ST3009 – Statistics
2018 Exam Solutions

Question 1

> 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]

Increase = p
Decrease = 1-p

Sell if gain €2

i) In order to satisfy criteria to sell the share price must reach €12. For this to happen exactly 4 days after purchasing it, the following must occur:

- Decrease -> Increase -> Increase -> Increase
- Increase -> Decrease -> Increase -> Increase

The probability of either of these scenarios occurring is as follows:

$$2 * [(p)^3 * (1-p)^2]$$

ii) In order to sell our share at least 4 days after you buy it we can do 1-(Prob Sell on Day 1 + Prob Sell on Day 2 + Prob Sell on Day 3).

- Prob Sell on Day 1 = 0
- Prob Sell on Day 2 = (Increase -> Increase) = p^2
- Prob Sell on Day 3 = 0

Therefore the probability of us selling our share at least 4 days after we buy it is:

$$1 - p^2$$

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 $P(E|F_3)$ 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]

iii) **Conditional probability** is the probability that event E will occur given that event F has already been observed.

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

iv) Suppose we have mutually exclusive events F_1, F_2 and F_3 which together equal the entire sample space S, then:

$$P(E) = P(E|F_1)P(F_1) + P(E|F_2)P(F_2) + P(E|F_3)P(F_3)$$

v) Probability of D increasing on a given day is 0.5. Let C be the probability that company C's stock increases:

$$P(C) = P(C|D)P(D) + P(C|D')P(D')$$

$$= 0.2(0.5) + 0.1(1 - 0.5)$$

$$= 0.3$$

Question 2

2.	Suppose you play a game where four 6-sided fair dice are rolled. Let X be equal to the minimum 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 and you win €X.		
	(i)	Calculate P(X≥k) as a function of k=1,2,...,6.	[5 marks]
	(ii)	Assuming you know P(X≥k) for k=1,2,...,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]

i)

- $P(X \geq 1) = 1$
- $P(X \geq 2) = \left(\frac{5}{6}\right)^4 = 0.4823$
- $P(X \geq 3) = \left(\frac{4}{6}\right)^4 = 0.1975$
- $P(X \geq 4) = \left(\frac{3}{6}\right)^4 = 0.0625$
- $P(X \geq 5) = \left(\frac{2}{6}\right)^4 = 0.0123$
- $P(X \geq 6) = \left(\frac{1}{6}\right)^4 = 0.0008$

ii) We can use the above values to calculate the PMF as follows:

- $P(X = 6) = P(X \geq 6) = 0.0008$
- $P(X = 5) = P(X \geq 5) - P(X \geq 6) = 0.0123 - 0.0008$
- $P(X = 4) = P(X \geq 4) - P(X \geq 5) = 0.0625 - 0.0123$
- $P(X = 3) = P(X \geq 3) - P(X \geq 4) = 0.1975 - 0.0625$
- $P(X = 2) = P(X \geq 2) - P(X \geq 3) = 0.4823 - 0.1975$
- $P(X = 1) = P(X \geq 1) - P(X \geq 2) = 1 - 0.4823$

iii) The definition of the expected value of a discrete random variable X taking values in {x1, x2, …, xn} is defined to be:

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i)$$

iv) For our given dice game we can calculate the expected value as follows:

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i)$$

$$E[X] = 1 * P(X = 1) + 2 * P(X = 2) \dots$$

v) The expected profit for the game is as follows:

$$E[Profit] = E[X] - 2$$

If we were to play the game N times, with N being substantially large we could expect a profit of:

$$N * (E[X] - 2)$$

In order for the cost to play to enable us to break even:

$$E[X] - CostToPlay = 0$$

$$CostToPlay = E[X]$$

Therefore, the cost to play must equal the expected value in order for us to break even.

Question 3

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 an 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]

i) Using the linearity of the expected value we can write an expression for E[Z] as follows:

$$E[Z] = E\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} E[Y_i] = n * E[Y_i]$$

ii) Using the definition of expectation we can prove the above by:

$$E[Z/n] = \sum_{i=1}^{n} \frac{x_i}{n} P(Z = x_i) = \frac{\sum_{i=1}^{n} x_i P(Z = x_i)}{n} = \frac{E[Z]}{n}$$

iii)

$$E[Z/n] = \frac{E[Z]}{n} = \frac{n * E[Y_i]}{n} = E[Y_i] = \mu$$

$$Var\left(\frac{Z}{n}\right) = Var\left(\frac{1}{n} * Z\right) = \frac{1}{n^2} * Var(Z)$$

$$Var(Z) = Var\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} Var(Y_i) = n * Var(Y_i)$$

$$\therefore Var\left(\frac{Z}{n}\right) = \frac{n}{n^2} * Var(Y_i) = \frac{1}{n} * Var(Y_i)$$

Using the above formulae we can then plug then into Chebyshev's inequality:

$$P\left(\left|\frac{Z}{n} - \mu\right| \geq k\right) \leq \frac{Var(Y_i)}{n} * k^2$$

From this, we can infer that for any value of k>0, as n goes to infinity then the RHS goes to 0.

iv) A confidence interval, is an interval [a, b] within which a random variable X lies with a specified probability e.g with a probability of at least 0.95. This can be written as:

$$P(a \ \leq X \leq b) \geq 0.95$$

In the case of Z/n, as an estimate of μ we might consider the interval:

$$P\left(\mu - \epsilon \leq \frac{Z}{n} \leq \mu + \epsilon\right)$$

From part iii) we know that the above probability tends to 1 for any $\epsilon > 0$ as n grows large. Thus, as we increase the number of samples our confidence in stating the that X lies within any given interval tends to 1.

v) From the observed data, Y_i and i=1, 2, …, n, draw a sample of m points uniformly at random with replacement. Using this sample calculate an estimate for Z/n. Repeat to obtain a multiple number of estimates. From the distribution of these estimates we can then estimate a confidence interval for Z/n.

Question 4

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]

i) The log-likelihood of the observed marked data with the variables:

$$S_{ij}$$

$$i = 1, \dots, 200$$

$$j = 1, \dots, 10$$

Is:

$$P\left(S_{ij} = s_{ij}, i = 1, \dots, 200, j = 1, \dots, 10 \,\middle|\, a_i, d_j, i = 1, \dots, 200, j = 1, \dots, 10\right)$$

$$= \prod_{i=1}^{200} \prod_{j=1}^{10} P\left(S_{ij} = s_{ij} \,\middle|\, a_i, d_j\right)$$

The log-likelihood is:

$$L = Log \prod_{i=1}^{200} \prod_{j=1}^{10} P\left(S_{ij} = s_{ij} \,\middle|\, a_i, d_j\right)$$

$$L = \prod_{i=1}^{200} \prod_{j=1}^{10} \log \left[P\left(S_{ij} = s_{ij} \,\middle|\, a_i, d_j\right)\right]$$

$$L = -\prod_{i=1}^{200} \prod_{j=1}^{10} \log \left[1 + \exp\left(-s_{ij}(a_i - d_j)\right]\right.$$

ii) The gradient descent can be used to create an estimate which can select the parameters a_i and d_j to maximise the likelihood L. Starting from an initial estimate, these values van be found iteratively by updating the estimates such that L decreases after each update until the decrease in L becomes small enough. We can find updates that decrease L by local search or by taking a step in the direction of the derivatives of L wrt a_i and d_j.

iii) For random events E and F, Bayes Rule states:

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

- $P(F|E)$ is the Likelihood
- $P(E)$ is the Prior
- $P(E|F)$ is the Posterior

iv) In a *maximum a posteriori (MAP)* estimate the parameter values are selected to maximise the posterior probability P(parameters|data) rather than the likelihood P(data|parameters).

v) By Bayes, the posterior is proportional to:

$$P\left(S_{ij} = s_{ij}, i = 1, \dots, 200, j = 1, \dots, 10 \,\middle|\, a_i, d_j, i = 1, \dots, 200, j = 1, \dots, 10\right)P(a_i, i = 1, \dots, 200)$$

Therefore, the MAP estimate of the $a_i, i = 1, \dots, 200$ maximises this value.