STAB22F: 2002 Midterm Exam Solutions

$*$ = correct

Time: 2 hours
Instructions: Only use No. 2 black lead pencil on the Answer Sheet. First fill-in your student number and name on the Answer Sheet. Indicate your answer to each question by thoroughly filling in the appropriate circle on the Answer Sheet. Each question has the same weight. The exam paper must be turned in at the end of the exam.
Aids: The exam is open book. Students may use any notes, books and calculators (no laptops) in writing this exam.

1) Incomes in a certain software manufacturing company were found to have a distribution with a long right tail. Which of the following statements is most correct?

a) The mean is a good indication of the 'typical income'.
b) Less than 50% of the incomes are below the mean.
c)* Median and mean estimate different quantities.
d) Variation in the income is best indicated by the standard deviation.

2) For a normally distributed random variable X with mean 8 and standard deviation 5, the value of C such that $P(X < C) = 0.7$ can be determined using which of the following Minitab commands:?

a) `cdf 0.7; normal 0 1.`
b) `cdf 0.7; normal 8 5.`
c) `invcdf 0.7; normal 0 1.`
d)* `invcdf 0.7; normal 8 5.`

3) Let $P(A) = 0.3$, $P(B) = 0.4$ and $P(A \cap B) = 0.7$. Which of the following statements is always true?
a)* The events A and B are mutually exclusive but not independent.
b) The events A and B are both mutually exclusive and independent.
c) The events A and B are not mutually exclusive but independent.
d) The events A and B are not mutually exclusive and not independent.

4) The following summary statistics were obtained for a data set.
$\overline{X} = 80.5 \qquad \text{median} = 84 \qquad s = 10.5 \qquad Q_1 = 75.5 \qquad Q_3 = 96.0.$
Then the percentage of observations
a) below 96.0 is 25%.
b)* above 75.5 is 75%.
c) in the interval 75.5 to 96.0 is 75%.
d) in the interval 75.5 to 84 is 50%.

5) Suppose in a math course the class average and standard deviation were $\overline{X} = 60$, $s = 20$ respectively. A student obtained a $z-$score of 1.2, Then the actual raw score would be approximately

a) 61.
b) 36.
c) 58.
d)* 84.

6) During the early development of a medicated skin patch to help smokers break the habit, a test was conducted with 112 volunteers. Half the volunteers were given an unmedicated patch. The other subjects were assigned the medicated one. At the end of the study, the number of persons in each group who were abstinent and those who were smoking were recorded. Select the correct statement.

a) No placebo was used in the study.
b) The study is an observational one.
c)* Subjects, were randomly assigned into the two groups using the Minitab command

```
sample  112   c1   c2
```

d) None of the above.

7) A census is carried out on a population of size 720 and the values of two categorical variables X and Y are measured. Suppose that the following table of counts is obtained.

	$X = 1$	$X = 2$	$X = 3$
$Y = 1$	60	60	100
$Y = 2$	80	120	200
$Y = 3$	20	30	50

Which one of the following statements is correct?

a)* A relationship exists between X and Y.
b) Whether or not a relationship exists depends upon which variable we select as the predictor and which variable we select as the response.
c) No relationship exists between X and Y.
d) We really need to do a statistical test to conclude anything.

8) The normal probability plot is one of the most frequently used statistical methods in regression analysis. Which one of the following statements is correct for this plot?

a) It checks whether the residuals are randomly distributed.
b) It checks whether the estimated regression coefficients are randomly distributed.
c)* It checks whether the errors came from a $N(\mu, \sigma)$ distribution.
d) It checks whether sample data gives a linear regression line.

9) Suppose the probability of raining today is 0.30, the probability of raining tomorrow is 0.25, and the probability of raining both today and tomorrow is 0.20. Then the probability of raining tomorrow given that it rains today is

a) 0.35
b) 0.55
c)* 0.67
d) 0.80

10) A mall has two elevators. Let A be the event that the first elevator is busy, and let B be the event that the second elevator is busy. Assume that $P(A) = 0.3,\ P(B) = 0.4$, and $P(A \cup B) = 0.6$. Then the probability that both the elevators are busy is

a)* 0.10
b) 0.70
c) 0.60
d) 0.50

11) Why is it important that we use density histograms rather than count histograms when plotting the distributions of quantitative variables?

a) Because density is easier to compute.
b) Because density is expressed in appropriate units.
c) Because distributions require density.
d)* Because density leads to areas under the graph equaling proportions.

12) A mechanical system is composed of 3 components, all of which must work in order for the system to function. The probability that each component works is 0.9. What is the probability that the system will FAIL to function?

a) $(0.9)^3$
b) $(0.1)^3$
c)* $1 - (0.9)^3$
d) $1 - 3(0.9)$

13) An agronomist wishes to investigate whether the level of humidity in the environment (X) affects the growth rate (Y) of fungus. To determine this one should examine

a) the linear regression of Y with X.
b) a measure of association between Y and X given by the coefficient of linear correlation (r).
c)* the conditional distribution of Y given X for various values of X.
d) none of the above.

14) The number of days, X, that it takes the post office to deliver a letter between city A and city B has the probability distribution.

x	2	3	4	5
$f(x)$	.1	.4	.3	.2

Then the expected number of days it takes to deliver a letter is

a) 3.
b)* 3.6.
c) 4.
d) 4.6.
e) none of the above.

15) Based on **14)** the median number of days he takes to deliver a letter is

a) 4.5.
b) 4.
c) 3.5.
d)* 3.
e) none of the above.

16) The annual rainfall (in inches) in a city is normally distributed with mean $\mu = 50$ inches and standard deviation $\sigma = 5$ inches. To compute the probability that the rainfall next year will exceed 59.5 inches the appropriate Minitab command would be

a) `pdf  59.5;  normal  50  5.`
b) `cdf  59.5;  normal  50  5.`
c) `invcdf  59.5;  normal 50  5.`
d)* none of the above.

17) Let X be a random variable with $E[X] = 3$ and standard deviation 0.85. For a new random variable $Y = 2X + 5$, the correct value for the expected value and standard deviation is:

a) 11, 6.7
b)* 11, 1.7
c) 3, 6.7
d) 3, 1.7
e) none of the above.

18) An exponential model of the form

$$Y = a\exp(bX)$$

for $a > 0$, $b > 0$ was fit to data for the growth of motor vehicle registration with time. Using the log transformation, the least squares line fitted to the data is

$$\ln(y) = -30.62 + 0.016(x).$$

Then the fitted model for motor vehicle registration is

a) $\widehat{Y} = -30.62 \exp(0.016X)$.
b) $\widehat{Y} = \log_e(-30.62) \exp(0.016X)$.
c)* $\widehat{Y} = \exp(-30.62) \exp(0.016X)$.
d) none of the above.

19) Based on **18)** above which of the following is true?

a) A plot of the number of vehicles registered against time will be linear with a positive correlation.
b) A plot of the number of vehicles registered against time will be linear with a negative correlation.
c) A plot of logarithm of the number of vehicles registered against time is linear with a negative correlation.
d)* A plot of logarithm of the number of vehicles registered against time is linear with a positive correlation.

20) In a factory, a defective article is required to pass through two independent inspections before being packaged. Experience shows that one inspector will miss 4% of the defective articles, whereas the second inspector will miss 5% of them. The probability that a defective article will be missed by both inspectors is

a) 0.02
b) 0.09
c) 1.0
d)* none of the above.

21) The correlation coefficient between X and Y is either $r_{xy} = 0.6$ or $r_{xy} = -0.9$. Which of the following is true?

a)* One can predict Y from X with a higher ability when $r_{xy} = -0.9$ than when $r_{xy} = 0.6$
b) The % of variation in Y accounted for by the knowledge of X is greatest when $r_{xy} = 0.6$
c) Linearly predicting Y from X for $r_{xy} = -0.9$ gives poor results since we have a weak negative correlation.
d) None of the above.

22) Suppose that we flip a fair coin five independent times and record whether or not exactly 3 heads are observed. What is the long-run proportion of times that we observe 3 heads in repeated performances?

a) 1/16
b) 1/2
c) 1/4
d)* 5/16
e) None of the above.

23) While a simple linear regression model may be appropriate, it may not be very useful for predictive purposes. This is because

a)* the population variance for the error term may be so large as to dominate the total amount of variation observed in the response.
b) there are too many variables in the model.
c) the existence of random variation never permits prediction to be effectively carried out.
d) the underlying relationship between the predictors and the response may be too complicated.

24) The sampling distribution of a statistic calculated from a sample of size n from a population refers to

a) the method for selecting the n population members.
b) the fact that the n measurements obtained typically vary.
c) the distribution of the quantity being measured.
d)* the pattern of variation exhibited by the statistic when taking repeated samples of size n.
e) the fact that the normal distribution is often an appropriate distribution for a statistic.

25) The importance of simulation as a computational technique is that

a) it is fairly simple to do.
b)* it allows us to compute characteristics of a sampling distribution when sampling from a particular population distribution.
c) it allows us to avoid the use of tables of probability values.
d) none of the above.

26) The following Minitab commands

```
MTB > random 1000 c1-c4;
SUBC> normal 0 3.
MTB > let c5=rmedian(c1,c2,c3,c4)
MTB > let k1=mean(c5)
MTB > print k1
Data Display
K1 -0.00508136
```

a)* generate 1000 samples of size 4 from a $N(0, 3)$ distribution and estimate the mean of the sampling distribution of the sample median.
b) generate 4 samples of size 1000 from a $N(0, 3)$ distribution and estimate the mean of the sampling distribution of the sample median.
c) generate 1000 samples of size 4 from a $N(0, 3)$ distribution and estimate the mean of the sampling distribution of the sample mean.
d) generate 4 samples of size 1000 from a $N(0, 3)$ distribution and estimate the mean of the sampling distribution of the sample mean.

27) Suppose that X is a random variable that follows a $N(-6, 2)$ distribution. If the .975 quantile of a $N(0, 1)$ distribution is given by 1.96 then the .975 quantile of the distribution of X is given by

a) 1.96.
b) -6.
c) $2(1.96)$.
d)* $-6 + 2(1.96)$.
e) none of the above.

28) To determine whether or not a relationship exists we must conduct at least

a) an observational study.
b)* a sampling study.
c) an experiment.

29) In an observational study an apparent relationship

a) may not be real.
b)* may not apply to the population of interest.
c) may be quadratic rather than linear.
d) is a cause-effect relationship.

30) In an observational study an apparent relationship

a) may not be real.
b)* may not apply to the population of interest.
c) may be quadratic rather than linear.
d) is a cause-effect relationship.

31) The following output was obtained from regressing the data in c2 on the data in c1 in a Minitab worksheet.

```
MTB > regress c2 1 c1
Regression Analysis:  C2 versus C1
The regression equation is
C2 = 0.778 + 0.889 C1
Predictor    Coef    SE Coef     T      P
Constant   0.7775    0.4071   1.91  0.093
C1         0.88896   0.06561 13.55  0.000
S = 0.5959 R-Sq = 95.8% R-Sq(adj) = 95.3%
```

What is the slope of the least squares line?

a) .778
b)* .889
c) .06561
d) 95.8%

32) Using the output in **30)** what proportion of the variation in c2 is explained by changes in c1 via the linear dependence?

a) .5959.
b)* .958.
c) .88896.
d) .7775.

33) Using the output in **30)** what is the predicted value of c2 when c1=1?

a)* 1.666
b) .958.
c) .88896.
d) .7775.

34) Suppose you produce a boxplot of a variable in a data set and find several data points beyond the outer values. What does this tell you?

a) We are extrapolating at these values.
b) There are some highly influential observations in the data set.
c)* There are some outliers in the data set.
d) The variable is approximately normally distributed.

35) Why is it important that we use the same scales on graphs whenever we are comparing two data sets via plots?

a) Because there is no way the data can be normally distributed otherwise.
b) This also us to determine if a cause-effect relationship exists between the variables.
c)* Different scales can change the appearance of a plot considerably and render comparisons meaningless.
d) Otherwise one data set may be more variable than the other and the plots won't detect this.