ECE2 – Concours blanc 1

MATHÉMATIQUES 2-TYPE ESSEC 2

La présentation, la lisibilité, l'orthographe, la qualité de la rédaction, la clarté et la précision des raisonnements entreront pour une part importante dans l'appréciation des copies.
Les candidats sont invités à ***encadrer*** *dans la mesure du possible les résultats de leurs calculs.*
Il ne doivent faire usage d'aucun document. ***L'utilisation de toute calculatrice et de tout matériel électronique est interdite.*** *Seule l'utilisation d'une règle graduée est autorisée.*
Si au cours de l'épreuve, un candidate repère ce qui lui semble être une erreur d'énoncé, il la signalera sur sa copie et poursuivra sa composition en expliquant les initiatives qu'il sera amené à prendre.

Lorsque l'on effectue des sondages, de nombreux biais statistiques peuvent apparaître : on peut par exemple avoir considéré un échantillon non-représentatif de la population, il peut y avoir un biais dans les réponses des personnes sondées... On va s'intéresser dans ce problème à ce que l'on appelle le biais par la taille : il provient du fait que si l'on choisit une personne au hasard dans la population, celle-ci a plus de chances de faire partie d'une catégorie nombreuse de la population.
Le biais par la taille est la source de nombreux « paradoxes » probabilistes, comme le fait que les gagnants du loto vivent en moyenne plus longtemps (parce que les gagnants sont ceux qui ont pu jouer au loto plus longtemps) ou le fait que vos amis ont en moyenne plus d'amis que vous (car les gens qui ont un très grand nombre d'amis font sûrement partie de vos amis). On verra ici comment formaliser le biais par la taille, et l'utiliser dans différents contextes.
Toutes les variables aléatoires intervenant dans le problème sont définies sur un espace probabilisé $(\Omega, \mathscr{A}, \mathrm{P})$.
Pour toute variable aléatoire X, on notera E(X) son espérance (resp. V(X) sa variance) lorsqu'elles existent.

Partie I. Biais par la taille, exemples discrets.

1. Lorsqu'on choisit une famille française au hasard, on suppose que le nombre d'enfants est modélisé par une variable aléatoire X. Pour connaître la loi de X, une idée serait d'interroger les élèves d'une école pour connaître le nombre d'enfants dans leur famille.
   On va voir que cette approche introduit un biais, en considérant une situation particulière. Supposons que X suit la loi binomiale de paramètres $n = 10$ et $p = \frac{1}{5}$. On note $p_k = \mathrm{P}(\mathrm{X} = k)$ pour $k \in \{0, 1, \ldots, 10\}$.
   (a) i. Rappeler l'expression de p_k pour $k \in \{0, 1, \ldots, 10\}$.
   ii. Donner E(X), V(X) et en déduire $\mathrm{E}(\mathrm{X}^2)$.
   (b) Soit M_k le nombre de familles à k enfants et $\mathrm{M} = \sum_{k=0}^{10} \mathrm{M}_k$ le nombre total de familles.
   On a donc $p_k = \dfrac{\mathrm{M}_k}{\mathrm{M}}$.
   Soit N_k le nombre total d'enfants (c'est-à-dire dans toute la population) qui font partie d'une famille à k enfants et $\mathrm{N} = \sum_{k=0}^{10} \mathrm{N}_k$ le nombre total d'enfants de la population.
   i. Démontrer : $\mathrm{N}_k = k\, p_k\, \mathrm{M}$.
   ii. Démontrer : $\dfrac{\mathrm{N}}{\mathrm{M}} = \mathrm{E}(\mathrm{X})$.
   iii. Montrer que la proportion des enfants provenant d'une famille à k enfants est : $p_k^* = \dfrac{k\, p_k}{2}$.
   (c) On choisit une personne au hasard dans la rue, à qui l'on demande combien d'enfants ses parents ont eu (lui ou elle inclus). On note Y ce nombre d'enfants.
   i. Pour tout entier k élément de $\{1, 2, \ldots, 10\}$, démontrer : $\mathrm{P}(\mathrm{Y} = k) = \dfrac{k\, p_k}{2}$.
   ii. Démontrer : $\mathrm{E}(\mathrm{Y}) = \dfrac{\mathrm{E}(\mathrm{X}^2)}{\mathrm{E}(\mathrm{X})}$.
   iii. En déduire E(Y) et comparer à E(X).
2. Soit X une variable aléatoire à valeurs dans $\mathbb{N}$, non identiquement nulle et admettant une espérance non nulle. Pour tout entier $i > 0$, on pose : $q_i = \dfrac{i}{\mathrm{E}(\mathrm{X})}\, \mathrm{P}(\mathrm{X} = i)$.
   (a) Montrer que la suite $(q_i)_{i>0}$ définie ci-dessus définit une loi de probabilité.

On considère la variable aléatoire X^* dont la loi est donnée par les q_i, c'est-à-dire, pour tout i entier naturel non nul :

$$P(X^* = i) = \frac{i}{E(X)} P(X = i).$$

On dit que X^* suit *la loi de* X *biaisée par la taille.*

(b) On suppose que X admet un moment d'ordre 2. Démontrer : $E(X^*) = \frac{E(X^2)}{E(X)}$.

(c) En déduire que si $E(X^2)$ existe, on a : $V(X) = E(X)\big(E(X^*) - E(X)\big)$.

(d) Conclure que, si $E(X^2)$ existe, on a : $E(X^*) \geqslant E(X)$.

3. (a) Soit λ un réel strictement positif. On suppose que X est une variable aléatoire qui suit la loi de Poisson de paramètre λ. Soit X^* une variable aléatoire suivant la loi de X biaisée par la taille.

i. Donner la loi de X^*.

ii. Vérifier que X^* suit la même loi que $X + 1$.

(b) Réciproquement, on suppose que X est une variable aléatoire à valeurs dans $\mathbb{N}$ admettant une espérance non nulle, telle que X^* et $X + 1$ suivent la même loi.

i. Montrer que pour tout $k \geqslant 1$: $P(X = k) = \frac{E(X)}{k} P(X = k-1)$.

ii. Montrer que pour tout k entier naturel : $P(X = k) = \frac{\big(E(X)\big)^k}{k!} P(X = 0)$.

iii. En déduire la loi de X.

4. *Le paradoxe du temps d'attente du bus.*
Soit $n \geqslant 1$ un entier naturel et soit X une variable aléatoire à valeurs dans $\{1, \dots, n\}$ telle que pour tout $1 \leqslant k \leqslant n$, $P(X = k) > 0$. On suppose qu'à un arrêt de bus donné, les intervalles de temps entre deux bus consécutifs, exprimés en minutes, sont des variables aléatoires indépendantes, de même loi que X. Une personne arrive à cet arrêt à un instant aléatoire, et se demande combien de temps elle va attendre.

(a) Une première idée est que la personne arrive à un instant uniforme entre deux arrivées de bus, séparées par un intervalle de X minutes. On note T la variable aléatoire qui représente le temps d'attente (à valeurs dans $\{1, \dots, n\}$) et on suppose donc que pour tout entier k élément de $\{1, \dots, n\}$:

$$P_{[X=k]}(T = j) = \begin{cases} \frac{1}{k} & \text{si } j \in \{1, \dots, k\} \\ 0 & \text{si } j > k. \end{cases}$$

i. Montrer que pour tout entier $k \in \{1, \dots, n\}$, on a :

$$\sum_{j=1}^{n} j\, P_{[X=k]}(T = j) = \frac{k+1}{2}.$$

ii. En déduire :

$$\sum_{k=1}^{n} \sum_{j=1}^{n} j\, P(X = k)\, P_{[X=k]}(T = j) = \frac{E(X+1)}{2}.$$

iii. Démontrer : $E(T) = \frac{E(X)+1}{2}$.

(b) En réalité, en arrivant à l'arrêt de bus, on « tombe » dans un intervalle entre deux bus de manière proportionnelle à sa taille (plus l'intervalle est long, plus on a de chances de « tomber » dedans) : l'intervalle de temps est X^*, suivant la loi de X biaisée par la taille. Le temps d'attente T' vérifie donc en fait, pour tout $k \in \{1, \dots, n\}$:

$$P_{[X^*=k]}(T' = j) = \begin{cases} \frac{1}{k} & \text{si } j \in \{1, \dots, k\} \\ 0 & \text{si } j > k. \end{cases}$$

i. Justifier que : $E(T') = \frac{E(X^*)+1}{2}$.

ii. En déduire qu'on a : $E(T') \geqslant E(T)$.

Partie II. Biais par la taille, propriétés.

Dans cette partie, X est une variable aléatoire discrète à valeurs dans $\mathbb{N}$ admettant une espérance strictement positive.
On rappelle que l'on dit que X^* suit *la loi de* X *biaisée par la taille* si :

$$\forall i \in \mathbb{N}^*, \quad P(X^* = i) = \frac{i}{E(X)} P(X = i).$$

5. Dans cette question, f et g sont deux fonctions croissantes définies sur $\mathbb{R}_+$ à valeurs réelles. On suppose que $f(X)$, $g(X)$ et $f(X)g(X)$ possèdent une espérance.
   (a) Montrer que pour tout $(x_1, x_2) \in \mathbb{R}_+ \times \mathbb{R}_+$ on a :
   $$\big(f(x_1) - f(x_2)\big)\big(g(x_1) - g(x_2)\big) \geqslant 0.$$
   (b) Soient X_1 et X_2 deux variables aléatoires indépendantes et de même loi que X. Montrer :
   $$E\Big(\big(f(X_1) - f(X_2)\big)\big(g(X_1) - g(X_2)\big)\Big) = 2E\big(f(X)g(X)\big) - 2E\big(f(X)\big)E\big(g(X)\big).$$
   (c) En déduire : $E\big(f(X)g(X)\big) \geqslant E\big(f(X)\big)E\big(g(X)\big)$.
6. (a) Dans cette question, f et g sont deux fonctions définies sur $\mathbb{R}_+$ à valeurs réelles telles que :
   $$\forall x \in \mathbb{R}_+, \quad \big|f(x)\big| \leqslant g(x).$$
   Montrer que si $g(X)$ possède une espérance alors $f(X)$ aussi.
   (b) Dans cette question, h est une fonction bornée définie sur $\mathbb{R}_+$. Montrer que $h(X)$ possède une espérance.
   (c) Dans cette question, h est une fonction bornée définie sur $\mathbb{R}_+$. Montrer que $Xh(X)$ et $h(X^*)$ possèdent une espérance et que :
   $$E\big(h(X^*)\big) = \frac{1}{E(X)} E\big(X\,h(X)\big).$$
7. Dans cette question, on suppose qu'il existe un entier $m \geqslant 1$ tel que $E(X^{m+1})$ existe.
   (a) Soit p un entier naturel tel que $1 \leqslant p \leqslant m$.
      i. Montrer que pour tout réel $x \geqslant 0$, on a : $0 \leqslant x^p \leqslant 1 + x^{m+1}$.
      ii. En déduire que $E(X^p)$ existe.
   (b) Démontrer : $E(X^{m+1}) \geqslant E(X)\,E(X^m)$.
   (c) En déduire : $E\big((X^*)^m\big) \geqslant E(X^m)$.
8. Pour tout t réel positif, on définit la fonction g_t sur $\mathbb{R}_+$ par :
   $$\forall x \in \mathbb{R}_+, \quad g_t(x) = \begin{cases} 0 & \text{si } 0 \leqslant x \leqslant t \\ 1 & \text{si } x > t. \end{cases}$$
   (a) Soit $t > 0$. Montrer que la fonction g_t est croissante sur $\mathbb{R}_+$.
   (b) Montrer que pour tout t réel positif, $E\big(X\,g_t(X)\big)$ est bien défini et :
   $$E\big(X\,g_t(X)\big) \geqslant E(X)\,P(X > t).$$
   (c) En déduire que, pour tout t réel : $P(X^* > t) \geqslant P(X > t)$.
   On dit que X^* *domine stochastiquement* X.

Partie III. Généralisation aux variables aléatoires quelconques.

Soit X une variable aléatoire positive quelconque possèdant une espérance strictement positive.
Sur le modèle de la question 6.c, on dit que X^* suit la loi de X biaisée par la taille si, pour toute fonction $h : \mathbb{R}_+ \to \mathbb{R}$ bornée on a :

$$E(h(X^*)) = \frac{1}{E(X)} E(X h(X)).$$

On admet que cette propriété caractérise une unique loi de probabilité.
Soit $n \in \mathbb{N}^*$ et soient $X_1, \ldots, X_n$ des variables aléatoires positives, indépendantes, non nécessairement de même loi.
On suppose qu'elles admettent toutes une espérance strictement positive, et on note $\mu_i = E(X_i)$. De plus, on pose :

$$\mu = \sum_{i=1}^{n} \mu_i \quad \text{et} \quad S_n = \sum_{i=1}^{n} X_i.$$

9. Calculer $\mathrm{E}(\mathrm{S}_n)$.

10. Pour A un événement, on note $\mathbb{1}_\mathrm{A}$ la variable aléatoire définie par :

$$\forall \omega \in \Omega \quad \mathbb{1}_\mathrm{A}(\omega) = \begin{cases} 1 & \text{si } \omega \in \mathrm{A} \\ 0 & \text{sinon.} \end{cases}$$

   (a) Reconnaître la loi de $\mathbb{1}_\mathrm{A}$.

   (b) Soit $(\mathrm{A}_1, \ldots, \mathrm{A}_n)$ un système complet d'événements. Montrer que $\sum_{i=1}^{n} \mathbb{1}_{\mathrm{A}_i}$ suit une loi certaine que l'on précisera.

11. On considère X_1^*, …, X_n^* des variables aléatoires indépendantes, indépendantes de X_1, …, X_n telles que, pour tout entier i tel que $1 \leqslant i \leqslant n$, X_i^* suive la loi de X_i biaisée par la taille.
Soit J une variable aléatoire indépendante de $\mathrm{X}_1, \mathrm{X}_1^*, \ldots, \mathrm{X}_n, \mathrm{X}_n^*$ de loi donnée par :

$$\forall k \in \{1, \ldots, n\}, \quad \mathrm{P}(\mathrm{J} = k) = \frac{\mu_k}{\mu}.$$

On considère la variable aléatoire $\mathrm{X}_\mathrm{J} = \sum_{j=1}^{n} \mathrm{X}_j \mathbb{1}_{[\mathrm{J}=j]}$ et on définit $\mathrm{T}_n = \mathrm{S}_n - \mathrm{X}_\mathrm{J} + \mathrm{X}_\mathrm{J}^*$. Autrement dit, on choisit un indice aléatoire J et, dans la somme $\sum_{i=1}^{n} \mathrm{X}_i$, on remplace X_J par X_J^*.
Soit $h : \mathbb{R}_+ \to \mathbb{R}$ une fonction bornée.

   (a) i. Démontrer :

$$h(\mathrm{T}_n) = \sum_{i=1}^{n} h(\mathrm{T}_n)\, \mathbb{1}_{[\mathrm{J}=i]} = \sum_{i=1}^{n} h(\mathrm{S}_n - \mathrm{X}_i - \mathrm{X}_i^*)\, \mathbb{1}_{[\mathrm{J}=i]}.$$

   ii. En déduire :

$$\mathrm{E}\big(h(\mathrm{T}_n)\big) = \sum_{i=1}^{n} \mathrm{P}(\mathrm{J} = i)\, \mathrm{E}\big(h(\mathrm{S}_n - \mathrm{X}_i + \mathrm{X}_i^*)\big).$$

   (b) Pour $i \in \{1, \ldots, n\}$, démontrer :

$$\forall s \in \mathbb{R}_+, \quad \mathrm{E}\big(h(s + \mathrm{X}_i^*)\big) = \frac{1}{\mu_i}\, \mathrm{E}\big(\mathrm{X}_i\, h(s + \mathrm{X}_i)\big).$$

On **admettra** qu'on en déduit l'égalité : $\mathrm{E}\big(h(\mathrm{S}_n - \mathrm{X}_i + \mathrm{X}_i^*)\big) = \frac{1}{\mu_i}\, \mathrm{E}\big(\mathrm{X}_i\, h(\mathrm{S}_n)\big)$.

   (c) En déduire : $\mathrm{E}\big(h(\mathrm{T}_n)\big) = \dfrac{\mathrm{E}\big(\mathrm{S}_n\, h(\mathrm{S}_n)\big)}{\mathrm{E}(\mathrm{S}_n)}$.

   (d) Conclure que T_n suit la loi de S_n biaisée par la taille.

Partie IV. Applications en statistiques.

On s'intéresse maintenant au cas où le biais par la taille peut être utilisé en statistique, pour construire des estimateurs non biaisés. Une compagnie d'électricité possède n clients où n est un entier naturel non nul donné. Lors de l'année écoulée, le i-ième client a payé x_i euros ($x_i > 0$), mais a en réalité consommé une quantité d'électricité correspondant à y_i euros ($y_i > 0$). La compagnie sait combien ses clients ont payé, et elle souhaite estimer le rapport :

$$\theta = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

pour déterminer à quel point elle a mal facturé ses clients.
En pratique elle ne peut pas sonder tous ses clients : elle décide donc de choisir m ($m < n$) clients parmi les n pour effectuer les mesures.
Dans cette partie, on considère l'univers Ω constitué des parties de $\{1, \ldots, n\}$ à m éléments et on suppose que la compagnie choisit une partie à m éléments de manière uniforme. Ainsi :

$$\forall \mathrm{A} \in \Omega, \quad \mathrm{P}(\{\mathrm{A}\}) = \frac{1}{\binom{n}{m}}.$$

Pour $\mathrm{A} \in \Omega$, on définit :

$$\bar{x}_\mathrm{A} = \frac{1}{m} \sum_{i \in \mathrm{A}} x_i, \quad \bar{y}_\mathrm{A} = \frac{1}{m} \sum_{i \in \mathrm{A}} y_i, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Enfin, on définit deux variables aléatoires X et Y sur Ω par :

$$\forall A \in \Omega, \quad X(A) = \bar{x}_A \quad \text{et} \quad Y(A) = \bar{y}_A.$$

La compagnie décide d'utiliser $\theta_1 = \frac{Y}{X}$ comme estimateur de θ.

On admet que si l'on munit un univers fini Ω d'une probabilité P alors l'espérance d'une variable aléatoire T définie sur Ω est donnée par :

$$E(T) = \sum_{\omega \in \Omega} T(\omega) P(\{\omega\}).$$

12. (a) Que représentent $\bar{x}$ et $\bar{y}$?
    (b) Soit $A \in \Omega$. Que représentent $\bar{x}_A$ et $\bar{y}_A$?
13. (a) Démontrer : $E(X) = \binom{n}{m}^{-1} \sum_{A \in \Omega} \bar{x}_A$.
    (b) Soit $i \in \{1, \dots, n\}$. Combien y a-t-il de parties A à m éléments telles que $i \in A$?
    (c) En déduire :
    $$\sum_{A \in \Omega} \sum_{i \in A} x_i = \binom{n-1}{m-1} \sum_{i=1}^{n} x_i.$$
    (d) Conclure : $E(X) = \bar{x}$. On **admettra** que de même on a : $E(Y) = \bar{y}$.
    (e) Exprimer θ en fonction de E(X) et E(Y).
14. Soient W et Z sont deux variables aléatoires strictement positives, admettant un moment d'ordre deux. On note, pour tout $t \in \mathbb{R}$: $Q(t) = E((W + tZ)^2)$.
    (a) Montrer que Q est une fonction polynomiale de degré 2 et déterminer son tableau de signe.
    (b) En considérant le discriminant de Q en déduire :
    $$E(WZ) \leqslant \left(E(W^2)\right)^{\frac{1}{2}} \left(E(Z^2)\right)^{\frac{1}{2}}.$$
    (c) Montrer que l'inégalité précédente est une inégalité si et seulement si il existe un réel $\alpha > 0$ tel que $W = \alpha Z$ presque sûrement.
15. (a) Démontrer : $E\left(\frac{1}{X}\right) \geqslant \frac{1}{E(X)}$.
    (b) Montrer qu'il y a égalité si et seulement si X est une variable aléatoire constante, c'est-à-dire $X = E(X) = \bar{x}$.
    (c) Conclure que $E\left(\frac{1}{X}\right) = \frac{1}{E(X)}$ si et seulement si pour tout $i \in [\![1, n]\!]$, $x_i = \bar{x}$.
16. Si on suppose que X et Y sont indépendantes, montrer que $E(\theta_1) \geqslant \theta$, avec égalité si et seulement si pour tout $i \in [\![1, n]\!]$, $x_i = \bar{x}$.

Ainsi, $E(\theta_1)$ n'est pas forcément égal à θ : on dit alors que θ_1 est un estimateur *biaisé* de θ.
Ce problème peut être résolu en choisissant les m clients non de manière uniforme comme dans les questions précédentes, mais de manière biaisée par la taille. Par analogie avec la construction de T_n dans la question 11, on commence par choisir une variable aléatoire J à valeurs dans $\{1, 2, \dots, n\}$, dont la loi est donnée par :

$$\forall \in [\![1, n]\!], \quad P(J = i) = \frac{x_i}{\sum_{r=1}^{n} x_r}.$$

Ensuite, étant donné J, on choisit un groupe V de $m-1$ clients parmi les $n-1$ clients différents de J, de manière uniforme. Autrement dit, pour toute partie $A \in \Omega$ et tout $i \in A$, on a :

$$P_{[J=i]}(V = A \setminus \{i\}) = \frac{1}{\binom{n-1}{m-1}}.$$

Le groupe de clients examiné est alors : $A = V \cup \{J\}$.

17. On commence par déterminer, pour $A \in \Omega$, la probabilité p_A que A soit choisie avec le protocole précédent.
    (a) Démontrer :
    $$p_A = \sum_{i \in A} P(J = i) P_{[J=i]}(A \setminus \{i\}).$$
    (b) En déduire :
    $$p_A = \frac{1}{\binom{n}{m}} \frac{\bar{x}_A}{\bar{x}}.$$

18. On définit ainsi une nouvelle probabilité π sur Ω par :

$$\forall A \in \Omega, \quad \pi(\{A\}) = p_A = \frac{1}{\binom{n}{m}} \frac{\bar{x}_A}{\bar{x}}.$$

(a) Montrer que l'espérance de θ_1 pour la probabilité π vérifie :

$$E(\theta_1) = \frac{1}{\binom{n}{m}} \sum_{A \in \Omega} \frac{\bar{y}_A}{\bar{x}}.$$

(b) Conclure : $E(\theta_1) = \theta$. On a donc construit un estimateur non biaisé de θ.

• FIN •