ESSEC II 2020

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 $(\Omega, \mathscr{A}, \mathbb{P})$. Pour toute variable aléatoire X, on notera $\mathbb{E}(X)$ son espérance (resp. $\mathbb{V}(X)$ sa variance) lorsqu'elles existent.

Première partie : Biais par la taille, exemples discrets

1. On suppose que le nombre d'enfants dans une famille française est 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 suive la loi binomiale de paramètres $n = 10$ et $p = \frac{1}{5}$. On note $p_k = \mathbb{P}(\,[X = k]\,)$ pour $k \in \{0, 1, \ldots, 10\}$.

a) (i) Rappeler l'expression de p_k pour $k \in \{0, 1, \ldots, 10\}$.

(ii) Que vaut $\mathbb{E}(X)$?

(iii) Donner $\mathbb{V}(X)$, et en déduire $\mathbb{E}(X^2)$.

b) Soit M_k le nombre de familles à k enfants, $M = \sum_{k=0}^{10} M_k$ le nombre total de familles (donc $p_k = \dfrac{M_k}{M}$). 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) Démontrer : $N_k = k\, p_k\, M$.

(ii) Démontrer : $\dfrac{N}{M} = 2$.

(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 : $\mathbb{P}(\,[Y = k]\,) = \dfrac{k\, p_k}{2}$.

(ii) Démontrer : $\mathbb{E}(Y) = \dfrac{\mathbb{E}(X^2)}{\mathbb{E}(X)}$.

(iii) En déduire $\mathbb{E}(Y)$ et le comparer à $\mathbb{E}(X)$.

2. Soit X une variable aléatoire à valeurs dans $\mathbb{N}$, non identiquement nulle et admettant une espérance.
Pour tout entier $i > 0$, on pose : $q_i = \dfrac{i}{\mathbb{E}(X)}\ \mathbb{P}(\,[X = i]\,)$.

a) Calculer $\sum_{i=1}^{+\infty} q_i$.

La suite $(q_i)_{i>0}$ définie ci-dessus définit donc bien 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 :

$$\mathbb{P}(\,[X^* = i]\,) \ = \ \frac{i}{\mathbb{E}(X)}\ \mathbb{P}(\,[X = i]\,)$$

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 : $\mathbb{E}(X^*) = \dfrac{\mathbb{E}(X^2)}{\mathbb{E}(X)}$

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

d) Conclure : $\mathbb{E}(X^*) \geqslant \mathbb{E}(X)$

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 $k \geqslant 1$: $\mathbb{P}(\,[X = k]\,) = \dfrac{\mathbb{E}(X)}{k}\ \mathbb{P}(\,[X = k-1]\,)$.

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

(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 $1 \leqslant k \leqslant n$: $\mathbb{P}(\,[X = k]\,) > 0$. 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 $\{1, \dots, n\}$:

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

(i) Montrer que pour tout entier $k \in \{1, \dots, n\}$, on a : $\sum_{j=1}^{n} j\,\mathbb{P}_{[X=k]}(\,[T = j]\,) = \dfrac{k+1}{2}$.

(ii) En déduire : $\sum_{k=1}^{n}\sum_{j=1}^{n} j\,\mathbb{P}(\,[X = k]\,)\,\mathbb{P}_{[X=k]}(\,[T = j]\,) = \dfrac{\mathbb{E}(X+1)}{2}$.

(iii) Démontrer : $\mathbb{E}(T) = \sum_{j=1}^{n}\sum_{k=1}^{n} j\,\mathbb{P}(\,[X = k]\,)\,\mathbb{P}_{[X=k]}(\,[T = j]\,)$.

(iv) Démontrer : $\mathbb{E}(T) = \dfrac{\mathbb{E}(X+1)}{2}$.

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 $k \in \{1, \dots, n\}$:

$$\mathbb{P}_{[X^*=k]}(\,[T^* = j]\,) \;=\; \begin{cases} \dfrac{1}{k} & \text{si } j \in \{1, \dots, k\} \\ 0 & \text{si } j > k \end{cases}$$

(i) Montrer que pour tout entier $k \in \{1, \dots, n\}$, on a : $\displaystyle\sum_{j=1}^{n} j\,\mathbb{P}_{[X^*=k]}(\,[T^* = j]\,) = \frac{k+1}{2}$.

(ii) Démontrer : $\displaystyle\mathbb{E}(T^*) = \sum_{j=1}^{n}\sum_{k=1}^{n} j\,\mathbb{P}(\,[X^* = k]\,)\,\mathbb{P}_{[X^*=k]}(\,[T^* = j]\,)$.

(iii) Démontrer : $\mathbb{E}(T^*) = \dfrac{\mathbb{E}(X^*+1)}{2}$.

(iv) En déduire qu'on a : $\mathbb{E}(T^*) \geqslant \mathbb{E}(T)$.

Deuxième partie : Biais par la taille, propriétés

Dans cette partie, on démontre de nombreuses propriétés des variables aléatoires biaisées par la taille.

5. *Biais par la taille : le cas de variables à densité.*
Soit X une variable aléatoire **positive** de densité f et admettant une espérance $\mathbb{E}(X)$ strictement positive (donc $f(x) = 0$ pour tout x strictement négatif).
On définit la fonction g sur $\mathbb{R}$ par : $g : x \mapsto \dfrac{x}{\mathbb{E}(X)}\, f(x)$.

a) Montrer que g définit une densité d'une variable aléatoire positive.

Soit une variable aléatoire X^* dont une densité est g. On dit que X^* suit *la loi de X biaisée par la taille.*

b) Soit a un réel strictement positif.

(i) Montrer que la variable aléatoire $a\,X$ possède pour densité $x \mapsto \dfrac{1}{a}\, f\left(\dfrac{x}{a}\right)$.

(ii) En déduire que $(a\,X)^*$ et $a \cdot X^*$ possèdent la même loi.

c) *Une propriété importante.*
Soit $h : [0, +\infty[\to \mathbb{R}$ une fonction bornée et continue sauf éventuellement en un nombre fini de points. Montrer que $\mathbb{E}\big(X\,h(X)\big)$ est bien défini et :

$$\mathbb{E}\big(h(X^*)\big) \;=\; \frac{1}{\mathbb{E}(X)}\,\mathbb{E}\big(X\,h(X)\big)$$

On pose alors la définition suivante (**que la variable X soit à densité ou non**) : si X est une variable aléatoire réelle positive d'espérance $\mathbb{E}(X)$ strictement positive, on dit que **la variable aléatoire positive Y suit la loi de X biaisée par la taille** si on a :

$$\mathbb{E}\big(h(Y)\big) \;=\; \frac{1}{\mathbb{E}(X)}\,\mathbb{E}\big(X\,h(X)\big)$$

pour toute fonction $h : [0, +\infty[\to \mathbb{R}$ bornée et continue sauf éventuellement en un nombre fini de points.

6. Dans cette question, on se fixe $f : \mathbb{R} \to \mathbb{R}$ et $g : \mathbb{R} \to \mathbb{R}$ deux fonctions **croissantes**. Soit X une variable aléatoire telle que les espérances $\mathbb{E}\big(f(X)\big)$, $\mathbb{E}\big(g(X)\big)$ et $\mathbb{E}\big(f(X)\,g(X)\big)$ sont bien définies.

a) Montrer que quels que soient les réels x_1 et x_2, 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, de même loi que X. Démontrer :

$$\mathbb{E}\Big(\big(f(X_1) - f(X_2)\big)\,\big(g(X_1) - g(X_2)\big)\Big) \;=\; 2\,\mathbb{E}\big(f(X)\,g(X)\big) - 2\,\mathbb{E}\big(f(X)\big)\,\mathbb{E}\big(g(X)\big)$$

c) En déduire : $\mathbb{E}\big(f(X)\,g(X)\big) \geqslant \mathbb{E}\big(f(X)\big)\,\mathbb{E}\big(g(X)\big)$.

7. Dans cette question, on suppose que X est une variable aléatoire positive d'espérance strictement positive, et telle que $\mathbb{E}(X^{m+1})$ existe pour un entier $m \geqslant 1$ donné.

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) Montrer que $\mathbb{E}(X^p)$ existe.

b) Démontrer : $\mathbb{E}(X^{m+1}) \geqslant \mathbb{E}(X)\,\mathbb{E}(X^m)$.

c) En déduire : $\mathbb{E}\left((X^*)^m\right) \geqslant \mathbb{E}(X^m)$.

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

$$\mathbb{1}_A : \omega \mapsto \begin{cases} 1 & \text{si } \omega \in A \\ 0 & \text{sinon} \end{cases}$$

Pour tout t réel, on définit la fonction g_t sur $\mathbb{R}$ par : $g_t : x \mapsto \mathbb{1}_{]t,+\infty[}(x)$.

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

b) Soit X une variable aléatoire positive, admettant une espérance. Montrer que pour tout t réel, $\mathbb{E}\big(X\,g_t(X)\big)$ est bien défini et : $\mathbb{E}\big(X\,g_t(X)\big) \geqslant \mathbb{E}(X)\,\mathbb{P}(\,[X > t]\,)$.

c) Démontrer, pour tout t réel : $\mathbb{P}(\,[X^* > t]\,) \geqslant \mathbb{P}(\,[X > t]\,)$.

On dit que X^* *domine stochastiquement* X.

9. Soit $X_1, \ldots, X_n$ des variables aléatoires positives, indépendantes, non nécessairement de même loi. On suppose qu'elles admettent toutes une espérance strictement positive, et on note $\mu_i = \mathbb{E}(X_i)$. De plus, on pose : $\mu = \sum_{i=1}^{n} \mu_i$ et $S_n = \sum_{i=1}^{n} X_i$.

a) Donner $\mathbb{E}(S_n)$.

b) Soit J une variable aléatoire à valeurs dans $\{1, \ldots, n\}$, de loi $\mathbb{P}(\,[J = k]\,) = \dfrac{\mu_k}{\mu}$. Quelle est la loi de J si les variables aléatoires X_i sont de même loi ?

On considère $X_1^*, \ldots, X_n^*$ des variables aléatoires indépendantes, indépendantes de $X_1, \ldots, X_n$ 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 aussi J une variable aléatoire de loi $\mathbb{P}(\,[J = k]\,) = \dfrac{\mu_k}{\mu}$, indépendante de X_1, X_1^*, …, X_n, X_n^*.
On considère la variable aléatoire $X_J = \sum_{j=1}^{n} X_j\,\mathbb{1}_{[J=j]}$ 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^*.

c) Soit $h : [0, +\infty[\to \mathbb{R}$ une fonction bornée et continue sauf éventuellement en un nombre fini de points.

(i) Démontrer : $h(T_n) = \sum_{i=1}^{n} h(T_n)\,\mathbb{1}_{[J=i]} = \sum_{i=1}^{n} h(S_n - X_i - X_i^*)\,\mathbb{1}_{[J=i]}$.

(ii) En déduire : $\mathbb{E}\big(h(T_n)\big) = \sum_{i=1}^{n} \mathbb{P}(\,[J=i]\,)\,\mathbb{E}\big(h(S_n - X_i + X_i^*)\big)$.

d) Pour $i \in \{1, \dots, n\}$, démontrer, pour tout réel s : $\mathbb{E}\big(h(s + X_i^*)\big) = \dfrac{1}{\mu_i}\,\mathbb{E}\big(X_i\,h(s + X_i)\big)$.

On admettra qu'on en déduit l'égalité : $\mathbb{E}\big(h(S_n - X_i + X_i^*)\big) = \dfrac{1}{\mu_i}\,\mathbb{E}\big(X_i\,h(S_n)\big)$.

e) En déduire : $\mathbb{E}\big(h(T_n)\big) = \dfrac{\mathbb{E}\big(S_n\,h(S_n)\big)}{\mathbb{E}(S_n)}$.

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

Troisième partie : Applications en Statistique

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 $i^{\text{ème}}$ 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.

10. Soit m un entier fixé tel que $1 \leqslant m \leqslant n$. On note $\mathcal{P}_m$ l'ensemble des parties $A \subset \{1, \dots, n\}$ de cardinal m. On considère une variable aléatoire R, à valeurs dans $\mathcal{P}_m$ et de loi uniforme, c'est-à-dire telle que pour toute partie $A \in \mathcal{P}_m$: $\mathbb{P}(\,[R = A]\,) = \dfrac{1}{\binom{n}{m}}$.

On souhaite écrire un programme pour choisir l'ensemble R au hasard.

a) On considère la procédure suivante : on prend un premier élément s_1 uniformément dans $\{1, \dots, n\}$, puis un deuxième élément s_2 uniformément dans $\{1, \dots, n\} \setminus \{s_1\}$, etc. puis un $m^{\text{ème}}$ élément s_m uniformément dans $\{1, \dots, n\} \setminus \{s_1, \dots, s_{m-1}\}$. On note $S = (s_1, \dots, s_m)$, qui est un m-uplet aléatoire.

(i) Montrer que pour tout m-uplet $(a_1, \dots, a_m)$ d'entiers distincts de $\{1, \dots, n\}$, on a :

$$\mathbb{P}(\,[S = (a_1, \dots, a_m)]\,) \;=\; \frac{(n-m)!}{n!}$$

(ii) On note $R = \{s_1, \dots, s_m\}$ l'*ensemble* des entiers tirés lors de la procédure décrite plus haut (l'ordre dans lequel ils ont été tirés n'importe plus). Montrer que pour tout ensemble $A = \{a_1, \dots, a_m\} \subset \{1, \dots, n\}$ de cardinal m, on a : $\mathbb{P}(\,[R = A]\,) = \dfrac{m!\,(n-m)!}{n!}$.
En déduire que l'ensemble R a été choisi uniformément dans $\mathcal{P}_m$.

b) Pour un réel x, on note $\lfloor x \rfloor$ sa partie entière, c'est-à-dire le plus grand entier naturel inférieur ou égal à x. Montrer que si U suit la loi uniforme sur $[0, 1[$, alors $X = 1 + \lfloor n\,U \rfloor$ suit la loi uniforme sur $\{1, \dots, n\}$.

c) On rappelle que la fonction `rand()` renvoie un nombre aléatoire de loi uniforme sur $[0,1[$, et que `floor(x)` renvoie la partie entière de `x`. Écrire une fonction `Uniforme` en **Scilab** qui prend en argument un entier n, et renvoie un nombre (aléatoire), uniforme sur $\{1,\dots,n\}$.

```
function x = Uniforme(n)
    ...
endfunction
```

d) Écrire une fonction `Selection`, qui prend en argument un vecteur `V` et renvoie un élément `x` de `V` pris de manière aléatoire parmi tous les éléments de `V`, ainsi que le vecteur `W`, égal au vecteur `V` auquel on a enlevé l'élément `x`. L'instruction `length(V)` renvoie le nombre d'éléments du vecteur `V`.

```
function [x, W] = Selection(V)
    n = length(V)
    ...
endfunction
```

e) Compléter le programme suivant, qui prend en argument deux entiers n et m avec $m \leqslant n$, et renvoie un vecteur `R` de m entiers distincts, pris uniformément dans $\{1,\dots,n\}$:

```
function R = Choix(m, n)
    V = 1:n
    R = []
    for i = 1:m
        ...
    end
endfunction
```

11. Pour une partie $A \in \mathcal{P}_m$, on définit :

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

La compagnie décide d'utiliser $\theta_R = \dfrac{\bar{y}_R}{\bar{x}_R}$ comme estimateur de θ.

a) On définit deux variables aléatoires $X = \bar{x}_R = \dfrac{1}{m}\displaystyle\sum_{i\in R} x_i$ et $Y = \bar{y}_R = \dfrac{1}{m}\displaystyle\sum_{i\in R} y_i$, qui correspondent aux montants moyens payés et consommés par les m clients du groupe tiré au hasard.

(i) Démontrer : $\mathbb{E}(X) = \dbinom{n}{m}^{-1} \displaystyle\sum_{A\in\mathcal{P}_m} \bar{x}_A$.

(ii) Soit $1 \leqslant i \leqslant n$ un entier naturel. Calculer le nombre de parties $A \in \mathcal{P}_m$ telles que $i \in A$.

(iii) En déduire :

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

(iv) Conclure : $\mathbb{E}(X) = \bar{x}$. On **admettra** que de même on a : $\mathbb{E}(Y) = \bar{y}$.

(v) Exprimer θ en fonction de $\mathbb{E}(X)$ et $\mathbb{E}(Y)$.

b) Démontrer : $\mathbb{E}(\theta_R) = \mathbb{E}\left(\dfrac{Y}{X}\right)$.

c) On donne l'inégalité de Cauchy-Schwarz : si W et Z sont deux variables aléatoires strictement positives, admettant un moment d'ordre deux : $\mathbb{E}(W\,Z) \leqslant \left(\mathbb{E}(W^2)\right)^{\frac{1}{2}} \left(\mathbb{E}(Z^2)\right)^{\frac{1}{2}}$, avec égalité si et seulement s'il existe un $\alpha > 0$ tel que $W = \alpha\, Z$.

(i) Démontrer : $\mathbb{E}\left(\dfrac{1}{X}\right) \geqslant \dfrac{1}{\mathbb{E}(X)}$.

(ii) Montrer qu'il y a égalité si et seulement si X est une variable aléatoire constante, c'est-à-dire $X = \mathbb{E}(X) = \bar{x}$.

(iii) Conclure que $\mathbb{E}\left(\dfrac{1}{X}\right) = \dfrac{1}{\mathbb{E}(X)}$ si et seulement si pour tout $i \in [\![1, n]\!]$, $x_i = \bar{x}$.

d) Si on suppose que X et Y sont indépendantes, montrer que $\mathbb{E}(\theta_R) \geqslant \theta$, avec égalité si et seulement si pour tout $i \in [\![1, n]\!]$, $x_i = \bar{x}$.

Ainsi, $\mathbb{E}(\theta_R)$ n'est pas forcément égal à θ : on dit alors que θ_R est un estimateur *biaisé* de θ.

12. Ce problème peut être résolu en choisissant les m clients non de manière uniforme comme dans la question ***10.***, mais de manière biaisée par la taille. Par analogie avec la construction de T_n dans la question ***9.***, on commence par choisir une variable aléatoire J à valeurs dans $\{1, 2, \ldots, n\}$, dont la loi est donnée par : $\mathbb{P}(\,[J = i]\,) = \dfrac{x_i}{\displaystyle\sum_{r=1}^{n} x_r}$. 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 $A \in \mathcal{P}_m$, et tout $i \in A$, on a :

$$\mathbb{P}_{[J=i]}(\,[V = A \setminus \{i\}]\,) \;=\; \frac{1}{\binom{n-1}{m-1}}$$

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

a) On commence par déterminer $\mathbb{P}(\,[R = A]\,)$, pour $A \in \mathcal{P}_m$ donné.

(i) Démontrer :

$$\mathbb{P}(\,[R = A]\,) \;=\; \sum_{i \in A} \mathbb{P}(\,[J = i]\,)\, \mathbb{P}_{[J=i]}(\,[V = A \setminus \{i\}]\,)$$

(ii) En déduire :

$$\mathbb{P}(\,[R = A]\,) \;=\; \frac{1}{\binom{n}{m}}\, \frac{\bar{x}_A}{\bar{x}}$$

13. Une fois choisi le groupe de clients R (par la procédure de la question ***12.***), on définit : $\hat{\theta}_R = \dfrac{\bar{y}_R}{\bar{x}_R}$.

a) Démontrer :

$$\mathbb{E}\left(\hat{\theta}_R\right) \;=\; \frac{1}{\binom{n}{m}} \sum_{A \in \mathcal{P}_m} \frac{\bar{y}_A}{\bar{x}}$$

b) Conclure : $\mathbb{E}\left(\hat{\theta}_R\right) = \theta$. On a donc construit un estimateur non biaisé de θ.