FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

END OF SEMESTER EXAMINATION

MODULE CODE: MA4413

SEMESTER: Autumn 2013

MODULE TITLE: Statistics for Computing

DURATION OF EXAM: 2.5 hours

LECTURER: Kevin O'Brien

GRADING SCHEME: 100 marks
70% of total module marks

EXTERNAL EXAMINER: Prof. Brendan Murphy

INSTRUCTIONS TO CANDIDATES

This paper is comprised of five questions, each worth 25 marks. Attempt any four questions.
Scientific calculators approved by the University of Limerick can be used.
Formula sheet and statistical tables are provided.

Question 1 [25 Marks]

(a) ***Probability (6 Marks)***
An electronics assembly subcontractor receives its entire supply of resistors from two suppliers. Company A provides 70% of the subcontractor's resistors, while company B supplies the remainder. The additional information has also been made available:

- 2% of the resistors provided by company A failed the final test;
- 3% of company B's resistors also fail the final test.

Answer the following questions:

(i) (3 Marks) What is the probability that a resistor fails the final test?

(ii) (3 Marks) What is the probability that a resistor fails the final test given that the resistor in question came from company A?

(b) ***Descriptive Statistics (5 Marks)***
Consider the following data set of seven numbers:

```
4, 18, 2, 7, 18, 3, 4
```

For this sample, compute the following descriptive statistics:

(i) (1 Mark) the mean,

(ii) (1 Mark) the median,

(iii) (2 Marks) the variance,

(iv) (1 Mark) the standard deviation.

(c) ***Discrete Random Variables (6 Marks)***
The probability distribution of discrete random variable X is tabulated below. There are five possible outcome of X, i.e. 2, 5, 10, 15 and 25.

x_i	2	5	10	15	25
$P(x_i)$	0.25	0.25	0.15	k	0.10

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

(ii) (2 Marks) Determine the expected value $E(X)$.

(iii) (2 Marks) Evaluate $E(X^2)$.

(iv) (1 Mark) Compute the variance of random variable X.

Please turn over for the remaining sections of Question 1.

(d) ***Sampling without Replacement (4 Marks)***
An urn contains 10 disks, 6 white and 4 red. Two disks are selected, without replacement, from the urn. Calculate the following probabilities:

(i) (2 Marks) at least one disk is white;

(ii) (2 Marks) exactly one disk chosen is white.

(e) ***Independent Events (4 Marks)***
Suppose A and B are two events, with P(A), the probability that A occurs, equal to 0.4 and P(B), the probability that B occurs, equal to 0.5. You may assume that A and B are independent events.

(i) (2 Marks) Calculate $P(A \cap B)$, the probability of both A and B occuring.

(ii) (2 Marks) Calculate $P(A \cup B)$, the probability of either A or B (or both) occuring.

Question 2 [25 Marks]

(a) ***Probability Distributions (9 Marks)***
Telephone calls arrive at a switchboard at the rate of 40 per hour. Assume that the telecentre operators take 3 minutes to deal with a customer query. Calculate the probability of :

(i) (3 Marks) 2 or more calls arriving in any 3 minute period.

(ii) (2 Marks) No phone calls arriving in a 3 minute period,

(iii) (3 Marks) Exactly one phone call arriving in any 3 minute period,

(iv) (1 Marks) What is the average and standard deviation of the number of phone calls arriving in a 3 minute? period.

(b) ***Probability Distributions (8 Marks)***
For a digital communication channel, the probability of a bit being received in error is 5%. Consider the case where 100 bits are transmitted. Answer the following questions.

(i) (3 marks) What is the probability that the number of bits received in error is 5?

(ii) (3 marks) What is the probability that the number of bits received in error is greater than 10?

(iii) (2 marks) What is the probability that the number of bits received in error does not exceed 12?

(c) ***Probability Distributions (5 Marks)***
On average, six people per hour use an electronic teller machine during the prime shopping hours in a department store. Therefore it is assumed that the expected time until the next customer will arrive will be 10 minutes. You may assume that the distributions of waiting times can be described by the exponential probability distribution.

(i) (3 Marks) What is the probability that at least 10 minutes will pass between the arrival of two customers?

(ii) (2 Marks) What is the probability that after a customer leaves, another customer does not arrive for at least 20 minutes?

(d) ***Poisson Approximation of the Binomial Distribution (3 Marks)***

(i) (2 Marks) Describe how the Poisson distribution can be used to approximate the binomial distribution.

(ii) (1 Mark) Explain the circumstances in which this approximation may be used in preference to the binomial distribution.

Question 3 [25 marks]

(a) ***Normal Probability Distribution (8 Marks)***
A well known IT retailer has determined that the number of computer accessories sold, on a weekly basis in its largest shop, is normally distributed with a mean of 1000 and standard deviation of 50.

(i) (2 Marks) Estimate the proportion of weeks that the company will sell more than 950 products.

(ii) (3 Marks) Estimate the proportion of weeks that the company will sell less than 975 products.

(iii) (3 Marks) Estimate the proportion of weeks that the company will sell between 950 products and 1025 products?

(b) ***Theory of Statistical Inference (6 Marks)***
Answer the following questions on the theory of statistical inference.

(i) (2 Marks) What is a $p-$value?

(ii) (2 Marks) Briefly describe how $p-$value is used in hypothesis testing.

(iii) (1 Mark) What is meant by a "Type I error"?

(iv) (1 Mark) What is meant by a "Type II error"?

(c) ***Confidence Intervals (4 Marks)***
A press release for a broadband provider stated that 90% percent of customers are very satisfied with the standard of service. To test this claim, the local chamber of commerce surveyed 110 randomly selected customers. Among the sampled customers, 89 stated that they are very satisfied.

(i) (1 Mark) Compute the appropriate value for the standard error for a confidence interval.

(ii) (2 Marks) Compute the 95% confidence interval for π, the true proportion.

(iii) (1 Mark) What is your conclusion for this claim made by the press release? Justify your answer.

Please turn over for the remaining sections of Question 3.

(d) ***Inference Procedures with*** `R` ***(4 Marks)***

Consider the following inference procedure performed on data set X.

```
> shapiro.test(X)

        Shapiro-Wilk normality test

data:  X
W = 0.9619, p-value = 0.6671
```

(i) (1 Mark) Describe what is the purpose of this statistical procedure.

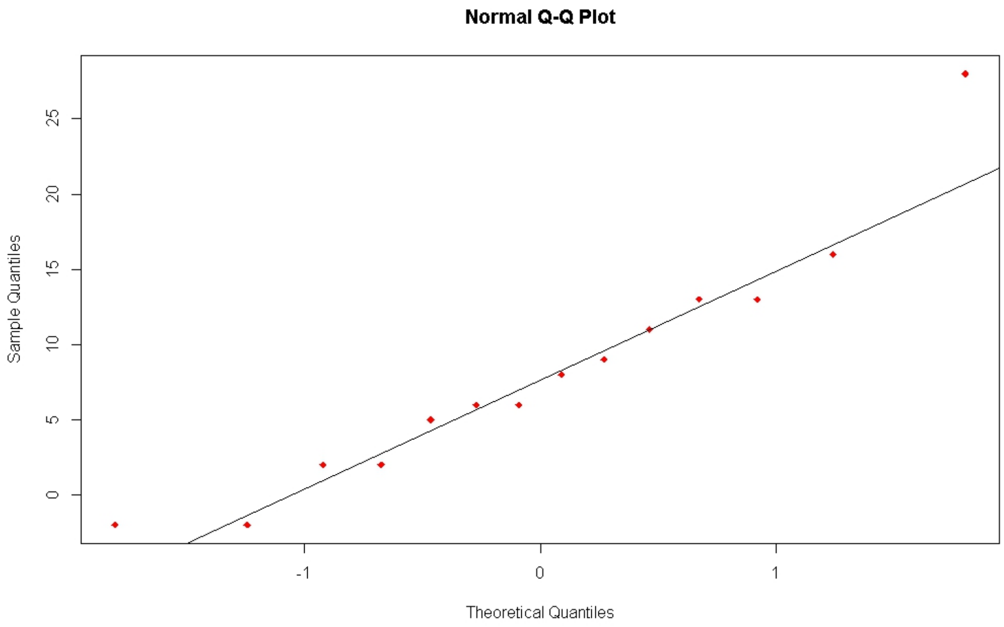
(ii) (2 Marks) What are the null and alternative hypotheses?

(iii) (1 Mark) Write the conclusion that follows from the code output, displayed above.

(e) ***Graphical Procedures (3 Marks)***

(i) (2 Marks) The graph below depicts a normal probability plot. Describe what this plot is used for and how to interpret one.

(ii) (1 Mark) What is your conclusion for the data used to construct the normal probability plot below?



Question 4 [25 marks]

(a) ***Binary Classification (6 Marks)***
For following binary classification outcome table, calculate the following appraisal metrics.

(i) (1 Mark) accuracy;

(ii) (1 Mark) recall;

(iii) (1 Mark) precision;

(iv) (1 Mark) F-measure.

	Predict Negative	Predict Positive
Observed Negative	9530	10
Observed Positive	300	160

(v) (2 Marks) Explain why the F-measure is considered a more informative measure of performance than the Accuracy score.

(b) ***Inference Procedures (10 Marks)***
Two IT training companies, *XtraTech* and *YourSkills*, offer an exam preparation course for a well-known computer industry certification. A study was carried out to compare the results from the most recent group of students from both companies.

- 30 students from the *XtraTech* course have completed the test. The average score for these students was 910 marks with a standard deviation of 48 marks.
- 25 students from the *YourSkills* course have completed the test. Their average score was 950 marks with a standard deviation of 42 marks.

Test the hypothesis that the both sets of students perform equally well on average. You may use a significance level of 5%. You may assume that both samples are normally distributed and have equal variance.

(i) (2 Marks) Formally state the null and alternative hypotheses for this procedure.

(ii) (2 Marks) Compute the point estimate for the difference in means of the results from both courses.

(iii) (2 Marks) Compute the appropriate value for standard error for this test. Clearly show your workings.

(iv) (2 Marks) Compute the test statistic.

(v) (2 Marks) What is your conclusion for this procedure?

Please turn over for the remaining sections of Question 4.

(c) ***Inference Procedures (9 Marks)***
A study finds that 42% of IT users out of a random sample of 450 in a large community preferred one web browser to all others. In another large community, 34% of IT users out of a random sample of 350 prefer the same web browser.

(i) (2 Marks) Compute the point estimate for the difference in proportions of IT users who prefer this particular web browser.

(ii) (4 Marks) Compute a 95% confidence interval for this difference in proportions.

(iii) (3 Marks) Based on this confidence interval, test the hypothesis that the proportion of IT users using this web browser is the same for both communities. State your null and alternative hypotheses clearly.

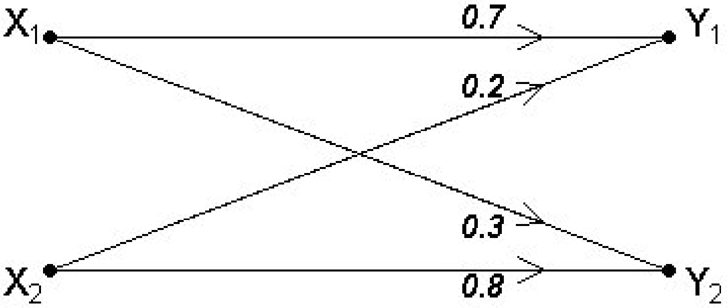
Question 5 [25 marks]

(a) ***Huffman Coding (8 Marks)***
A discrete memoryless source X has five symbols $\{x_1, x_2, x_3, x_4, x_5\}$ with probabilities $P(x_1) = 0.45$, $P(x_2) = 0.20$, $P(x_3) = 0.16$, $P(x_4) = 0.14$ and $P(x_5) = 0.05$.

(i) (5 Marks) Construct a Huffman code for X.

(ii) (3 Marks) Calculate the efficiency of the code.

(b) ***Binary Channels (8 Marks)***
Consider the binary channel in the figure below.



(i) (2 Marks) Determine the channel matrix of the channel

(ii) (3 Marks) Find $P(Y_1)$ and $P(Y_2)$ when $P(X_1) = 0.65$ and $P(X_2) = 0.35$.

(iii) (3 Marks) Find the joint probabilities $P(X_1, Y_1)$ and $P(X_2, Y_2)$ when $P(X_1) = 0.65$ and $P(X_2) = 0.35$.

Please turn over for the remaining sections of Question 5.

(c) ***Rate of Information (5 Marks)***
A high-resolution TV picture consists of about 2×10^6 picture elements (symbols) and 16 different brightness levels.

Pictures are repeated at a rate of 32 per second. All picture elements are assumed to be independent, and all levels have equal likelihood of occurrence.

(i) (5 Marks) Calculate the average rate of information conveyed by this TV picture source.

(d) ***Communication Channels (4 Marks)***
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	1/8	0	0	0
y=b	0	1/4	1/8	0
y=c	0	1/16	1/8	0
y=d	1/16	0	0	1/4

(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 $H(Y)$.

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 $= H/E(L)$
- $I(X;Y) = H(X) - H(X|Y)$
- $P(C[r]) = \sum_{j=1}^{m} P(C[r]|Y = d_j)P(Y = d_j)$
- $R = rH(X)$ (b/second)

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

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

Standard deviation of case-wise differences

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

Binary Classification

- $F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$