STAB22F:1999 Final Exam

Time: 3 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 is worth 3 marks. There are 102 marks on the exam but your mark will be taken out of 100. This 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 in writing this exam.

1. Suppose we have two variables X and Y defined on a population Π and we want to know if there is a relationship existing between these variables. Which one of the following statements is correct?

(a) If the least-squares line between X and Y has a nonzero slope then a relationship exists.
(b) If the conditional distributions of X given Y change as we change the value of Y that we condition on then a relationship exists.
(c) If the values of X can be controlled; i.e. assigned to population members, then we know that a relationship exists.

2. To assert that a cause effect relationship exists between two variables X and Y defined on a population, we need to be able to assign the values of X to population members and then observe Y as this ensures that no other variables could be causing X and Y to change simultaneously.

(a) True
(b) False

3. The correlation coefficient measures the degree to which a relationship exists between two quantitative variables no matter what form this relationship takes.

(a) True
(b) False

4. Suppose that a random variable X follows the distribution given by

X	-2	-1	0	1	2
$P(X = x)$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$

The value of $P(X^2 > 1)$ is

(a) $\frac{1}{8}$
(b) $\frac{1}{4}$
(c) $\frac{1}{2}$

5. In repeated sampling the long-run average value of X in **4** is

(a) 0
(b) $\frac{1}{4}$
(c) 1

6. Suppose that a value of X in **4** has occurred and that the value is an even number. Then the probability that the value of X is 0 is

(a) $\frac{1}{4}$
(b) 0
(c) $\frac{1}{2}$

7. The distribution function of X in **4** evaluated at 1 is

(a) $\frac{1}{4}$
(b) 1
(c) $\frac{7}{8}$

8. Suppose that a sample of 2 is taken from the distribution in **4** obtaining X_1 and X_2. Then $Var[X_1 + 2X_2]$ equals

(a) 1.5
(b) 7.5
(c) 4.5

9. Suppose we want to determine the probability of obtaining a head from a particular coin and we toss the coin n times. Which is the smallest sample size below that guarantees we will know the true probability of a head to within .1 with virtual certainty on the basis of these tosses?

(a) 100
(b) 196
(c) 225
(d) 500

10. Suppose that there are 25 chips in a bowl with 10 chips white and 15 chips black. If a sample of size 10 is drawn then the number of white chips in the sample is distributed according to the Binomial$(10, .4)$ distribution.

(a) False
(b) True

11. Suppose that $X \sim$ Binomial$(10, .5)$ statistically independent of $Y \sim$ Binomial$(20, .5)$. Then $W = X + Y$ is distributed

(a) Binomial$(30, 1)$.
(b) Binomial$(10, .5)$.
(c) Binomial$(30, .5)$.
(d) Normal$(5, 4)$.

12. To generate samples from a distribution using Minitab we use the

(a) **sample** command.
(b) **random** command.
(c) **normal** command.
(d) **discrete** command.

13. We use randomization to select individual items from a population or to assign treatments to experimental units because this ensures that we are obtaining a sample of values from the true population distribution of whatever quantity we are measuring.

(a) True
(b) False

14. The Minitab **tinterval** command is used to calculate confidence intervals for a population mean

(a) when the sample size n is smaller than 30.
(b) when the **zinterval** command is not appropriate.
(c) when we feel confident that we have a sample from a normal distribution.

15. Distribution-free methods don't require very specific distributional assumptions or large sample sizes to be appropriate but

(a) the calculations are often difficult to implement.
(b) they aren't very powerful.
(c) they don't allow us to make inferences about population quantities of direct interest.

16. Suppose we know that $X \sim N(2, 3)$. To compute the point that has .75 of the area under the density curve to the left of it we use the Minitab commands

(a)
```
MTB >invcdf .75;
SUBC>normal 2 3.
```
(b)
```
MTB >cdf .75;
SUBC>normal 2 3.
```
(c)
```
MTB >pdf .75;
SUBC>normal 2 3.
```

17. We estimate the .27 quantile of a distribution we are sampling from by

(a) calculating the sample mean $\bar{x}$ and sample standard deviation s and then using an approximate $N(\bar{x}, s)$ distribution together with the **invcdf** command in Minitab.
(b) via a simulation.
(c) we calculate the empirical distribution function based on the sample and determine the point in the sample which has the value of this function closest to .27.

18. The difference between statistical significance and practical significance

(a) is that a practically significant result affects our daily lives while a statistically significant result is only concerned with statistics.
(b) is negligible.
(c) a statistically significant result only indicates that the data indicate that there is evidence against a null hypothesis not whether the deviation from the null hypothesis is of any practical importance.

19. The power of a test refers to its ability to detect that the null hypothesis is true when it is indeed true.

(a) True
(b)False

20. Suppose we have two quantitative variables X and Y defined on a single population Π and we want to test if the mean of X equals the mean of Y. The reason we prefer to take a single sample from Π and base our inferences on the differences $X - Y$ as opposed to taking two samples and comparing sample means is because

(a) it is cheaper to take only a single sample.
(b) the central Limit Theorem is more likely to apply with a single sample.
(c) when taking two samples a person could appear in both samples.
(d) our inferences are more accurate as we have removed the effects of individual variation in the comparisons.

21. Suppose we have two categorical variables X and Y defined on a single population Π and we want to assess whether or not there is a relationship between X and Y. If X takes 6 values and Y takes 4 values then the degrees of freedom in the Chisquare test of the null hypothesis of no relationship equals

(a) 24.
(b) 23.
(c) 15.
(d) 2.

22. For the variables X and Y in **21** the null hypothesis of no relationship is equivalent to saying

(a) X and Y are statistically independent.
(b) the correlation of X and Y is 0.
(c) the conditional distributions of X given Y are similar.

23. For the variables X and Y in **21** if we decide that there is a relationship between these variables then a good way to depict the relationship graphically is

(a) in a scatter plot.
(b) in a stem-and-leaf plot..
(c) present barplots of the conditional distributions of one variable given the other.

24. Below is a 2×2 table summarizing the counts obtained from a single categorical variable (columns) defined on two populations (rows)

23	27
15	35

This table was placed in the first two rows of columns C1 and C2 of a Minitab worksheet. The following command

```
MTB >chisquare c1 c2
```

was given which produced the output

```
Chi-Square Test:  C1, C2
Expected counts are printed below observed counts
      C1     C2    Total
 1    23     27    50
      19.00  31.00
 2    15     35    50
      19.00  31.00
Total 38     62    100
Chi-Sq = 0.842 + 0.516 +
 0.842 + 0.516 = 2.716
DF = 1, P-Value = 0.099
```

The expected number of observations in category 1 of population 1 under the null hypothesis is

(a) 27.
(b) 19.
(c) 23

25. For the data and output in **24** the value of the Chisquare statistic is

(a) .099.
(b) 0.842.
(c) 2.716.
(d) 1

26. For the data and output in **24** we reject the null hypothesis of no difference between the population distributions at the 5% level.

(a) True
(b) False

27. In a regression model the linear assumption refers to the fact that we are fitting a line.

(a) True
(b) False

28. Suppose we fit the model $Y = \beta_1 + \beta_2 U + \beta_3 V + \epsilon$, where U, V are quantitative predictor variables and $\epsilon \sim N(0, \sigma)$, to a set of data. The first thing we do in our analysis is

(a) test the null hypothesis of no effect due to U and V.
(b) calculate confidence intervals for the β_i.
(c) test the model assumptions by looking at various plots of the residuals.
(d) test the model assumptions by looking at various plots of the standardized residuals.

29. By standardizing residuals we mean

(a) make the residuals all look the same.
(b) divide each residual by its standard error.
(c) make sure the residuals sum to 0.

30. For the model in **28** suppose we fit this using Minitab with the response Y in C3, U in C1, V in C2 and as part of our output obtained the plot in Figure 1. What is your conclusion?

(a) Looks like there is a relationship between C3 and C2..
(b) There is a problem with the model assumptions as it looks like there might be a correlation between C1 and C2.
(c) There is a problem with the model assumptions as it looks like there might be a more complicated relationship between Y and V.
(d) There are no problems with the model.

31. For the model of 28 suppose we gave the command

```
MTB >regress c3 2 c1 c2
```

and obtained the output

```
The regression equation is
C3 = 159168 - 18451 C1 + 455 C2
Predictor   Coef     SE Coef       T       P
Constant   159168    17608       9.04 0.000
C1        -18451.4     804.7   -22.93 0.000
C2         454.501       7.720  58.88 0.000
S = 57524 R-Sq = 99.6% R-Sq(adj) = 99.6%
Analysis of Variance
Source          DF      SS              MS            F     P
Regression       2  7.42706E+13  3.71353E+13   11222.34  0.000
Residual Error 97  3.20978E+11    3309053618
Total           99  7.45916E+13
```

From this we can conclude that

(a) over 99% of the variation in Y can be explained by the changes in U and V through the model.
(b) the model doesn't allow us to predict Y from U and V very well.
(c) the number of degrees of freedom for regression seems too small to allow for accurate inferences.

32. From the output in **31** we conclude that there is a positive relationship between Y and U.

(a) True
(b) False

33. From the output in **31** we estimate σ by.

(a) 3.20978E+11
(b) 3309053618
(c) 57524
(d) 7.45916E+13

34. The problem of multiple comparisons refers to the fact that if we carry out many tests of significance, where we decide a difference exists if the P-value is less than 5%, then even if all the null hypotheses are true we will conclude that approximately 5% of them are false.

(a) True
(b) False