Name: _____
Student No.: _____

University of Alberta

Department of Mathematical and Statistical Sciences

Statistics 252 - Midterm Examination Version A Solution

Date: March 6, 2013

Instructor: Alireza Simchi

Time: 9:00 p.m. - 9:50 a.m.

Instructions: (READ ALL INSTRUCTIONS CAREFULLY.)

1. This is a closed book exam. You are permitted to use a non-programmable calculator. Please turn off your cellular phones or pagers.

2. The exam consists of **six** parts. In the parts one to five, there are 19 multiple-choice questions. For each multiple-choice question choose the answer that is closest to being correct. Circle one of the letters (a)-(e) **on the second page** corresponding to your chosen answer for each question. All answers will be graded right or wrong (no partial credit) in this part. Each single question is worth 1 point. All numerical answers are rounded. Question **20** is a long-answer question. **Show all your work to get full credit**. In fact, answers must have adequate justification. If you run out of space, use the back of any page for answers as needed. Clearly direct the marker to answers that you provide on the back of a page.

3. This exam has **7** pages including this cover. Please ensure that you have all pages and write your name and your student ID at the top of each page.

4. The statistical tables and formula sheet are provided in a separate booklet.

5. The exam is graded out of a total of **25** points.

6. **When referring to "log", I am always referring to the natural log.**

Name: _____
Student No.:_____

Circle one answer for each question on the following table. Each question is worth 1 mark.

Question	Answer					Question	Answer				
1	a	b	c	**d**	e	**11**	**a**	**b**	**c**	**d**	**e**
2	**a**	**b**	**c**	**d**	**e**	12	a	b	**c**	d	e
3	**a**	b	c	d	e	**13**	**a**	**b**	**c**	**d**	**e**
4	**a**	**b**	**c**	**d**	**e**	14	a	b	c	d	**e**
5	a	**b**	c	d	e	**15**	**a**	**b**	**c**	**d**	**e**
6	**a**	**b**	**c**	**d**	**e**	16	**a**	b	c	d	e
7	a	**b**	c	d	e	**17**	**a**	**b**	**c**	**d**	**e**
8	**a**	**b**	**c**	**d**	**e**	18	a	b	c	d	**e**
9	a	**b**	c	d	e	**19**	**a**	**b**	**c**	**d**	**e**
10	**a**	**b**	**c**	**d**	**e**						

PART 1

The following ANOVA table is for the Simple Linear Regression (SLR) model relating average shelf life of cough syrup to the storage temperature.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	?	?	?	?	?[b]
	Residual	?	?	209.511		
	Total	228035.6	39			

a. Dependent Variable: LIFE

b. Predictors: (Constant), TEMP

1. What is the absolute value of the t-test, rounded to the nearest integer, for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 < 0$, where β_1 is the slope of population regression line?

a) 2 **b)** 12 **c)** 13 **d)** 32 **e)** 33

2. Refer to question **1**, the range of p-value can be describe as:

a) less than 0.0005
b) between 0.0005 and 0.001
c) between 0.001 and 0.01
d) between 0.01 and 0.05
e) greater than 0.05

PART 2

In a study of the effects of marijuana use during pregnancy, measurements on babies of mothers who used marijuana during pregnancy were compared to measurements on babies of mothers who did not. A 95% confidence interval for the difference in mean head circumference was 0.04 to 0.46 cm. Suppose this interval is based on sample sizes of 21 from each group. Suppose the confidence interval is based on pooling samples and you were to carry out a test for any mean difference using a t-test with pooling? Questions **3** to **5** are related to this confidence interval. .

3. What is the standard error of $\bar{y}_1 - \bar{y}_2$?

a) 0.1 **b)** 0.2 **c)** 0.3 **d)** 0.4 **e)** 0.5

4. What is (approximately) the value of the test statistic for testing $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 > \mu_2$?

a) 0.34 **b)** 0.96 **c)** 1.34 **d)** 1.96 **e)** 2.41

5. Refer to question **4**, the range of p-value can be describe as:

a) less than 0.01
b) between 0.01 and 0.02
c) between 0.02 and 0.05
d) between 0.05 and 0.10
e) greater than 0.1

PART 3

Suppose there were 9 sections of Statistics 151 in winter 2010. There were three instructors and each instructor teaches three sections. Each section has 100 students enrolled.

6. To determine if any instructor has different marks between their own sections, what is the distribution of the test statistic under null hypothesis?

a) F(3, 897) **b)** F(6, 894) **c)** F(3, 891) **d)** F(6, 891) **e)** F(8, 891)

PART 4

Consider comparing average corn yield for four different corn varieties. You collect random sample of 8 observations for each variety. The ANOVA table for comparing four corn varieties is given in the following table:

Source of Variation	Sum of Squares	d.f.	Mean Square	F-Statistic	p-value
Between	8.25	3	2.750	10.26	
Within	7.50	28	0.268		
Total	15.75	31			

Suppose you wish to compare **all** means two at a time (a total of 6 pair-wise comparisons). Therefore, we decided to use Bonferroni method to calculate 6 simultaneous 88% confidence interval for the difference in the average corn yield for different varieties. Questions **7** and **9** are related to this method.

7. What is the best estimate for the common standard deviation of the above 4 groups?

a) 0.27 **b)** 0.52 **c)** 1.42 **d)** 1.65 **e)** 2.75

8. What is (approximately) the critical value for 88% family-wise confidence intervals?

a) 1.3 **b)** 1.7 **c)** 2.1 **d)** 2.5 **e)** 2.7

9. What is (approximately) the margin of error for all 88% family-wise confidence intervals?

a) 0.26 **b)** 0.64 **c)** 1.26 **d)** 1.64 **e)** 2.15

PART 5

Three replicated water samples were taken at each of four locations in a river to determine whether the quantity of dissolved oxygen, a measure of water pollution, varied from one location to another (the higher the level of pollution, the lower the dissolved oxygen reading). Location 1 was adjacent to the waste-water point for a certain industrial plant, and location 2, 3, and 4 were selected at points 10, 20 and 30 miles downstream from this discharge point. The summary statistics of data appear in the accompanying table.

Location	N	Mean	Std. Deviation	Std. Error
1	3	4.7333	.90738	.52387
2	3	6.5667	.50332	.29059
3	3	8.2667	.97125	.56075
4	3	9.3000	.95394	.55076
Total	12	7.2167	1.95301	.56379

Use the following SPSS results to answer following questions:

ANOVA Table (Comparing all Four Groups):

ANOVA

Oxygen

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	36.097	3	12.032	16.426	.001
Within Groups	5.860	8	.732		
Total	41.957	11			

ANOVA Table (Comparing First Location with the Last Three Locations):

ANOVA

Oxygen

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	24.668	1	24.668	14.268	.004
Within Groups	17.289	10	1.729		
Total	41.957	11			

Linear contrast:

Contrast Coefficients

Contrast	Location			
	1.00	2.00	3.00	4.00
1	0	1	1	-2
2	0	1	2	-3
3	1	1	1	-3
4	-3	-1	1	3
5	-1	-1	1	1
6	1	-1	-1	1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Oxygen	Assume equal variances	1	-3.7667	1.21037	-3.112	8	.014
		2	-4.8000	1.84887	-2.596	8	.032
		3	-8.3333	1.71172	-4.868	8	.001
		4	15.4000	2.20983	6.969	8	.000
		5	6.2667	.98826	6.341	8	.000
		6	-.8000	.98826	-.809	8	.442

Suppose it was conjectured that the mean dissolved oxygen content for the first location is different from other locations, because its location was adjacent to the waste-water point. Suppose we would like to determine if there are any differences in mean dissolved oxygen content among the last three locations. Questions **10** to **13** are related to this question.

10. What is the distribution of the test statistic under the null hypothesis?
a) F(1,8) **b)** F(1, 10) **c)** F(2, 8) **d)** F(3,8) **e)** F(8, 10)

11. What is (approximately) the extra sum of squared (reduction in residual sum of squares in the model under the null hypothesis vs. the model under the alternative hypothesis)?
a) 5.2 **b)** 11.4 **c)** 24.7 **d)** 36.1 **e)** 46.4

12. What is (approximately) the value of the test statistic?

a) 1.5 **b)** 4.2 **c)** 7.8 **d)** 14.3 **e)** 16.4

13. The range of p-value can be describe as:

a) less than 0.001 **b)** between 0.001 and 0.01
c) between 0.01 and 0.05 **d)** between 0.05 and 0.1
e) greater than 0.1

Suppose it was conjectured that the locations caused increase in dissolved oxygen content and that this increase accumulated as the location was farther from the discharge point. Suppose we would like to test the conjecture using a contrast , γ, that defines a measure of the linear relationship between location (the four equally spaced distance 0, 10, 20, and 30 miles downstream from the plant) and the mean of dissolved oxygen content. Questions **14** and **15** related to this contrast.

14. What is (approximately) the value of the standard error of the estimate of the contrast? Just choose one from table for contrast tests in the SPSS outputs.

a) 0.998 **b)** 1.210 **c)** 1.711 **d)** 1.849 **e)** 2.209

15. The range of p-value can be describe as:

a) less than 0.001 **b)** between 0.001 and 0.01 **c)** between 0.01 and 0.02

d) between 0.02 and 0.05 **e)** greater than 0.05

Consider a contrast , γ, that defines the effect of location (the four equally spaced distance 0, 10, 20, and 30 miles downstream from the plant) on the mean of dissolved oxygen content. Questions **16** to **19** related to this contrast.

16. What is the estimate of the contrast?

a) 0.15 **b)** 1.5 **c)** 15.3 **d)** 55.3 **e)** 220.3

17. What is (approximately) the standard error of the estimate of the contrast?

a) 0.02 **b)** 0.49 **c)** 1.02 **d)** 1.49 **e)** 7.97

18. What is (approximately) the absolute value of the test statistic for testing $H_0 : \gamma = 0$ versus $H_1 : \gamma \neq 0$, rounded to the nearest integer?

a) 0 **b)** 1 **c)** 2 **d)** 3 **e)** 7

19. Refer to question **18**, the range of p-value can be describe as:

a) less than 0.001 **b)** between 0.001 and 0.01
c) between 0.01 and 0.05 **d)** between 0.05 and 0.10
e) greater than 0.10

PART 6

Previous studies suggest that vegetarians may not receive enough zinc in their diets. As the zinc requirement is particularly important during pregnancy, researchers conducted a study to determine whether vegetarian pregnant women are at greater risk from low zinc levels than are non-vegetarian pregnant women. Twenty-nine women were monitored: twelve vegetarians who were pregnant, six non-vegetarians who were pregnant, five vegetarians who were not pregnant, and six non-vegetarians who were not pregnant. None of these women were smokers and none of the non-pregnant women were taking oral contraceptives. The zinc content in hair was measured for each woman.

Define:

μ_1 : Average zinc content for non-vegetarians pregnant women (NV-P)

μ_2 : Average zinc content for vegetarians and pregnant women (VP)

μ_3 : Average zinc content for vegetarians and non-pregnant women (V-NP)

μ_4 : Average zinc content for non-vegetarians and non-pregnant women (NV-NP)

Table 1: The ANOVA table for the comparison of average zinc content for the 4 groups:

ANOVA

Zinc

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2918.416	3	972.805	3.067	.046
Within Groups	7928.550	25	317.142		
Total	10846.966	28			

Table 2: The ANOVA table for the comparison of average zinc content between pregnant and non-pregnant women (ignoring vegetarian status):

ANOVA

Zinc

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2724.142	1	2724.142	9.055	.006
Within Groups	8122.823	27	300.845		
Total	10846.966	28			

Table 3: The ANOVA table for the comparison of average zinc content between vegetarians and non-vegetarians women (ignoring pregnancy status):

ANOVA

Zinc

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	436.519	1	436.519	1.132	.297
Within Groups	10410.446	27	385.572		
Total	10846.966	28			

20. (6 Marks) Does there appear to be a significant vegetarian effect within either pregnancy status? In other words, is there a significant difference in zinc content between vegetarian and non-vegetarian women with the same type of pregnancy status? Carry out **a single overall test** to determine if there is a difference in average zinc content between vegetarian and non-vegetarian women who have the same pregnancy status (either pregnant or non-pregnant). In terms of the parameters defined earlier, state clearly the models in the null and alternative hypothesis. Also, identify the residual sum of squares and degrees of freedom for the models in the null and alternative hypotheses. Calculate the test statistic, the *p*-value (use the F-tables), and identify the distribution of the test statistic under the null hypothesis. What do you conclude?

Solution:

If there is no vegetarian effect for pregnant women, then we have $\mu_1 = \mu_2$.

If there is no vegetarian effect for non-pregnant women, then we have $\mu_3 = \mu_4$.

$$\begin{cases} H_0 : \mu_1 = \mu_2 \quad and \quad \mu_3 = \mu_4 \quad (\text{Reduced model : Two means model only for pregnant and/or nonpregnant women}) \ \textbf{(1.25 marks)} \\ H_1 : \mu_1, \mu_2, \mu_3, \mu_4 \ (\text{Full model : Four means model}) \end{cases}$$

From table 1:

$SS_{\text{Res}}(\text{Four means model}) = 7928.550$ and $df_{\text{Res}}(\text{Four means model}) = 25$ **(0.75 marks)**

From table 2:

$SS_{\text{Res}}(\text{Two means model}) = 8122.823$ and $df_{\text{Res}}(\text{Two means model}) = 27$ **(0.75 marks)**

Hence, Extra SS = 8122.823 – 7928.550 = 194.273 and Extra df = 27 -25 = 2. Hence, the value of the test statistic is $TS = \dfrac{Extra \quad SS / Extra \quad df}{MS_{\text{Res}}(\text{Full} \quad \text{model})} = \dfrac{194.273/2}{7928.550/25} = 0.306$. **(1 mark)**

If the null hypothesis is true, then the test statistic has an F-distribution with degrees of freedom $df_1 = 2, df_2 = 25$ **. (0.75 marks)**

So p-value is:

$$p - value = P(TS > 0.306) \Rightarrow p - value > 0.10$$ **(0.5 marks)**

Conclusion: The p-value greater than 0.10 indicates weak evidence against null hypothesis. Therefore, there is not enough evidence to conclude that there is a difference in average zinc content between vegetarians and non-vegetarians women who have the same pregnancy status (either pregnant or non-pregnant). **(1 mark)**