KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DEPARTMENT OF MATHEMATICS

STAT 212 BUSINESS STATISTICS II
Semester 222, Final Exam
Saturday May 27, 2023

Time allowed 120 minutes.

Name:_____ ID#: _____

Section#:_____ Serial#: _____

Important Notes:

- Formula sheet will be provided to you in exam. You are not allowed to bring, with you, formula sheet or any other printed/written paper.
- Make sure you have 15 pages of exam paper (including this title page) and 24 questions.
- If exact answer is not in the given options then go for the nearest option.
- Students are not allowed to enter the exam hall without either KFUPM ID or Saudi ID/ Iqama ID.
- Students must take the exam in the place assigned to them.
- Students are not allowed to carry mobiles, smart watches, or electronic devices to the exam halls/rooms.
- Violations of these rules will result in a penalty decided by the chairman of Math Department.

1. An auditor for a government agency is assigned the task of evaluating reimbursement for office visits to physicians paid by Medicare. The audit was conducted on a sample of 25 of the reimbursements, with the following results:

The amount of reimbursement was $\bar{X} = \$93$, $S = \$34.5$ and population is normally distributed. At the 0.05 level of significance, is there evidence that the mean of incorrect reimbursements in the population was greater than 100?
The test statistic for this test equals to

a. -1.0145
b. 1.7321
c. 1.4174
d. 1.0145
e. -1.7321

2. The owner of a gasoline station wants to study gasoline purchasing habits of motorists at his station. He selects a random sample of 40 motorists during a certain week, with the following results:
Fifteen motorists purchased premium-grade gasoline. If we want to test that more than 20% of the motorists at the station purchased premium-grade gasoline, then the ***p*-value** of the test is

a. 0.0028
b. 0.9972
c. 0.9780
d. 0.9890
e. 0.489

3. A problem with a telephone line that prevents a customer from receiving or making calls is upsetting to both the customer and the telephone company. The table below contains samples of 20 problems reported to two different offices of a telephone company and the time to clear these problems (in minutes) from the customers' lines:

	Sample Mean	**Sample Variance**
Central Office I Time (min.)	2.2140	2.9517
Central Office II Time (min.)	2.0115	3.5786

Use α = 0.05 and assume that the population variances from both offices are equal. If the company wants to test a difference in the **variance** of the waiting time between the two offices, then the **test statistic** is

a. 1.212
b. 0.825
c. 1.101
d. 0.908
e. 2.211

4. A study was conducted on the use of cell phones for accessing news. The study reported that 47% of users under age 50 accessed news on their cell phones (population 1) and 15% of users age 50 and over accessed news on their cell phones (population 2). Suppose that the survey consisted of 1,000 users under age 50, of whom 470 accessed news on their cell phones, and 640 users age 50 and over, of whom 185 accessed news on their cell phones.
At 0.05 level of significance, if we want to test whether the proportion of users under the age 50 who access news on their cell phones are **more** than the proportion of users age 50 and above who access news on their cell phones . Then, the **test statistic is**:

a. 7.5770
b. 0.3202
c. -7.2964
d. -0.3202
e. 1.3110

5. The weight of the teabags should be as consistent as possible. In the past, the standard deviation of the weight of the teabags has been 0.05 grams (g). A sample of 20 tea bags produced during an eight-hour shift gave a mean of 5.3975 g and a standard deviation of 0.0635 g. If you want to test (using $\alpha = 0.01$) that, the standard deviation of the amount of tea per bag is greater than 0.05 g, then the **final conclusion** of the test is

a. There is no evidence that the standard deviation of the amount of tea per bag is greater than 0.05 g.
b. There is no evidence that the standard deviation of the amount of tea per bag is equal to 0.05 g.
c. There is evidence that the standard deviation of the amount of tea per bag is greater than 0.05 g.
d. There is no evidence that the standard deviation of the amount of tea per bag is at most 0.05 g.
e. There is evidence that the standard deviation of the amount of tea per bag is at most 0.05 g.

6. A computer used by a 24-hour banking service is supposed to randomly assign each transaction to one of 5 memory locations. A check at the end of a day's transactions gave the counts shown in the table to each of the 5 memory locations, along with the number of reported errors.

Memory Location	1	2	3	4	5	Total
Number of Transactions	82	100	74	92	102	
Number of Reported Errors	11	12	6	9	10	

Which test would be used to properly analyze the data in this experiment?

a. Chi square test for difference among more than two proportions.
b. Chi square test of independence of two variables.
c. McNemar test for the difference between two proportions.
d. McNemar test for the difference among more than two proportions.
e. Z test for the difference between two proportions.

A company that holds the DVD distribution rights to movies previously released only in theaters wants to estimate sales revenue of DVDs based on box office success. The box office gross (in $ millions) for each of **22** movies in the year that they were released and the DVD revenue (in $ millions). Here y denotes the DVD revenue (in $ millions) and x denotes the box office gross (in $ millions). The summary of this data is given below:
$\sum y = 900.40, \quad \sum x = 1745.21, \quad \sum xy = 86195.43, \quad \sum y^2 = 47250.47, \quad \sum x^2 = 176913.38,$ and $SSR = 5669.7847$
Using the data above answer the following 2 questions:

7. The estimated slope, of the predicted regression equation, is:

a. 0.3839
b. 2.6048
c. 92.3001
d. 0.9882
e. 2.3092

8. A 98% confidence interval for the population slope is given by:

a. [0.1857,0.5821]
b. [-0.5821, 0.1857]
c. [-0.001, 0.001]
d. [0.1857, 1.3900]
e. [-0.1857, 0.5821]

A mail-order catalog business selling personal computer supplies, software, and hardware maintains a centralized warehouse. Management is cur- rently examining the process of distribution from the ware- house. The business problem facing management relates to the factors that affect warehouse distribution costs. Currently, a small handling fee is added to each order, regardless of the amount of the order. Data collected over the past 24 months indicate the warehouse distribution costs (in thousands of dollars), the sales (in thousands of dollars), and the number of orders received.

Use the following outputs to answer the following two questions:

Statistics

Variable	Mean	StDev	Minimum	Maximum
Cost	71.26	12.93	52.46	94.44
Sales	456.5	81.5	301.0	623.0
Orders	4393	737	2921	5735

Regression Analysis: Cost versus Sales, Orders

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	3368.1	1684.04	74.13	0.000
Error	21	477.0	22.72		
Total	23	3845.1			

Model Summary

S	R-sq	R-sq(adj)
4.76617	87.59%	86.41%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	-2.73	6.16	-0.44	0.662
Sales	0.0471	0.0203	2.32	0.031
Orders	0.01195	0.00225	5.31	0.000

Regression Equation

Cost = -2.73 + 0.0471 Sales + 0.01195 Orders

9. Predict the monthly warehouse distribution cost when sales are $400,000 and the number of orders is 4,300.

a. $67,495.00

b. $18,888.66

c. $16.16

d. $16.62

e. $16,623.85

10. A regression model to predict distribution cost that includes sales, orders, and the interaction of sales and orders is developed:

Regression Analysis: Cost versus Sales, Orders

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	3391.2	1130.40	49.81	0.000
Error	20	453.9	22.70		
Total	23	3845.1			

Model Summary

S	R-sq	R-sq(adj)
4.76406	88.19%	86.42%

Coefficients

Term	Coef	SE Coef	T-Value
Constant	31.6	34.5	0.91
Sales	-0.0296	0.0787	-0.38
Orders	0.00410	0.00809	0.51
Sales*Orders	0.000017	0.000017	1.01

Regression Equation

Cost = 31.6- 0.0296 Sales + 0.00410 Orders + 0.000017 Sales*Orders

Which of the following is true about the interaction term at 0.05 level of significance?

a. The interaction term does not make a significant contribution to the model.
b. The interaction term makes a significant contribution to the model.
c. Since T-value for the variable "Orders" is very small, the interaction term makes a significant contribution to the model.
d. Since T-value for the variable "Sales" is very small, the interaction term makes a significant contribution to the model.
e. Since coefficient for the variable "Orders" is very small, the contribution of the interaction term is not significant to the model.

The business problem facing a consumer products company is to measure the effectiveness of different types of advertising media in the promotion of its products. Specifically, the company is interested in the effectiveness of radio advertising in thousands of dollars (X_1) and newspaper advertising in thousands of dollars (X_2) on the sales in thousands of dollars (Y). Data were collected from a sample of 22 cities. The following is the ANOVA table for the regression model:

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2028032.69	1014016.345	40.15823	
Residual	19				
Total	21				

Also, $SSR_{(X_1)}$ =1216940 and $SSR_{(X_2)} = 632259.4483$

From the data above answer the following **3** questions:

11. At the 0.05 level of significance, we want to test whether there is evidence that the newspaper advertising makes a significant contribution to the regression model. **The test statistic is**:

a. 32.12
b. 55.28
c. 45.33
d. 31.07
e. 39.09

12. At the 0.05 level of significance, we want to test whether there is evidence that the newspaper advertising makes a significant contribution to the regression model. **The critical value is**:

a. 4.38
b. 248.0
c. 5.92
d. 60.13
e. 72.91

13. At the 0.05 level of significance, we want to test whether there is evidence that the newspaper advertising makes a significant contribution to the regression model. **The conclusion is**:

a. Reject H_o: The addition of the variable (newspaper advertising) significantly improves a regression model that already contains the variable (Radio advertising).
b. Don't reject H_o: The addition of the variable (newspaper advertising) significantly improves a regression model that already contains the variable (Radio advertising).
c. Don't reject H_o: The addition of the variable (newspaper advertising) does not significantly improves a regression model that already contains the variable (Radio advertising).
d. Reject H_o: The addition of the variable (newspaper advertising) does not significantly improves a regression model that already contains the variable (Radio advertising).
e. None of the above.

14. The following is the ANOVA summary table for a multiple regression model with two independent variables:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression		60			
Error	18	120			
Total	20	180			

If $SSR_{(X1)} = 45$ and $SSR_{(X2)} = 25$

To test whether the overall model is significant at the 0.05 level of significance, the test statistic is

a. 4.5
b. 0.5
c. 3.75
d. 6.75
e. 0.22

15. A researcher wants to develop the most appropriate multiple regression model to predict a store's monthly sales using the following variables:

Y = Sales, X_1 = Age, X_2 = Growth, X_3 = Income X_4 = HS, X_5 = College.

The best-subset approach yielded the following models:

Best Subsets Regression: Y versus X1, X2, X3, X4, X5

Model#	Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	X1	X2	X3	X4	X5
1	1	24.1	21.9	17.2	-1.8	802004				X	
2	1	14.7	12.3	4.0	2.1	849860			X		
3	2	24.3	20.0	8.7	0.0	811809	X			X	
4	2	24.2	19.9	14.0	0.1	812572				X	X
5	3	24.4	17.8	5.4	2.0	823193	X			X	X
6	3	24.4	17.7	0.0	2.0	823626	X	X		X	
7	4	24.4	15.3	0.0	4.0	835508	X		X	X	X
8	4	24.4	15.3	0.0	4.0	835563	X	X		X	X
9	5	24.4	12.6	0.0	6.0	848433	X	X	X	X	X

Using the best-subset approach, the candidate model(s) using the Cp Mallows:

a. All models are candidate models except model#2.
b. Only Models #: 1,2,3,4,5,6
c. Only Models #: 7, 8, 9
d. Only Models #: 7, 8, 9
e. Only model #1.

16. Using the following outputs for the full model

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2270706	3696533	-0.61	0.543	
X1	-27384	93046	-0.29	0.770	1.32
X2	2084	44098	0.05	0.963	1.44
X3	2.4	30.5	0.08	0.937	3.79
X4	62735	35090	1.79	0.083	3.52
X5	-5702	28359	-0.20	0.842	2.74

Using the stepwise algorithm, the first variable that will be included in the model is:

a. X4
b. X2
c. X1
d. X3
e. X5

17. A manager of a chain that sells sporting goods wanted to investigate which variables affect the monthly sales. He decided to use the following variables as predictors: Age, growth, income, average number of highschoolers who shop there monthly, and the average number of college students who shop there monthly. He built different models using different number of variables and the following table shows different measures (such $as\ R^2, R^2_{adj}, C_p$) for different models using different number of variables.

$\boldsymbol{Variables\ used}$	$\boldsymbol{R^2}$	$\boldsymbol{R^2_{adj}}$	$\boldsymbol{Mallow's\ C_p}$
$\boldsymbol{Age}$	0.0017	-0.0260	8.283
$\boldsymbol{Age, and\ Highschool}$	0.2435	0.2002	0.044
$\boldsymbol{Growth}$	0.0126	-0.0148	7.822
$\boldsymbol{Highschool, and\ College}$	0.1607	0.1128	3.548
$\boldsymbol{Income, and\ College}$	0.1540	0.1056	3.833
$\boldsymbol{Age, growth, and\ income}$	0.1663	0.0928	5.310
$\boldsymbol{Age, growth, income, and\ college}$	0.1690	0.0683	7.196

Using R^2_{adj}, the model that should be taken into consideration contains:

a. Age and Highschool, since it has the highest R^2_{adj}.
b. Highschool and college, since it has the highest R^2.
c. Income and college, since the R^2_{adj} exceeds 10%.
d. Age only, since it has the lowest R^2_{adj}.
e. Growth only, since it has R^2_{adj} very close to 0.

18. If the coefficient of determination between two independent variables is 0.72, is there a problem of collinearity?

a. No, since the $VIF = 3.571$.
b. Yes, since the $VIF = 3.571$.
c. Yes, since the coefficient of determination is more that 50%.
d. Yes, since the $VIF = -3.571$.
e. No, since the $VIF = -3.571$.

19. An agronomist designed a study in which tomatoes were grown using twelve different amounts of fertilizers. These fertilizer application rates were then randomly assigned to plots of land to predict the yield of tomatoes (in pounds). The following is the quadratic regression model:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
$Intercept$	6.642857	2.132874	3.114511	0.012428
X	0.895	0.100312		
X^2	-0.00411	0.000963		

At the 0.02 level of significance, we want to test whether there is evidence that the quadratic effect of the fertilizer is significant to the regression model. **The p-value is**:

a. $0.002 < p - value < 0.005$
b. $0.001 < p - value < 0.0025$
c. $0.0005 < p - value < 0.001$
d. $p - value < 0.0005$
e. $0.025 < p - value < 0.05$

An economist develop a multiplicative time-series model to forecast the petrol prices in the USA in future quarters, using quarterly data of petrol prices (in dollars per gallon) from January 2006 to April 2010. The following is the resulting regression equation:

$$\log \hat{y} = 2.623 + 0.00039\, X_i - 0.220\, Q1 + 0.233\, Q2 + 0.436\, Q3$$

Where:

- $\hat{y}$: $is\ the\ estimated\ price\ of\ petrol\ in\ quarter$
- X_i: $is\ the\ code\ quarterly\ value\ with\ X = 0\ in\ the\ first\ quarter\ of$ 2006
- $Q1$: $dummy\ variable\ equal\ to$ **1** $in\ the\ first\ quarter\ of\ a\ year\ and$ **0** $other\ wise$
- $Q2$: $dummy\ variable\ equal\ to$ **1** $in\ the\ second\ quarter\ of\ a\ year\ and$ **0** $other\ wise$
- $Q3$: $dummy\ variable\ equal\ to$ **1** $in\ the\ third\ quarter\ of\ a\ year\ and$ **0** $other\ wise$

Using the data above answer the following **3** questions:

20. To obtain a forecast for the **fourth** quarter of 2008 using the model, which of the following sets of values should be used in the regression equation?

a. $X = 11, Q1 = 0, Q2 = 0, Q3 = 0$
b. $X = 12, Q1 = 0, Q2 = 0, Q3 = 0$
c. $X = 16, Q1 = 0, Q2 = 0, Q3 = 0$
d. $X = 11, Q1 = 1, Q2 = 0, Q3 = 0$
e. $X = 12, Q1 = 1, Q2 = 0, Q3 = 0$

21. The estimated quarterly compound growth rate in the price is:

a. 0.09%
b. 4.19%
c. -0.09%
d. 1.28%
e. 2.44%

22. Using the regression equation, the forecast for the prices in the third quarter of 2010 is:

a. 1164.18$
b. 729.49$
c. 707.98$
d. 1005.38$
e. 1250.09$

The following table presents the mean prices (in dollars) and the estimated quantities (in pounds) for three fruit items for selected periods from 1980 to 2010: **(Index numbers, Example 16.4 in the online source)**

Fruits	Years							
	1995		2000		2005		2010	
	Price	Quantity	Price	Quantity	Price	Quantity	Price	Quantity
Apples	0.835	18.9	0.927	17.5	0.966	16.0	1.141	15.6
Bananas	0.490	27.4	0.509	28.5	0.485	26.8	0.586	25.6
Oranges	0.625	12.0	0.638	11.7	0.838	10.6	0.899	9.7

Using the above information answer the following **2** questions:

23. The unweighted aggregate index number for the prices of fruits of 2010 based on 1995:

a. 134.66%
b. 74.26%
c. 65.38%
d. 79.10%
e. 121.62%

24. Laspeyres index number for the prices of fruits of 2005 based on 2010:

a. 85.77%
b. 116.59%
c. 238.89%
d. 58.77%
e. 131.20%