UNIVERSITY of LIMERICK

OLLSCOIL LUIMNIGH

FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

END OF SEMESTER ASSESSMENT PAPER

MODULE CODE: MA4413

SEMESTER: Autumn 2014

MODULE TITLE: Statistics for Computing

DURATION OF EXAM: 2.5 hours

LECTURER: Dr. Kevin Burke

GRADING SCHEME: 100 marks (60% of module)

INSTRUCTIONS TO CANDIDATE

- **Attempt four** of the six questions (each one carries 25 marks).
- All work must be shown *clearly and logically* using appropriate symbols and probability notation. Failure to do so will *lose marks*.
- Write down the formula you intend to use at each stage *before* filling it in with numbers.
- Formula sheets are provided at the back of this exam paper.
- Statistical tables are available from the invigilators.

Question 1 **(25 Marks)**

a) We wish to investigate the battery life for two types of laptop. Eight Type A and eight Type B laptops were fully charged and then run until the batteries reached a critical level. In each case the time (in hours) was recorded and the results were as follows:

Type A	2.7	2.9	4.1	4.3	5.7	6.6	9.0	15.0
Type B	1.0	1.0	1.8	2.3	2.3	3.3	3.5	5.5

i) Find the values of Q_1, Q_2 and Q_3 for each type. **(3 marks)**

ii) Identify any outliers in each group. **(2 marks)**

iii) Draw the boxplots for each type side by side. **(3 marks)**

iv) Comment on the boxplots. **(2 marks)**

---

b) Identify the data type for each of the following quantities:

i) laptop battery life ; **(1 mark)**

ii) age group (“18 – 22”, “23 – 30”, “30+”) ; **(1 mark)**

iii) the number of lines of code in a program ; **(1 mark)**

iv) gender ; **(1 mark)**

v) temperature. **(1 mark)**

---

c) A bottle-filling machine is programmed to put 500ml into each bottle. To test if the machine is working correctly, a sample of 40 bottles was selected and, for this sample, the average content was 501.5ml and the standard deviation was 3.05 ml.

i) What is the parameter here? (provide symbol and value) **(2 marks)**

ii) What is the statistic here? (provide symbol and value) **(2 marks)**

iii) Calculate a 95% confidence interval for the parameter; does the evidence suggest that the machine is working as programmed? **(3 marks)**

iv) How large a sample is required to reduce the margin of error in the previous confidence interval to ± 0.5ml? **(3 marks)**

Question 2 **(25 Marks)**

a) Consider the following sample of incomes (in thousands) of 40 individuals living in a particular area:

14	14	15	15	15	15	15	15	15	15
15	15	16	18	19	19	19	20	21	21
22	22	24	24	24	26	27	28	29	33
36	38	38	39	41	44	44	45	48	71

i) Construct a frequency table with 6 classes (note: use 12 as the first breakpoint when setting up the intervals). **(4 marks)**

ii) Draw the histogram. **(3 marks)**

iii) What measure of centrality is appropriate for this data? Calculate its value. **(3 marks)**

---

b) A manufacturer wants to compare two designs of CPU in terms of clock speed. Two small samples are selected and the results are as follows:

	Design 1	Design 2
sample size	8	12
mean	1.211 Ghz	0.870 Ghz
variance	0.05 Ghz^2	0.04 Ghz^2

i) Before comparing means, `var.test` was carried out using `R`:

```
 F test to compare two variances
F = 1.243, num df = 7, denom df = 11, p-value = 0.7168
alternative hypothesis:  true ratio of variances is not equal to 1
```

State H_0 and H_a for this test. Based on the above output, provide your conclusion (this impacts your calculations for part (ii)). **(3 marks)**

ii) Formally test the hypothesis that there is no difference in the mean clock speeds for the two CPU designs using the 1% level of significance:

- Write down H_0 and H_a.
- Compute the test statistic (equal or non-equal variance approach?).
- Compare this test statistic to the appropriate critical value.
- Conclusion: statistical *and* non-statistical language. **(10 marks)**

iii) The t-test requires the data to be approximately normally distributed. What plot is used to check this? Draw a rough picture of what such a plot looks like. **(2 marks)**

Question 3 **(25 Marks)**

a) A sample of employees was randomly selected. Each of them was assigned the same task of programming a procedure using C++. The number of lines of code used in each case was recorded:

13	9	8	10	5

Calculate the following for the above sample:

i) the mean; **(1 mark)**

ii) the variance; **(2 marks)**

iii) the standard deviation; **(1 mark)**

iv) a 90% confidence interval for the true mean. **(3 marks)**

---

b) Let $\Pr(A) = 0.4$, $\Pr(B) = 0.8$ and $\Pr(A \cap B) = 0.3$.

Calculate the following:

i) $\Pr(A \cup B)$; **(1 mark)**

ii) $\Pr(B \mid A)$; **(1 mark)**

iii) $\Pr(A^c \cup B^c)$. **(1 mark)**

---

c) Assume that a manufacturer of laptops sources processors from two different companies: C_1 and C_2. Specifically, 80% of stock comes from C_1 and the rest comes from C_2.

Let X be the temperature of a CPU after one hour of moderate use.

For a C_1 processor the temperature is $\text{Normal}(\mu_1 = 30, \sigma_1 = 1)$.

For a C_2 processor the temperature is $\text{Normal}(\mu_2 = 29, \sigma_2 = 5)$.

i) Show that (rounding to two decimal places):
- $\Pr(X > 31 \mid C_1) = 0.16$ and
- $\Pr(X > 31 \mid C_2) = 0.34$. **(4 marks)**

ii) Calculate $\Pr(X > 31 \cap C_1)$ and $\Pr(X > 31 \cap C_2)$. **(4 marks)**

iii) Calculate $\Pr(X > 31)$ using the law of total probability. **(3 marks)**

iv) Calculate $\Pr(C_1 \mid X < 31)$. **(4 marks)**

Question 4 **(25 Marks)**

---

a) Consider a RAID-1 (redundant array of inexpensive disks) system constructed using two hard disks that work/fail *independently* of each other. Let H_1 = "hard disk 1 works" and H_2 = "hard disk 2 works" and, furthermore, $\Pr(H_1) = \Pr(H_2) = 0.2$, i.e., these hard disks are of a very low quality (20% chance of working).

  i) Calculate Pr(RAID-1 fails). Note that a RAID-1 system will only fail if *both* hard disks fail. **(2 marks)**

  ii) How many hard disks are required to reduce the failure probability, Pr(RAID-1 fails), to 0.01? **(3 marks)**

---

b) Assume that 10% of all audio cables produced by a particular company are defective in some way. Thus, if defects occur independently of one another, the number of defective cables in a shipment of size n has a binomial distribution, i.e., $X \sim \text{Binomial}(n, p)$.

Calculate:

  i) the probability that there are at least 2 defective cables in a shipment of size 10 ; **(3 marks)**

  ii) the probability that there are less than 15 defective cables in a shipment of size 100 ; **(3 marks)**

  iii) the expected number of defective cables in a shipment of size 20 and the corresponding standard deviation. **(3 marks)**

---

c) Emails arrive at a rate of 4 per hour according to a Poisson distribution.

Calculate:

  i) the probability of receiving exactly 4 emails in a one-hour period ; **(2 marks)**

  ii) the probability of receiving between 2 and 4 emails in *half* an hour ; **(3 marks)**

  iii) the probability of receiving between 10 and 20 emails in a five-hour period ; **(3 marks)**

  iv) the probability that the *waiting time* until the next email is greater than 30 minutes. **(3 marks)**

Question 5 **(25 Marks)**

a) Consider the following probability distribution:

x	1	3	5	10
$\Pr(X = x)$	??	0.2	0.2	0.1

Calculate:

i) $\Pr(X = 1)$; **(1 mark)**

ii) the expected value, $E(X)$, and explain what it means; **(2 marks)**

iii) the standard deviation, $Sd(X)$. **(2 marks)**

b) A soft drinks company is working on a new recipe for its best-selling drink. The company intends to carry out a study where participants will taste both flavours (current and new) and then answer the question:

"Do you prefer the new flavour?"

It is assumed that the *current* recipe is superior, i.e., that *less than or equal to* 50% of people prefer the new drink ($p \leq 0.5$).

We wish to test the hypothesis that $p \leq 0.5$.

i) What type of data will be collected in this study? **(2 marks)**

ii) State the null and alternative hypotheses. **(2 marks)**

iii) From a sample of 65 people, we find that 43 people prefer the new recipe. Calculate the test statistic and, hence, the p-value. **(4 marks)**

iv) Based on the evidence (i.e., the p-value), state your conclusion in both stastical and non-statistical language. **(2 marks)**

c) Let $X \sim \text{Normal}(\mu = 20, \sigma = 3)$.

Calculate the following:

i) $\Pr(X < 25)$; **(3 marks)**

ii) the value x such that $\Pr(X > x) = 0.1$; **(3 marks)**

iii) $\Pr(\overline{X} > 19.5)$ where $\overline{X}$ is the sample mean for a group of $n = 40$. **(4 marks)**

Question 6 (25 Marks)

---

a) Customers arrive to a service counter at a rate of $\lambda_a = 30$ per hour; the average service time is $E(T_s) = 0.025$ hours. Assume that this is an $M/M/1$ system, i.e., the number of arrivals per hour is $X_a \sim \text{Poisson}(\lambda_a)$ and the service time is $T_s \sim \text{Exponential}(\lambda_s)$. Also define:

- T = time spent in the whole system ;
- N = number of customers in the whole system ;
- T_q = time spent in the queue component ;
- N_q = number of customers in the queue component.

Calculate:

i) the service rate ; **(2 marks)**

ii) $E(T)$, $E(N)$, $E(T_q)$ and $E(N_q)$; **(5 marks)**

iii) the utilisation factor and interpret its value ; **(2 marks)**

iv) the probability that a customer spends less than 12 minutes in the system ; **(3 marks)**

v) the probability that more than 7 customers exit the system in a 10 minute period. **(3 marks)**

---

b) A source file contains only four unique characters as indicated below.

x	a	b	c	d
$p(x)$	0.5	0.25	0.2	0.05

i) Calculate the entropy for this file. **(3 marks)**

ii) Construct a Huffman code for the characters $\{a, b, c, d\}$. **(3 marks)**

iii) Calculate the expected length of this Huffman code and, hence, its efficiency. **(4 marks)**

Useful Formulae: Page 1

Histogram:

- class width $= \dfrac{\max(x) - \min(x)}{\text{number of classes}}$

Numerical Summaries:

- $\bar{x} = \dfrac{\sum x_i}{n}$
- $s^2 = \dfrac{\sum x_i^2 - n\,\bar{x}^2}{n-1}$
- Position of Q_k : $\dfrac{n+1}{4} \times k$
- $IQR = Q_3 - Q_1$
- $LF = Q_1 - 1.5 \times IQR$
- $UF = Q_3 + 1.5 \times IQR$

Probability:

- $\Pr(A^c) = 1 - \Pr(A)$
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
- $\Pr(E_1 \cup E_2 \cup \cdots \cup E_k) = \Pr(E_1) + \Pr(E_2) + \cdots + \Pr(E_k)$ (if mutually exclusive)
- $\Pr(A \cap B) = \Pr(A)\ \Pr(B \mid A) = \Pr(B)\ \Pr(A \mid B)$
- $\Pr(E_1 \cap E_2 \cap \cdots \cap E_k) = \Pr(E_1)\ \Pr(E_2)\ \cdots\ \Pr(E_k)$ (if independent)
- $\Pr(A \mid B) = \dfrac{\Pr(A \cap B)}{\Pr(B)} = \dfrac{\Pr(A)\ \Pr(B \mid A)}{\Pr(B)}$
- $$\begin{aligned}\Pr(B) &= \Pr(B \cap E_1) + \Pr(B \cap E_2) + \cdots + \Pr(B \cap E_k) \\ &= \Pr(E_1)\ \Pr(B \mid E_1) + \Pr(E_2)\ \Pr(B \mid E_2) + \cdots + \Pr(E_k)\ \Pr(B \mid E_k)\end{aligned}$$

  (if $E_1, \ldots, E_k$ are mutually exclusive & exhaustive)

Useful Formulae: Page 2

Counting Techniques:

- $n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$
- $\binom{n}{k} = \dfrac{n!}{k!\,(n-k)!}$

Random Variables:

- $E(X) = \sum x_i \; p(x_i)$
- $E(X^2) = \sum x_i^2 \; p(x_i)$
- $Var(X) = E(X^2) - [E(X)]^2$
- $Sd(X) = \sqrt{Var(X)}$

Distributions:

• $X \sim \text{Binomial}(n, p)$	• $X \sim \text{Poisson}(\lambda)$	• $T \sim \text{Exponential}(\lambda)$
• $\Pr(X = x) = \binom{n}{x} p^x \, (1-p)^{n-x}$	• $\Pr(X = x) = \dfrac{\lambda^x}{x!} \, e^{-\lambda}$	• $\Pr(T > t) = e^{-\lambda t}$
• $x \in \{0, 1, 2, \ldots, n\}$	• $x \in \{0, 1, 2, \ldots, \infty\}$	• $t \in [0, \, \infty)$
• $E(X) = n\,p$	• $E(X) = \lambda$	• $E(T) = \dfrac{1}{\lambda}$
• $Var(X) = n\,p\,(1-p)$	• $Var(X) = \lambda$	• $Var(T) = \dfrac{1}{\lambda^2}$

Note: the normal distribution is shown on the next page

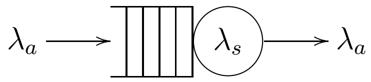
Useful Formulae: Page 3

Queueing Theory:

- $E(N) = \lambda_a \, E(T)$

- $\rho = \dfrac{\lambda_a}{\lambda_s}$

- $M/M/1$ System:



$$\Rightarrow \; T \sim \text{Exponential}(\lambda_s - \lambda_a)$$

(where T is the total time in the system)

Normal Distribution:

- $X \sim \text{Normal}(\mu, \sigma)$

- $E(X) = \mu$

- $Var(X) = \sigma^2$

- $(1 - \alpha)100\%$ of the Normal(μ, σ) distribution lies in $\mu \pm z_{\alpha/2} \, \sigma$

- $\Pr(X > x) = \Pr\left(Z > \dfrac{x - \mu}{\sigma}\right)$

- $\Pr(Z < -z) = \Pr(Z > z)$

- $\Pr(Z > -z) = \Pr(Z < z) = 1 - \Pr(Z > z)$

- If $X_1 \sim \text{Normal}(\mu_1, \sigma_1)$ and $X_2 \sim \text{Normal}(\mu_2, \sigma_2)$

$$\Rightarrow \quad \text{Sum:} \quad X_1 + X_2 \sim \text{Normal}\left(\mu_1 + \mu_2, \; \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

$$\Rightarrow \quad \text{Difference:} \quad X_1 - X_2 \sim \text{Normal}\left(\mu_1 - \mu_2, \; \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

- For $X_1, \ldots, X_n \sim$ any distribution with $\mu = E(X)$ and $\sigma = Sd(X) = \sqrt{Var(X)}$

$$\Rightarrow \quad \text{Sample mean:} \quad \overline{X} \sim \text{Normal}\left(\mu, \; \frac{\sigma}{\sqrt{n}}\right) \quad \text{if } n > 30$$

Useful Formulae: Page 4

Statistics and Standard Errors:

Parameter	Statistic	Standard Error	Samples	Details
μ	$\bar{x}$	$\dfrac{s}{\sqrt{n}}$	large / small	$\nu = n - 1$
p	$\hat{p}$	$\sqrt{\dfrac{\hat{p}\,(1-\hat{p})}{n}}$	large	confidence interval
		$\sqrt{\dfrac{p_0\,(1-p_0)}{n}}$	large	hypothesis test
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$	large / small	$\nu = \dfrac{(a+b)^2}{\frac{a^2}{n_1-1} + \frac{b^2}{n_2-1}}$ $a = \dfrac{s_1^2}{n_1}, \quad b = \dfrac{s_2^2}{n_2}$
		$\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}$ where $s_p^2 = \dfrac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2}$	small	$\nu = n_1 + n_2 - 2$ assuming $\sigma_1^2 = \sigma_2^2$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\dfrac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}}$	large	confidence interval
		$\sqrt{\dfrac{\hat{p}_c\,(1-\hat{p}_c)}{n_1} + \dfrac{\hat{p}_c\,(1-\hat{p}_c)}{n_2}}$ where $\hat{p}_c = \dfrac{x_1+x_2}{n_1+n_2}$	large	hypothesis test

Confidence Intervals:

- Large sample: statistic $\pm\ z_{\alpha/2} \times$ standard error
- Small sample: statistic $\pm\ t_{\nu,\alpha/2} \times$ standard error

Useful Formulae: Page 5

Hypothesis Testing:

- $$z = \frac{\text{statistic} - \text{hypothesised value}}{\text{standard error}}$$

- $$\text{p-value} = \begin{cases} 2 \times \Pr(Z > |z|) & \text{if } H_a: \mu \neq \mu_0 \\ \Pr(Z < z) & \text{if } H_a: \mu < \mu_0 \\ \Pr(Z > z) & \text{if } H_a: \mu > \mu_0 \end{cases}$$

- $$F = \frac{\text{larger variance}}{\text{smaller variance}} = \frac{s^2_{\text{larger}}}{s^2_{\text{smaller}}}$$

  $$\nu_1 = n_{\text{top}} - 1, \quad \nu_2 = n_{\text{bottom}} - 1$$

- $$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

Goodness-of-fit: $e_i = \text{total} \times p(x_i)$, $\nu = n_f - 1 - k$

Independence: $e_{ij} = \dfrac{r_i \times c_j}{\text{total}}$, $\nu = (n_r - 1) \times (n_c - 1)$

Information Theory:

- $h(x) = -\log_2[p(x)]$

- $H(X) = E[h(X)] = \sum h(x_i)\, p(x_i)$

- $l(x_i)$ = code-length for character x_i

- $E(L) = \sum l(x_i)\, p(x_i)$

- $$e = \frac{H(X)}{E(L)}$$

- $\sum 2^{-l(x_i)} \leq 1$