Name _____ Fall 2015

Applied Statistics Comprehensive Examination

- Calculators are permitted on this examination.
- When you are asked to construct a confidence interval, always interpret the interval in terms of the problem.
- When you are asked to perform a hypothesis test, always write down the null and alternative hypotheses and write the conclusions in terms of the problem.
- There are 200 points for the entire examination.

1. (20 pts) One hundred political experts are independently asked to predict the winner for three governor races. For all three races, there are only two candidates to choose from. For each expert, the number of correct predictions is recorded, with the following summary of results:

Number of Correct Predictions	0	1	2	3
Frequency	10	41	35	14

Can we conclude at the 0.05 level that the experts performed any differently than would have been expected by chance (i.e. by having each expert flip a fair coin to decide the winners)?

2. (30 pts) A video game company wanted to assess whether there is a difference in how men and women rate the quality of violent computer games. The company also wanted to assess whether the quality rating varied significantly among 10 specific popular computer games labeled as violent by the Entertainment Software Rating Board. Four individuals of each sex were randomly assigned to play each of the 10 games, so that a total of 40 men and 40 women participated in the experiment. Each individual was given game instructions and then asked to play the game for 5 hours in a controlled environment, after which the individual provided a quality rating for the game on a scale from 1 to 100 (100 being the highest quality).

(a) (10 pts) Write down the analysis of variance model. Clearly define all terms used and state all necessary assumptions.
(b) (10 pts) Complete the ANOVA table.
(c) (10 pts) Conduct all appropriate inferences at the 0.05 level.

Source	df	SS	MS	F
Sex		4000		
Game		7300		
Sex × Game		1700		
Residual				
Total		22000		

3. (30 pts) Suppose a completely randomized experiment with four treatments and five observations per treatment is planned. But inadvertently, an observation is lost from the third treatment. Assume an effects model and let G be the grand total of all the observations and T_i be the total of the observations in the i^{th} treatment group.

(a) (10 pts) Write the normal equations in matrix form.

(b) (10 pts) Write the normal equations in matrix form using the "set to zero" restriction.

(c) (10 pts) Write the normal equations in matrix form using the "sum to zero" restriction.

4. (20 pts) A regional restaurant chain recently opened a new location and wants to investigate whether it is worthwhile to offer dessert choices on the menu at this new location. Based on the other restaurants in the chain, they know that more than 8% of customers need to purchase a dessert to make it profitable to offer dessert choices.

(a) (10 pts) A random sample of 214 customers at the new restaurant is obtained, and 30 of these customers purchase dessert. Find a 95% confidence interval for the true proportion of customers buying dessert at the restaurant. What does the interval indicate regarding whether it would be profitable to offer dessert?

(b) (10 pts) If in reality 10% of the restaurant's customers buy dessert, based on a sample of 214 customers and an α-level of 0.05, what is the power for concluding the research hypothesis that the true proportion is greater than 8%?

5. (25 pts) To study the relationship between the number of employed people (in million) (x) and the gross domestic product (GDP, in billions of US dollars) (y), a researcher has collected data from 12 countries. Use the following information to answer the questions:

$$\sum_{i=1}^{12} x_i = 581 \quad \sum_{i=1}^{12} (x_i - \bar{x})^2 = 377 \quad \sum_{i=1}^{12} (x_i - \bar{x})(y_i - \bar{y}) = 64$$

$$\sum_{i=1}^{12} y_i = 53 \quad \sum_{i=1}^{12} y_i^2 = 267$$

Source	df	SS	MS	F
Regression				
Error		22.08		
Total				

(a) (5 pts) Write out the fitted linear regression line.

(b) (5 pts) State the assumptions that are necessary for this simple linear regression.

(c) (10 pts) Complete the ANOVA table.

(d) (5 pts) What hypothesis is being tested in the above ANOVA table? State and conduct the test at the significance level of 0.05.

6. (20 pts) Consider an experiment to compare the effectiveness of two different pesticides, A and B say, both applied in two different forms: spray (A_1 and B_1) and powder (A_2 and B_2). A control treatment (i.e., no pesticide), C say, is also included in the experiment in order to establish any effectiveness of the pesticides at all. Thus, altogether there are 5 treatments (A_1, A_2, B_1, B_2 and C) and each treatment is randomly applied to r uniformly infested plots of land. Obtain a complete set of meaningful orthogonal contrasts and describe what each contrast is testing.

7. (10 pts) A pharmaceutical company conducted an experiment comparing four different blood pressure medications. For the experiment, 20 patients with high systolic blood pressure were selected and randomly assigned to the four medications. The blood pressure of each patient was measured at the beginning of the experiment and then again a year later. The results for the systolic blood pressure reductions are given below:

Med 1	Med 2	Med 3	Med 4
8	8	20	11
-12	-2	7	27
-2	-1	17	16
-3	14	6	9
4	9	11	15
$\bar{y}_1 = -1.0$	$\bar{y}_2 = 5.6$	$\bar{y}_3 = 12.2$	$\bar{y}_4 = 15.6$
$s_1 = 7.6$	$s_2 = 6.9$	$s_3 = 6.1$	$s_4 = 7.0$

The F-test for whether there is a difference in mean blood pressure reductions among the four medications has p-value 0.008, indicating there are statistically significant differences. Use Fisher's LSD procedure to determine all significant pairwise differences among the means at the 0.05 level.

8. (25 pts) Suppose a realtor wants to model the appraised price (in thousands of euros) of an apartment in Vitoria, Spain, as a function of two predictors: living area (in square meters) and the apartment's energy efficiency (in two levels, high and low). Use the SAS output below to help answer the following questions.

(a) (5 pts) Conduct the global hypothesis test for the model at the significance level of 0.05.

(b) (5 pts) Calculate and interpret the value of R^2 for this model.

(c) (5 pts) Based on the fitted model in the output, write down separate prediction equations for each energy efficiency category.

(d) (10 pts) Test for equality of the slopes of the models in part (c) at the significance level of 0.05.

Multiple Regression with interaction

The GLM Procedure

Dependent Variable: price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	739411.007	246470.336	174.26	<.0001
Error	214	302682.981	1414.406		
Corrected Total	217	1042093.988			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
area	1	682389.1004	682389.1004	482.46	<.0001
energy	1	31069.8909	31069.8909	21.97	<.0001
area*energy	1	25952.0161	25952.0161	18.35	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
area	1	400141.5499	400141.5499	282.90	<.0001
energy	1	11849.6175	11849.6175	8.38	0.0042
area*energy	1	25952.0161	25952.0161	18.35	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	22.31	B	13.135	1.70	0.0909
area	2.97	B	0.143	20.73	<.0001
energy low	73.09	B	25.250	2.89	0.0042
energy high	0.00	B	.	.	.
area*energy low	-1.21	B	0.283	-4.28	<.0001
area*energy high	0.00	B	.	.	.

9. (20 pts) Suppose the College Board has developed a new SAT-type test where test scores among college bound students are distributed continuous uniform over the range 0 to 2400. Recall that the variance of this uniform distribution is $\frac{1}{12}(2400\text{-}0)^2$=480,000

(a) (5 pts) What is the probability that 40 randomly chosen students would have an average score over 1300?

(b) (15 pts) Suppose that 30 college bound students randomly chosen from those committed to Villanova score an average of 1320. Can we conclude at the 0.01 level that the average score for Villanova-bound students is higher than 1200?