(x)$
    B. $y = a + \log(x^b)$
    C. $\log\left(y\right) = a + bx$
    D. $y = \log a + \log(bx)$
    E. $y = \log(a \cdot b)x$

87. Suppose you make a scatterplot for the population of the United States every decade from 1780 to 1990. The graph is exponential, so to transform this relationship you can take the natural logarithms of the response variable to find the regression equation for these data. The resulting regression equation, r and r^2 is:

    $$\ln y = -35.4173 + 0.02078x; r = 0.9835; r^2 = 0.9672$$

    This equation would predict the population for 1915 to be:

    A. $10^{4.3764}$ people.
    B. 4.3764 million people.
    C. $e^{4.3764}$ million people.
    D. $e^{4.3764}$ people.
    E. (4.3764)(10) million people.

Questions 89 and 90 refer to the following data:
Some researchers were interested in whether the number of crimes committed during the summer is related to the outdoor temperature. The results of a survey of 150 municipal police departments revealed the following:

	Crime Rate		
Temp.	Below	Normal	Above
Below	12	8	5
Normal	35	41	24
Above	4	7	14

88. The marginal distribution for crime rate (below, normal, above) is:
    A. 0.34, 0.37, 0.29
    B. 25, 100, 25
    C. 51, 56, 43
    D. 0.17, 0.67, 0.17
    E. There isn't enough information to determine the marginal distribution for crime rate.

89. The conditional distribution of temperature by above average crime rate is:
    A. 0.08, 0.13, 0.33
    B. 0.17, 0.67, 0.17
    C. 0.14, 0.73, 0.13
    D. 0.12, 0.56, 0.33
    E. 0.24, 0.69, 0.08

90. **(Skip this question!)** Imagine a study of surgery survival rates in two hospitals: Hospital A and Hospital B. Surgery patients in these hospitals are classified as being in either good condition or poor condition. Which of the following would be an example of Simpson's paradox?
    A. When looking separately at survival rates for each kind of patient (good condition and poor condition), Hospital B has lower rates than Hospital A for both kinds, and Hospital A has better doctors.
    B. When looking separately at survival rates among patients in good condition and in poor condition, Hospital B has higher rates for both groups. Yet when the two groups are combined, Hospital A has a higher rate.
    C. Hospitals A and B have the same survival rates for patients in poor condition and good condition, even though the conditions of the patients are different.
    D. The two kinds of patients confound the doctors, because Hospital A has a higher surgery survival rate overall due to some kind of lurking variable. However, Hospital B has a more rigorous research program.
    E. Doctors at both hospitals are confounded to a varying degree because of lurking surgery survivors.

91. You're given some bivariate (x,y) data. You use your calculator to find the linear regression line for the transformed data (ln x, ln y). Your calculator reports the equation of the least squares regression line to be Y = 0.2 + 0.4x. When x = 2, the correct predicted value for y is:
   A. 1.6117
   B. 0.4773
   C. 1.000
   D. 2.972
   E. 1.0570

<u>Problems 92 – 95</u> refer to this table:
The following table shows hypothetical data for two experimental treatments (A and B) against two stages of cancer (Early and Advanced). "Remission" indicates a successful treatment. "Rate is the proportion of remissions.

	Early			**Advanced**		
	Remissions	Death	Rate	Remissions	Death	Rate
A	5	1	0.833	4	6	0.400
B	10	4	0.714	1	4	0.200

92. Which table correctly shows the aggregation of the data?

A.

	Remissions	Death	Rate
A	9	7	0.563
B	11	8	0.579

B.

	Remissions	Death	Rate
A	9	7	0.600
B	11	8	0.579

C.

	Remissions	Death	Rate
A	10	7	0.588
B	8	8	0.500

D.

	Remissions	Death	Rate
A	10	7	0.600
B	8	8	0.500

E. None of these.

93. Which of the following is true?
   A. It's unclear whether Treatment A or Treatment B has a higher success rate.
   B. When success rates for early and advanced cancers are kept separate, Treatment A has a higher success rate. But when data for the two stages are combined, Treatment B has a higher rate.
   C. When the success rates for the two stages of cancer are kept together, Treatment A has a higher success rate for each stage of cancer. But when the two stages of cancer are separated, Treatment B has a higher success rate.
   D. Treatment A has a higher success rate no matter how the data are combined.
   E. Treatment B has a higher success rate no matter how the data are combined.

94. What's the lurking variable when the data for early and advanced cancers are combined?
   A. Treatment method (A or B)
   B. Number of treatments
   C. Stage of cancer (Early or Advanced)
   D. Quality of treatment (Good or Poor)
   E. Quality of hospital (Good or Poor)

95. Which treatment is actually more effective?
   A. It depends on whether you're treating separate groups or one large group.
   B. It depends on the patient's condition.
   C. Treatment A
   D. Treatment B
   E. Both treatments should be combined.

96. Which of the situations below probably does not have a lurking variable operating in some way?
   A. Neighborhoods with more station wagons tend to have more playgrounds.
   B. Beaches with more sand than rocks tend to be older.
   C. Small dogs bite more people than large dogs nationwide, but in both rural areas and urban areas, large dogs are more likely to bite people.
   D. Towns that have more teachers have higher sales of floor wax and cat litter.
   E. None of the above.

97. Which of the following is a characteristic of a census?
   A. It's based on anecdotal evidence.
   B. It's generally more accurate than a sample.
   C. It uses secondary data.
   D. It's always part of an experiment.
   E. It gathers data from every member of a population.

98. An observational study based on survey data concluded that individuals who took more vitamin C were able to recover from the flu faster. You want to replicate this study using an experimental approach. The treatment in this experiment might be:
   A. how long it takes to recover from the flu, in days.
   B. whether an individual took vitamin C in a pill form or liquid form.
   C. the amount of vitamin C taken per day: 0 mg, 1000 mg, 2000 mg, or 3000 mg.
   D. the change in body temperature over the period of the experiment.
   E. All are acceptable treatments

99. Many statisticians say that the US Census, which attempts to count every population member directly, is significantly less accurate than a count estimated by random sampling. Why might a count estimated from random samples be more accurate than a census?
   A. Random samples are scientific whereas censuses are not.
   B. A true census takes so long that by the time all population members are counted, the population has changed.
   C. A census often can't find every population member, so some groups (such as the homeless) are often under-represented.
   D. A census is a haphazard sample, and census takers may be bribed by households.
   E. A census is an old method written into law before anyone knew anything about statistics.

100. Which of the following is the best representative sample of the adult population in the United States?
   A. Simple random sample of 10000 adults from different city phone books
   B. Simple random sample of 10000 voters from across the country
   C. Sample of 50000 individuals at the Super Bowl (which draws from all over the country)
   D. Simple random sample of 1000 adults from across the country
   E. Sample of 50000 members of AARP (American Association of Retired Persons)

101. A block is best described as:
   A. the use of chance to divide experimental units into groups.
   B. a design in which neither the experimenter nor the subject knows who is in the treatment group and who is in the control group
   C. a group of subjects that are similar in some way known to affect the response to the treatment
   D. the policy of repeating an experiment on different subjects to reduce chance variation and to determine the generalizability of the findings.
   E. the tendency of subjects to respond favorably to any treatment.

102. Replication is best described as:
   A. the use of chance to divide experimental units into groups.
   B. a design in which neither the experimenter nor the subject knows who is in the treatment group and who is in the control group
   C. a group of subjects that are similar in some way known to affect the response to the treatment
   D. the policy of repeating an experiment on different subjects to reduce chance variation and to determine the generalizability of the findings.
   E. the tendency of subjects to respond favorably to any treatment.

103. Double-blind is best described as:
   A. the use of chance to divide experimental units into groups.
   B. a design in which neither the experimenter nor the subject knows who is in the treatment group and who is in the control group
   C. a group of subjects that are similar in some way known to affect the response to the treatment
   D. the policy of repeating an experiment on different subjects to reduce chance variation and to determine the generalizability of the findings.
   E. the tendency of subjects to respond favorably to any treatment.

104. The placebo effect is best described as:
   A. the use of chance to divide experimental units into groups.
   B. a design in which neither the experimenter nor the subject knows who is in the treatment group and who is in the control group
   C. a group of subjects that are similar in some way known to affect the response to the treatment
   D. the policy of repeating an experiment on different subjects to reduce chance variation and to determine the generalizability of the findings.
   E. the tendency of subjects to respond favorably to any treatment.

105. An experiment is conducted in which a series of tests are performed on pairs of identical twins who were raised separately. A comparison of scores on each pair of twins is used for analysis. This is best described as:
   A. a matched-pairs procedure
   B. a double-blind procedure
   C. a randomized treatment group procedure
   D. a procedure with two explanatory variables
   E. an example of the placebo effect

106. You're going to test two new varieties of fish food vs. a commonly used fish food. You set up an experiment as follows: 60 fish are randomly assigned to each of three different tanks. One tank is randomly selected to receive one of the new foods, another to receive the other new food, and the third tank to receive the common food. Fish growth is measured over time. This is an example of:
   A. a randomized block design
   B. a double-blind matched pairs test
   C. a completely randomized design with no control group
   D. a comparative block design
   E. a completely randomized design with a control group

107. Let's say you're interested in the effects on boys of different dosage levels of a new drug for the treatment of ADD. You set up an experiment to consider the factor of dosage with two levels (300 mg vs. 500 mg). What would be the different treatment groups of the experiment within each block?
   A. Three groups: no drug/300 mg of new drug/500 mg of new drug
   B. Three groups: placebo drug/300 mg of new drug/500 mg of new drug
   C. Two groups: 300 mg of new drug/500 mg of new drug
   D. Four groups: no drug/placebo drug/300 mg of new drug/500 mg of new drug
   E. Two groups: placebo drug/either 300 or 500 mg of new drug

108. You've read a story in *The New York Times* claiming that individuals who engage in aerobic exercise for at least an hour a day demonstrate fewer symptoms of depression. You read that an experiment was conducted in which a researcher first administered a survey on depression and self-esteem to 100 individuals and then taught them some proper techniques of aerobic exercise. The 100 individuals were then sent off to exercise at least one hour a day. After two months, the depression and self-esteem survey was administered again and showed that depression symptoms declined and self-esteem increased. The experimental design used here is a:

A. matched pairs design in which subjects are matched on a set of characteristics before random assignment.
B. matched pairs before and after design.
C. completely randomized design.
D. randomized block design, blocked on form of exercise and gender.
E. quasi design.

109. For a semester project, a student needs to select a random sample of 10 students from his senior class of 250. He carefully numbers the class list from 000 to 249 and then uses a random number generator to obtain 3-digit random numbers. The 10 unique numbers are his sample. He notices that they all belong to the same AP Calculus class. Another student claims that this could not be a random sample. Which of the following is true?

A. The sample drawn is so unlikely that it could not be considered a random sample.
B. Since the selected students are not representative of the entire senior class, this is not a random sample.
C. Whether a sample is a random sample or not is determined by the sampling method, not the results. The method used here is valid.
D. A sample size of 10 is too small to be a random sample of 250.
E. The class should have been numbered from 001 to 250 rather than from 000 to 249 to produce a better random sample.

110. As primary research for one of her books, Shere Hite distributed 100,000 questionnaires to women's groups: 4,500 women responded. Hite found that 96% of the women felt they give more emotional support to than they get from their husbands or boyfriends. Which of the following best describes her sample?

A. Representative sample
B. Voluntary response sample
C. Stratified random sample
D. Systematic sample
E. Cluster sample

111. An important reason a market researcher collects data using a stratified random sample rather than a simple random sample is:

A. to collect data at a lower cost.
B. to eliminate, or at least reduce, bias.
C. to make a representative sample more likely than one produced by simple random sampling.
D. convenient data collection.
E. to have a systematic way of obtaining the data.

112. Which of the following would generate an SRS of 50 integers from 5 to 25 on the TI-83?

A. randNorm(5,25,50)
B. randInt(50,5,25)
C. rand(5,25,50)
D. 50randInt(5,25)
E. randInt(5,25,50)

113. True or False: Blocking in an experiment and stratifying in a survey accomplish the same thing. They control the amount of variation in key characteristics within the total sample, such as race or gender, that are likely to be important to the outcome of interest.

114. All of the following are examples of a sample that suffers from an undercoverage bias except:

A. an estimate of an election outcome based on a sample of voters' names taken from automobile registrations.
B. a poll on Missouri dentists' attitudes, based on questionnaires sent to a registry of attendees at the largest dental convention in the state.
C. the mean income of US citizens, based on Census data.
D. a random sample of 5,000 US adults with survey data collected via e-mail.
E. a simple random sample of licensed drivers, taken from driver-licensing records.

115. Consider a survey in which 79% of the respondents said yes to the following question:

> Agree or disagree: Since our economy can't sustain itself without a healthy environment, it's important that Congress pass laws to protect the environment.

What's a reasonable reaction to the survey results?

A. They're not valid, due to undercoverage.
B. They're not valid, due to voluntary response bias.
C. They're not valid, due to a leading question.
D. I totally agree. We need laws that protect the environment.
E. I disagree. We don't need laws that protect the environment.

116. A firm administers a survey in the state of New York. The survey is mailed out to a random set of 5,000 households throughout the state, and responses are received from 200 households. The two key questions on the survey are:

1) Whom do you plan to vote for in the upcoming Senate election?
2) Do you agree with the idea of imposing term limits on the amount of time a Senator can spend in office?

What is a possible problem with this survey?

A. Voluntary response bias
B. Undercoverage bias
C. Leading questions
D. Stratified original sample
E. All of the above

117. A survey is conducted at a local mall. An interviewer stops every 20th person that walks by and asks them to answer some questions about attitude on transportation use. Of the 200 individuals who complete the survey (all of the people asked completed the survey), 80 say they prefer the bus as a transportation option. From this analysis, you have reason to conclude that:

A. About 40% of the sample responded to a representative survey
B. About 40% of the local residents prefer the bus as a transportation option
C. About 80% of the mall goers prefer the bus as a transportation option
D. About 40% of the mall goers prefer the bus as a transportation option
E. About 40% of those who ride the bus shop at the local mall

118. If subjects in an experiment are separated into groups of whites and non-whites prior to the random assignment stage, we should think of race as a (an):

A. bias
B. blocking variable
C. matched pair
D. explanatory variable
E. response variable

<u>Questions 119 – 122</u> refer to the following data:
The table below shows the results of a hypothetical survey on participation in sports activities by men and women. Those surveyed answered yes to activities they had done at least twice in the previous 12 months. Assume that the sample is representative of the overall population.

	Men	**Women**
Total	109,059	115,588
Aerobic Exercising	3,717	19,535
Baseball	12,603	2,974
Hunting	18,512	2,343
Softball	11,535	8,541
Exercise Walking	25,146	46,286

119. If you randomly select one man, what's the probability that he hunts, expressed as a percentage?

A. 0.1697
B. 16.97%
C. 16.02%
D. 0.1602
E. 8.24%

120. If you randomly select one woman, what's the probability that she does aerobics?

A. 0.034
B. 0.169
C. 0.201
D. 0.087
E. 5.917

121. If you randomly select one man, what the probability that he doesn't do exercise walking?
A. 0.231
B. 1.300
C. 0.400
D. 0.600
E. 0.769

122. If you randomly select one person from the study, what would be the probability that he or she plays baseball?
A. 0.1156
B. 0.0257
C. 0.0693
D. 0.0707
E. 0.0347

123. You have the following information about voters attending a political meeting:

Hometown	Pop. Size	Probability of Democrat	Probability of Republican	Probability of Independent
Washington, DC	800,000	0.810	0.115	0.045
Tacoma, WA	150,000	0.432	0.258	0.278
Ogden, UT	45,000	0.135	0.832	0.021
Ashland, OR	26,000	0.649	0.222	0.083

You randomly choose one attendee to speak with about his or her political views. Of the individuals listed below, this individual is most likely a(n):

A. Democrat from Tacoma, WA
B. Independent from Tacoma, WA
C. Democrat from Ashland, OR
D. Republican from Ogden, UT
E. Republican from Washington, DC

124. Which of the following are true?
I. Two events are mutually exclusive if they can't both occur at the same time.
II. Two events are independent if they have the same probability.
III. An event and its complement have probabilities that always add to 1.

A. I only
B. II only
C. III only
D. I and II only
E. I and III only

Questions 125 – 128 refer to the following information:
You have two specially created dice. Each die has six sides. The first die has the numbers {1,3,5,7,9,11) on the sides. The second die has the numbers {2,4,8,10,11,15} on the sides.

125. If you roll the two dice, how many possible outcomes (not necessarily unique) are there?
   A. 6
   B. 12
   C. 24
   D. 36
   E. 40

126. You're about to roll only the second die. Estimate the probability of getting an odd number on a single roll of this die.
   A. 0.167
   B. 0.333
   C. 0.056
   D. 0.500
   E. 0.000

127. Which of the following would not represent an instance of independent events when using the two specially created dice?
   A. Getting an odd number on the first die AND getting an odd number on the second die.
   B. Getting a 9 on the first die AND getting a sum of 19 on the two die combined.
   C. Getting a 2 on the first die AND getting a value greater than 10 on the second die.
   D. Getting an even number on the first die AND getting a 15 on the second die.
   E. All are examples of independent events.

128. Calculate the probability that the sum of the two dice is greater than or equal to 20.
   A. 0.000
   B. 0.056
   C. 0.250
   D. 0.108
   E. 0.194

129. One card is randomly selected from a standard 52-card deck. Which of the following gives the probability that the card is a black ace?
   A. $\frac{1}{2}+\frac{1}{13}-\frac{2}{52}$
   B. $\left(\frac{1}{2}\right)\left(\frac{1}{13}\right)$
   C. $\frac{1}{52}$
   D. $\left(\frac{1}{2}\right)\left(\frac{1}{13}\right)-\frac{2}{52}$
   E. $\frac{1}{2}+\frac{1}{13}$

Questions 130 – 132 refer to this table:
This two-way table gives frequencies of the combined simple events of color and size of marbles in a bag.

	Blue	**Yellow**	**White**	**Black**	
Small	27	34	16	10	87
Medium	7	19	53	13	92
Large	3	0	11	42	56
	37	53	80	65	235

130. What's the probability that a randomly selected marble will be blue or white?
   A. 0.34
   B. 0.16
   C. 0.50
   D. 0.18
   E. None of the above

131. Find the probability of a randomly drawn marble being yellow, given that the marble is small.
   A. 0.39
   B. 0.64
   C. 0.14
   D. 0.37
   E. 0.23

132. Find the probability of a randomly drawn marble being large or small, given that the marble is black.
   A. 0.35
   B. 0.28
   C. 0.22
   D. 0.20
   E. 0.80

133. Which of the following statements is correct?
   A. The probability of any given outcome for a discrete random variable is 0.
   B. The area between any two x-values under a density curve is 1.
   C. A random variable can have either numbers or categories for its values.
   D. A random variable must have a number as its value.
   E. A uniform distribution can't have a density curve.

134. The following table is for a discrete probability distribution.

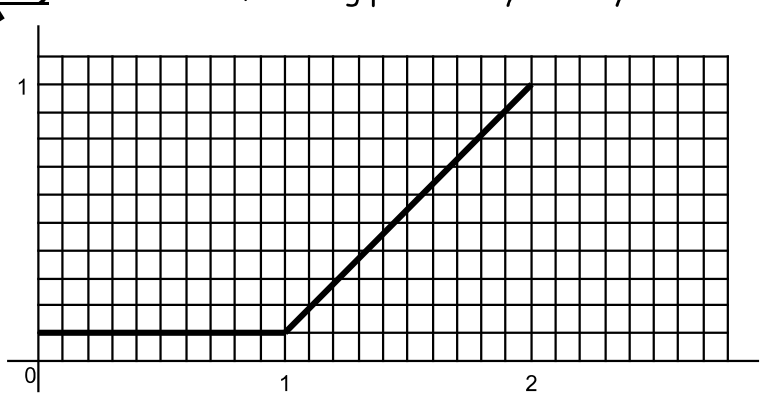
x	7	3	9	11	15
p	0.05	0.10	0.15	?	0.25

A. 0.45
B. 1
C. 0.6
D. 0.2
E. The value cannot be calculated based on the information given.

135. Which of the following are discrete random variables?
I. The enrollment at state universities
II. The outcomes of a loaded die
III. The set of all random numbers between 1 and 10

A. I only
B. II only
C. III only
D. I and II only
E. II and III only

136. **(Skip this question!)** Consider the following probability density curve:



What is P(.5 < x < 2)?
A. 1
B. 0.60
C. 0.73
D. 0.92
E. 0.78

137. Randomly selected students are asked whom they're going to vote for in the next student body election. The sample space is {Joan, Harvey, Anita, Luis}. Which of the following represents a legitimate assignment of probabilities for this sample space?

A. {0.2, 0.3, 0.4, 0.2}
B. {0.1, 0.2, 0.3, 0.4}
C. {0.3, 0.4, 0.5, -0.2}
D. {0.6, 0.1, 0.1, 0.1}
E. None of the above.

138. For a uniform distribution defined on ($0 \le x \le 20$), find $P(x < 2 \text{ or } x \ge 15)$.

A. 0.1
B. 0.25
C. 0.35
D. 0.7
E. 1

139. Which of the following examples would constitute a discrete random variable?

I. Total number of points scored in a football game
II. Height of the ocean's tide at a given location
III. Number of near collisions of aircraft in a year

A. I and II
B. I and III
C. I only
D. II only
E. I, II and III

140. Consider a random experiment with S = {a,b,c,d,e,f,g}. The event B is composed of simple events {c,e,g}. Which of the events below is mutually exclusive with event B?

A. {a,b,d,f}
B. {a,b}
C. {b,d,f}
D. All of the above.
E. None of the above.

141. Which of the following are requirements of a discrete probability distribution?

I. $0 \le p(x) \le 1$
II. $p(x) - 1 = 0$
III. $\Sigma p(x) = 1$

A. I only
B. II only
C. I and II
D. I and III
E. II and III

142. If two events, A and B, are complementary, what else must be true?
   A. They are independent.
   B. They are mutually exclusive.
   C. The have the sample simple events.
   D. They have different sample spaces.
   E. All of the above.

<u>Questions 143 and 144</u>:
Consider the following probability distribution for a discrete random variable X:

X	-1	1	10
P(X)	25/36	10/36	1/36

143. What is the expected value of X?
   A. -5/36
   B. 1
   C. 1.93
   D. 5/36
   E. 45/36

144. What is the variance of X?
   A. -5/36
   B. 1.93
   C. 0
   D. 3.73
   E. 4.12

<u>Questions 145 and 146</u> refer to the following variables:
Random variable X has 12 values with a mean of 8 and a standard deviation of 3. Random variable Y also has 12 values, but it has a mean of 11 and a standard deviation of 2.
A new random variable Z is created as follows: $Z = 2X + 3Y$.

145. What is the mean of Z?
   A. 24
   B. 12
   C. 49
   D. 60
   E. 19

146. What is the standard deviation of Z?
   A. 24
   B. $\sqrt{72}$
   C. 49
   D. 60
   E. There isn't enough information to know the standard deviation of Z.

Questions 147 and 148 refer to the following variables:
Consider two discrete, independent, random variables X and Y with $\mu_X = 3$, $\sigma_X^2 = 1$, $\mu_Y = 5$, $\sigma_Y^2 = 1.3$.

147. Find σ_{3+6X}^2.
   A. 9
   B. 6
   C. 39
   D. 36
   E. 81

148. Find μ_{X+Y} and σ_{X+Y}.
   A. 8, 2.14
   B. 15, 1.3
   C. 8, 1.52
   D. 15, 2.3
   E. The mean is 8, but there isn't enough information to determine σ.

149. Which of the following indicates the value of $\binom{8}{2}(.3)^2(.7)^6$?
   A. binomcdf(8,0.3,2)
   B. binompdf(8,0.7,2)
   C. binompdf(8,0.3,2)
   D. 1 - binompdf(8,0.3,2)
   E. binomcdf(8,0.7,2)

150. You take a 100-question multiple-choice test. Each question has five choices, and you guess at each question. Which of the following calculator commands would give you the probability of getting at least 30 questions correct?
   A. 1 - binompdf(100,0.2,29)
   B. binomcdf(100,0.2,29)
   C. binomcdf(100,0.2,30)
   D. 1 - binomcdf(100,0.2,30)
   E. 1 - binomcdf(100,0.2,29)

151. Which of the following are not essential characteristics of a binomial event?
   I. Each outcome must be independent.
   II. The sample size must be at least 20.
   III. A trial can have only two possible outcomes.

   A. I only
   B. I and II only
   C. II and III only
   D. II only
   E. III only

152. The probability that a given 80-year-old person will die in the next year is 0.27. What's the probability that exactly 10 of 40 80-year-olds will die in the next year?

A. 0.8615
B. 0.4685
C. 0.1385
D. 0.1208
E. 0.00000000031795

153. The probability that a given 80-year-old person will die in the next year is 0.27. What's the probability that between 10 and 15 (inclusive) of 40 80-year-olds will die in the next year?

A. 0.9191
B. 0.4685
C. 0.6191
D. 0.4806
E. 0.6454

154. Which of the following situations satisfies all the conditions of a binomial setting?

A. A jar contains five balls: three red and two white. Two balls are randomly selected from the jar without replacement, and the number X of red balls is recorded.
B. A jar contains 500 balls – 300 red and 200 white. Ten balls are randomly selected from the jar with replacement, and the number X of red balls is recorded.
C. A meteorologist in Chicago records the number of days of rain from August through March.
D. A market research firm hires operators to conduct telephone surveys. The computer goes down an alphabetical list of names, dialing the number for each name. When a caller answers the phone, an operator asks the respondent if they have time to answer some questions.
E. In a class of 40 AP Statistics students who've just begun their course, how many will pass?

155. Suppose you roll a fair, standard six-sided die 10 times. What's the probability of getting exactly three fives in those 10 rolls?

A. 0.60
B. 0.30
C. 0.155
D. 0.000618
E. 0.930

156. Under what conditions can you use a normal distribution to approximate a binomial distribution?

A. $np \geq 10$ and $n(1 - p) \geq 10$
B. $n \geq 10$
C. $p = 0.5$
D. You can't. They're different distributions.
E. When you use the continuity correction.

157. For which of these values of n and p can you use the normal approximation to the binomial distribution?

A. $n = 100, p = 0.02$
B. $n = 10, p = 0.7$
C. $n = 15, p = 2/3$
D. $n = 20, p = 0.6$
E. $n = 60, p = 0.4$

158. In a binomial distribution with sample size $n = 65$, and probability of success $p = 0.8$, what would be the approximate mean of the distribution?

A. 52
B. 65
C. 10.4
D. 3.22
E. 0.8

159. In a binomial distribution with $n = 140$ and $p = 0.62$, what is the expected standard deviation of the distribution, to the nearest hundredth?

A. 32.98
B. 0.04
C. 86.80
D. 5.74
E. 0.62

<u>Notes/Important Things to Know</u>

- Thirty-five of these questions will be on your final exam (not necessarily in any order and answer choices may change order).
- In addition, your exam will contain five free-response questions, one of which will be an experimental design question and the other a simulation. The topics in this review packet should assist you with the other three free-response questions.