Student Name: _____ANSWERS_____

McGill ID: _____

Course Number: AEMA 310

Course Name: Statistical Methods 1

Examination Date: December 5, 2013

Examination Time: 9:00-12:00

Value of Exam: /100 (52.5% of the course grade)

Name of Examiner: Valérie GRAVEL

Name of co-examiner: Pierre DUTILLEUL

Type of Exam: Final **X** Take Home ☐ Multiple Choice ☐ Deferral ☐ Special Arrangement ☐

Answer Directly on Exam Paper **X**

Answer in Exam Booklet(s) ☐

Calculators are permitted: Yes **X** No ☐ If yes, specify i.e. (Graphical, Statistical): **The simpler, the better – with no special statistical functions and no storage of text or equations**

Definition Dictionaries are allowed: Yes **X** No ☐

Answer of Scantron: indicate short (1-120 responses) ☐

or long (121-240 responses) ☐

Electronic devices (i.e. beepers, cellular phones)
are not permitted in the examination room.

READ THE INSTRUCTIONS CAREFULLY BEFORE YOU START TO WRITE:

1. Place your valid student ID card on your desk.

2. No material may be brought into the exam, except writing instruments and calculators. **Memories of calculators may not contain course material.** The invigilators reserve the right to inspect and/or confiscate any material brought into the exam.

3. Make sure that you have **13** pages (including this one) labelled **1** through **13**. Print your last name on each page in the space provided. Do not separate pages. **No marks will be given for missing pages.**

4. Except for the multiple-choice and true/false questions with boxes, **solutions must appear in the spaces provided between questions**. If you must include a part of your answer elsewhere, **indicate clearly** where your answer continues and this can be only on the reverse of one of the **13** pages.

5. Write only in **blue** or **black pen**, or in **black lead pencil**. If you write in pencil, the lead must be sharp and dark.

6. Carry **at least three decimal places** in all your calculations.

7. Show all formulas and steps used in your calculations, except for the multiple-choice and true/false questions.

8. The mathematical expressions and formulas that you did not have to learn by heart are on the reverse of the last page of statistical tables.

9. **Read all questions carefully.**

Part I: Multiple-choice and true/false questions

Multiple-choice questions (**2** pts. each): Indicate clearly the most appropriate answer by filling in the box to its left.

No mark will be given if two boxes or more are checked.

1) A researcher wants to know how many lettuces are required to detect with a probability of 0.90 that the contamination of lettuce heads with a certain bacteria is above the accepted standard of 10 per µl, when it actually is 3 above the standard. (α=0.01 and σ^2=4). Which of the equations below is correct? (Note: The Central Limit Theorem can be used for the normality distribution assumption on the sample mean.)

☐ $0.99 = ((10-13)/\sqrt{(4/n)}) + 2.326$ ☐ $1.282 = ((10-13)/\sqrt{(4/n)}) + 2.326$

. ☐ $0.99 = ((10-13)/\sqrt{(4/n)}) + 1.282$ ■ $-1.282 = ((10-13)/\sqrt{(4/n)}) + 2.326$

2) The same researcher as in question 1) is told that if contamination of lettuce heads exceeds the accepted standard, they cannot be sold at the market otherwise consumers will get sick. In that case, which of the following statements is **not** correct?

■ Increasing β will reduce the risk of consumers buying contaminated lettuce.
☐ Increasing α will reduce the risk of consumers buying contaminated lettuce.
☐ Making a Type I error could be costly for the market.
☐ Increasing the power of the test will reduce the risk of consumers buying contaminated lettuce.

3) When performing a simple linear regression analysis with the model $Y = a + bx + \varepsilon$, which of the following statements is **correct**?

☐ The expected value of Y for a given x can be calculated from the linear regression equation $E(Y) = a + bx + \varepsilon$.
■ The ratio statistic $\hat{b}/S_{\hat{b}}$ has a "Student" t distribution with n – 2 degrees of freedom.
☐ The estimate of the intercept must be equal to zero for the fitted linear regression model to be significant.
☐ The value of the estimated slope is comprised between -1 and 1 inclusively.

4) In the ANOVA for an RCBD with more than 1 observation per treatment per block (>1 replicate), which of the following statements is **not** correct?

☐ When the treatment factor is considered fixed, the sum of squared treatment main effects is assumed to be zero under the null hypothesis.
☐ If the interaction is not significant, a new error term must be calculated.
■ The mean square of the new error term is calculated by adding the values of the initial mean squares for the interaction and the error term.
☐ The block main effect B_j follows $N(0,\sigma_B^2)$.

5) True/false questions (**1** pt. each): Fill in one of the boxes to the left of each statement. If you believe the statement is **T**rue, then fill in the box below the **T**; if you believe the statement is **F**alse, then fill in the box below the **F**. If you want to change your choice, erase your former choice and indicate clearly which choice is your final choice.

T	F	
☐	■	In a chi-square test of homogeneity for a 3 x 3 contingency table, H_0 is rejected at α=0.05 when the observed value of the test statistic is greater than $\chi^2_{0.95}$ (3).
☐	■	In a simple linear regression model, ε denotes the random error which is assumed to be normally distributed with a population variance of zero.
☐	■	When testing H_0: p = 0.50 against H_1: p > 0.50, H_0 is rejected at significance level α if the observed value of the test statistic t_{obs} is greater than $t_{1-\alpha}(n-1)$.
☐	■	If some H_0 is rejected against some H_1 at α = 0.01, then the same H_1 is automatically rejected against the same H_0 at α = 0.05.
☐	■	A biased estimator necessarily has a high variance.
☐	■	For paired observations (when two observations are collected on the same individual for the same random variable X), (X_1, X_2) is normally distributed so the difference $X_1 - X_2 = D \sim Bi(\mu_D, \sigma^2_D)$.
■	☐	In an ANOVA for an RCBD with 1 replicate per treatment per block, the number of degrees of freedom used to calculate the LSD is (nb of blocks – 1)(nb of treatments – 1).
☐	■	The ANOVA model for an RCBD with >1 replicate per treatment per block is: $X_{ij} = \mu + a_i + B_j + \varepsilon_{ij}$, where μ is the overall population mean, a_i is the main effect of the treatment, B_j is the main effect of the block, and ε_{ij} is the experimental error.
☐	■	Decreasing the significance level always increases the power of a statistical test.
☐	■	A statistically significant negative linear correlation between X and Y implies that when X decreases, Y is expected to respond by decreasing.
☐	■	When performing a t-test for the difference between two means in the case of independent random samples (H_0: $\mu_A = \mu_B$), an effective number of degrees of freedom needs to be calculated when a pooled sample variance needs to be used.
■	☐	When testing H_0: $\mu = \mu_0$ against H_1: $\mu > \mu_0$ with a known σ^2, one cannot reject the null hypothesis when z_{obs} is less than $-z_{1-\alpha}$.
■	☐	If the number of snow storms in Montréal is 3 on average every winter, then the number of snow storms (X) in 5 years follows a Poisson distribution with E(X) = Var(X) = 15.
■	☐	In the estimation of a 95% confidence interval, increasing the sample size will result in a narrower interval.
☐	■	A Type II error is made when a false H_0 is accepted whereas the power of the test is the probability of accepting a true H_0.

Part II: Episodes of "Dr. Small's Investigation of Unwanted Pathogens"

Dr. Small is a plant pathologist with a specific interest in tomato late blight, a destructive disease. The following problems describe various approaches he uses to address issues related to disease development and how to overcome them…

FIRST EPISODE:

Dr. Small is working on developing thresholds to determine the most appropriate time to apply fungicide treatments. His team performs surveys within tomato fields to assess the level of disease development (number of diseased leaves).

6-a) Assume that Dr. Small's research team would count, on average, 2 diseased leaves in a 1-m^2 quadrat in any part of a given tomato field. In order to have a representative assessment of the infestation, Dr. Small and his team usually evaluate areas of 5 m^2. What is the probability that they will spot at least 5 diseased leaves in a usual field evaluation?

/5

Poisson distribution
$P(X \geq 5) = 0.971$

6-b) In his research, Dr. Small has found that at least 5 leaves need to be diseased in an area of 5 m^2, to justify applying a fungicide treatment. Dr. Small wants to know which of the three independent fields that he is evaluating presently will need to be treated. He screens one area of 5 m^2 in each of the three fields. Using your work to answer question 6-a), calculate the probability that he will have to apply a fungicide treatment in two fields.

/5

Binomial distribution
$P(X=2) = 0.082$

Part II (continued)

SECOND EPISODE :

Dr. Small is also interested in testing the differences between conventional fungicides and biofungicides. More specifically, he wants to know if tomato plants treated with biofungicides provide a yield different from that of tomato plants grown in similar conditions and treated with conventional fungicides. To investigate this, he measured the yield (kg) of five tomato plants randomly sampled for each type of fungicide treatment. He obtained the following data:

	1	2	3	4	5	Sample Mean	Sample Variance
Biofungicides	5	4	6	3	5	4.6	1.04
Conventional fungicides	5	6	7	6	5	5.8	0.56

7) Use this data to perform the appropriate test in order to answer Dr. Small's research question (α=0.05).

/8

H0: $\sigma^2_B = \sigma^2_C$
H1: $\sigma^2_B \neq \sigma^2_C$
Fobs = 1.8571
CV = 9.60
AH0: pooled S^2

H0: $\mu_B = \mu_C$
H1: $\mu_B \neq \mu_C$
tobs = -2.1213
CV = -2.31 and 2.31
AH0: no difference between the fungicides

Part III: Episodes of "Caroline's Quest to Study Human Behavior"

FIRST EPISODE:

Caroline is a Ph.D. student working on human behavior. In her study, she wants to know if the amount of coffee consumed in one day (no coffee, 1 cup of coffee, more than 1 cup of coffee) affects the ability to concentrate on a specific task, measured as the time it takes to solve a puzzle (the less time it takes, the more the person is able to concentrate). Unfortunately, she could find only 18 persons willing to participate in her study: 6 McGill students, 6 working professionals and 6 retirees. Each group of participants is randomly divided into 3 subgroups and each subgroup is told the amount of coffee to drink every day. After one week, the time it takes each participant to solve the same puzzle is measured (min). Caroline obtained the following data.

		Amount of coffee			$\bar{X}_{.j.}$
		No coffee	1 cup	>1 cup	
Participants	McGill students	(10, 15) $\bar{X}_{ij.} = 12.5$	(7, 7) $\bar{X}_{ij.} = 7$	(15, 12) $\bar{X}_{ij.} = 13.5$	11
	Working professionals	(12, 8) $\bar{X}_{ij.} = 10$	(7, 10) $\bar{X}_{ij.} = 8.5$	(10, 15) $\bar{X}_{ij.} = 12.5$	10.33
	Retirees	(10,10) $\bar{X}_{ij.} = 10$	(8, 9) $\bar{X}_{ij.} = 8.5$	(12, 15) $\bar{X}_{ij.} = 13.5$	10.67
$\bar{X}_{i..}$		10.83	8	13.17	$\bar{X}_{...} = 10.67$

8-a) Construct the ANOVA table and test whether there is a main effect of the treatment (α=0.05).
(NOTE: There is room on the next page to continue your work.)

Source	Df	SS	MS
Amount of coffee	2	80.427	40.213
Groups (blocks)	2	1.347	0.674
Amount x Groups	4	11.334	2.834
Error	9	46.892	5.210
total	17	140.000	

H0: Diff. are constant from block to bock
H1: not constant
Fobs = 0.5438
CV = 3.63
AH0: constant so we calculate the New error

Part III (continued)

8-a) (continued)

/10

Source	Df	SS	MS
New Error	13	58.226	4.479

H0: No difference between treatments
H1: at least 2 are diff.
Fobs = 8.978
CV = 3.81
RH0: at least 2 are diff.

8-b) Based on the reported data and your statistical analyses of them (above and here), how much coffee (if any) would you recommend to drink in one day to someone who needs to concentrate (for example, someone who is about to write the Stats 1 Final)?

/4

LSD = 2.639

Recommend 1 cup of coffee

Part III (continued)

SECOND EPISODE:

Caroline is also interested in the way sleep habits affect the ability to concentrate on a specific task (again measured as the time it takes to solve a puzzle). As a first study on this, she wants to know if there is a relationship between the number of hours of sleep and the time it takes to solve a specific puzzle. To accomplish this, she randomly selects 7 persons and asks them how many hours they slept the previous night (X; h) before timing them until they solve the puzzle (Y; min). She obtained the following data.

Hours of sleep (X_i) (h)	Time to solve puzzle (Y_i) (min)	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
4	10	8.18	11.76	–9.81
4	12	8.18	29.48	–15.53
6	8	0.74	2.04	1.24
7	7	0.02	0.18	0.06
8	4	1.30	6.60	–2.93
9	3	4.58	12.74	–7.64
10	2	9.86	20.88	–14.35
$\sum = 48$	$\sum = 46$	$\sum = 32.86$	$\sum = 83.68$	$\sum = -48.96$
$\Sigma/7 = 6.86$	$\Sigma/7 = 6.57$	$\Sigma/6 = 5.48$	$\Sigma/6 = 13.95$	$\Sigma/6 = -8.16$

9) Perform a statistical test (α=0.01) that will help Caroline determine if there is a relationship between the amount of sleep and the ability to concentrate.

/8

H0: ρ=0
H1: ρ≠0
r = -0.9333
tobs = -5.8117
CV = -4.03 et 4.03
RH0: correlation is significant

Part III (continued)

THIRD EPISODE:

In the last year of her Ph.D. studies, Caroline also examined consumer behavior. She wanted to know if going to the supermarket when hungry affects what is being purchased ("regular food" versus "junk food"). To find out, she surveyed 75 customers doing their grocery shopping just before dinner (assumed to be hungry) and 65 customers doing their grocery shopping after dinner (assumed not to be hungry anymore). As a result, she found out that 40 customers shopping before dinner bought mainly "junk food", whereas 25 customers shopping after dinner bought mainly "junk food"

10) Help Caroline answer her research question by defining the appropriate hypotheses and performing the appropriate statistical test (α=0.05).

/8

H0: $p_h = p_{nh}$
H1: $p_h \neq p_{nh}$
zobs = 1.7594
CV = -1.96 and 1.96
AH0: no difference whether you are hungry or not

Part IV: Episodes of "Dr. Campbell's Goat Saga"

FIRST EPISODE:

Dr. Campbell thinks that his goats would produce better milk (it would have higher protein content) if they were "free-range" goats (the opposite of being kept inside pens in a barn). To evaluate that, Dr. Campbell planned an experiment in which he first measured the milk protein content (%) from five goats kept inside and subsequently, he measured their milk protein content (for the same five goats) after they had spent 1 month outside (they were "free-range"). He obtained the following data.

	1	2	3	4	5	Sample Mean	Sample Variance
Kept inside	3.3	3.0	3.2	3.3	3.1	3.18	0.10
Free-range	3.5	3.1	3.3	3.2	3.0	3.22	0.04

11) Based on this data, test if Dr. Campbell is right (the milk protein content is higher after his goats have spent 1 month outside) (α=0.05).

/8

H0: $\mu_D = 0$
H1: $\mu_D > 0$
Diff.= (Free-range) – (Inside)
tobs = 0.6667
CV = 2.13
AH0: no difference

Part IV (continued)

SECOND EPISODE:

Dr. Campbell believes that the Québec Goat Industry would really take off if the marketing of goat-derived products was improved. Thus, he decided to perform a survey in order to know how often consumers buy goat-derived products and if their living place has an effect on the frequency. Therefore, he randomly sampled 200 persons and asked them where they live (City, Suburb or Rural area) and how often they buy goat-derived products (Never, Once in a while or Every month). He obtained the following results.

	City	Suburb	Rural area
Never	45	55	15
Once in a while	25	25	10
Every week	10	10	5

12) Using Dr. Campbell's data, test whether there is an association between the living place of a person and how often he/she buys goat-derived products (α=0.01).

/8

Test of independance
H0: variables are independant
H1: not independant
X^2obs = 1.3758
Cv = 13.3
AH0: 2 variables are independant

Part IV (continued)

THIRD EPISODE:

Dr. Campbell is convinced that feeding goats with too much grain mix reduces the amount of milk they produce. To test this, Dr. Campbell assigned daily portions of grain mix to 5 goats (the five goats were producing similar volumes of milk at the beginning of the experiment), from very small (100 g/day) to very large (500 g/day). Three months later, Dr. Campbell measured the volume of milk produced by each goat and obtained the following results.

Assigned daily portions x_i (g/day)	Milk production Y_i (L/day)
100	0.5
200	1
300	3.5
400	4.5
500	6
Mean = 300	Mean = 3.1

Note: $s_x^2 = 25000$, $s_Y^2 = 5.425$, $s_{xY} = 362.5$.

13) Dr. Campbell thinks that the simple linear regression should be statistically significant. Assist him in performing the test with $\alpha=0.05$.

/8

H0: b=0
H1: b≠0
â = -1.25
^b = 0.0145
tobs = 9.6667
CV = - 3.18 et 3.18
RH0: regression is significant

Part IV (last page)

FOURTH EPISODE:

According to Dr. Campbell, there is a linear correlation between the height (cm) and the weight (lb) of a goat. His theory is that if such a relationship existed, he would not have to weigh his goats to know if they are gaining weight; all he would have to do is measure their height (which is, according to him, easier to do). Therefore, he collected height and weight data for 19 randomly sampled goats, entered them in SAS and obtained the following output.

The CORR Procedure

2 Variables: HEIGHT WEIGHT

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
HEIGHT	19	61.94737	5.19052	1177	51.00000	72.00000
WEIGHT	19	99.84211	22.81876	1897	50.00000	150.00000

Pearson Correlation Coefficients, N = 19
Prob > |r| under H0: Rho=0

	HEIGHT	WEIGHT
HEIGHT	1.00000	0.87800 <.0001
WEIGHT	0.87800 <.0001	1.00000

14) Is the linear correlation between the height (cm) and the weight (lb) a goat significant (α=0.01)? State clearly the hypotheses involved in your test of significance; indicate the numerical information in the SAS output that supports either the acceptance or the rejection of the null hypothesis; describe **briefly** the type of relationship found and what Dr. Campbell should conclude from it.

/5

H0: ρ=0
H1: ρ≠0
r = 0.87800
Prob = <0.0001 is less than 0.01 so RH0: correlation is significant

Positive linear correlation between height and weight (as height increase, weight increase)