Name _____

Self-Assessment. Of the original 10 practice questions, I answered _____ correctly on the first attempt. I spent approximately _____ hours preparing for this midterm examination.

Select your grading option:
_____ 1 2 out of 3 problems on first page, 140 pts multiple choice
_____ 2 all 3 problems on first page, 105 pts multiple choice
_____ 3 all 3 problems on first page, 140 pts multiple choice

Bayes' Theorem application. Suppose A = heart attack and B = positive stress test. Suppose, moreover, that P(A) = 0.10, P(B|A) = 0.80, and P({not B}|{not A}) = 0.95. By complements, P({not A}) = _____ and P(B|{not A}) = _____ . By Bayes' Theorem, P(A|B) = _____ .

Descriptive statistics. Considering the data set {64 73 77 84 88 91 97 103}, 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	64	80				
2	73	90				
3	77					
4	84					
5	88					
6	91					
7	97					
8	103					

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 0.159.
=1 - NORMDIST(1,0,1,TRUE)
Based on that result, complete the following sentence: The probability is 0.159 that a normal random variable will exceed _____.

Suppose I type the following in D1 and drag it down through D8.
=IF(A1<90,IF(A1>80,1,0),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<90,IF(A1>80,1,0),0)
Print in E1 through E8 what will be the results.

Name _____

Multiple Choice. Write your answer as a large capital letter to the left of the four options presented. Print "W" for any question you wish to waive. Grading option 1 permits you to waive five questions; options 2 and 3 permit you to waive ten questions. If I can't distinguish your answer, it's wrong. Each highlighted, italicized question was a practice item and should be disregarded.

1. If an emergency room director calculated the median wait time for patients as a measure of central tendency instead of the mean, what measure of variability might she have calculated ?

A. Standard deviation
B. Variance
C. Interquartile range
D. Mode

2. For what reason might lack of double blinding cause a physician-scientist to over-estimate the benefit of an experimental treatment ?

A. Evaluation of a patient's health requires the subjective visual assessment of an image, and a physician may "see what he wants to see".

B. Knowing that a patient is on placebo, a physician may be more likely to prescribe that patient another drug versus a patient receiving the experimental treatment.

C. Double blinding may be compromised if patients were randomized in blocks of size two.
D. None; lack of double blinding may cause under-estimation but not over-estimation of benefit.

3. 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

4. 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 7.2. The 7.2 is an estimate of

A. a parameter.
B. a population.
C. a sample.
D. a statistic.

5. Suppose a physician-scientist asks her patients whether they wish to watch a 30-minute video whose purpose is to motivate improvement in dietary and exercise habits. Her intent is to compare subsequent weight loss among those who watch the video versus those who do not. What threatens the validity of this comparison ?

A. A patient may choose to watch the video because he/she wants to improve dietary and exercise habits.
B. A patient may choose not to watch the video because he/she is pressed for time.
C. Both A. and B. may threaten validity.
D. While both statements may be true, neither A. nor B. actually threatens validity.

Name _____

6. 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 an explanatory variable.
C. Both A. and B. are correct.
D. Neither A. nor B. is correct.

7. Suppose that a mean is calculated. Which of the following may be true ?

A. The variable in question is interval but not ratio.
B. The variable in question is ordinal but not interval.
C. The variable in question is nominal but not ordinal.
D. Both B. and C. may be true.

8. 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.

9. Suppose the formula =B$2/C$2 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.

10. Suppose the formula =B2/C2 appears in cell D2. Which of the following modifications will guarantee that cell E2 contains the same numerical contents as cell D2, if you drag cell D2's formula to the right ?

A. No modification is necessary.
B. Change it to =B$2/C$2.
C. Change it to =$B2/$C2.
D. Both B. and C. will work.

11. Suppose cell F10 contains the number 8. What will be the result of =(F10 > 7)+0 ?

A. 0
B. 1
C. FALSE
D. TRUE

12. Suppose cell F10 contains the number 8. What will be the result of =IF(F10>7, 7, IF(F10 < -2, -2, F10)) ?

A. -2
B. 7
C. 8
D. None of the above.

Name _____

13. Suppose cell F10 contains the number 3. What will be the result of =IF(F10>7, 7, IF(F10 < -2, -2, F10)) ?
    A. -2
    B. 0
    C. 7
    D. None of the above.

14. Suppose an emergency room director wants to explore graphically whether the number 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.

15. If we take care to distinguish between $\bar{X}$ and $\bar{x}$, then the former is _____ and the latter is _____.
    A. a random quantity; a fixed number.
    B. a fixed number; a random quantity.
    C. what we have after observing the data; what we have before observing the data.
    D. Both B. and C. are correct.

16. Human beings who have a difficult time mimicking stochastic processes should
    A. avail themselves of RAND() or similar capabilities in Excel or other software.
    B. make sure their data avoid clusters.
    C. Both A. and B. are correct.
    D. watch more television.

17. 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.

18. The RAND function in Excel is useful for
    A. randomized treatment allocation in a cohort.
    B. randomized treatment allocation within strata.
    C. creating permuted blocks.
    D. all of the above.

19. Why are permuted blocks of size six a sound randomization strategy ?
    A. Breaking the blind on one patient does 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 _____

20. What is the probability that the RAND function yields a number between 0.4 and 0.9 ?
    A. 0.1
    B. 0.4
    C. 0.9
    D. None of the above

21. Why is working with primary data potentially advantageous over working with secondary data ?
    A. You have more opportunity to mitigate or eliminate confounding.
    B. You can usually complete the study more quickly.
    C. You can make sure that the data are recorded properly as you collect them.
    D. Both A. and C. are correct.

22. If you had measurements on systolic blood pressure, diastolic blood pressure, and heart rate, what sort of error would likely not be detected using the MIN and MAX functions on these variables ?
    A. Someone's DBP of 90 accidentally recorded as 9.
    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.

23. Suppose that you have a spreadsheet with 1000 records but that data on a critical variable are missing for 2 records, whence case deletion would leave you with 998 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 998 records.
    B. Abandon data analysis; the situation cannot be salvaged with 2 missing records.
    C. Perform a complicated Monte Carlo imputation on account of the 2 missing records.
    D. Enroll in his upcoming statistics class.

24. How would a distribution of annual incomes best be described ?
    A. Bell-shaped.
    B. Left skewed.
    C. Right skewed.
    D. Uniform.

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

26. In general, how can we tell whether the distribution of measurements in a population is normal ?
    A. Make a histogram of measurements in the population.
    B. Make a pie chart of measurements in the population.
    C. Make a histogram of measurements in a sample drawn from the population.
    D. Make a pie chart of measurements in a sample drawn from the population.

Name _____

27. Considering 100 families with two children each, suppose that 40 of the 200 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. 3
C. 8
D. 100

28. If P(C | D) denotes positive predictive value, then which of the following is specificity ?

A. P(D | C)
B. P({not D}|C)
C. P({not D}|{not C})
D. P({not C}|{not D})

29. Suppose that 30% of patients visiting a clinic are sedentary, 50% have poor dietary habits, and 60% 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.

30. Suppose that 30% of patients visiting a clinic are sedentary, 50% have poor dietary habits, and 60% have at least one of these two problems. What fraction of sedentary patients have poor dietary habits ?

A. 1/2
B. 3/5
C. 2/3
D. This cannot be determined from the information given.

31. Let X denote the number of true emergencies at a small rural hospital this afternoon, given that there are 10 visitors. Which of the following is the most reasonable probabilistic model for X ?

A. Binomial
B. Poisson
C. Normal
D. Uniform

32. Continuing from the preceding item, what can we say about the probability that $X \geq 8$, if the probability is 0.8 that any particular visit is a true emergency ?

A. Between 0% and 10%
B. Between 10% and 30%
C. Between 30% and 80%
D. Between 80% and 100%

Name _____

33. Among the next four 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. 35%
B. 41%
C. 59%
D. 65%

34. The probability that a standard normal random variable equals +1 is approximately

A. 0%
B. 2.4%
C. 16%
D. None of the above

35. 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

36. 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 variance may be ?

A. 12
B. 60
C. 144
D. 3600

37. 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 left.
D. Both B and C are correct.

38. Suppose hospital administrators are polled regarding lost reimbursements for hospital-acquired infections. If I report the median instead of the mean, I did so for which of the following reasons ?

A. Amount of lost reimbursement is an ordinal but not an interval variable.
B. I perceived that the distribution of lost reimbursements was skewed and, accordingly, thought that the mean was not a good measure of central tendency.
C. The median minimizes a sum of squared differences.
D. Both B and C are correct.

Name _____

39. If (approximately) 25% of the values in a sample are above 80 and the interquartile range is reported as the single number 40, then *below* what number do (approximately) 25% of the values lie ?

A. 40
B. 60
C. 80
D. 120

40. I learn from Excel that a Normal variable with mean 30 and standard deviation 15 has probability 34% of being below 24. What else can I say ?

A. This variable has probability 34% of being above 36.
B. This variable has probability 66% of being above 36.
C. This variable has probability 34% of being above 45.
D. This variable has probability 66% of being above 45.

41. I learn from Excel that a Normal variable with mean 50 has 84% chance of being below 70. What must be its standard deviation ?

A. 1
B. 10
C. 20
D. This cannot be determined from the information given.

42. Suppose that the number of emergency room arrivals is described by a Poisson process with rate 6 per hour. What is the probability that no one arrives during the next ten minutes ?

A. Less than 1%.
B. About 37%.
C. Nearly 100%.
D. This cannot be determined from the information given.

43. Suppose the sample mean is 40 and the sample variance is 100. Between what two numbers will we find about 68% 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

44. When the pivot table capability of Excel is used to examine two categorical variables, we can obtain

A. joint empirical probabilities.
B. marginal empirical probabilities.
C. conditional empirical probabilities.
D. all of the above.

45. 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. I'm getting tired of changing last year's questions !