FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

END OF SEMESTER ASSESSMENT

MODULE CODE: MA4413	SEMESTER: Autumn 2012
MODULE TITLE: Statistics for Computing	DURATION OF EXAM: 2.5 hours
LECTURER: Mr. Kevin O'Brien	GRADING SCHEME: 100 marks 85% of module grade

INSTRUCTIONS TO CANDIDATES

Scientific calculators approved by the University of Limerick can be used.
Formula sheet and statistical tables provided at the end of the exam paper.
Students must attempt 5 questions from 7.

Question 1. (20 marks)

(A) Competitors A and B fire at their respective targets. The probability that A hits a target is 1/3 and the probability that B hits a target is 1/5. Find the probability that:

i. (2 marks) A does not hit the target,

ii. (2 marks) both hit their respective targets,

iii. (2 marks) only one of them hits a target,

iv. (2 marks) neither A nor B hit their targets.

(b) On completion of a programming project, three programmers from a team submit a collection of subroutines to an acceptance group.

The following table shows the percentage of subroutines each programmer submitted and the probability that a subroutine submitted by each programmer will pass the certification test based on historical data.

Programmer	A	B	C
Proportion of subroutines submitted	0.40	0.35	0.25
Probability of acceptance	0.75	0.95	0.85

i. (3 marks) What is the proportion of subroutines that pass the acceptance test?

ii. (3 marks) After the acceptance tests are completed, one of the subroutines is selected at random and found to have passed the test. What is the probability that it was written by Programmer A?

(c) Two manufacturing plants produce specialized devices for telecommunications networks. Plant A produces 1000 parts, of which 60 are defective. Plant B produces 2000 parts, of which 70 are found to be defective.

i. (2 marks) Present this information using a contingency table.

ii. (2 marks) What is the probability that a randomly selected component is defective?

iii. (2 marks) A part is selected at random and found to be defective. What is the probability that it came from plant B ?

Question 2. (20 marks)

(a) The probability distribution of discrete random variable X is tabulated below. There are 6 possible outcome of X, i.e. 0, 1, 2, 4 ,8 and 10.

x_i	0	1	2	4	8	10
$P(x_i)$	0.25	0.15	0.25	0.15	k	0.10

i. (1 marks) Compute the value for k.

ii. (3 marks) Determine the expected value $E(X)$.

iii. (3 marks) Evaluate $E(X^2)$.

iv. (3 marks) Compute the variance of random variable X.

(b) The design of an online database gives a mean time to process a query from a central server of 280 milliseconds with a standard deviation of 40 milliseconds. It can be assumed that the query times are normally distributed.

i. (2 Mark) What proportion of query times will be greater than 340 milliseconds?

ii. (2 Mark) What proportion of query times will be less than 310 milliseconds?

iii. (3 Mark) What proportion of query times will be between 270 milliseconds and 340 milliseconds?

iv. (3 Mark) Estimate the query time above which 5% of query times will be.

Use the `R` code on the following page to assist you with your calculations.

```
> Zscore=4:50/20
> Zscore
 [1] 0.20 0.25 0.30 0.35 0.40
 [6] 0.45 0.50 0.55 0.60 0.65
[11] 0.70 0.75 0.80 0.85 0.90
[16] 0.95 1.00 1.05 1.10 1.15
[21] 1.20 1.25 1.30 1.35 1.40
[26] 1.45 1.50 1.55 1.60 1.65
[31] 1.70 1.75 1.80 1.85 1.90
[36] 1.95 2.00 2.05 2.10 2.15
[41] 2.20 2.25 2.30 2.35 2.40
[46] 2.45 2.50
>
> pnorm(Zscore)
 [1] 0.5792597 0.5987063 0.6179114 0.6368307 0.6554217
 [6] 0.6736448 0.6914625 0.7088403 0.7257469 0.7421539
[11] 0.7580363 0.7733726 0.7881446 0.8023375 0.8159399
[16] 0.8289439 0.8413447 0.8531409 0.8643339 0.8749281
[21] 0.8849303 0.8943502 0.9031995 0.9114920 0.9192433
[26] 0.9264707 0.9331928 0.9394292 0.9452007 0.9505285
[31] 0.9554345 0.9599408 0.9640697 0.9678432 0.9712834
[36] 0.9744119 0.9772499 0.9798178 0.9821356 0.9842224
[41] 0.9860966 0.9877755 0.9892759 0.9906133 0.9918025
[46] 0.9928572 0.9937903
```

Question 3. (20 marks)

(a) Suppose that a student is taking a multiple-choice exam in which each question has four possible answers to choose from. There are ten questions in the exam. Suppose that she has no knowledge of the correct answer to any of the questions. Furthermore, suppose that she selects one of the possible choices at random as her answer.

i. (3 marks) What is the probability that she will answer four questions correctly.

ii. (3 marks) What is the probability that she will get at least three correct answers?

iii. (2 marks) What is the probability that she will answer none of the questions correctly?

iv. (2 marks) What is the probability that she will answer at least one question correctly?

You may use the appropriate block of `R` code output. Write your answer to no more than four decimal places.

```
> pbinom(0:5,size = 5, prob = 0.10)
[1] 0.59049 0.91854 0.99144 0.99954 0.99999
[6] 1.00000
>
> pbinom(0:5,size = 10,prob = 0.25)
[1] 0.05631351 0.24402523 0.52559280
[4] 0.77587509 0.92187309 0.98027229
>
> pbinom(0:5,size = 4, prob = 0.10)
[1] 0.6561 0.9477 0.9963 0.9999 1.0000 1.0000
>
> dbinom(0:5,size = 5, prob = 0.10)
[1] 0.59049 0.32805 0.07290 0.00810 0.00045
[6] 0.00001
>
> dbinom(0:5,size = 10,prob = 0.25)
[1] 0.05631351 0.18771172 0.28156757
[4] 0.25028229 0.14599800 0.05839920
>
> dbinom(0:5,size = 4, prob = 0.10)
[1] 0.6561 0.2916 0.0486 0.0036 0.0001 0.0000
```

(b) An IT telephone help-line receives calls at a rate of 3 per hour, on average. Answer the following questions, using the `R` code provided. Write your answer to no more than four decimal places.

   i. (3 marks) What is the probability that the help-line receives exactly two calls in one particular hour?

   ii. (3 marks) What is the probability that the help-line receives at least two calls in one particular hour?

   iii. (2 marks) How many calls should the help-line expect in an 8 hour working day?

   iv. (2 marks) What is the probability that the help-line receives more than eighteen calls in an 8 hour period?

```
> ppois(0:6, lambda = 1/3 )
[1] 0.7165313 0.9553751 0.9951824 0.9996054 0.9999740
[6] 0.9999986 0.9999999
>
> dpois(0:6, lambda = 3 )
[1] 0.04978707 0.14936121 0.22404181 0.22404181 0.16803136
[6] 0.10081881 0.05040941
>
> ppois(0:6, lambda = 3 )
[1] 0.04978707 0.19914827 0.42319008 0.64723189 0.81526324
[6] 0.91608206 0.96649146
>
>
> dpois(10:18, lambda = 12 )
[1] 0.10483726 0.11436792 0.11436792 0.10557038 0.09048890
[6] 0.07239112 0.05429334 0.03832471 0.02554981
>
> ppois(10:18, lambda = 12 )
[1] 0.3472294 0.4615973 0.5759652 0.6815356 0.7720245
[6] 0.8444157 0.8987090 0.9370337 0.9625835
>
>
> dpois(15:20, lambda = 24 )
[1] 0.01457476 0.02186214 0.03086420 0.04115227 0.05198181
[6] 0.06237817
>
> ppois(15:20, lambda = 24 )
[1] 0.03440009 0.05626224 0.08712644 0.12827870 0.18026051
[6] 0.24263869
```

Question 4. (20 marks)

(a) The operating life span of a new electronic component produced by Deltatech is assumed to be approximately normally distributed.

A sample of 60 components was tested by Deltatech's design department. The key findings of this study are as follows: the mean life span for the sample components was found to be 8950 hours, with standard deviation of 600 hours.

Deltatech designed the component to have a mean life span of 8800 hours. Perform an appropriate significance test for the hypothesis that this specification was met.

i. (2 marks) Clearly state the null and alternative hypothesis for this procedure.

ii. (3 marks) Compute the appropriate test statistic.

iii. (3 marks) What is your conclusion for this procedure? Justify your answer.

iv. (2 marks) Provide a 95% confidence interval for the population mean.

You may use the following `R` code to assist you in answering the questions.

```
> PCs=c(0.90,0.95,0.975,0.99,0.995)
> qnorm(PCs)
[1] 1.281552 1.644854 1.959964 2.326348 2.575829
>
```

(b) A study was carried out to determine the proportion of students who owned mobile devices, such as i-phones in a number of European countries.

Out of 150 Austrian students who took part in the survey, 100 stated that they owned mobile devices.

Out of 100 Irish students who took part in the survey, 70 stated that they owned mobile devices. With reference to the `R` code output on the following page, answer the following questions.

i. (2 marks) Provide an estimate for the proportion of students who owned mobile devices in both Austria and Ireland.

ii. (1 mark) State the 95% confidence intervals for these estimates.

iii. (1 marks) Hence determine an estimate for the difference in population proportions for both countries.

iv. (1 mark) State the 95% confidence intervals for the estimate for difference in proportions.

v. (2 Marks) Formally state the null and alternative hypothesis for a test of significance for this difference in proportion.

vi. (3 Marks) With reference to both the p-value and confidence interval, state your conclusions about this test of significance. You may assume a significance level of 5%.

```
> prop.test(100,150,0.7)
        1-sample proportions test
data:  100 out of 150, null probability 0.7
X-squared = 0.6429, df = 1, p-value = 0.4227
alternative hypothesis: true p is not equal to 0.7
95 percent confidence interval:
 0.5844682 0.7401791
sample estimates:
        p
....
```

```
> prop.test(70,100,0.8)
        1-sample proportions test
data:  70 out of 100, null probability 0.8
X-squared = 5.6406, df = 1, p-value = 0.01755
alternative hypothesis: true p is not equal to 0.8
95 percent confidence interval:
 0.5989396 0.7854574
sample estimates:
  p
....
```

```
> prop.test(c(100,70),c(150,100))
        2-sample test for equality of proportions
data:  c(100, 70) out of c(150, 100)
X-squared = 0.1723, df = 1, p-value = 0.678
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.15896165  0.09229499
sample estimates:
   prop 1    prop 2
............
```

Question 5. (20 marks)

(a) Answer the following questions on the theory of statistical inference.

i. (4 marks) Briefly describe the central limit theorem.

ii. (2 marks) Provide a brief description of the standard error.

iii. (3 marks) In the context of hypothesis testing, explain what a p-value is, and how it is used. Support your answer with a simple example.

iv. (2 marks) What is meant by Type I error and Type II error?

(b) The standard deviations of data sets `X` and `Y` are given in the `R` code below. An inference procedure was carried out to assess whether or not `X` and `Y` can be assumed to have equal variance.

i. (1 marks) Formally state the null and alternative hypothesis.

ii. (2 marks) The test statistic has been omitted from the `R` code output. Compute the value of the test statistic.

iii. (2 marks) What is your conclusion for this procedure? Justify your answer.

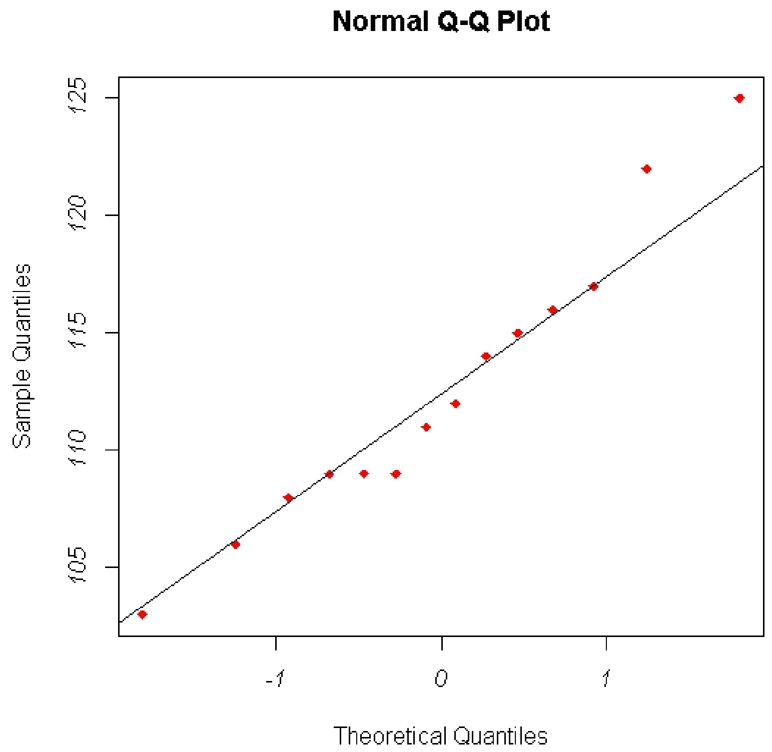
iv. (1 Marks) Explain how a conclusion for this procedure can be based on the 95% confidence interval.

```
> sd(X)
[1] 10.8557
>
> sd(Y)
[1] 5.58845
>
> var.test(X,Y)

        F test to compare two variances

data:  X and Y
F = ......, num df = 13, denom df = 13, p-value = 0.02312
alternative hypothesis: true variances ratio is not 1
95 percent confidence interval:
  1.21135 11.75427
sample estimates:
ratio of variances
          .......
```

(c) The data set `X` and `Y` are both assumed to be normally distributed. A graphical procedure was carried out to assess whether or not this assumption of normality is valid for data set `Y`. Consider the Q-Q plot in the figure below.



i. (2 marks) Provide a brief description on how to interpret this plot.

ii. (1 marks) What is your conclusion for this procedure? Justify your answer.

Question 6. (20 marks)

(a) The input source to a noisy communication channel is a random variable X over the four symbols $\{a, b, c, d\}$. The output from this channel is a random variable Y over these same four symbols. The joint distribution of these two random variables is as follows:

	x=a	x=b	x=c	x=d
y=a	3/16	0	0	1/16
y=b	0	1/4	1/16	0
y=c	0	1/16	5/16	0
y=d	1/64	0	0	3/64

i. (2 marks) Write down the marginal distribution for X and compute the marginal entropy $H(X)$.

ii. (2 marks) Write down the marginal distribution for Y and compute the marginal entropy $H(Y)$.

iii. (4 marks) What is the joint entropy $H(X, Y)$ of the two random variables?

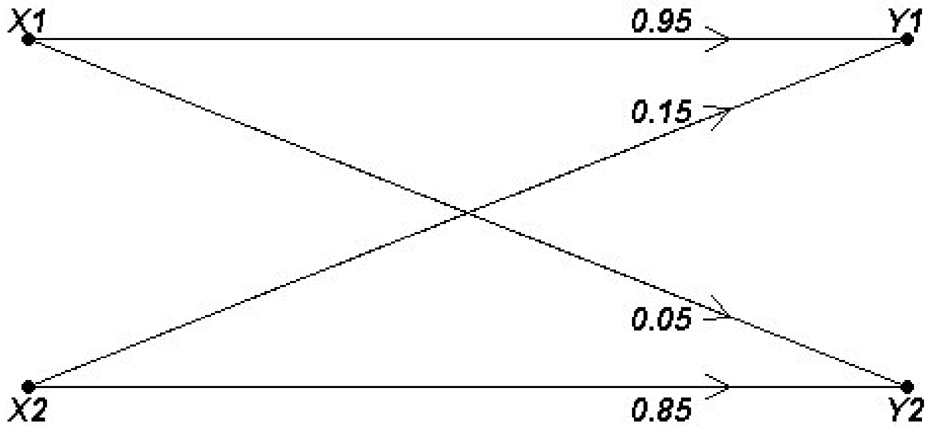
iv. (4 marks) What is the conditional entropy $H(Y|X)$?

v. (4 marks) What is the conditional entropy $H(X|Y)$?

vi. (4 marks) What is the mutual information $I(X; Y)$ between the two random variables?

Question 7. (20 marks)

(a) Consider the binary channel in the figure below.



i. (3 marks) Determine the channel matrix of the channel

ii. (4 marks) Find $P(Y_1)$ and $P(Y_2)$ when $P(X_1) = 0.7$ and $P(X_2) = 0.3$

iii. (3 marks) Find the joint probabilities $P(X_1, Y_1)$ and $P(X_2, Y_2)$.

(b) A discrete memoryless source X has five symbols $\{x_1, x_2, x_3, x_4, x_5\}$ with probabilities $P(x_1) = 0.40$, $P(x_2) = 0.25$, $P(x_3) = 0.15$, $P(x_4) = 0.12$ and $P(x_5) = 0.08$.

i. (4 marks) Construct a Huffman code for X.

ii. (4 marks) Calculate the efficiency of the code.

iii. (2 marks) Calculate the redundancy of the code.

Formulae

Probability

- Conditional probability:

$$P(B|A) = \frac{P\,(A \text{ and } B)}{P\,(A)}.$$

- Bayes' Theorem:

$$P(B|A) = \frac{P\,(A|B) \times P(B)}{P\,(A)}.$$

- Binomial probability distribution:

$$P(X = k) =^n C_k \times p^k \times (1-p)^{n-k} \qquad \left(\text{where} \qquad {}^nC_k = \frac{n!}{k!\,(n-k)!}.\right)$$

- Poisson probability distribution:

$$P(X = k) = \frac{m^k \mathrm{e}^{-m}}{k!}.$$

Information Theory

- $I(p) = -log_2(p) = log_2(1/p)$

- $I(pq) = I(p) + I(q)$

- $H = -\sum_{i=1}^{m} p_i \; log_2(p_i)$

- $E(L) = \sum_{i=1}^{m} l_i p_i$

- Efficiency $= H/E(L)$

- $I(X;Y) = H(X) + H(Y) - H(X,Y)$

- $I(X;Y) = H(X) - H(X|Y)$

- $P(C[r]) = \sum_{j=1}^{m} P(C[r]|Y = d_j)P(Y = d_j)$

Confidence Intervals

One sample

$$\begin{aligned} S.E.(\bar{X}) &= \frac{\sigma}{\sqrt{n}}. \\ S.E.(\hat{P}) &= \sqrt{\frac{\hat{p} \times (100 - \hat{p})}{n}}. \end{aligned}$$

Two samples

$$\begin{aligned} S.E.(\bar{X}_1 - \bar{X}_2) &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \\ S.E.(\hat{P}_1 - \hat{P}_2) &= \sqrt{\frac{\hat{p}_1 \times (100 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (100 - \hat{p}_2)}{n_2}}. \end{aligned}$$

Hypothesis tests

One sample

$$\begin{aligned} S.E.(\bar{X}) &= \frac{\sigma}{\sqrt{n}}. \\ S.E.(\pi) &= \sqrt{\frac{\pi \times (100 - \pi)}{n}} \end{aligned}$$

Two large independent samples

$$\begin{aligned} S.E.(\bar{X}_1 - \bar{X}_2) &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \\ S.E.(\hat{P}_1 - \hat{P}_2) &= \sqrt{(\bar{p} \times (100 - \bar{p}))\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \end{aligned}$$

Two small independent samples

$$\begin{aligned} S.E.(\bar{X}_1 - \bar{X}_2) &= \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \\ s_p^2 &= \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}. \end{aligned}$$

Paired sample

$$S.E.(\bar{d}) \quad = \quad \frac{s_d}{\sqrt{n}}.$$

Standard deviation of case-wise differences

$$s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n-1}}.$$