Concours Blanc 1 – Maths 2
ESSEC2, 2020

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. Ils 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 candidat repère ce qui lui semble être une erreur d'énoncé, il le signalera sur sa copie et poursuivra sa composition en expliquant les raisons des 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 $(\Omega, \mathscr{A}, \mathbf{P})$ (ou $(\Omega, \mathscr{A}, \pi)$ dans la dernière question du sujet). Pour toute variable aléatoire X, on notera $\mathbf{E}(X)$ son espérance et $\mathbf{V}(X)$ sa variance lorsqu'elles existent.

Partie I. Biais par la taille, exemples discrets.

1. On suppose que le nombre d'enfants dans une famille française est modélisé par une variable aléatoire X (lorsque l'on choisit une famille au hasard dans l'ensemble de toutes les familles françaises). 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. On suppose que X suit la loi binomiale de paramètres $n = 10$ et $p = 1/5$. On note $p_k = \mathbf{P}(X = k)$ pour $k \in \{0, 1, \dots, 10\}$.

   (a) i. Rappeler l'expression de p_k pour $k \in \{0, 1, \dots, 10\}$.
   ii. Donner $\mathbf{E}(X)$ et $V(X)$, et en déduire $\mathbf{E}(X^2)$.

   (b) Soit M_k le nombre de familles à k enfants dans la population.
   Soit de plus $M = \sum_{k=0}^{10} M_k$ le nombre total de familles dans la population. On suppose que les proportions observées sur les familles des élèves de l'école s'identifient à la loi de probabilité de X : on a donc $p_k = \frac{M_k}{M}$. 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 $N = \sum_{k=0}^{10} N_k$ le nombre total d'enfants de la population.
   i. Montrer que : $N_k = k p_k M$.
   ii. Montrer que : $\frac{N}{M} = \mathbf{E}(X)$.
   iii. Montrer que la proportion des enfants provenant d'une famille à k enfants est : $p_k^* = \frac{k p_k}{2}$.

   (c) On suppose toujours que les proportions observées sur les familles des élèves de cette école s'identifient à celles de la population générale. On choisit un élève au hasard dans l'école, 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 k élément de $\{1, 2, \dots, 10\}$, justifier que : $\mathbf{P}(Y = k) = \frac{k p_k}{2}$.
   ii. Montrer que : $\mathbf{E}(Y) = \frac{\mathbf{E}(X^2)}{\mathbf{E}(X)}$.

iii. En déduire la valeur de $\mathbf{E}(Y)$ et la comparer à $\mathbf{E}(X)$.

2. Soit X une variable aléatoire à valeurs dans $\mathbb{N}$, non identiquement nulle et admettant une espérance non nulle. Pour tout entier naturel non nul i, on pose :

$$q_i = \frac{i}{\mathbf{E}(X)}\mathbf{P}(X = i)$$

(a) Montrer que la suite $(q_i)_{i\geqslant 1}$ définit une loi de probabilité.
On considère la variable aléatoire X^* dont la loi est donnée par les q_i, c'est-à-dire :

$$\forall i \in \mathbb{N}^*, \mathbf{P}\left(X^* = i\right) = \frac{i}{\mathbf{E}(X)}\mathbf{P}(X = i)$$

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

(b) On suppose que X admet un moment d'ordre 2.

i. Montrer que: $\mathbf{E}(X^*) = \dfrac{\mathbf{E}\left(X^2\right)}{\mathbf{E}(X)}$.

ii. En déduire que $\mathbf{E}(X^*) \geqslant \mathbf{E}(X)$.

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. Expliciter 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 :

$$\forall k \in \mathbb{N}, \mathbf{P}(X = k+1) = \frac{\mathbf{E}(X)}{k+1}\mathbf{P}(X = k)$$

ii. Montrer que :

$$\forall k \in \mathbb{N}, \mathbf{P}(X = k) = \frac{(\mathbf{E}(X))^k}{k!}\mathbf{P}(X = 0)$$

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 $\{1,2,\ldots,n\}$ telle que pour $1 \leqslant k \leqslant n$, $\mathbf{P}(X = k) > 0$. 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 de bus 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 $\{1,\ldots,n\}$) et on suppose que pour tout entier k élément de $\{1,\ldots,n\}$:

$$\mathbf{P}_{[X=k]}(T = j) = \begin{cases} \dfrac{1}{k} & \text{si } j \in \{1,\ldots,k\} \\ 0 & \text{si } j > k \end{cases}$$

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

$$\sum_{j=1}^{n} j\mathbf{P}_{[X=k]}(T = j) = \frac{k+1}{2}$$

ii. En déduire que :

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

iii. En déduire que :

$$\mathbf{E}(T) = \frac{\mathbf{E}(X)+1}{2}$$

(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 à considérer est ainsi X^*, suivant la loi de X biaisée par la taille. Le temps d'attente T′ correspondant à cette situation vérifie donc en fait, pour $k \in \{1,\dots,n\}$:

$$\mathbf{P}_{[X^*=k]}\left(T' = j\right) = \begin{cases} \dfrac{1}{k} & \text{si } j \in \{1,\dots,k\} \\ 0 & \text{si } j > k \end{cases}$$

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

ii. En déduire que $\mathbf{E}\left(T'\right) \geqslant \mathbf{E}(T)$.

(c) On souhaite vérifier les résultats théoriques précédents à l'aide d'une simulation numérique en Python. On suppose ici que $X = Y + 1$, où Y suit la loi binomiale $\mathscr{B}\left(10, \dfrac{1}{2}\right)$.

i. Calculer $\mathbf{E}(T)$ et $\mathbf{E}\left(T'\right)$ dans cette situation et vérifier que $\mathbf{E}\left(T'\right) - \mathbf{E}(T) \simeq 0{,}2$.

ii. On suppose que les bus opèrent entre 8 h 01 et 18 h 00 et que la personne souhaitant prendre un bus arrive à un instant aléatoire t_0 entre 8 h 01 et 18 h 00 .
Justifier pourquoi la fonction Python suivante permet de simuler la variable aléatoire donnant le temps d'attente de la personne à l'arrêt de bus :

```
def T():
    t0 = rd.randint(1,601)
    s = 0
    while s <= t0:
        X = rd.binomial(10,1/2)+1
        s = s+X
    return s-t0
```

iii. Écrire un programme utilisant la fonction précédente qui calcule et affiche la moyenne du temps d'attente de la personne lorsque l'on réalise 10000 fois l'expérience.

iv. Trois exécutions successives de ce programme renvoient les valeurs `3.6824`, `3.7169` et `3.6855`. Qui de T ou de T′ semble répondre au mieux à la situation considérée ?

(d) Pourquoi l'hypothèse d'indépendance des intervalles de temps entre deux bus consécutifs du modèle décrit en introduction de la question 4 paraît trop simplificatrice par rapport à la réalité ?

Partie II. Propriétés du biais par la taille dans le cas discret.

Dans cette partie, X est une variable aléatoire discrète à valeurs dans $\mathbb{N}$ admettant une espérance $\mathbf{E}(X) > 0$. On rappelle que l'on dit que X^* suit la loi de X biaisée par la taille si :

$$\forall i \in \mathbb{N}^*, \mathbf{P}\left(X^* = i\right) = \frac{i}{\mathbf{E}(X)}\mathbf{P}(X = i)$$

5. Dans cette question, on se fixe $f : \mathbb{R}_+ \to \mathbb{R}$ et $g : \mathbb{R}_+ \to \mathbb{R}$ deux fonctions croissantes.
On suppose de plus que les espérances $\mathbf{E}(f(X)), \mathbf{E}(g(X))$ et $\mathbf{E}(f(X)g(X))$ existent.

(a) Montrer que quels que soient les réels positifs x_1 et x_2, on a :

$$\left(f(x_1) - f(x_2)\right)\left(g(x_1) - g(x_2)\right) \geqslant 0$$

(b) Soient X_1 et X_2 deux variables aléatoires indépendantes de même loi que X.
Montrer que :

$$\mathbf{E}\left(\left(f(X_1) - f(X_2)\right)\left(g(X_1) - g(X_2)\right)\right) = 2\mathbf{E}(f(X)g(X)) - 2\mathbf{E}(f(X))\mathbf{E}(g(X))$$

(c) En déduire que :

$$\mathbf{E}(f(X)g(X)) \geqslant \mathbf{E}(f(X))\mathbf{E}(g(X))$$

6. (a) Dans cette question, f et g sont deux fonctions définies sur $\mathbb{R}_+$ telles que :

$$\forall x \in \mathbb{R}_+, \quad |f(x)| \leqslant g(x)$$

Montrer que si $g(\mathrm{X})$ admet une espérance, alors $f(\mathrm{X})$ admet une espérance.

(b) Dans cette question, h est une fonction bornée définie sur $\mathbb{R}_+$.

i. Montrer que $h(\mathrm{X})$ admet une espérance.

ii. Montrer que $\mathrm{X}h(\mathrm{X})$ admet une espérance et que :

$$\mathbf{E}\left(h\left(\mathrm{X}^*\right)\right) = \frac{1}{\mathbf{E}(\mathrm{X})}\mathbf{E}(\mathrm{X}h(\mathrm{X}))$$

7. Dans cette question, on suppose qu'il existe un entier $m \geqslant 1$ tel que $\mathbf{E}\left(\mathrm{X}^{m+1}\right)$ 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 $\mathbf{E}\left(\mathrm{X}^p\right)$ existe.

(b) Montrer que: $\mathbf{E}\left(\mathrm{X}^{m+1}\right) \geqslant \mathbf{E}(\mathrm{X})\mathbf{E}\left(\mathrm{X}^m\right)$.

(c) Montrer que: $\mathbf{E}\left((\mathrm{X}^*)^m\right) \geqslant \mathbf{E}\left(\mathrm{X}^m\right)$.

8. Pour tout t réel positif, on définit la fonction g_t sur $\mathbb{R}_+$ par :

$$\forall x \in \mathbb{R}_+, g_t(x) = \begin{cases} 0 & \text{si } 0 \leqslant x \leqslant t \\ 1 & \text{si } x > t \end{cases}$$

(a) Montrer que la fonction $x \mapsto g_t(x)$ est croissante sur $\mathbb{R}_+$.

(b) Montrer que pour tout t réel positif, $\mathbf{E}\left(\mathrm{X}g_t(\mathrm{X})\right)$ est bien définie et :

$$\mathbf{E}\left(\mathrm{X}g_t(\mathrm{X})\right) \geqslant \mathbf{E}(\mathrm{X})\mathbf{P}(\mathrm{X} > t)$$

(c) En déduire que: $\forall t \in \mathbb{R}, \mathbf{P}\left(\mathrm{X}^* > t\right) \geqslant \mathbf{P}(\mathrm{X} > t)$
On dit que X^* domine stochastiquement X.

Partie III. Généralisation au cas de variables aléatoires quelconques.

Soit X est une variable aléatoire positive quelconque, non nécessairement discrète.
On suppose que X admet une espérance $\mathbf{E}(\mathrm{X}) > 0$.
Sur le modèle de la question 6(b)ii, on dit que X^* suit la loi de X biaisée par la taille si :

$$\text{Pour toute fonction } h : \mathbb{R}_+ \to \mathbb{R} \text{ bornée, } \mathbf{E}\left(h\left(\mathrm{X}^*\right)\right) = \frac{1}{\mathbf{E}(\mathrm{X})}\mathbf{E}(\mathrm{X}h(\mathrm{X}))$$

On admet que cette propriété caractérise une unique loi de probabilité.
Soit n un entier naturel non nul et soient $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$ des variables aléatoires positives quelconques, indépendantes, non nécessairement de même loi.
On suppose que pour tout $i \in \{1, 2, \ldots, n\}$, X_i admet une espérance strictement positive et on note $\mu_i = \mathbf{E}(\mathrm{X}_i)$.
On pose de plus :

$$\mu = \sum_{i=1}^{n} \mu_i \text{ et } \mathrm{S}_n = \sum_{i=1}^{n} \mathrm{X}_i$$

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

10. Pour A un événement, on note $\mathbb{1}_\mathrm{A}$ la variable aléatoire définie par $\mathbb{1}_\mathrm{A}(\omega) = 1$ si $\omega \in \mathrm{A}$, et $\mathbb{1}_\mathrm{A}(\omega) = 0$ sinon.

(a) Montrer que $\mathbb{1}_\mathrm{A}$ suit la loi de Bernoulli de paramètre $\mathbf{P}(\mathrm{A})$.

(b) Montrer que si $(\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n)$ est un système complet d'événements, alors $\sum_{i=1}^{n} \mathbb{1}_{\mathrm{A}_i}$ est la variable aléatoire constante égale à 1.

11. On considère des variables aléatoires indépendantes $X_1^*, X_2^*, \ldots, X_n^*$, indépendantes des X_i, telles que pour tout $i \in \{1,2,\ldots,n\}$, X_i^* suit la loi de X_i biaisée par la taille.
Soit J une variable aléatoire indépendante de $X_1, X_1^*, X_2, X_2^*, \ldots, X_n, X_n^*$, de loi donnée par :

$$\forall i \in \{1,2,\ldots,n\}, \mathbf{P}(J=i) = \frac{\mu_i}{\mu}$$

On considère la variable aléatoire $X_J = \sum_{i=1}^{n} X_i \mathbb{1}_{[J=i]}$ et on définit $T_n = S_n - X_J + X_J^*$.

Autrement dit, on choisit un indice aléatoire J et, dans la somme $\sum_{i=1}^{n} X_i$, on remplace X_J par X_J^*.

Soit enfin $h : \mathbb{R}_+ \to \mathbb{R}$ une fonction bornée.

(a) i. Montrer que :

$$h(T_n) = \sum_{i=1}^{n} h\left(S_n - X_i + X_i^*\right) \mathbb{1}_{[J=i]}$$

ii. En déduire que :

$$\mathbf{E}(h(T_n)) = \sum_{i=1}^{n} \mathbf{P}(J=i)\mathbf{E}\left(h\left(S_n - X_i + X_i^*\right)\right)$$

(b) Pour $i \in \{1,2,\ldots,n\}$, démontrer que :

$$\forall s \in \mathbb{R}_+, \mathbf{E}\left(h\left(s + X_i^*\right)\right) = \frac{1}{\mu_i}\mathbf{E}(X_i h(s+X_i))$$

On admettra qu'on en déduit l'égalité : $\mathbf{E}\left(h\left(S_n - X_i + X_i^*\right)\right) = \frac{1}{\mu_i}\mathbf{E}(X_i h(S_n))$

(c) En déduire que :

$$\mathbf{E}(h(T_n)) = \frac{\mathbf{E}(S_n h(S_n))}{\mathbf{E}(S_n)}$$

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

Partie IV. Une application 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-ème 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 tout le monde : elle décide donc de choisir m clients parmi les n pour effectuer ses mesures.
On considère ainsi dans cette partie l'univers $\Omega = \mathscr{P}_m$ constitué des parties à m éléments de $\{1,2,\ldots,n\}$.
Pour tout $A \in \mathscr{P}_m$, on note $\{A\}$ l'événement : « On choisit la partie A de $\{1,2,\ldots,n\}$ » . Si l'on suppose que pour sonder les clients on choisit une partie de $\{1,2,\ldots,n\}$ de manière uniforme dans $\mathscr{P}_m$, on a alors :

$$\forall A \in \mathscr{P}_m, \mathbf{P}(\{A\}) = \frac{1}{\binom{n}{m}}$$

Pour $A \in \mathscr{P}_m$, on notera :

$$\overline{x_A} = \frac{1}{m}\sum_{i \in A} x_i \text{ et } \overline{y_A} = \frac{1}{m}\sum_{i \in A} y_i$$

On note également :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ et } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

On définit enfin deux variables aléatoires X et Y sur $\Omega = \mathscr{P}_m$ de la manière suivante :

$$\forall \mathrm{A} \in \mathscr{P}_m, \mathrm{X}(\mathrm{A}) = \overline{x_{\mathrm{A}}} \quad \text{et} \quad \mathrm{Y}(\mathrm{A}) = \overline{y_{\mathrm{A}}}$$

La compagnie d'électricité décide d'utiliser $\theta_1 = \dfrac{\mathrm{Y}}{\mathrm{X}}$ comme estimateur de θ.
On admet que si l'on munit un univers fini Ω de la probabilité $\mathbf{P}$, alors l'espérance d'une variable aléatoire T définie sur Ω est donnée par:

$$\mathbf{E}(\mathrm{T}) = \sum_{\omega\in\Omega} \mathrm{T}(\omega)\mathbf{P}\left(\{\omega\}\right)$$

12. (a) Que représentent les nombres $\overline{x}$ et $\overline{y}$?
    (b) Que représentent les variables aléatoires X et Y ?

13. (a) Montrer que:

$$\mathbf{E}(\mathrm{X}) = \frac{1}{\binom{n}{m}} \sum_{\mathrm{A}\in\mathscr{P}_m} \overline{x_{\mathrm{A}}}$$

    (b) Soit $i \in \{1,2,\ldots,n\}$ un entier fixé.
    Combien y a-t-il de parties $\mathrm{A} \in \mathscr{P}_m$ telles que $i \in \mathrm{A}$?

    (c) En déduire que :

$$\sum_{\mathrm{A}\in\mathscr{P}_m} \sum_{i\in\mathrm{A}} x_i = \binom{n-1}{m-1} \sum_{i=1}^{n} x_i$$

    (d) Conclure que $\mathbf{E}(\mathrm{X}) = \overline{x}$. On admettra que de même on a $\mathbf{E}(\mathrm{Y}) = \overline{y}$.
    (e) Exprimer θ en fonction de $\mathbf{E}(\mathrm{X})$ et $\mathbf{E}(\mathrm{Y})$.

14. Soient W et Z deux variables aléatoires strictement positives admettant un moment d'ordre 2.
    On note, pour tout $t \in \mathbb{R}$: $\mathrm{Q}(t) = \mathbf{E}\left((\mathrm{W} + t\mathrm{Z})^2\right)$

    (a) Montrer que Q est un polynôme du second degré et déterminer le signe de Q sur $\mathbb{R}$.
    (b) En considérant le discriminant de Q, en déduire que :

$$(\mathbf{E}(\mathrm{WZ}))^2 \leqslant \mathbf{E}\left(\mathrm{W}^2\right)\mathbf{E}\left(\mathrm{Z}^2\right)$$

    (c) Montrer que l'inégalité de la question précédente est une égalité si et seulement si il existe un réel $\alpha > 0$ tel que $\mathrm{W} = \alpha\mathrm{Z}$ presque sûrement.

15. (a) Montrer que: $\mathbf{E}\left(\dfrac{1}{\mathrm{X}}\right) \geqslant \dfrac{1}{\mathbf{E}(\mathrm{X})}$.
    (b) Montrer qu'il y a égalité si et seulement si X est une variable aléatoire presque sûrement constante égale à $\overline{x}$.
    (c) Conclure que $\mathbf{E}\left(\dfrac{1}{\mathrm{X}}\right) = \dfrac{1}{\mathbf{E}(\mathrm{X})}$ si et seulement si $x_i = \overline{x}$ pour tout $i \in \{1,2,\ldots,n\}$.

16. Si l'on suppose que X et Y sont indépendantes, montrer que $\mathbf{E}(\theta_1) \geqslant \theta$, avec égalité si seulement si $x_i = \overline{x}$ pour tout $i \in \{1,2,\ldots,n\}$.

Ainsi $\mathbf{E}(\theta_1)$ n'est pas forcément égal à θ : on dit que θ_1 est un estimateur biaisé de θ.
Ce problème peut être résolu en choisissant la partie A contenant les m clients étudiés non de manière uniforme comme dans les questions 13 à 16, mais de manière biaisée par la taille. Par analogie avec la construction de T_n dans la question 11, on commence par choisir un indice aléatoire J à valeurs dans $\{1,2,\ldots,n\}$, dont la loi est donnée par :

$$\forall i \in \{1,2,\dots,n\}, \mathbf{P}(J=i) = \frac{x_i}{\sum_{r=1}^{n} x_r} \quad (**)$$

Ensuite, sachant $[J=i]$, on choisit un groupe V de $m-1$ clients parmi les $n-1$ clients différents de i, de manière uniforme:

$$\mathbf{P}_{[J=i]}(\{V\}) = \frac{1}{\binom{n-1}{m-1}}$$

La partie choisie pour effectuer le sondage est alors: $A = V \cup \{i\}$.

17. On souhaite programmer en Python un tirage aléatoire de la variable J, connaissant les valeurs des x_i.

    (a) On rappelle qu'en Python, la commande `rd.random()` renvoie un tirage d'une variable X équirépartie dans $[0,1]$: en particulier, pour tout $p \in [0,1]$, $\mathbf{P}(X \leqslant p) = p$.
    Montrer que si $0 \leqslant x \leqslant y \leqslant 1$, $\mathbf{P}(X \in]x,y]) = y - x$.

    (b) On note, pour $k \in [\![0,n]\!]$: $r_0 = 0$, et $r_k = \frac{\sum_{r=1}^{k} x_i}{\sum_{r=1}^{n} x_r}$ si $k \geqslant 1$. On a en particulier $r_n = 1$.
    Un tirage de la variable X décrite en $17a$ étant effectué, on note j l'unique entier de $[\![1,n]\!]$ tel que $r_{j-1} < X \leqslant r_j$ (on ne demande pas de justifier l'existence et l'unicité de j).
    Montrer que l'indice j défini ainsi est un tirage de la variable aléatoire J dont la loi est donnée par $(**)$.

    (c) Coder une fonction Python `def tirageJ(X)` qui prend en argument un `np.array` $X = [x_1,\dots,x_n]$ et renvoie un tirage de la variable aléatoire J.

    (d) On suppose codée une fonction `partie_aleatoire(L,m)` qui prend en argument une liste L de taille p, et un entier $m \leqslant p$, et renvoie une liste de m élements de L tirée aléatoirement de manière uniforme.
    Programmer, à l'aide d'un code qui appellera les deux fonctions précédentes, une fonction `tirage_taille(n,m,X)` qui permet de tirer une partie $A \subset [\![1,n]\!]$ de cardinal `m`, de manière biaisée par la taille, `X` étant la liste des x_i.

    Rappel : si `L` est une liste de longueur n, et $0 \leqslant a \leqslant b \leqslant n$, `L[a:b]` est la liste `[L[a],L[a+1],...,L[b-1]]`.
    De plus, si `L` et `M` sont deux listes, la commande `L+M` renvoie la concaténation de ces 2 listes.

18. Dans cette question, on détermine pour $A \in \mathscr{P}_m$, la probabilité p_A que A soit choisie avec le protocole précédent.

    (a) Montrer que :

    $$p_A = \sum_{i \in A} \mathbf{P}(J=i)\mathbf{P}_{[J=i]}(\{A \setminus \{i\}\})$$

    (b) En déduire que :

    $$p_A = \frac{1}{\binom{n}{m}} \frac{\overline{x_A}}{\overline{x}}$$

19. Dans cette dernière question, on munit à présent l'univers $\Omega = \mathscr{P}_m$ d'une nouvelle probabilité π définie par :

    $$\forall A \in \mathscr{P}_m, \pi(\{A\}) = \frac{1}{\binom{n}{m}} \frac{\overline{x_A}}{\overline{x}}$$

    Avec cette nouvelle distribution de probabilité, on reprend la variable aléatoire θ_1 telle que :

    $$\forall A \in \mathscr{P}_m, \theta_1(A) = \frac{Y(A)}{X(A)} = \frac{\overline{y_A}}{\overline{x_A}}$$

    (a) Montrer que :

    $$\mathbf{E}(\theta_1) = \sum_{A \in \mathscr{P}_m} \pi(\{A\})\theta_1(A) = \frac{1}{\binom{n}{m}} \sum_{A \in \mathscr{P}_m} \frac{\overline{y_A}}{\overline{x}}$$

    (b) Conclure que $\mathbf{E}(\theta_1) = \theta$.
    En choisissant l'échantillon A à étudier de manière biaisée par la taille, θ_1 est cette fois un estimateur non biaisé de θ.