UGANDA MARTYRS UNIVERSITY

FACULTY OF SCIENCE

DEPARTMENT OF MATHEMATICS AND STATISTICS

University Examinations 2012/2013

Examination for the Degree of Bachelor of Financial Mathematics and for Bachelor of Science General

Friday, December 14, 2012

STA 3101 STATISTICAL INFERENCE AND DATA ANALYSIS

Time allowed: 3 hours

Instructions

(i) Answer **five** questions.

(ii) Write both sides of the paper, properly number your work and begin a new question on a fresh page.

(iii) Only approved basic scientific calculators may be used in this examination.

(iv) Use graph papers where necessary. Statistical tables are allowed.

1 (a) (i) State the difference between each of the following terms:

- correlation and regression,
- simple regression and multiple regression.

(3 marks)

(ii) What do you understand by the *Coefficient of determination*. Explain how you would use the it to make a decision on whether a model is a good predictor of the dependent variable.

(3 marks)

(b) A medical researcher wishes to describe the relationship between the prescription cost of a brand name drug and its generic equivalent. The data (in dollars) are shown.

Brand name x	96	93	59	80	44	47	15	56
Generic y	42	31	17	16	8	12	6	22

Do a complete regression analysis by performing the following steps:

(i) Draw a scatter plot.

(ii) Compute the correlation coefficient and state the hypotheses,

(iii) Test the hypotheses at $\alpha = 0.05$.

(iv) Determine the regression line equation.

(v) Determine the coefficient of determination.

(vi) Would the equation be considered a good predicator of generic drug price.

(vii) Plot the regression line on the scatter plot and summarize the results.

(14 marks)

2 (a) Define the following terms:

- critical value,
- critical region,
- non-critical region.

(3 marks)

(b) The *P-value* is one of the methods used for hypothesis testing.

(i) What to you is the P-value?

(2 marks)

(ii) Outline the five steps used in P-value method,

(3 marks)

(iii) Suppose the P-value method is used with level of significance $\alpha = 0.01$. When do you reject the null hypothesis and when don't you reject it.

(2 marks)

(c) A researcher estimates that the average revenue of the largest businesses in the USA is greater than \$24 billion. A sample of 50 companies is selected, and revenues (in billions of dollars) are shown. At $\alpha = 0.05$, is there enough evidence to support the researcher's claim? Use the p-value method.

178	122	91	44	35	61	56	46	20	32
30	28	28	20	27	29	16	16	19	15
41	38	36	15	25	31	30	19	19	19
24	16	15	19	25	25	18	14	15	24
23	17	17	22	22	21	20	17	20	15

(10 marks)

3 (a) Write the general form of the multiple regression equation.

(1 mark)

(b) Explain what you understand by the following terms as applied in Regression.

- Expected variation,
- Unexpected variation,
- Standard error of estimate,
- Coefficient of multiple determination.

(6 marks)

(c) The nursing instructor wishes to see whether a student's grade point average and age are related to the student's score on the state board nursing examination. She selects five students and obtain the following data.

The multiple regression equation obtained from data above is

$$y' = 57.742 + 60.667x_1 + 13.308x_2$$

Student	GPA(x_1)	Age(x_2)	State Board Score (y)
A	3.2	22	550
B	2.7	27	570
C	2.5	24	525
D	3.4	28	670
E	2.2	23	490
F	4.3	29	688
G	2.8	26	577
H	3.9	30	690

The formula for *multiple correlation coefficient* R applicable to the data above is

$$R = \sqrt{\frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1}.r_{yx_2}.r_{x_1x_2}}{1 - r^2_{x_1x_1}}}$$

(i) In the formula for R above, define variables r_{yx_1}, r_{yx_2} and $r_{x_1x_2}$.

(ii) Given that $r_{yx_1} = 0.918$, $r_{yx_2} = 0.900$, $r_{x_1x_1} = 0.720$. Find R and use the F-test to check the significance of R at $\alpha = 0.01$.

(iii) Give an interpretation of the model above, stating what each coefficient implies in the multiple regression equation.

(iv) Determine the Adjusted R^2. Why is the Adjusted R^2 sometimes preferred compared R^2.

(8 marks)

(d) A study was conducted, and a significant relationship was found among the number of hours a teeneger watches television per day x_1, the number of hours the teenager talks on the telephone per day x_2, and the teenager's weight y. The regression equation is

$$y' = 78.7 + 11.89x_1 + 6.51x_2$$

(i) State any one assumption made in multiple regression analysis.

(ii) Predict a teenager's weight if she averages 3 hours of TV and 1.5 hours on the phone per day.

(iii) If it was determined that the adjusted R^2 is 0.459, would the equation be considered a good predictor of the teenager's weight?

(5 marks)

4 (a) (i) When is the *t-test* most appropriate for testing hypotheses involving mean compared to a *z-test*?

(1 mark)

(ii) State two differences and two similarities between the t-distribution and the standard normal distribution.

(3 marks)

(b) The data below represent a sample of the number of home fires started by candles for the past several years in Uganda provided by a police department. Find the 99% confidence interval for the mean number of home fires started by candles each year.

5460 5900 6090 6310 7160 8440 9930

(7 marks)

(c) (i) Describe the *chi-square distribution.*

(2 marks)

(ii) Define the terms *type I error* and *type II error.*

(2 marks)

(iii) A cigarette manufacturer wishes to test the claim that the variance of the nicotine content of its cigarettes is 0.644. Nicotine content is measured in milligrams, and assume that it is normally distributed. A sample of 20 cigarettes has a standard deviation of 1.00 milligram. At $\alpha = 0.05$, is there enough evidence to reject the manufacturers claim?

(5 marks)

5 (a) Explain the following terms as applied to hypothesis testing.

- Test value
- Two-tailed test
- Level of significance

(4 marks)

(b) You recently recieved a job with a Company as Company Statistian. The company manufactures an automobile antitheft device. To conduct an advertising campaign for your product, you need to make a claim about the number of automobile thefts per year. Since the population of various towns in Uganda varies, you decide to use rates per 10,000 people (the rates are based on the number of people living in towns). Your boss said that last year the theft rate per 10,000 people was 44 vehicles. You want to see if it has changed. The following are rates per 10,000 people for 36 randomly selected locations in Uganda.

55	42	125	62	134	73	39	69	23	94	73	24
51	55	26	66	41	67	15	53	56	91	20	78
70	25	62	115	17	36	58	56	33	75	20	16

Use the information above to answer the following questions.

(i) What are the hypotheses that you would use?

(ii) Is the sample considered large or small?

(iii) Which probability distribution would you use?

(iv) Would you select a one- or two-tailed test? Why?

(v) What critical value(s) would you use?

(vi) Conduct a hypothesis test.

(vii) What is your decision.

(viii) What is your conclusion?

(11 marks)

(c) A recent survey found that 64.7% of the population own their own homes. In a random sample of 150 heads of households, 92 responded that they owned their homes. At 0.01 level of significance, does that indicate a difference from the national proportion?

(5 marks)

6 (a) (i) What is the difference between a *point estimate* and an *interval estimate*?

(2 marks)

(ii) A good estimator of a population mean should be an *unbiased* estimator and a *consistent* estimator.

- When is an estimator said to be unbiased?
- What do you understand by a consistent estimator?

(2 marks)

(iii) What does the term *confidence level* mean to you?

(2 marks)