**Paper Code**

Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

**Faculty of Engineering, Mathematics and Science**

**School of Computer Science & Statistics**

**Integrated Computer Science Programme** **Hilary Term 2018**
**Year 3**

**ST3009: Statistical Methods for Computer Science**

**DD MMM YYYY** **Venue** **00.00 – 00.00**

**Doug Leith**

**Instructions to Candidates:**

Attempt **all** questions.

You may not start this examination until you are instructed to do so by the invigilator.

**Materials Permitted for this examination:**

Non-programmable calculators are permitted for this examination – please indicate the make and model of your calculator on each answer book used.

1. You buy one share of stock in company C for €10. Each day the price of C either increases by €1 with probability p or decreases by €1 with probability 1-p. These changes from day to day are statistically independent. You decide to sell your share if it gains €2 (i.e. reaches a price of €12).

   (i) What is the probability that you will sell your share exactly 4 days after you buy it ? [5 marks]

   (ii) What is the probability that you sell your share at least 4 days after you buy it ? [5 marks]

   Suppose now that the daily change in the price of stocks in company C is observed to be related to the change in price of stocks in company D. Namely, the probability that stock in C increases by €1 is equal to 0.2 when the price of stock in company D increases that day, and is equal to 0.1 otherwise.

   (iii) State the definition of conditional probability. [5 marks]

   (iv) Describe how marginalisation can be used to calculate the probability of an event E based on knowledge of the conditional probabilities $P(E|F_1)$, $P(E|F_2)$ and $P(E|F_3)$ plus the probabilities $P(F_1)$, $P(F_2)$ and $P(F_3)$ when events $F_1$ , $F_2$ , $F_3$ are mutually exclusive and $F_1 \cup F_2 \cup F_3$ equals the sample space. [5 marks]

   (v) Suppose that the probability that stock in company D increases on a given day is 0.5. Calculate the probability that stock in company C increases that day. [5 marks]

2. Suppose you play a game where four 6-sided fair dice are rolled. Let X be equal to the <u>minimum</u> of the four values rolled (it is ok if more than one dice has the minimal value). It costs €2 to play the game and you win €X.

   (i) Calculate $P(X \geq k)$ as a function of $k=1,2,\ldots,6$. [5 marks]

   (ii) Assuming you know $P(X \geq k)$ for $k=1,2,\ldots,6$, show how to calculate the PMF of X. [5 marks]

   (iii) State the definition of the expected value. [5 marks]

   (iv) Calculate E[X]. [5 marks]

   (v) If you play the game many times do you expect to make a profit (win more than you pay to play the game) ? Explain your reasoning. What is the amount cost to play that would make you break even (i.e. have an expected profit of zero) ? [5 marks]

3. A survey is carried out by selecting n people from the population and asking each person to answer either “yes” or “no” to a question. Let random variable $Y_i$ take value 1 when the i’th respondent answers “yes” and 0 otherwise. The random variables $Y_i$ i=1,2,..,n are independent and identically distributed with $E[Y_i]= \mu$.

   (i) Let random variable $Z = \sum_{i=1}^{n} Y_i$. Write an expression for E[Z] in terms of $E[Y_i]$. Explain your answer. Hint: use the linearity of the expected value.
   [5 marks]

   (ii) Using the definition of expectation prove that E[Z/n]=E[Z]/n for n>0.
   [5 marks]

   (iii) Using Chebyshev’s inequality explain the weak law of large numbers and the behaviour of $|Z/n - \mu|$ as n becomes large. Recall that for random variable X Chebyshev’s inequality is: $P(|X - \mu| \geq k) \leq E[(X- \mu)^2]/k^2$ for any k and μ.
   [5 marks]

   (iv) Explain what a confidence interval is, using Z/n as an estimate of μ as an example. [5 marks]

   (v) Describe how to use bootstrapping to estimate a confidence interval for Z/n.
   [5 marks]

4. Suppose we mark the answers of 200 students to each of 10 exam questions. Let $S_{ij}$ be an indicator variable which is 1 if student i answered question j correctly and -1 otherwise. You observe all of the answers for all students. Assume that

$$P(S_{ij}=y \mid a_i, d_j) = 1/(1+\exp(-y(a_i-d_j))$$

where $a_i$ is a parameter that represents the students ability and $d_j$ is a parameter which represents the questions difficulty.

   (i) Give an expression for the log-likelihood of this exam data (the data consisting of the answers by all 200 students). Hint: this is an example of a logistic regression model. [5 marks]

   (ii) Outline how gradient descent might be used to find the maximum likelihood estimates for the unknown parameters $a_i$ and $d_j$. [5 marks]

   (iii) With reference to Bayes Rule explain what is meant by the likelihood, prior and posterior. [5 marks]

   (iv) Explain how the maximum a posteriori (MAP) estimate of a parameter differs from the maximum likelihood estimate. [5 marks]

   (v) How could you incorporate knowledge of the prior probability distribution of parameters $a_i$ into the above model to obtain a MAP estimate ? [5 marks]