Name _____

Self-Assessment. I spent approximately _____ hours preparing for this midterm examination. On average, I spend approximately _____ hours on each team project.

Bayes' Theorem application. Suppose A = heart attack and B = positive stress test. Suppose, moreover, that P(A) = 0.05, P(B|A) = 0.90, and P({not B}|{not A}) = 0.98. By complements, P({not A}) =_____and P(B|{not A}) =_____. By Bayes' Theorem, P(A|B) =

_____.

Descriptive statistics. Considering the data set {108 117 125 128 135 138 146 155}, calculate each of the following quantities.

Mean _____   Median _____   Interquartile Range _____
Variance _____   Standard Deviation _____

Excel operations. Consider the following spreadsheet.

	A	B	C	D	E	F
1	108	120				
2	117	140				
3	125					
4	128					
5	135					
6	138					
7	146					
8	155					

Suppose I highlight C1 through C3, type the following in the formula bar, and press Ctrl-Shift-Enter.

=FREQUENCY(A1:A8,B1:B2)

Print in C1 through C3 what will be the results.

Suppose I type the following in C6 and Excel returns the result of 0.242.

=NORMDIST(100,85,10,1) – NORMDIST(90,85,10,1)

Based on that result, complete the following sentence: The probability is 0.242 that a normal random variable with mean _____ and standard deviation _____ falls between _____ and _____.

Suppose I type the following in D1 and drag it down through D8.

=IF(A1>120,IF(A1>140,2,1),0)

Print in D1 through D8 what will be the results.

Suppose I type the following in E1 and drag it down through E8.

=IF($A1>120,IF($A1>140,0,1),0)

Print in E1 through E8 what will be the results.

Multiple choice questions. Print "W" for any question you wish to waive; you may waive up to five questions.

1. _____	6. _____	11. _____	16. _____	21. _____	26. _____	31. _____
2. _____	7. _____	12. _____	17. _____	22. _____	27. _____	32. _____
3. _____	8. _____	13. _____	18. _____	23. _____	28. _____	33. _____
4. _____	9. _____	14. _____	19. _____	24. _____	29. _____	34. _____
5. _____	10. _____	15. _____	20. _____	25. _____	30. _____	35. _____

Name _____

1. If an emergency room director calculated the mean wait time for patients as a measure of central tendency, what measure of variability might she have calculated ?
      A. Standard deviation
      B. Range
      C. Interquartile range
      D. Mode

2. Which of the following schemes is most time-efficient, if you are visiting households to interview people ?
      A. cluster sampling, with neighbourhood block as a cluster
      B. systematic sampling, based on an alphabetical list of property owners
      C. simple random sampling, based on an alphabetical list of property owners
      D. census, where you visit every household

3. Suppose that a questionnaire is administered to 10 randomly selected LHD directors in a state and that their average score on a 0-to-10 scale of cultural and linguistic competence is 6.5. The 6.5 is
      A. a parameter.
      B. a population.
      C. a sample.
      D. a statistic.

4. Suppose that a linear regression model is fit to ascertain whether variable X predicts variable Y. What may we say about variable Y ?
      A. Variable Y is continuous (or finely enough discretized to be treated as continuous).
      B. Variable Y is appropriately called a response variable.
      C. Both A. and B. are correct.
      D. Neither A. nor B. is correct.

5. Suppose that a mean is calculated. Which of the following may be true ?
      A. The variable in question is interval.
      B. The variable in question is ratio.
      C. The variable in question is ordinal but not interval or ratio.
      D. Both A. and B. may be true.

6. Suppose you want to define a variable in column D of an Excel spreadsheet to be the quotient of the variable in column B by the variable in column C. Assuming that your first row contains headers and that you have several hundred rows of data, why is using a formula like =B2/C2 better than typing in actual numbers, as in (for example) =140/100 ?
      A. Using a formula will save time in populating column D.
      B. Using a formula will allow column D to adjust automatically if column B or column C entries are
         altered.
      C. Both A. and B. are correct.
      D. Neither A. nor B. is correct.

7. Suppose the formula =$B2/$C2 appears in cell D2. If you drag this formula to the right, what will be the numerical contents of cell E2 ?
      A. You'll get D2.
      B. You'll get C2 divided by D2.
      C. You'll get the square of C2 divided by B2.
      D. Both B. and C. are correct.

Name _____

8. Suppose cell F10 contains the number 5. What will be the result of =(F10 > 7)+0 ?
        A. 0
        B. 1
        C. FALSE
        D. TRUE

9. Suppose cell F10 contains the number 5. What will be the result of =IF(F10>7,7,IF(F10<-2,-2,F10)) ?
        A. -2
        B. 7
        C. 5
        D. None of the above.

10. Suppose cell F10 contains the number -4. What will be the result of =IF(F10>7, 7, IF(F10 < -2, -2, F10)) ?
        A. -2
        B. 7
        C. -4
        D. None of the above.

11. Suppose an emergency room director wants to explore graphically whether the proportion of true emergencies varies by day of the week. Which of the following Excel chart types is most suited for this purpose ?
        A. Scatter plot.
        B. Pie chart.
        C. 100% stacked column chart.
        D. Clustered column chart.

12. To generate a random number between 0 and 9, you should
        A. type =FLOOR(10*RAND(),1) in Excel.
        B. type =RAND() in Excel.
        C. ask someone you don't know to select a number between 1 and 10, then subtract 1.
        D. read the last number off of a digital watch.

13. If I type NORMINV(RAND(), 0, 1) in Excel, the result will be
        A. a fixed number between 0 and 1.
        B. a fixed number which need not be between 0 and 1.
        C. a number between 0 and 1, which will change with further operations on the spreadsheet.
        D. a number which need not be between 0 and 1, which will change with further operations.

14. Why are permuted blocks of size six a sound randomization strategy ?
        A. Breaking the blind on one patient will not break the blind on other patients.
        B. Group sizes will be approximately equal even if the study ends early.
        C. Both A. and B. are correct.
        D. Neither A. nor B. is correct.

Name _____

15. What is the probability that =RAND() yields a number between 0.5 and 0.9 ?

       A. 0.4
       B. 0.5
       C. 0.9
       D. None of the above

16. Why is a study with primary data potentially advantageous over a study with secondary data ?
       A. You have the opportunity to ensure that variables of interest are measured.
       B. You can usually complete the study more quickly.
       C. You will not need to interact directly with study subjects.
       D. Both A. and B. are correct.

17. If you had measurements on systolic blood pressure, diastolic blood pressure, and heart rate, what sort of error would likely be detected using the MIN and MAX functions on these variables ?
       A. Someone's DBP of 90 accidentally recorded as 99.
       B. Someone's SBP of 140 accidentally recorded as 1400.
       C. Accidental transposition of DBP and HR in the spreadsheet.
       D. All of the above.

18. Suppose that you have a spreadsheet with 1000 records but that data on a critical variable are missing for 3 records, whence case deletion would leave you with 997 records for data analysis. Assuming that you have no means by which to actually obtain the missing values, what do you think a competent statistician who is not excessively self-interested will advise you to do ?
       A. Just proceed with data analysis using 997 records.
       B. Abandon data analysis; the situation cannot be salvaged with 3 missing records.
       C. Perform a complicated Monte Carlo imputation on account of the 3 missing records.
       D. Enroll in his upcoming statistics class.

19. How would a distribution of human lifetimes best be described ?
       A. Bell-shaped.
       B. Left skewed.
       C. Right skewed.
       D. Uniform.

20. A cumulative frequency distribution will feature numbers between
       A. 0 and the sample size.
       B. 0 and 1.
       C. 0% and 100%.
       D. Both B. and C. are correct.

21. In general, how might we tell whether the distribution of measurements in a population is normal ?
       A. Use the SKEW command in Excel on measurements in a sample drawn from the population.
       B. Make a histogram of measurements in a sample drawn from the population.
       C. Make a histogram of measurements in the population.
       D. Both A. and B. may be employed.

Name _____

22. Considering 1000 families with two children each, suppose that 600 of the 2000 children are overweight.

Supposing that a child is more likely to be overweight if his sibling is overweight, which of the following is plausible for the number of families in which both children are overweight ?
     A. 0
     B. 120
     C. 300
     D. 600

23. If $P(C \mid D)$ denotes sensitivity, then which of the following is specificity ?
     A. $P(D \mid C)$
     B. $P(\{not\ D\} \mid C)$
     C. $P(\{not\ D\} \mid \{not\ C\})$
     D. $P(\{not\ C\} \mid \{not\ D\})$

24. Suppose that 40% of patients visiting a clinic are sedentary, 40% have poor dietary habits, and 64% have at least one of these two problems. Are these two problems independent ?
     A. No, sedentary people are disproportionately likely to have poor dietary habits.
     B. Yes.
     C. No, sedentary people are disproportionately unlikely to have poor dietary habits.
     D. This cannot be determined from the information given.

25. Suppose that 40% of patients visiting a clinic are sedentary, 40% have poor dietary habits, and 64% have at least one of these two problems. What fraction of sedentary patients have poor dietary habits ?
     A. 30%
     B. 40%
     C. 50%
     D. This cannot be determined from the information given.

26. Let X denote the number of true emergencies at a small rural hospital this afternoon, given that there are 15 visitors in total. Which of the following is the most reasonable probabilistic model for X ?
     A. Binomial
     B. Poisson
     C. Normal
     D. Uniform

27. Let X denote the number of true emergencies at a small rural hospital this afternoon. Which of the following is the most reasonable probabilistic model for X ?
     A. Binomial
     B. Poisson
     C. Normal
     D. Uniform

28. Among the next five patients admitted to a hospital, let Z be the number whose lengths of stay exceed three days. If I am willing to assume a Binomial probabilistic model with p = 0.3, then what is the approximate probability that $Z \leq 1$?
     A. 36%
     B. 53%
     C. 16%
     D. 19%

Name _____

29. The probability that a standard normal random variable equals -1 is approximately
      A. 0%
      B. 2.4%
      C. 16%
      D. None of the above

30. The probability that a standard normal random variable is less than -1 is approximately
      A. 0%
      B. 2.4%
      C. 16%
      D. None of the above

31. I have a small data set whose minimum value is 100 and whose maximum value is 160. Without any additional information, which is the most reasonable guess for what the standard deviation may be ?
      A. 12
      B. 60
      C. 144
      D. 3600

32. Suppose that several LHD's are polled regarding their numbers of employees (FTE). If the sample mean is 20 and the sample standard deviation is also 20, what do you conclude ?
      A. About 95% of LHD's have between -20 and +60 employees.
      B. The distribution is not normal.
      C. The distribution is skewed to the right.
      D. Both B and C are correct.

33. If (approximately) 25% of the values in a sample are below 80 and the interquartile range is reported as the single number 40, then above what number do (approximately) 25% of the values lie ?
      A. 40
      B. 80
      C. 120
      D. None of the above

34. Suppose the sample mean is 40 and the sample standard deviation is 10. Between what two numbers will we find about 95% of the data, if the data are close to normally distributed ?
      A. -60 and 140
      B. -160 and 240
      C. 30 and 50
      D. 20 and 60

35. Why is the pivot table capability of Excel used to examine a nominal categorical variable instead of the FREQUENCY command ?
      A. The FREQUENCY command does not accept non-numeric designations for categories.
      B. The pivot table capability automatically calculates the mean and standard deviation.
      C. I don't like having to remember to press Ctrl-Shift-Enter.
      D. Your instructor is getting tired of editing last year's questions !