

MAPSI – Examen réparti 1 – 18/11/25 – 61 pts

Durée : 2h

Seuls documents autorisés : 1 feuille A4 recto-verso manuscrite, calculatrice
 – Barème indicatif –

Exercice 1 (7pts) – Probabilités conditionnelles

Soient A et B deux variables aléatoires discrètes. $A \in \{1, 2\}$, $B \in \{1, 2, 3\}$

Soit le tableau ci-contre, partiellement rempli, correspondant à la distribution conditionnelle $\mathbb{P}(B|A)$. On sait, en outre, que

$$\forall k \in \{1, 2, 3\}, \mathbb{P}(B = k|A = 2) \propto k$$

Q 1.1 Compléter le tableau.

Q 1.2 Par ailleurs, on sait que la distribution $\mathbb{P}(A)$ est uniforme.

Déterminer le tableau de la loi jointe $\mathbb{P}(A, B)$. Calculer la marginale en $\mathbb{P}(B)$.

Q 1.3 Les deux variables A et B sont-elles indépendantes selon \mathbb{P} ?

		$\mathbb{P}(B A)$	
		A 1	A 2
B	1	0.1	?
	2	0.2	?
	3	?	?

Exercice 2 (13 pts) – Max de vraisemblance

Un grand conservatoire de musique a collecté les données d'apprentissage sur un échantillon de 1000 étudiants, avec l'idée de distinguer 3 phases d'apprentissage, sachant que l'arrivée au 3^e stade représente une étape importante dans la maîtrise des techniques musicales enseignées dans ce conservatoire.

Temps pour arriver au stade 3 (en années)	1	2	3	4	5
Effectif	50	350	100	250	250

Q 2.1 Modélisation avec une loi géométrique.

Le statisticien en charge de l'étude propose de modéliser chaque année comme une expérience de Bernoulli indépendante puis de considérer que l'arrivée en phase 3 correspond à la première réussite de l'épreuve.

Rappel : Pour une épreuve de Bernoulli de paramètre p , la variable X qui vaut le nombre d'expériences à réaliser jusqu'à la première expérience réussie, a pour distribution une loi géométrique :

$$\forall k \geq 0, p(X = k) = (1 - p)^{(k-1)} p$$

Montrer comment trouver la valeur optimale du paramètre p au sens du maximum de vraisemblance à partir des données recueillies. Quelle est l'estimation fournie pour ce paramètre ? Quel est le nombre moyen d'années nécessaires pour arriver en phase 3 ?

Q 2.2 Un autre collègue statisticien fait remarquer que le recueil des données a pu être légèrement biaisé, les données n'ayant été recueillies que dans un seul conservatoire. Il propose d'ajouter un a priori sous la forme d'une loi Gamma de densité :

$$g(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \text{ avec } \alpha = 5, \beta = 1$$

La fonction Γ est une fonction utilisée communément, qui prolonge la fonction factorielle à l'ensemble des nombres complexes. En particulier, si n entier strictement positif, $\Gamma(n) = (n-1)!$

Montrer comment trouver la valeur optimale du paramètre au sens du maximum a posteriori à partir des données recueillies. Quelle est l'estimation fournie pour ce paramètre ?

Q 2.3 Quelle serait l'estimation du paramètre p par les 2 méthodes précédentes si aucune donnée n'avait été recueillie ? Que pensez-vous du choix de l'a priori ?

Exercice 3 (12 pts) – Tests et tumeurs bénignes

Les tumeurs bénignes peuvent se développer à partir de tous les tissus de l'organisme. Elles n'engagent pas le pronostic vital à la différence des tumeurs malignes, sauf dans certains cas si elles sont mal placées ou trop volumineuses.

Un service dermatologique suit une cohorte (groupe) de 100 patients qui se répartissent en 4 catégories correspondant au diamètre moyen de leurs tumeurs bénignes (sur la peau) respectifs de 3, 5, 8 et 12 mm.

Diamètre moyen (en mm)	3	5	8	12
Nombres de patients	30	20	30	20

Q 3.1 La personne en charge de l'analyse commence par considérer que toutes les classes sont grossièrement équiprobables. Qu'en pensez-vous du point de vue statistique, avec un niveau de confiance de 95% ?

Q 3.2 L'année précédente, le diamètre moyen des tumeurs bénignes était de 7.49 mm sur l'ensemble de la cohorte. Durant cette année, les dermatologues ont tenté un nouveau traitement non chirurgical. Peut-on conclure que le diamètre moyen des tumeurs bénignes a diminué (avec une confiance de 95%, en prenant un écart type sur le diamètre de 3 mm) ?

Q 3.3 Une statisticienne membre de l'équipe remarque qu'il peut y avoir un biais sur le diamètre D , selon les hôpitaux H d'où viennent les patients.

L'échantillon considéré vient de 3 hôpitaux, et l'on a calculé la répartition des 3 hôpitaux parmi la population de même taille moyenne de tumeur :

Diamètre moyen (en mm)	3	5	8	12
Hôpital 1	0.80	0.30	0.20	0.00
Hôpital 2	0.10	0.30	0.20	0.50
Hôpital 3	0.10	0.40	0.60	0.50

Q 3.3.1 En terme de probabilité, à quoi correspond la distribution du tableau ci-dessus par rapport aux variables aléatoires 'hôpital' H et 'diamètre' D ?

Q 3.3.2 Est-il possible de calculer le diamètre moyen des tumeurs par hôpital à partir des informations présentes dans les questions précédentes ? Dans l'affirmative, faites le calcul, sinon, justifier de l'impossibilité de le faire.

Q 3.4 D'après vous, et toujours à un niveau de confiance de 95%, y a-t-il indépendance entre l'hôpital d'origine et le diamètre moyen des tumeurs ?

Exercice 4 (14pts) – Classes de composants

Un fabricant d'un certain composant électronique C très fragile voudrait connaître la durée de vie de ce composant C . A priori, il estime qu'elle est d'environ 2.5 mois. Un échantillon de 1000 composants C a été testé, et les durées de vie constatées (en mois) ont été reportées dans le tableau ci-dessous :

durée (en mois)	0.5	1	2	2.5	3	3.5
nombre de composants C	200	200	150	50	250	150

Q 4.1 Classiquement, on modélise la durée de vie d'un composant par une loi exponentielle de densité

$$p(t|\lambda) = \lambda e^{-\lambda t}$$

Q 4.1.1 Estimer le paramètre λ par maximum de vraisemblance en vous appuyant sur l'échantillon ci-dessus. Vous détaillerez vos calculs.

Q 4.1.2 En se rappelant que, pour une variable T suivant une loi exponentielle, $E[T] = 1/\lambda$, expliquer pourquoi ce résultat pouvait être attendu.

Q 4.2 Un expert de l'entreprise intervient alors pour vous expliquer qu'il existe en fait deux gammes de produits : la gamme basique, qui représente approximativement 60% des ventes (π_B), et la gamme premium, plus résistante, qui représente 40% des ventes (π_P). Les deux gammes sont modélisables par des lois exponentielles, respectivement de paramètres λ_B et λ_P (Basique et Premium).

Même si l'échantillon des 1000 composants ne distingue pas les composants basiques et premium, nous souhaiterions tout de même utiliser cet échantillon pour estimer ces deux paramètres λ_B et λ_P .

Q 4.2.1 En appelant $C \in \{B, P\}$ la variable aléatoire indicatrice de la classe du composant, montrer comment on peut modéliser ce problème de façon à pouvoir le résoudre à l'aide de l'algorithme EM. Donner les 4 paramètres à estimer.

Q 4.2.2 En utilisant les données de l'énoncé, identifier des valeurs initiales pour ces 4 paramètres.

Note : On pourra se souvenir que, en première approximation, les composantes de durées de vie les plus basses sont de type B (basique) et les composantes de durées de vie les plus hautes sont de type P (premium).

Q 4.2.3 Donner les formules permettant de calculer $Q_i(j)$, la probabilité d'appartenance de composant i à une gamme j sachant sa durée de vie et les paramètres actuels (i.e. la probabilité des variables cachées sachant les observations).

Note : donner la formulation sans faire les applications numériques

Q 4.2.4 Rappeler la formalisation du problème d'optimisation pour la mise à jour des paramètres des deux modèles en détaillant le calcul de la log-vraisemblance en fonction des $Q_i(j)$.

En notant qu'on peut réduire le nombre de paramètres à 3 (en utilisant $\pi_B + \pi_P = 1$), donner la formule des nouveaux paramètres optimaux en fonction des $Q_i(j)$.

Q 4.2.5 Donner l'algorithme général et proposer un critère d'arrêt.

Exercice 5 (15 pts) – Voiture markovienne

Une voiture automatique se déplace régulièrement entre deux bâtiments A et B d'une entreprise pour permettre aux employés de transiter entre A et B.

- Tous les quarts d'heure, systématiquement, la voiture se déplace de A à B ou de B à A et ne fait aucun autre trajet.
- En particulier, la voiture se déplace même si il n'y a pas d'employé à transporter.
- Quatre employés travaillent dans ces 2 bâtiments et doivent, de temps en temps, transiter d'un bâtiment à l'autre. Ils n'utilisent que la voiture automatique pour cela.
- Lors d'un trajet d'un bâtiment à l'autre, la voiture a la possibilité de prendre un unique employé.
- Elle prend toujours un employé qui serait en attente de transit.
- Lorsque la voiture stationne à un bâtiment, on suppose que la probabilité qu'un employé (dans ce bâtiment) veuille transiter au prochain trajet est constante et vaut p .

Q 5.1 Peut-on modéliser cette dynamique comme une chaîne de Markov ? En remarquant que la seule information pertinente pour décrire l'état de ce système est le nombre de personnes qui sont dans le bâtiment d'où va partir la voiture lors de son prochain déplacement, représenter, si possible, la matrice de transition, et le graphe de transition.

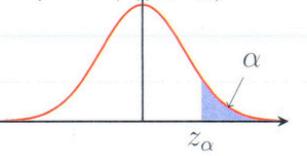
Q 5.2 Existe-t-il un régime permanent pour ce problème ? Si possible, calculer la distribution de probabilité de convergence dans ce régime permanent.

Q 5.3 Supposons que $p = 0.7$, quelle est la probabilité que la voiture transporte quelqu'un dans le prochain déplacement ?

Q 5.4 Combien d'employés faudrait-il pour faire descendre le risque d'un trajet "à vide" à moins de 1% ?

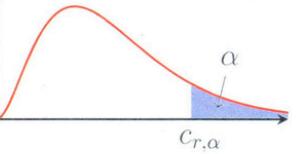
Extraits des tables de la loi normale et du χ^2

Tableau des z_α tels que $P(Z \geq z_\alpha) = \alpha$ avec $Z \sim \mathcal{N}(0, 1)$



z_α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2297	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0859	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0466	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233

Tableau des $c_{r,\alpha}$ tels que $P(Z \geq c_{r,\alpha}) = \alpha$ avec $Z \sim \chi^2_{(r)}$



$r \setminus \alpha$	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
1	0,0000393	0,000157	0,000982	0,00393	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21	10,6
3	0,0717	0,115	0,216	0,352	0,584	6,25	(7,81)	9,35	11,3	12,8
4	0,207	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3	14,9
5	0,412	0,554	0,831	1,15	1,61	9,24	11,1	12,8	15,1	16,7
6	0,676	0,872	1,24	1,64	2,20	10,6	12,6	14,4	16,8	18,5
7	0,989	1,24	1,69	2,17	2,83	12,0	14,1	16,0	18,5	20,3
8	1,34	1,65	2,18	2,73	3,49	13,4	15,5	17,5	20,1	22,0
9	1,73	2,09	2,70	3,33	4,17	14,7	16,9	19,0	21,7	23,6
10	2,16	2,56	3,25	3,94	4,87	16,0	18,3	20,5	23,2	25,2
11	2,60	3,05	3,82	4,57	5,58	17,3	19,7	21,9	24,7	26,8
12	3,07	3,57	4,40	5,23	6,30	18,5	21,0	23,3	26,2	28,3
13	3,57	4,11	5,01	5,89	7,04	19,8	22,4	24,7	27,7	29,8
14	4,07	4,66	5,63	6,57	7,79	21,1	23,7	26,1	29,1	31,3
15	4,60	5,23	6,26	7,26	8,55	22,3	25,0	27,5	30,6	32,8