# UNIVERSITY *of* LIMERICK

### O L L S C O I L   L U I M N I G H

# FACULTY OF SCIENCE AND ENGINEERING

## DEPARTMENT OF MATHEMATICS AND STATISTICS

## END OF SEMESTER ASSESSMENT PAPER

MODULE CODE: MA4413            SEMESTER: Autumn 2015

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) Let $\Pr(A) = 0.7$, $\Pr(B) = 0.6$ and $\Pr(A \cap B) = 0.5$.

Calculate the following:

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

    ii) $\Pr(A^c \cup B^c)$;                                             **(2 marks)**

    iii) $\Pr(A \mid B)$ and, hence, determine whether or not $A$ and $B$ are independent events.                 **(2 marks)**

---

b) The operating temperature of a particular design of CPU is 22°C. A sample of CPUs were run and the temperature was measured on each as follows:

| 17 | 19 | 24 | 24 | 24 | 26 | 29 | 32 | 32 | 33 | 34 | 34 |
|----|----|----|----|----|----|----|----|----|----|----|----|

    i) Find the values of the first, second and third quartiles.     **(3 marks)**

    ii) Identify any outliers.                                     **(2 marks)**

    iii) Draw a boxplot for the above sample.                **(3 marks)**

    iv) Based on the boxplot, does the CPU appear to be working as designed.                          **(2 marks)**

---

c) Consider the following output from the `t.test` function in `R`:

```
 One Sample t-test
t = 4.3297, df = 11, p-value = 0.001195
alternative hypothesis:  true mean is not equal to 2
95 percent confidence interval:
2.360548 3.106119
```

Answer the questions below based on the above output.

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

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

    iii) What is the conclusion of this test based on the p-value?     **(3 marks)**

    iv) Interpret the 95% confidence interval and, hence, explain *briefly* how this provides us with the same conclusion as the p-value.     **(3 marks)**

# Question 2 (25 Marks)

a) Consider the following sample of 50 measurements:

| 7 | 9 | 10 | 11 | 11 | 12 | 12 | 12 | 12 | 13 |
|---|---|----|----|----|----|----|----|----|----|
| 13 | 13 | 13 | 13 | 13 | 13 | 13 | 14 | 14 | 14 |
| 14 | 14 | 15 | 15 | 15 | 15 | 15 | 15 | 15 | 15 |
| 15 | 15 | 15 | 16 | 16 | 17 | 17 | 17 | 17 | 17 |
| 18 | 18 | 18 | 18 | 19 | 20 | 20 | 22 | 22 | 23 |

   i) Construct a frequency table with 7 classes (note: let "5" be the lower limit of the first interval). **(4 marks)**

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

   iii) Comment on the shape of the histogram and, hence, suggest (but do not calculate) an appropriate measure of centrality. **(2 marks)**

---

b) Answer the short questions below; keep your answers **brief**.

   i) Both the inter-quartile range and standard deviation are measures of dispersion. Explain specifically what each of these quantities measure. **(2 mark)**

   ii) In what situation might the interquartile range be used instead of the standard deviation? **(2 mark)**

   iii) In terms of population parameters of interest, what is the purpose of calculating statistics such as the sample mean or sample proportion? **(2 mark)**

---

c) Assume that there are two routes to college ($R_1$ and $R_2$) where you take $R_1$ 30% of the time and $R_2$ the rest of the time. Furthermore, there is a 15% chance of being late if you take $R_1$ and a 4% chance of being late if you take $R_2$. Note: let $L$ represent the event of being late.

Calculate the following:

   i) $\Pr(L \cap R_1)$ and $\Pr(L \cap R_2)$; **(4 marks)**

   ii) $\Pr(L)$ ; **(2 marks)**

   iii) $\Pr(R_1 \mid L^c)$ where $L^c$ is the event of being on time. **(4 marks)**

# Question 3 (25 Marks)

a) Consider the following sample of data:

| 5 | 2 | 2 | 3 | 1 | 3 |
|---|---|---|---|---|---|

For this sample, calculate the following:

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

   ii) the standard deviation. **(3 marks)**

b) A market researcher believes that 40% of individuals use a particular brand of mobile device. A sample of 168 individuals were contacted and it was found that 50 of these used the brand in question.

   i) What type of data was collected on each individual? **(1 mark)**

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

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

   iv) Calculate a 99% confidence interval for the parameter; does this support the researcher's belief? **(4 marks)**

   v) What sample size is required to reduce the margin of error in the previous confidence interval to $\pm 0.03$? **(3 marks)**

c) A games developer wanted to test the hypothesis that two games have the same mean gameplay time. Thus, 80 individuals were randomly selected where 40 individuals played Game 1 and the other 40 individuals played Game 2 (until completion). All individual completion time were recorded and can be summarised as follows:

|  | Game 1 | Game 2 |
|---|---|---|
| sample size | 40 | 40 |
| mean | 83.1 hrs | 80.1 hrs |
| variance | 30.6 hrs$^2$ | 18.5 hrs$^2$ |

   i) Formally state the null and alternative hypotheses. **(2 marks)**

   ii) Compute the test statistic and compare this to the appropriate critical value (use the 1% level of significance). **(4 marks)**

   iii) State your conclusion in both statistical *and* non-statistical language. **(3 marks)**

# Question 4 (25 Marks)

a) Consider the following probability distribution:

| $x$ | 0 | 3 | 6 | 9 |
|---|---|---|---|---|
| $\Pr(X = x)$ | 0.1 | 0.4 | $k$ | 0.2 |

Calculate:

   i) the value of $k$; **(1 mark)**

   ii) the expected value, $E(X)$; **(2 marks)**

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

b) Assume that 4% of all individuals have some non-contagious disease. Assuming that occurrences of this disease are independent of one another, the number of individuals with the disease in a sample of size $n$ has a binomial distribution, i.e., $X \sim \text{Binomial}(n, p)$.

Calculate the following:

   i) the expected number of individuals with the disease in a sample of size 80; **(2 marks)**

   ii) the probability that between 2 and 5 individuals (inclusively) have the disease in a sample of size 15; **(3 marks)**

   iii) the probability that more than 8 individuals have the disease in a sample of size 100; **(3 marks)**

   iv) and, explain *briefly* why assuming a binomial distribution may not be appropriate for contagious diseases or for individuals within the same family. **(3 marks)**

c) Tweets arrive at a rate of 7 per hour according to a Poisson distribution.

Calculate:

   i) the probability of receiving 3 or more tweets in *half* an hour; **(3 marks)**

   ii) the probability of receiving between 15 and 25 tweets (inclusively) in a *three*-hour period; **(3 marks)**

   iii) the probability that the *waiting time* until the next tweet is less than 5 minutes. **(3 marks)**

# Question 5　(25 Marks)

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

| $x$ | a | b | c | d |
|---|---|---|---|---|
| $p(x)$ | 0.40 | 0.10 | 0.35 | 0.15 |

  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)**

---

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

  Calculate the following:

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

  ii) $\Pr(23.5 < X < 28.4)$; **(3 marks)**

  iii) the value $x$ such that $\Pr(X > x) = 0.35$; **(3 marks)**

  iv) $\Pr(\overline{X} > 20.8)$ where $\overline{X}$ is the sample mean for a group of size $n = 45$. **(3 marks)**

  v) $\Pr(X_1 + X_2 > 45.7)$ where $X_1$ and $X_2$ are both $\text{Normal}(\mu = 20, \sigma = 3)$ random variables. **(3 marks)**

# Question 6     (25 Marks)

a) An online retailer wants test the hypothesis that there is no difference between two website designs in terms of how much money a customer spends on average. In order to achieve this, two samples of individuals were randomly selected to avail of beta-versions of these websites.

  • A sample of 8 individuals used the first design: their mean spend was calculated as $\$7.96$ and the standard deviation was $\$0.73$.

  • A sample of 7 individuals used the second design: their mean spend was calculated as $\$6.83$ and the standard deviation was $\$2.36$.

  i) Formally state the null and alternative hypothesis.     **(2 marks)**

  ii) Calculate a 95% confidence interval for the difference in the mean customer spend (do **not** assume equal variances).     **(5 marks)**

  iii) State your conclusion in both statistical *and* non-statistical language.
    **(3 marks)**

---

b) Customers arrive to a service counter at a rate of $\lambda_a = 15$ per hour where the average service time is $E(T_s) = 0.05$ 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 (per hour);     **(2 marks)**

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

  iii) $E(T)$ and $E(T_q)$ (give your answer in minutes);     **(2 marks)**

  iv) $E(N)$ and $E(N_q)$;     **(3 marks)**

  v) the probability that a customer spends more than 45 minutes in the system;     **(3 marks)**

  vi) the service rate required to reduce $E(T)$ to 5 minutes and, for this service rate, the corresponding value of $E(N)$.     **(3 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 $Q_k$ : $\quad \dfrac{n+1}{4} \times k$

- $IQR = Q_3 - Q_1$

- $LF = Q_1 - 1.5 \times IQR$

- $UF = Q_3 + 1.5 \times IQR$

**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)

- $\Pr(A \mid B) = \dfrac{\Pr(A \cap B)}{\Pr(B)} = \dfrac{\Pr(A)\,\Pr(B \mid A)}{\Pr(B)}$

- $\Pr(B) = \Pr(B \cap E_1) + \Pr(B \cap E_2) + \cdots + \Pr(B \cap E_k)$

  $= \Pr(E_1)\,\Pr(B \mid E_1) + \Pr(E_2)\,\Pr(B \mid E_2) + \cdots + \Pr(E_k)\,\Pr(B \mid E_k)$

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

# Useful Formulae: Page 2

**Counting Techniques:**

- $n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$

- $\dbinom{n}{k} = \dfrac{n!}{k!\,(n-k)!}$

**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:**

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

**Note: the normal distribution is shown on the next page**

# Useful Formulae: Page 3

**Queueing Theory:**

$$\bullet \quad E(N) = \lambda_a \, E(T)$$

$$\bullet \quad \rho = \frac{\lambda_a}{\lambda_s}$$

$\bullet \quad M/M/1$ System: $\quad \lambda_a \longrightarrow$ ⬚⬚⬚⬚ $\left(\lambda_s\right) \longrightarrow \lambda_a$

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

$$\left(\text{where } T \text{ is the total time in the system}\right)$$

**Normal Distribution:**

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

$$\bullet \quad E(X) = \mu$$

$$\bullet \quad Var(X) = \sigma^2$$

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

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

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

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

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

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

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

$\bullet \quad$ 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 |
|:---:|:---:|:---:|:---:|:---:|
| $\mu$ | $\bar{x}$ | $\dfrac{s}{\sqrt{n}}$ | 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 | $\nu = \dfrac{(a+b)^2}{\frac{a^2}{n_1-1} + \frac{b^2}{n_2-1}}$ $a = \dfrac{s_1^2}{n_1},\ \ b = \dfrac{s_2^2}{n_2}$ |
| | | $\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}$ 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:**

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

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

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

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

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

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

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

**Information Theory:**

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

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

- $l(x_i) = \text{code-length for character } x_i$

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

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

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