STAB22F: 2000 (summer) 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 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 in writing this exam.

1. Suppose that an urn contains 14 black balls and 7 white balls. If 5 balls are drawn, with replacement after each draw, then the probability of obtaining 2 white balls is:

(a) $5\frac{7}{20}\frac{13}{20}$
(b) $\left(\frac{7}{20}\right)^2\left(\frac{13}{20}\right)^3$
(c)* $10\left(\frac{7}{20}\right)^2\left(\frac{13}{20}\right)^3$
(d) $5\left(\frac{7}{20}\right)^2\left(\frac{13}{20}\right)^3$

2. Suppose that an urn contains 14 black balls and 7 white balls. If 5 balls are drawn, without replacement after each draw, then the number of white balls obtained is approximately binomially distributed.

(a) True
(b)* False

3. Suppose that a random variable X has the following probability distribution:

X	1	2	3	4	5
$P(X = x)$	.25	.5	.05	.1	.1

It makes sense to approximate this distribution via a normal density curve.

(a) True
(b)* False

4. For the probability distribution specified in **3** the mean of X is given by:

(a) 1.4
(b) 1.6
(c)* 2.3
(d) 4.0

5. Suppose that a sample X_1, X_2 is generated from the probability distribution given in **3**. Then $P(X_1 + X_2 = 3)$ equals:

(a)* 1/4
(b) 3/4
(c) 1/10

6. For the probability distribution specified in **3,** the distribution function at 4 is given by:

(a) .2
(b) .1
(c)* .9

7. For the probability distribution specified in **3,** the third quartile is given by:

(a)* 2
(b) 3
(c) 4

8. Suppose that $X \sim N(2.2, 1)$ and $Y = X^3 + X$. Then the appropriate Minitab commands for generating a sample of 100 from the distribution of Y are:

(a)
```
MTB >let k1=100**3
MTB >random k1 c1;
SUBC>normal 2.2 1.1.
```
(b)
```
MTB >random 100 c1;
SUBC>normal (2.2)**3 (1.1)**3.
```
(c)*
```
MTB >random 100 c1;
SUBC>normal 2.2 1.1.
MTB >let c2=c1**3 + c1
```

9. If the .8 quantile of a $N(2,1)$ distribution is 2.8416 then the .2 quantile of a $N(1,1)$ distribution is:

(a) 1.8416
(b) -1.8416
(c)* 0.158400
(d) -0.158400

10. An experimenter has 30 experimental units and 3 treatments each of which is going to applied 10 times. After labelling the experimental units $1, \ldots, 30$ which of the following Minitab commands can the experimenter use to make the assignment of treatments to experimental units.

(a)

```
MTB >set c1
DATA>1:30
DATA>end
MTB >discrete 30 c1 c2
```

(b)

```
MTB >random 30 c1;
SUBC>integer 1 30.
```

(c)*

```
MTB >set c1
DATA>1:30
DATA>end
MTB >sample 30 c1 c2
```

11. Suppose that we have a sample $X_1, \ldots, X_{10}$ from a distribution and we compute the estimate $T(X_1, \ldots, X_{10})$ of a quantity associated with the population distribution we are sampling from. Which statement below is false?

(a) The sampling distribution of T gives an indication of how accurate $T(X_1, \ldots, X_{10})$ is as an estimate.
(b) When T is an average we can sometimes use a normal distribution to approximate the distribution of T.
(c)* The sample standard deviation gives us an indication of how accurate the estimate $T(X_1, \ldots, X_{10})$ is.
(d) If we are sampling from a discrete population distribution then the Central Limit Theorem can sometimes be used to approximate the sampling distribution of T.

12. The following Minitab commands

```
MTB >random 1000 c1-c12;
SUBC>binomial 4 .3.
MTB >rsum c1-c12 c13
MTB >rssq c1-c12 c14
MTB >let c15=c13/c14
```

(a) generate a sample of size 12 from the distribution of $\sum_{i=1}^{1000} X_i / \sum_{i=1}^{1000} X_i^2$ when $X_1, \ldots, X_{1000}$ is a sample from the Binomial$(4, .3)$ distribution.
(b) generate a sample of size 1000 from the distribution of $\sum_{i=1}^{4} X_i / \sum_{i=1}^{4} X_i^2$ when $X_1, \ldots, X_4$ is a sample from the Binomial$(12, .3)$ distribution.
(c) generate a sample of size 12 from the distribution of $\sum_{i=1}^{4} X_i / \sum_{i=1}^{4} X_i^2$ when $X_1, \ldots, X_4$ is a sample from the Binomial$(1000, .3)$ distribution.
(d)* generate a sample of size 1000 from the distribution of $\sum_{i=1}^{12} X_i / \sum_{i=1}^{12} X_i^2$ when $X_1, \ldots, X_{12}$ is a sample from the Binomial$(4, .3)$ distribution.

13. When we carry out a simulation to approximate the sampling distribution of a statistic, the convergence of the proportion of simulated values of the statistic falling in any given interval, as we increase the simulation size, is justified by the Central Limit Theorem.

(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. Suppose we are tossing a possibly biased coin. How many tosses just guarantees that the interval given by the proportion of heads $\pm$ a standard error is shorter than .1?

(a) 25
(b) 0
(c)* 100
(d) 1000

16. Suppose you were to conduct a study to investigate which of two detergents was better. Choose the most appropriate study.

(a) A random sample of 200 individuals is selected and 100 are randomly assigned to each detergent. Each participant was then asked to score the relative cleanliness of a soiled surface after observing the detergent in action.
(b)* A random sample of 200 individuals is selected and each participant scores both detergents, with respect to their ability to clean a solied surface, where the order of observation is random.

17. Suppose you are sampling from a $N(\theta, 2)$ distribution where θ is unknown and we want to test the null hypothesis $H_0 : \theta = 3$ versus $H_a : \theta > 3$. If we have decided to reject H_0 whenever the P-value is less than .05 then the power of this test at $\theta = 4$ is

(a)* the probability of rejecting H_0 when $\theta = 4$.
(b) .95.
(c) impossible to determine from these ingredients.

18. A P-value is a measure of

(a)* how surprising the observed value of a test statistic is when compared to its sampling distribution.
(b) the amount of probability in the tail of a distribution.
(c) how practically significant an observed deviation from a null hypothesis is.

19. Suppose we wish to test whether or not a particular coin is biased or not. The results of 1000 tosses are observed and we reject the null hypothesis that the probability of a head on a single toss is .5. Which of the following is the most appropriate concern in this context?

(a) The test is not powerful enough.
(b) The sample size is too small if we want accurate results.
(c)* A very small deviation from .5 may have been detected while for practical purposes the coin is fair.

20. The primary reason that we construct a confidence interval for an unknown population quantity is that

(a) it makes for a convenient summary of all the information in the sample about the population unknown.
(b)* we use the length of the confidence interval to assess the accuracy of an estimate of the population quantity.
(c) we wish to avoid the effects of biased selection.
(d) they are less complicated than tests of significance.

21. Suppose that we obtained a sample from a distribution and we wanted to draw inferences about the mean of this distribution. Suppose further that a normal probability plot indicated that we could not assume that the population distribution is normal. Our next step is to

(a) make use of the large sample inferences via the Central Limit Theorem.
(b) use the sign test and confidence interval.
(c)* look for a transformation that allows us to use the normality assumption for the transformed variable.

22. Suppose that a sample of 10 is drawn from a $N(\theta, 1)$ distribution where θ is unknown and suppose we obtained the values $\bar{X} = 5.2, S = 1.5$. Which of the following statements is the most accurate.

(a) The true value of θ lies in the interval $5.2 \pm 3 \left(1.5/\sqrt{10}\right)$.
(b)* The true value of θ lies in the interval $5.2 \pm 3 \left(1/\sqrt{10}\right)$.
(c) The true value of θ lies in the interval $5.2 \pm 1.96 \left(1.5/\sqrt{10}\right)$ with confidence .95.

23. Suppose that we have samples of a normally distributed variable X from each of two populations and we want to compare the means of X in these populations. The reason, in general, that we do not use the two sample t-test with pooling for testing whether or not the means are equal is that

(a) the test with pooling is not as sensitive as the test without pooling.
(b)* pooling requires that the standard deviation of X be the same in both populations.
(c) pooling is not valid unless the sample sizes are very large.

24. Suppose that we conduct a census in a population measuring the values of two binary valued variables X and Y and the following counts are obtained.

Variables	$X = 1$	$X = 2$
$Y = 0$	40	60
$Y = 1$	280	420

Are X and Y related?

(a) Yes
(b)* No

25. Suppose that a sample of size 100 is taken from each of two populations and the values of a categorical variable X are measured obtaining the following counts.

	$X = 1$	$X = 2$	$X = 3$
Population 1	20	40	40
Population 2	40	20	40

The following was obtained using Minitab.

```
MTB >chisquare c1-c3
Chi-Square Test:  C1, C2, C3
Expected counts are printed below observed counts
      C1     C2     C3      Total
  1   20     40     40       100
      30.00  30.00  40.00
  2   40     20     40       100
      30.00  30.00  40.00
Total 60   60     80       200
Chi-Sq = 3.333 + 3.333 + 0.000 +
 3.333 + 3.333 + 0.000 = 13.333
DF = 2, P-Value = 0.001
```

Choose the most appropriate statement.

(a) The degrees of freedom for this test are only 2 so it is not a very powerful test.
(b) The fact that some cells contribute 0 to the Chi-Square statistic indicates that there might be some problems with the assumptions.
(c)* The null hypothesis of no difference amongst the conditional distributions of X given population is rejected at the 1% level.

26. For the data and output in **25** the value 3.333 for the cell in the upper left-hand corner equals

(a)* (count - expected count)^2/(expected count).
(b) the residual for this cell.

27. Suppose that after carrying out a Chi-square test to determine whether or not two categorical variables defined on a population are related we decide that they are related. Our next step is to

(a)* examine the conditional distributions of one variable given the other to see how they change.
(b) see which cells contribute large residuals and so make the Chi-square statistic large.
(c) transform the data to try and eliminate the relation.

28. Suppose we fit the model $Y = \beta_1 + \beta_2 X + \epsilon$, where X is a quantitative predictor variable and $\epsilon \sim N(0, \sigma)$, to a set of data and obtained the following plot. What is the most appropriate conclusion?

(a) The predictor variable seems to be having an effect.
(b) There are clearly outliers in this data set and so the model is not appropriate.
(c)* There is a discernible pattern and so the model is not appropriate.
(d) The model would appear to be a reasonable one for this data set.

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. Suppose that we fit the model $Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \epsilon$, where X_1, X_2 are quantitative predictor variables and $\epsilon \sim N(0, \sigma)$, to a set of data and decided that the model is appropriate to that we can make inference about the unknowns. The following output was obtained.

```
The regression equation is
C4 = - 0.580 + 1.16 C1 + 1.99 C2
Predictor     Coef    SE Coef        T      P
Constant   -0.5801    0.8123     -0.71  0.485
C1          1.1569    0.1782      6.49  0.000
C2          1.99085   0.00824   241.59  0.000

S = 1.092 R-Sq = 100.0% R-Sq(adj) = 100.0%

Analysis of Variance
Source          DF      SS            MS          F        P
Regression       2 1297156        648578 544038.41    0.000
Residual Error  17      20             1
Total           19 1297176
```

Which of the following statements is correct?

(a) There is no relationship between the response and the predictors.
(b) The F-statistic for testing if there is any relationship takes the value 648578.
(c)* The degrees of freedom for the F-test are 2 and 17..
(d) The variance is extremely large.

31. From the output in **30** we can determine that the least-squares estimate of the regression coefficient β_3 is:

(a) -0.5801
(b)* 1.99085
(c) 0.1782
(d) 241.59

32. From the output in **30** we conclude that the most accurate inferences can be made about:

(a) β_1
(b) β_2
(c)* β_3

33. From the output in **30** we estimate σ by:

(a) 1.000
(b)* 1.092
(c) 648578
(d) 544038.41

34. In a regression analysis we use an F-test to test whether or not regression coefficients can be taken to be 0.

(a)* True
(b) False

35. Suppose as part of a fitting a regression model with quantitative predictors in C1 and C2 we obtain the following partial output from the describe command in Minitab.

```
Variable    Minimum      Maximum
C1            -2.859       2.013
C2          -17.80        35.39
```

If we wish to predict the value of the response at the combination C1 = 0, C2 = 40 then this is an example of an

(a) interpolation.
(b) estimation.
(c)* extrapolation.

36. Suppose that we fit a regression model and on the basis of the residual analysis we decide that the model is appropriate. Further as part of our subsequent analysis we decide that all the predictors are related to the response. If we obtained $R^2 = .08$ then we also conclude that

(a) the model doesn't fit the data.
(b)* the model won't work very well for prediction purposes.
(c) the model provides excellent predictions.
(d) the data used to fit this model wasn't very variable.

37. The way we assess whether or not an observation in a data set is influential in a regression analysis is to

(a) look at residual plots.
(b) look at standardized residual plots.
(c) look at normal probability plots.
(d)* redo the regression analysis without that data value and see if the results change markedly.

38. Suppose we have a quantitative response variable in C5 and a categorical predictor in C6 and we obtained the following plot as part of our analysis.

What is the most appropriate conclusion to draw?

(a) There appears to be some evidence of a difference in location amongst the four conditional distributions of the response given the predictor.
(b)* There appears to be some evidence of a difference in location and scale amongst the four conditional distributions of the response given the predictor.
(c) There appears to be no evidence of a difference in location amongst the four conditional distributions of the response given the predictor.
(d) This plot indicates that the model is not appropriate.

39. Suppose in a 1-way ANOVA we used a P-value of .05 to determine if two means are different or not. This gave rise to a family error rate of .201. Choose the most appropriate statement.

(a) The probability of detecting a single difference between two means, when there are no differences, is .201.
(b)* The probability of detecting at least one difference between two means, when there are no differences, is .201.
(c) The probability of detecting a single difference between two means is .201.
(d) The probability of detecting at least one difference between two means is .201.

40. We can reduce a family error rate when carrying out multiple comparisons by

(a) increasing the sample size.
(b) carrying out a more sensitive test.
(c)* decreasing the P-value at which we declare means to be different.
(d) including more variables in the analysis.