# Examen Réparti 1 - MAPSI

Durée : 2 heures

*Seuls documents autorisés : Calculatrice, antisèche recto-verso, tables de lois.*
*– Barême indicatif –*

## Exercice 1 – Indépendence

Soit deux dés à six faces non pipés, un de couleur blanc et un de couleur noir. Les deux sont jetés une fois. On définit les événement suivants :

- le dé blanc donne 1, 2 ou 3.
- le dé blanc donne 2, 3 ou 6.
- la somme des deux dés est égal à 9.
- les deux dés donnent deux nombres égaux, dont la somme est inférieure à 9.

**Q 1.1 (0.5 pt)** Quel est la probabilité des ces événements ?

L'espace est donnée par toutes les couples $(1,1)$, ..., $(6,6)$, soit $6^2 = 36$ événements atomiques.

- A : $(1,\_)$, $(2,\_)$, $(3,\_)$. $p(A) = \frac{6*3}{36}$ ou $\frac{3}{6} = 0.5$
- B : $(2,\_)$, $(3,\_)$, $(6,\_)$. $p(B) = 0.5$
- C : $(3,6), (6,3), (4,5), (5,4)$. $p(C) = \frac{4}{36} = \frac{1}{9}$
- D : $(1,1)$, ..., $(4,4)$. $p(D) = \frac{4}{36} = \frac{1}{9}$

**Q 1.2 (1 pt)** Quels événements sont deux-à-deux indépendants ?

- $P(A,B) = \frac{2}{6} = \frac{1}{3} \neq P(A) \cdot P(B)$. No.
- $P(A,C) = \frac{1}{36} \neq P(A) \cdot P(C)$. No.
- $P(A,D) = \frac{3}{36} = \frac{1}{12} \neq P(A) \cdot P(D)$. No.
- $P(B,C) = \frac{2}{36} = \frac{1}{18} = P(B) \cdot P(C)$. Oui.
- $P(B,D) = \frac{2}{36} = \frac{1}{18} = P(B) \cdot P(D)$. Oui.
- $P(C,D) = 0 \neq P(C) \cdot P(D)$. No.

**Q 1.3 (0.5 pt)** Sont-ils mutuellement indépendants ? Si non, trouvez les groupes (à trois ou quatre variables) qui sont mutuellement indépendants.

- $P(A,B,C,D) = 0 \neq P(A) \cdot P(B) \cdot P(C) \cdot P(D)$ No.
- $P(A,B,C) = \frac{1}{36} = P(A) \cdot P(B) \cdot P(C)$ Oui.
- $P(A,B,D) = \frac{2}{36} \neq P(A) \cdot P(B) \cdot P(C)$ no.

## Exercice 2 – Une page top news ?

Les gérants d'un site internet de news veulent garder l'intérêt des internautes. Pour cela, ils ont pensé contrôler en temps réel les articles qui sont en première page par le nombre de clics qu'ils attirent. Si le nombre de clics par heure descend en dessous d'un certain seuil donné (par catégorie), un article n'est plus considéré comme intéressant et est enlevé de la première page pour être archivé sur le site. Pour évaluer les effets de cette règle on veut étudier statistiquement le comportement du site.
On a trois catégories d'articles (Politique, Culture, Sport). On suppose que la décroissance de l'intérêt en fonction du temps suit une loi exponentielle de paramètre $\lambda$ dont la fonction de densité est $f(x|\lambda) = \lambda e^{-\lambda x}$, pour $x \geq 0$, et que chaque catégorie suit une loi d'intérêt différente ($\lambda_P$, $\lambda_C$, $\lambda_S$).
Le tableau suivant recense le temps d'affichage en première page (en heures) pour un échantillon de 13 articles :

| P | C | C | P | S | S | S | C | P | P | C | C | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 35 | 95 | 75 | 23 | 13 | 25 | 35 | 59 | 55 | 34 | 73 | 50 | 14 |

**Q 2.1 (1.5 pt)** On suppose que la distribution de l'attention sur les articles est effectivement une loi exponentielle. Estimer par maximum de vraisemblance la valeur de $\lambda$ pour chaque catégorie.

Pour chaque catégorie $C$, la vraisemblance est égale à :

$$L(\mathbf{x}, \theta) = \pi(\mathbf{x}|\theta) = \prod_{i=1}^{\#C} p(x_i|\theta) = \prod_{i=1}^{\#C} \theta e^{-\theta x_i}$$

En calculant la log-vraisemblance :

$$\ln L(\mathbf{x}, \theta) = \sum_{i=1}^{\#C} \ln(\theta e^{-\theta x_i}) = \#C \ln \theta - \theta \sum_{i=1}^{\#C} x_i$$

Si on dérive, on a :

$$\frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta} = \frac{\#C}{\theta} - \sum_{i=1}^{\#C} x_i = 0$$

Alors :

$$\theta = \frac{\#C}{\sum_{i=1}^{\#C} x_i}$$

On a donc : $\theta_P = \frac{4}{147} = 0.027$, $\theta_C = \frac{5}{257} = 0.019$, $\theta_S = \frac{4}{87} = 0.046$

**Q 2.2 (0.5 pt)** Après un certain moment, on s'est aperçu que le mécanisme de mise à jour ne permet pas de garder l'hypothèse que ces paramètres soient fixes. Partagez-vous cette idée ? Pourquoi ?

Le nombre des articles dans la page est variable, donc intuitivement, les lecteurs distribuent l'attention différemment dans le temps.

## Exercice 3 – Production des oeufs

Dans un élevage de poules pondeuses, on veut calculer le poids moyen des oeufs, en sachant qu'il suit une loi normale (en grammes) $\mathcal{N}(\mu, 6^2)$. Le tableau ci-dessous donne le poids de 9 oeufs pris au hasard :

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 43 | 67 | 52 | 66 | 46 | 71 | 65 | 48 | 64 |

**Q 3.1 (0.5 pt)** Donnez une estimation ponctuelle du poids moyen d'un oeuf dans cet élevage, et la médiane de cet échantillon.

L'estimation ponctuelle est donnée simplement par la moyen empirique. $\mu = 58$. La médiane est : $M = 64$.

**Q 3.2 (1.5 pt)** Le responsable de l'élevage vous dit que le poids moyen normalement est de 64 gr, mais la moyenne de cet échantillon a baissé. En utilisant un test de niveau de confiance de 95%, pouvez vous confirmer que la moyenne donnée par le résponsable n'a pas baissée ?

Soit $X$ la variable aléatoire "poids d'un oeuf", et soit $\overline{X}$ le poids moyen de l'échantillon. Le test d'hypothèse est donc $H_0 : \mu = 64$ vs $H_1 : \mu < 64$. Pour le théorème du limit central, on sait que $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{\overline{X}-64}{6/\sqrt{9}} = \frac{\overline{X}-64}{2} \sim \mathcal{N}(0,1)$.
Pour construir un test, il faut qu'on considère une seuil $c$ pour laquelle, si $\overline{x} < c$, $H_1$ peut être accepté. On cherche alors un $c$ tel que :

$$P(\frac{\overline{X}-64}{2} < \frac{c-64}{2} | \frac{\overline{X}-64}{2} \sim \mathcal{N}(0,1)) = P(Z < \frac{c-64}{2}) = 0.05 = P(Z < -1.6)$$

On a donc :

$$\frac{c-64}{2} \approx -1.6 \rightarrow c \approx 60.8$$

Car $\overline{x} = 58$, on peut déduire que la moyenne a baissé.

## Exercice 4 – Test d'indépendance

Le responsable des opérations d'une usine de production veut déterminer s'il y a des différences concernant la qualité entre deux groupes de travail qui travaillent alternativement. Il sélectionne au hasard un certain nombre de produits pour chaque groupe et les inspecte avec soin, en suivant un protocole qui classifie chaque produit comme défectueux ou satisfaisant. Ce tableau donne une synthèse des données recueillies :

| | Défectueux | Satisfaisant |
|---|---|---|
| Groupe 1 | 6 | 102 |
| Groupe 2 | 5 | 83 |

Il se demande maintenant, au vu des donnés, si il existe une différence de qualité entre les deux groupes.

**Q 4.1 (0.25 pt)** Construisez le tableau de contingence avec les effectifs reels par catégorie et marginaux.

On calcule les marginaux comme sous-totaux dans le tableau de contingence :

| | Défectueux | Satisfaisant | Totale |
|---|---|---|---|
| Groupe 1 | 6 | 102 | 108 |
| Groupe 2 | 5 | 83 | 88 |
| Totale | 11 | 185 | 196 |

**Q 4.2 (0.75 pt)** Calculez les effectifs attendus, en supposant que les deux variables aléatoires *groupe* et *qualité* sont indépendantes.

On calcule les effectifs attendus par $\frac{n_{i\cdot}.n_{\cdot j}}{n_{ij}}$ :

| | Défectueux | Satisfaisant |
|---|---|---|
| Rotation 1 | 6.06 | 101.94 |
| Rotation 2 | 4.94 | 83.06 |

**Q 4.3 (1.25 pt)** Tester, au sens du chi2, si l'hypothèse d'indépendance est correcte, avec un niveau de confiance du 95 %.

L'hypothèse nulle $H_0$ est que les deux variables sont indépendantes. Nous calculons l'écart entre les valeurs empiriques et les valeurs théoriques dans le cas d'indépendance :

$$d^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} = \frac{\left(n_{ij} - \frac{n_{i\cdot}.n_{\cdot j}}{n_{ij}}\right)^2}{\frac{n_{i\cdot}.n_{\cdot j}}{n_{ij}}}$$

La valeur empirique de $D^2_{(n)} \sim \chi^2_{(2-1)(2-1)} = \chi^2_1$. Si la valeur de l'écart calculée $d^2$ est inférieure de $d^2_\alpha$, donnée par la loi du $\chi^2$ avec $\alpha = 1 - 0.95 = 0.05$, l'hypothèse peut être accepté.
Dans notre cas, on a : $d^2 = 0.00145931055022 < d^2_\alpha = 3.84$
Ainsi, il y a indépendance.

## Exercice 5 – Publicité dynamique sur les sites web

Nous nous intéressons à la création d'un système de personnalisation de la publicité affichée sur un site d'information important (comme `www.lemonde.fr` par exemple). Le site en question a établi $K = 20$ catégories d'articles (e.g. sport, culture, économie nationale...) et utilise des cookies pour déterminer l'intérêt des utilisateurs pour telle ou telle catégorie. Ainsi, pour le site, un utilisateur $\mathbf{x} = [0, 1, 0, \ldots, 0]$ est un vecteur binaire de taille $K$.
La régie publicitaire du site distingue $M = 10$ types de publicités (articles de luxe, films, ...). Pour savoir si un utilisateur est intéressé par une publicité, ils enregistrent les *clics* effectués (afin d'économiser de la place, seuls les clics sont enregistrés et pas les simples affichages).

Pendant une longue période de référence, le site utilise une politique d'affichage aléatoire et enregistre les comportements des utilisateurs. Il constitue ainsi une base de donnée $X, Y$ de $N$ entrées où chaque ligne désigne un couple utilisateur/catégorie de la publicité cliquée : $\mathbf{x}_i = \{x_{i1}, \ldots, x_{ik}, \ldots, x_{iK}\}, x_{ik} \in \{0, 1\}, y_i \in \{1, \ldots, M\}$.

**Q 5.1 (0.5 pt) Modélisation.** Quelle loi utiliser pour modéliser chaque centre d'intérêt $k$ (connaissant la catégorie de publicité $m$) ? Montrer que la probabilité d'un centre d'intérêt $k$ pour un utilisateur $i$ peut s'écrire :

$$p(x_{ik}|y_m) = \theta_{k,m}^{x_{ik}}(1 - \theta_{k,m})^{(1-x_{ik})}$$

Pour un utilisateur, nous faisons l'hypothèse que ses centres d'intérêt sont indépendant les uns des autres. Exprimer $p(\mathbf{x}_i|y_m)$ en fonction des $\theta_{k,m}$.

Loi de Bernoulli
l'expression factorise la probabilité d'une observation suivant une loi de Bernoulli, que cette observation soit 0 ou 1

$$p(\mathbf{x}_i) = \prod_k \theta_k^{x_{ik}}(1 - \theta_k)^{(1-x_{ik})}$$

**Q 5.2 (0.5 pt)** Donner le protocole pour créer une politique d'affichage adaptée aux centres d'intérêt de l'utilisateur à partir de ces données. Donner le nombre de paramètres à apprendre pour entrainer notre modèle.

- Exprimer la vraisemblance de l'ensemble des données de chaque classe $m$.
- Maximiser la vraisemblance pour pour trouver les paramètres optimaux $\theta_{k,m}$ qui permettent de *coller* le mieux aux données
- Il y a donc $K \times M$ paramètres à apprendre $= 200$
- Au sens du maximum de vraisemblance (par exemple), il suffit ensuite de calculer les $p(\mathbf{x}|y_m)$ et de retourner la classe la plus vraisemblable.

**Q 5.3 (1.5 pt) Apprentissage.** Donner l'expression des paramètres maximisant la vraisemblance des données pour chaque catégorie de publicité.
Formaliser le problème et donner les détails des calculs.

Pour chaque classe de donnée $y_m$,

$$\mathcal{L}(X_m, p_{k,m}) = \prod_i \prod_k \theta_{k,m}^{x_{ik}}(1 - \theta_{k,m})^{(1-x_{ik})}$$

$$\theta_{k,m}^\star = \underset{\theta_{k,m}}{\text{Argmax}}\, \mathcal{L}(X_m, \theta_{k,m}) = \underset{\theta_{k,m}}{\text{Argmax}} \log \mathcal{L}(X_m, \theta_{k,m})$$

On enlève les $m$ pour plus de lisibilité :

$$\log \mathcal{L}(X, p_k) = \sum_{i,k} x_{ik} \log(\theta_k) + (1 - x_{ik}) \log(1 - \theta_k)$$

$$\frac{\partial \log \mathcal{L}}{\partial \theta_k} = \sum_i x_{ik} \frac{1}{\theta_k} + (1 - x_{ik}) \frac{-1}{1 - \theta_k} = 0$$

$$\iff \sum_i x_{ik}(1 - \theta_k) - (1 - x_{ik})\theta_k = 0 \iff \sum_i -\theta_k + x_{ik} = 0 \iff \theta_k = \frac{\sum_i x_{ik}}{N}$$

**Q 5.4 (1 pt) Inférence.** Pour un $\mathbf{x}$ donné, quel type de publicité la régie doit-elle afficher pour maximiser le taux de clic (ie, au sens du maximum a posteriori) ?
Donner la formulation détaillée du problème à poser et une estimation des paramètres manquant.
La rapidité du système est un facteur clé, le chargement des pages devant resté (quasi-)instantané...
Comment jugez-vous l'efficacité du système proposé ?

$$m^\star = \underset{m}{\text{Argmax}}\, p(y_m|\mathbf{x}) = \underset{m}{\text{Argmax}}\, \frac{p(\mathbf{x}|y_m)p(y_m)}{p(\mathbf{x})} = \underset{m}{\text{Argmax}}\, p(\mathbf{x}|y_m)p(y_m)$$

où :

$$p(\mathbf{x}|y_m) = \prod_k \theta_{k,m}^{x_k}(1 - \theta_{k,m})^{(1-x_k)}, \qquad p(y_m) = \frac{\#y_m}{N}$$

Le système est très efficace : les $p_{k,m}$ sont appris une fois pour toutes et l'inférence est très simple.
Le seul point compliqué réside dans l'extraction et la mise à jour des préférences utilisateurs...
Mais cela peut-être traité en tâche de fond.

**Q 5.5 (0.5 pt) Amélioration du système.** Afin de palier le risque de non-affichage de certaines catégories de publicité, nous utilisons les capacités génératives de notre modèle. Expliquer sur quel critère et comment tirer des publicités reflétant l'ensemble des goûts de l'utilisateur et pas seulement son meilleur score d'affinité.

- Critère : $p(y_m|\mathbf{x})$
- Moyen :
  - Calcul de somme cumulée des $p(y_m|\mathbf{x})$
  - Tirage d'un nombre aléatoire entre 0 et 1 (loi uniforme)
  - Récupération de la catégorie associée et affichage d'une pub de la catégorie
- Ainsi, on tire le plus souvent les catégories qui plaisent le plus à l'utilisateur... Mais on explore aussi les catégories minoritaires.

**Q 5.6 (0.5 pt)** L'évaluation de ce type d'approche est délicate, nous procédons en général par A-B *testing*, en comparant 2 systèmes mis en ligne en parallèle. Sur quel critère départager les deux systèmes ? Décrire brièvement l'implémentation de l'A-B *testing*.

- Le critère est le nombre de clics
- Implémentation :

— Séparation aléatoire des clients en deux ensembles de tailles respectives $\alpha_A, \alpha_B$
— Récupération du nombre de clics $n_A, n_B$
— Comparaison de $n_A/\alpha_A$ et $n_B/\alpha_B$

**Q 5.7 (0.25pt)** Quelle entreprise française a fait fortune grace à ce type d'approche ?

Criteo

## Exercice 6 – Modélisation des usagers RATP par des lois normales

Nous nous intéressons à la modélisation des usagers du métro parisien pour les départs au travail, sur la plage [6h 10h]. Les logs d'une population contiennent simplement les temps auxquels ont eu lieu les validations des usagers. Ils ont été constitués sur 13 semaines et le temps a été discrétisé par minutes (soit 240 intervalles entre 6h et 10h).

Nous modélisons cette situation à l'aide d'une loi normale $\mathcal{N}(\mu, \sigma)$, en prenant un temps discrétisé $t \in [0,\ 239]$

$$p(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(t-\mu)^2\right)$$

Note : pour simplifier, nous faisons l'hypothèse que la loi est toujours normalisée (même si $\mu$ se rapproche d'une des bornes et que la distribution est en réalité tronquée).

**Q 6.1 (2 pts)** Il y a deux façons de voir un usager :

1. Comme un ensemble de $N$ validations : $u = \{t_1, \ldots, t_N\}, \qquad t \in [0,\ 239]$
2. Comme le comptage de ces validations par créneau horaire : $u = \{c_0, \ldots, c_t, \ldots, c_{239}\}, \quad c_t \in \mathbb{N}$

Un usager avec une validation unique à $t = 10$ est donc représenté par : $u = \{t_1 = 10\}$ ou par $u = \{c_0 = 0, c_1 = 0, \ldots, c_9 = 0, c_{10} = 1, c_{11} = 0, \ldots, c_{239} = 0\}$.
Exprimer la vraisemblance d'un usager en explicitant l'hypothèse que vous faites pour simplifier le calcul. Vous exprimerez la vraisemblance en fonction des $c_t$. Passer ensuite la formule au log et développer l'expression.

$$\mathcal{L}_u = \prod_{n=1}^{N} p(t_n)$$

Hypothèse : les validations sont indépendantes.

$$\mathcal{L}_u = \prod_{t=1}^{239} p(t)^{c_t} \qquad \log \mathcal{L}_u = \sum_{t=1}^{239} c_t \left(-0.5\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}(t-\mu)^2\right)$$

**Q 6.2 (2 pts)** Pour distinguer les comptages de 2 utilisateurs $u$ et $u'$ au pas de temps $t$, nous introduisons un nouvel index : $c_{u,t}$ (qui est donc différent de $c_{u',t}$). Trouver les paramètres optimaux

de la loi normale pour un ensemble d'usager au sens du maximum de vraisemblance en fonction des $c_{u,t}$.
NB : donner la formulation du problème et les détails du calcul.

$$\log \mathcal{L} = \sum_u \sum_{t=1}^{239} c_{u,t} \left( -0.5 \log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}(t-\mu)^2 \right)$$

Formulation :

$$\mu^\star, \sigma^\star = \arg\max_{\mu,\sigma} \log \mathcal{L}$$

$$\frac{\partial \log \mathcal{L}}{\partial \mu} = \sum_{u,t} c_{u,t} \left( -\frac{1}{2\sigma^2}(-2)(t-\mu) \right) = 0$$

$$\iff \sum_{u,t} c_{u,t}(t-\mu) = 0 \iff \mu = \frac{\sum_{u,t} t c_{u,t}}{\sum_{u,t} c_{u,t}}$$

Pour sigma, c'est plus facile de dériver par rapport à $\sigma^2$ :

$$\frac{\partial \log \mathcal{L}}{\partial \sigma^2} = \sum_{u,t} c_{u,t} \left( -\frac{1}{2}\frac{1}{\sigma^2} - \frac{1}{2}(t-\mu)^2(-\frac{1}{\sigma^4}) \right) = 0$$

$$\iff \sum_{u,t} c_{u,t} \left( -1 + (t-\mu)^2 \frac{1}{\sigma^2} \right) = 0 \iff \sigma^2 = \frac{\sum_{u,t} c_{u,t}(t-\mu)^2}{\sum_{u,t} c_{u,t}}$$

**Q 6.3 (0.5 pt)** A partir des données précédentes, un expert a constitué 3 ensembles d'usagers correspondant à des classes distinctes : les *lève-tôt*, les *tardifs*, les *variables*. Donner le protocole expérimental pour créer un classifieur d'usager, au sens du maximum de vraisemblance, capable de ranger un nouvel utilisateur (identifié par ses logs) dans l'une des catégories.
Donner la formulation du critère de décision.

- 1 classifieur par classe, appris par max de vraisemblance sur les données de la classe.
- Les paramètres à estimer seront les $\mu_{Cl}, \sigma_{Cl}$ avec les formules vues dans la question précédente
- Pour un usager $u$, nous utiliserons le critère :

$$Cl^\star = \underset{Cl}{\text{Argmax}}(\log \mathcal{L}(u, Cl)), \qquad \mathcal{L}(u, Cl) = \sum_{t=1}^{239} c_t \left( -0.5 \log(2\pi) - \log(\sigma_{Cl}) - \frac{1}{2\sigma_{Cl}^2}(t-\mu_{Cl})^2 \right)$$

**Q 6.4 (0.5 pt)** Comment construire un classifieur au sens du maximum a posteriori à partir de ces mêmes données.

— Mesurer la taille des ensembles d'utilisateurs donnés par l'expert pour estimer les proba a priori $\pi_1, \pi_2, \pi_3$
— Ces paramètres seront des constantes qui ne remettent pas en cause l'estimation des paramètres des modèles
— Trouver la classe en utilisant le critère :

$$Cl^\star = \underset{Cl}{\text{Argmax}}(\log \mathcal{L}(u, Cl)\pi_{Cl}/p(u)) = \underset{Cl}{\text{Argmax}}(\log \mathcal{L}(u, Cl)\pi_{Cl})$$

$$\text{avec} \qquad \mathcal{L}(u, Cl) = \sum_{t=1}^{239} c_t \left( -0.5 \log(2\pi) - \log(\sigma_{Cl}) - \frac{1}{2\sigma_{Cl}^2}(t - \mu_{Cl})^2 \right)$$

**Q 6.5 (0.5 pt)** Est-il possible que notre modèle de classification de voyageur soit en désaccord avec certaines étiquettes proposées par l'expert sur le jeu de données original ? Dans l'affirmative, comment faire évoluer le modèle ?

Oui, c'est possible ponctuellement
On pourrait alors mettre en place une procédure itérative de type de EM pour ré-estimer les paramètres des modèles et modifier l'étiquetage de l'expert.