Examinat: x No:

UNIVERSITY OF COLOMBO, SRI LANKA

UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

Academic Year 2013/2014 – Second Year Examination – Semester II – ~~2015~~ 2014

SCS2111 – Laboratory II

TWO (2) HOURS

To be completed by the candidate

Examination Index No: _____

Important Instructions to candidates:

1. The medium of instruction and questions is **English**.
2. If a page or a part of this question paper is not printed, please inform the supervisor immediately.
3. Note that questions appear on both sides of the paper. If a page is not printed, please inform the supervisor immediately.
4. Write your index number on each and every page of the question paper.
5. This paper has **4** questions and **11** pages.
6. Answer **ALL** questions. All questions carry equal marks (**25** marks).
7. Any electronic device capable of storing and retrieving text including electronic dictionaries and mobile phones are **not allowed**.
8. Non-programmable calculators are **allowed.**

For Examiner's use only	
Question No	**Marks**
1	
2	
3	
4	
Total	

Examination Index No:

1. Circle the correct answer choice. [100 Marks]

(a) Some hotels ask their guests to rate the hotel's services as excellent, very good, good, and poor. This is an example of the

i. ordinal scale

ii. ratio scale

iii. nominal scale

iv. interval scale

(b) Quantitative data refers to data obtained with a(n)

i. ordinal scale

ii. nominal scale

iii. either interval or ratio scale

iv. only interval scale

(c) Qualitative data

i.must be numeric

ii.must be nonnumeric

iii.cannot be numeric

iv.may be either numeric or nonnumeric

(d) 15% of the students in a school of Business Administration are majoring in Economics, 20% in Finance, 35% in Management, and 30% in Accounting. The graphical representative(s) which can be used to present these data is (are)

i. a line graph

ii. only a bar graph

iii. only a pie chart

iv. both a bar graph and a pie chart

(e) A graphical presentation of the relationship between two variables is

i. a boxplot

ii. a histogram

iii. either a boxplot or a histogram, depending on the type of data

iv. a scatter diagram

Examination Index No:

(f) When a histogram has a longer tail to the right, it is said to be
i. symmetrical
ii. skewed to the left
iii. skewed to the right
iv. none of these alternatives is correct

(g) A numerical value used as a summary measure for a sample, such as sample mean, is known as a,
i. population parameter
ii. sample parameter
iii. sample statistic
iv. population mean

(h) Since the population size is always larger than the sample size, the sample statistic
i.can never be larger than the population parameter
ii.can never be equal to the population parameter
iii.can be smaller, larger, or equal to the population parameter
iv.can never be smaller than the population parameter

(i) Some statistics about the hourly wages of a sample of 130 system analysts are given below.
mean = 60 range = 20
mode = 73 variance = 324
median = 74
The coefficient of variation equals
i. 0.30%
ii. 30%
iii. 5.4%
iv. 54%

$S = \sqrt{324}$ $\bar{x} = 60$

$\frac{S}{\bar{x}} \times 100\% = \frac{\sqrt{324}}{60} \times 100 = \frac{18}{60} \times 100 = 30\%$

(j) The variance of a sample of 169 observations equals 576. The standard deviation of the sample equals
i. 13
ii. 24
iii. 576
iv. 28,461

(k) The value added and subtracted from a point estimate in order to develop an interval estimate of the population parameter is known as the

i. confidence level

ii. margin of error

iii. parameter estimate

iv. interval estimate

(l) Whenever the population standard deviation is **unknown** and the population has a normal or near-normal distribution, which distribution is used in developing interval estimation?

i. standard distribution

ii. z distribution

iii. alpha distribution

iv. t distribution

(m) A Type II error is committed when

i. a true alternative hypothesis is mistakenly rejected

ii. a true null hypothesis is mistakenly rejected

iii. the sample size has been too small

iv. not enough information has been available

(n) The level of significance in hypothesis testing is the probability of

i. accepting a true null hypothesis

ii. accepting a false null hypothesis

iii. rejecting a true null hypothesis

iv. None of these alternatives is correct.

(o) Your investment executive claims that the average yearly rate of return on the stocks she recommends is more than 10.0%. You plan on taking a sample to test her claim. The correct set of hypotheses is

i. H_0: $\mu < 10.0\%$ H_a: $\mu \geq 10.0\%$

ii. H_0: $\mu \leq 10.0\%$ H_a: $\mu > 10.0\%$

iii. H_0: $\mu > 10.0\%$ H_a: $\mu \leq 10.0\%$

iv. H_0: $\mu \geq 10.0\%$ H_a: $\mu < 10.0\%$

H_0: $\mu \leq 10.0\%$

H_a: $\mu > 10.0\%$

Examination Index No:

2.

(a) The amount of time that a sample of students spends watching television per day is given below.

Student	Time (in Minutes)
1	40
2	28
3	71
4	48
5	49
6	35
7	40
8	57

i. Compute the mean.

[10 Marks]

$\bar{X} = \frac{368}{8} = 46$ //

ii. Compute the median. 28, 35, 40, 40, 48, 49, 57, 71

[10 Marks]

Median $= \frac{8+1}{2} = 4.5^{th}$ value = Average of 4^{th} and 5^{th} value

$= \frac{(40+48)}{2} = 44$ //

iii. Compute the standard deviation.

[20 Marks]

Variance $= S^2 = [(40-46)^2 + (28-46)^2 + (71-46)^2 + (48-46)^2 + (49-46)^2 + (35-46)^2 + (40-46)^2 + (57-46)^2] / (8-1)$

$= \frac{1276}{7}$

$S.D = S = \sqrt{\frac{1276}{7}} = 13.501$ //

iv. Compute the 75th percentile.

[10 Marks]

$i = \frac{75}{100} \times 8 = 6$

75th Percentile = Average of 6th and 7th value

$= \frac{(49+57)}{2} = 53$ //

v. Compute the coefficient of variation.

[10 Marks]

$\frac{S}{\bar{X}} \times 100\% = \frac{13.501}{46} \times 100\%$

$= 29.35\%$ //

(b) State whether the following statements are True or False.

[40 Marks]

i. In cluster analysis, objects with larger distances between them are more similar to each other than are those at smaller distances. (……..)

ii. One method of assessing reliability and validity of clustering is to use different methods of clustering and compare the results. (……..)

Old Syllabus

iii. Agglomerative clustering is a clustering procedure where all objects start out in one giant cluster. Then, the clusters are formed by dividing this cluster into smaller and smaller clusters. (……..)

iv. Non-hierarchical clustering is a clustering procedure characterized by the development of a tree-like structure. (……..)

v. Divisive clustering is frequently referred to as *k*-means clustering. (……..)

vi. Matrix scatterplot is a graphical tool that illustrates each pair of variables plotted against each other. (……..)

Examination Index No:

3.

(a) Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At 0.05 level of significance, the researcher wants to test the hypothesis that the mean penguin weight in this year does not differ from that of the last year.

i. State null and alternative hypothesis.

[10 Marks]

$H_0: \mu = 15.4$

$H_a: \mu \neq 15.4$

ii. Following R-code is given for the analysis.

```
> xbar <-14.6
> mu0 <-15.4
> sigma <-2.5 deviation
> n <-35
> z <-(xbar-mu0)/(sigma/sqrt(n))
> z
 [1] -1.893146
> p <-pnorm(z)
> p
 [1] 0.029169
```

What is the correct p-value associated with the hypothesis test?

[15 Marks]

$2 \times 0.029169 = 0.058338$ //

iii. Give your decision and conclusion clearly.

[20 Marks]

Since p-value > 0.05, we do not reject null hypothesis at 5% significance level. So we can say that mean penguin weight in this year does not differ from the last year.

Examination Index No:

(b) An experiment was carried out by a researcher to compare yields (as measured by dried weight of plants) obtained under a control and a treatment conditions. He wishes to test whether the treatment gives a same yield as the control. The following R code and output are given.

```
> control<-weight[group=="ctrl"]
> treat<-weight[group=="trt1"]
> t.test(control,treat,var.equal=T)

        Two Sample t-test
data:  control and treat
t = 1.1913, df = 18, p-value = 0.249
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2833003  1.0253003
sample estimates:
mean of x mean of y
    5.032     4.661
```

i. Write down the null and alternative hypothesis.

[10 Marks]

$H_0: \mu_c - \mu_t = 0$

$H_a: \mu_c - \mu_t \neq 0$

ii. Mention the test statistic and the p-value for the problem.

[15 Marks]

Test statistic = 1.1913

p-value = 0.249

iii. Give your conclusion.

[15 Marks]

Since p-value > 0.05 we do not reject the null hypothesis. So we can say that yields obtained under control and treatment conditions are same

iv. Write down the 95% confidence interval and interpret it.

[15 Marks]

95% confidence interval is between -0.283303 and 1.0253003.

4.

The regression analysis is carried out between the number of years of college and the current annual income (in thousands) for a random sample of heavy equipment salespeople. The following R output is obtained.

```
> fit<-lm(income~years)
> summary(fit)

Call:
lm(formula = income ~ years)

Residuals:
   Min     1Q Median     3Q    Max
 -5.60  -2.60  -1.10   3.15   5.40

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   21.600      4.189   5.156 0.000868 ***
years          2.000      1.325   1.510 0.169559
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.189 on 8 degrees of freedom
Multiple R-squared: 0.2217,     Adjusted R-squared: 0.1244
F-statistic: 2.279 on 1 and 8 DF,  p-value: 0.1696
```

i. Which variable is the dependent variable?

[10 Marks]

income

Examination Index No:

ii. Determine the least squares estimated regression line.

[10 Marks]

$\hat{income} = 21.6 + 2.00 * years$

iii. Interpret the slope coefficient.

[10 Marks]

For each additional 1 year, the income rises by 2000.

iv. Predict the annual income of a salesperson with one year of college.

[15 Marks]

$\hat{income} = 21.6 + 2.00 \times 1$

$= 23.6$ (thousands) //

$= 23.6 \times 1000$

$= 23600$ //

v. Test if the relationship between years of college and income is statistically significant at the .05 level of significance.

[20 Marks]

Calculated t-test statistic is 1.510 with the corresponding p-value 0.169559.

Calculated F-test statistic is 2.279 with the corresponding p-value 0.1696.

In both p-value > 0.05, and hence the null hypothesis cannot be rejected at 5% level of significance.

Therefore, there is no significant relationship between income and years.

Examination Index No:

vi. State the coefficient of determination and interpret it.

[20 Marks]

The coefficient of determination is 0.2217, So 22% of the variability in income is explained by the regression fit. This cannot be considered as a good fit.

vii. Calculate the sample correlation coefficient between income and years of college. Interpret the value you obtain.

[15 Marks]

I think old syllabus

****