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

FACULTY OF SCIENCE, TECHNOLOGY, ENGINEERING & MATHEMATICS

SCHOOL OF ENGINEERING

Electronic & Electrical Engineering

Engineering **Semester 2, 2022**

Junior Sophister

Annual Examinations

3E3 Probability and Statistics (EEU33E03-1)

Date: 3rd May 2022 **Venue: RDS Main Hall** **Time: 09:30-11:30**

Anthony Quinn

Instructions to candidates:

ANSWER QUESTION 1, and any THREE of the remaining five questions.

Question 1 is worth 30 marks *in toto*. All remaining questions are worth equal marks (*i.e.* 70/3 marks each). The percentage division of marks within each question is indicated on the paper.

All symbols and acronyms have their usual meaning.

Materials Permitted for this Examination:

Standard mathematical tables are available from the invigilators, if required;

Graph paper is available from the invigilators, if required;

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

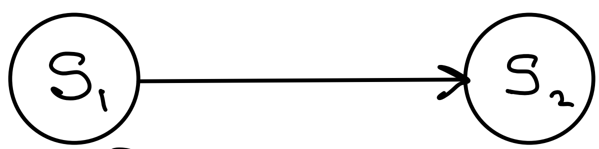
Q.1 [Compulsory]

Answer ALL the following FIVE questions.

(a) Consider three (Boolean) propositions, S_1, S_2 and S_3, with probabilities 0.3, 0.6, and 0.8, respectively. $S_1 \Rightarrow S_2$. Also, $S_1 \perp\!\!\!\perp S_3$ and $S_2 \perp\!\!\!\perp S_3$. Evaluate $\Pr[S_2 \mid \overline{S_1 \vee S_3}]$.

[20 %]

(b) Consider the complete graphical probability model in Fig. Q.1(b), involving two ternary nodes. Evaluate $\Pr[S_1 \in \{1, 2\} \mid S_2 \in \{1, 2\}]$.



$$\underline{P}_1 = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.4 \end{bmatrix} \begin{array}{c} 1 \\ 2 \\ 0 \end{array} \qquad \overline{\overline{T}}_{12} = \begin{array}{c} 1 \\ 2 \\ 0 \\ \\ \end{array} \begin{array}{c} \begin{bmatrix} 0.6 & 0.1 & 0.4 \\ 0.2 & 0.6 & 0 \\ 0.2 & 0.3 & 0.6 \end{bmatrix} \\ \begin{array}{ccc} 1 & \quad 2 & \quad 0 \end{array} \end{array}$$

Figure Q.1(b)

[20 %]

[Q.1 CONTINUES OVERLEAF]

Q.1 [Continued]

(c) Data storage devices are supplied to a university in batches of size 10. Each batch is supplied either by producer A or B, with equal probability. The probability that a device from A is faulty is 0.06, but is twice this for producer B. If a particular batch contains at least 2 faulty devices, what is the probability that it was supplied by B?

[20 %]

(d) Both lanes of a road are uncongested, with the average traffic rate being 12.7 cars per minute in lane 1, and 9.5 cars per minute in lane 2. Given that altogether 5 cars pass a point on this road in *either* lane in 15 s, what is the probability that (exactly) three of these pass in lane 1?

[20 %]

(e) With probability $\frac{2}{3}$, a biometric quantity (arbitrary units), X, is normally distributed with standard interval, 0.3 ± 0.4. Otherwise, $X = +1$. Evaluate the standard interval of X.

[20 %]

Q.2

(a) Consider the graphical probability model in Figure Q.2(a), comprising three uncertain Boolean nodes (i.e. each with two possible states, denoted 1 and 0, respectively). Complete this model with any consistent probability parameters. Hence,deduce each of the following:

(i) The marginal probabilities of the 3rd node. **[20 %]**

(ii) The probability that the first node is in state 0 if the 3rd node is also in state 0. **[25 %]**

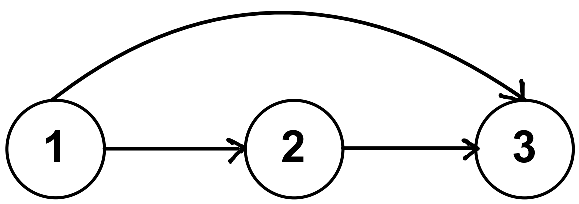


Figure Q.2(a)

(b) The reliability of a new rapid antigen test (A) and a new PCR test (P) are each to be assessed in a particular community affected by the SARS-CoV-2 virus. 55$\%$ of people in this community are virus-***free***, 20$\%$ have a ***low*** viral load, and the rest have a ***high*** viral load, respectively (this viral status (V) is recorded definitively via other means). Either test (A or P) can yield a ***negative***, ***inconclusive*** or ***positive*** result (i.e. 3 possible results), respectively. A statistical study provides the following conditional probability matrix estimates, relating viral load to the results of A and P, respectively:

$$\mathbf{T}_{V\to A} = \begin{bmatrix} 0.75 & 0.23 & 0.02 \\ 0.13 & 0.32 & 0.14 \\ 0.12 & 0.45 & 0.84 \end{bmatrix}; \quad \mathbf{T}_{V\to P} = \begin{bmatrix} 0.93 & 0.10 & 0.01 \\ 0.07 & 0.11 & 0.01 \\ 0.00 & 0.79 & 0.98 \end{bmatrix}.$$

Using this evidence, and stating any assumptions that you make, compute the following:

(i) The probability that the results of A and P differ for a member of this community. **[35 %]**

(ii) The probability that a member of this community is not virus-free, in the case evaluated in (i) (i.e. when the results of A and P differ). **[20 %]**

Q.3

(a) The transition probability matrix of a $q \geq 2$-state *homogeneous Markov chain* (HMC) is $\mathbf{T}$, and the initial probability vector is $\mathbf{p}_1$. Define the technical term in italics, and prove from first principles that the marginal probability vector at the nth stage ($n \geq 1$) of the HMC is given by

$$\mathbf{p}_n = \mathbf{T}^{n-1}\mathbf{p}_1.$$

[40 %]

(b) The carpark in front of a small shop comprises 3 parking spaces. The number of cars (i.e. occupied spaces) is recorded every 10 minutes, starting at 07:00 (opening time) when the number is zero. Using these data, the probability that the number of cars changes after 10 minutes is estimated to be three times the probability that the number does not change. Any change is at most ± 1, and equiprobable where there is a choice. The only exception to these dynamics is when the carpark is full. Then, the number of cars after 10 minutes is 3, 2 or 1, with probabilities 0.6, 0.3 and 0.1, respectively.

Note: these dyanamics imply that the long-run (i.e. stationary) probabilities are 0.2, 0.3, 0.3 and 0.2 (for 0, 1, 2 and 3 cars, respectively).

(i) Specify the computations needed to process 14 hours of these data into sufficient statistics for estimating $\mathbf{T}$, the transition matrix of a HMC model of the sequence. Provide an example of these statistics, consistent with the estimated dynamics above.

[25 %]

(ii) If the carpark first fills up at 07:40, what is the probability that there was one car at 07:20?

[20 %]

(iii) If the carpark is full at 19:00, what is the probability that it was not full 20 minutes earlier?

[15 %]

Q.4

(a) The prevalence of (i.e. percentage in a population with) COVID-19 is estimated, distinguishing between those with mild disease ($C = 1$), acute disease ($C = 2$) and no disease ($C = 0$) (three classes only), yielding the prior probability vector, $\mathbf{p}_\mathsf{C} = [p_1,\ p_2,\ p_0]$.

The reliability of an antigen test is studied. The conditional probabilities of a positive test result ($T = 1$, i.e. disease present) and a negative test result ($T = 0$), respectively, are estimated as follows:

$$\mathbf{T}_{\mathsf{C}\to\mathsf{T}} = \begin{bmatrix} 1-\epsilon_1 & 1-\epsilon_2 & \epsilon_0 \\ \epsilon_1 & \epsilon_2 & 1-\epsilon_0 \end{bmatrix}.$$

NB The study controls for (i.e. eliminates) all inconclusive test results.

(i) Write down an expression—in terms of the probability parameters above—for the total probability of an error in this antigen test, when used by a member of this population.

[15 %]

(ii) A member of this population performs the test r times (r odd), and uses a majority rule to decide whether or not she has COVID-19. Deduce an expression for the total probability of error in this decision, and demonstrate that this is a decreasing function of r. Derive from fundamentals the probability model you use. **[45 %]**

(b) Consider, again, the population in (a), with prior probability vector for COVID-19 class ($C \in \{1, 2, 0\}$ given by $\mathbf{p}_\mathsf{C}$ (i.e. ignoring, now, antigen test results). Write down an expression for each of the following (you do not need to derive the models you adopt):

(i) What is the probability that—among five people in this population—people with COVID-19 are in a majority, and, among those, a majority have mild disease? **[20 %]**

(ii) In a large sample of size n from this population, what is the most probable number having acute disease, and what is this probability? You should assume that p_2 is small. **[20 %]**

Q.5

(a) The fairness of a classifier is to be assessed in the context of a protected attribute, $A(\omega) = a \in 1,2,3$, with prior probabilities, $\mathbf{p}_A = [0.4,\ 0.5,\ 0.1]$ in a particular population. The conditional pdfs of a particular outcome, $X(\omega) = x \in \mathbb{R}$, for members of this population are estimated (i.e. learned) by the classifier from training data, as follows:

$$f(x|a=1) = \mathcal{N}(5.5, 2.1), \quad f(x|a=2) = \mathcal{N}(8.5, 3.4), \quad f(x|a=3) = \mathcal{N}(6.1, 3.6).$$

(i) Compute the correlation coefficient between A and X, and use it to comment on the fairness of the classifier. **[30 %]**

(ii) If $X > 6.0$ for a member of this population, compute the (posterior) probability that $A = 2$.

[20 %]

(b) Three ECG leads on a patient's chest yield the voltage signal vector, $\mathbf{V}$ (mV), modelled as multivariate Gaussian at a single instance of time, as follows:

$$\mathbf{V} = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma}),$$

where

$$\mathbf{m} = \begin{bmatrix} +2.7 \\ +1.8 \\ +0.6 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} +9.1 & -4.3 & +2.1 \\ -4.3 & +5.6 & +0.8 \\ +2.1 & +0.8 & +3.1 \end{bmatrix}.$$

Which pair of voltages is the most strongly correlated? For this pair, deduce both regression lines, and plot them on a single graph, equipping each with its standard interval.

[50 %]

[CONTINUED OVERLEAF]

A note on the bivariate normal distribution:

If

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma}), \quad \text{where} \quad \mathbf{m} = \begin{bmatrix} m_X \\ m_Y \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix},$$

with

$$\sigma_{XY} = \rho_{XY}\sigma_X\sigma_Y,$$

then

$$f(x|Y=y) = \mathcal{N}(m_{X|y},\, \sigma^2_{X|y}),$$

where

$$\begin{aligned} m_{X|y} &= m_X + \rho_{XY}\frac{\sigma_X}{\sigma_Y}(y - m_Y), \\ \sigma^2_{X|y} &= \sigma_X^2\left(1 - \rho_{XY}^2\right). \end{aligned}$$

Q.6

(a) The null and alternative hypotheses for a random sample, $\mathbf{x}_n = [x_1, \ldots, x_n]$, are as follows:

$$\mathcal{H}_0: \quad x_i \overset{\text{iid}}{\sim} \mathcal{N}(m_0, \sigma_0^2)$$

$$\mathcal{H}_1: \quad x_i \overset{\text{iid}}{\sim} \mathcal{N}(m_1, \sigma_1^2)$$

Derive a statistical test to choose between the null and alternative hypotheses for $\mathbf{x}_n$, in the case where $m_0 = m_1 = m$.

[40 %]

(b) The numbers of cars passing a point on a bridge are counted during non-overlapping intervals, each of duration 5 minutes. Under road condition $C = 1$, the average of these counts is 58.7 cars, and it is 44.2 cars under road condition $C = 2$. Later, the following 10 counts, n_i, are recorded during non-overlapping intervals, of variable durations, t_i (minutes), $i = 1, \ldots, 5$:

n_i	55	21	29	28	52	29	70	25	29	43
t_i	6.2	2.9	3.4	3.6	6.5	3.6	7.7	3.1	3.3	5.0

It is known that all 10 of these measurements are made under the same—but unknown—road condition, $C \in \{1, 2\}$.

(i) Estimate appropriate conditional probability models for the car counts under each road condition. Then, use binary hypothesis testing to choose one of them to model the displayed counts. **[30 %]**

(ii) Under your model choice in (i), compute the probability that the third-next car will pass in less than 20 seconds. **[20 %]**

Error Function Table

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

	Hundredths digit of x									
x	0	1	2	3	4	5	6	7	8	9
0.0	0.00000	0.01128	0.02256	0.03384	0.04511	0.05637	0.06762	0.07886	0.09008	0.10128
0.1	0.11246	0.12362	0.13476	0.14587	0.15695	0.16800	0.17901	0.18999	0.20094	0.21184
0.2	0.22270	0.23352	0.24430	0.25502	0.26570	0.27633	0.28690	0.29742	0.30788	0.31828
0.3	0.32863	0.33891	0.34913	0.35928	0.36936	0.37938	0.38933	0.39921	0.40901	0.41874
0.4	0.42839	0.43797	0.44747	0.45689	0.46623	0.47548	0.48466	0.49375	0.50275	0.51167
0.5	0.52050	0.52924	0.53790	0.54646	0.55494	0.56332	0.57162	0.57982	0.58792	0.59594
0.6	0.60386	0.61168	0.61941	0.62705	0.63459	0.64203	0.64938	0.65663	0.66378	0.67084
0.7	0.67780	0.68467	0.69143	0.69810	0.70468	0.71116	0.71754	0.72382	0.73001	0.73610
0.8	0.74210	0.74800	0.75381	0.75952	0.76514	0.77067	0.77610	0.78144	0.78669	0.79184
0.9	0.79691	0.80188	0.80677	0.81156	0.81627	0.82089	0.82542	0.82987	0.83423	0.83851
1.0	0.84270	0.84681	0.85084	0.85478	0.85865	0.86244	0.86614	0.86977	0.87333	0.87680
1.1	0.88021	0.88353	0.88679	0.88997	0.89308	0.89612	0.89910	0.90200	0.90484	0.90761
1.2	0.91031	0.91296	0.91553	0.91805	0.92051	0.92290	0.92524	0.92751	0.92973	0.93190
1.3	0.93401	0.93606	0.93807	0.94002	0.94191	0.94376	0.94556	0.94731	0.94902	0.95067
1.4	0.95229	0.95385	0.95538	0.95686	0.95830	0.95970	0.96105	0.96237	0.96365	0.96490
1.5	0.96611	0.96728	0.96841	0.96952	0.97059	0.97162	0.97263	0.97360	0.97455	0.97546
1.6	0.97635	0.97721	0.97804	0.97884	0.97962	0.98038	0.98110	0.98181	0.98249	0.98315
1.7	0.98379	0.98441	0.98500	0.98558	0.98613	0.98667	0.98719	0.98769	0.98817	0.98864
1.8	0.98909	0.98952	0.98994	0.99035	0.99074	0.99111	0.99147	0.99182	0.99216	0.99248
1.9	0.99279	0.99309	0.99338	0.99366	0.99392	0.99418	0.99443	0.99466	0.99489	0.99511
2.0	0.99532	0.99552	0.99572	0.99591	0.99609	0.99626	0.99642	0.99658	0.99673	0.99688
2.1	0.99702	0.99715	0.99728	0.99741	0.99753	0.99764	0.99775	0.99785	0.99795	0.99805
2.2	0.99814	0.99822	0.99831	0.99839	0.99846	0.99854	0.99861	0.99867	0.99874	0.99880
2.3	0.99886	0.99891	0.99897	0.99902	0.99906	0.99911	0.99915	0.99920	0.99924	0.99928
2.4	0.99931	0.99935	0.99938	0.99941	0.99944	0.99947	0.99950	0.99952	0.99955	0.99957
2.5	0.99959	0.99961	0.99963	0.99965	0.99967	0.99969	0.99971	0.99972	0.99974	0.99975
2.6	0.99976	0.99978	0.99979	0.99980	0.99981	0.99982	0.99983	0.99984	0.99985	0.99986
2.7	0.99987	0.99987	0.99988	0.99989	0.99989	0.99990	0.99991	0.99991	0.99992	0.99992
2.8	0.99992	0.99993	0.99993	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995	0.99996
2.9	0.99996	0.99996	0.99996	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998
3.0	0.99998	0.99998	0.99998	0.99998	0.99998	0.99998	0.99998	0.99999	0.99999	0.99999
3.1	0.99999	0.99999	0.99999	0.99999	0.99999	0.99999	0.99999	0.99999	0.99999	0.99999
3.2	0.99999	0.99999	0.99999	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000