DEVOIR NUMÉRO 3 VERSION B

ECG2 MATHS APPLIQUÉES

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

Dans tous les programmes Python, on suppose qu'on a importé les bibliothèques habituelles avec leurs alias courants.

```python
import numpy as np
import numpy.random as rd
```

Première partie : biais par la taille, exemple 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 $p = 1/5$. 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 $E(X)$?

(iii) Donner $\mathrm{Var}(X)$, et en déduire $E\left(X^2\right)$.

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

Date: 16 Novembre 2023 08h30-12h00.
http://louismerlin.fr.

(ii) Montrer que $N/M = 2$.

(iii) Montrer que la proportion des enfants provenant d'une famille à k enfants est $p_k^* = kp_k/2$.

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 $\{1, 2, \ldots, 10\}$, montrer que $P(Y = k) = kp_k/2$.

(ii) Montrer que $E(Y) = E\left(X^2\right)/E(X)$.

(iii) En déduire $E(Y)$ et le comparer à $E(X)$.

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 $q_i = \frac{i}{E(X)}\mathbb{P}([X = i])$.

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

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

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

b. On suppose que X admet un moment d'ordre 2. Montrer que $E(X^*) = E\left(X^2\right)/E(X)$.

c. En déduire que si $E\left(X^2\right)$ existe, on a $\mathrm{Var}(X) = E(X)(E(X^*) - E(X))$.

d. Conclure que $E(X^*) \geq E(X)$.

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

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

(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 $\{1, \ldots, n\}$ 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 $\{1, \ldots, n\}$) et on suppose donc que pour tout entier k élément de $\{1, \ldots, n\}, \mathbb{P}_{[X=k]}(T = j) = 1/k$ si $j \in \{1, \ldots, k\}$ et $\mathbb{P}_{[X=k]}(T = j) = 0$ si $j > k$.

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

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

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

(iv) Montrer que $E(T) = \frac{E(X+1)}{2}$.

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 $k \in \{1,\dots,n\}, \mathbb{P}_{[X^*=k]}(T^* = j) = 1/k$ si $j \in \{1,\dots,k\}$ et $\mathbb{P}_{[X^*=k]}(T^* = j) = 0$ si $j > k$.

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

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

(iii) Montrer que $E(T^*) = \frac{E(X^*+1)}{2}$.

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

Deuxième partie : biais par la taille, propriétés.

La partie précédente a fait apparaître la définition suivante :

Définition : Biais par la taille

Soit X une variable aléatoire réelle positive d'espérance $E(X)$ strictement positive. On dit que la variable aléatoire Y suit la loi de X biaisée par la taille si on a

$$E(h(Y)) = \frac{1}{E(X)}E(Xh(X)),$$

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

a. Remarque pour les cubes : cette définition est valable aussi lorsque X est à densité.

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

5. 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 $E(f(X)), E(g(X))$ et $E(f(X)g(X))$ sont bien définies.

a. Montrer que quels que soient les réels x_1 et x_2, on a $(f(x_1) - f(x_2))(g(x_1) - g(x_2)) \geq 0$.

b. Soient X_1, X_2 deux variables aléatoires indépendantes, de même loi que X. Montrer que

$$E((f(X_1) - f(X_2))(g(X_1) - g(X_2))) = 2E(f(X)g(X)) - 2E(f(X))E(g(X))$$

c. En déduire que $E[f(X)g(X)] \geqslant E(f(X))E(g(X))$.

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

b. Montrer que $E(X^{m+1}) \geqslant E(X)E(X^m)$.

c. En déduire que $E((X^*)^m) \geq E(X^m)$.

7. Pour A un événement, on note $\mathbb{I}_A$ la variable aléatoire définie par $\mathbb{I}_A(\omega) = 1$ si $\omega \in A$ et $\mathbb{I}_A(\omega) = 0$ sinon. Pour tout t réel, on définit la fonction $g_t(x) = \mathbb{I}_{]t,+\infty[}(x)$.

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, $E(Xg_t(X))$ est bien défini et que $E(Xg_t(X)) \geq E(X)P(X > t)$.

c. Montrer que pour tout t réel, $P(X^* > t) \geq P(X > t)$.
On dit que X^* domine stochastiquement X.

8. 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 $\mu_i = E(X_i)$. De plus, on pose $\mu = \sum_{i=1}^{n} \mu_i$, et $S_n = \sum_{i=1}^{n} X_i$.

a. Donner $E(S_n)$.

b. Soit J une variable aléatoire à valeur dans $\{1, \ldots, n\}$, de loi $P(J = k) = \mu_k/\mu$. 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 $1 \leq i \leq n, X_i^*$ suive la loi de X_i biaisée par la taille. Soit aussi J une variable aléatoire de loi $P(J = k) = \mu_k/\mu$, indépendante de $X_1, X_1^*, \ldots, X_n, X_n^*$. On considère la variable aléatoire $X_J = \sum_{j=1}^{n} X_j \mathbb{I}_{[J=j]}$ 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) Montrer que $h(T_n) = \sum_{i=1}^{n} h(T_n)\, \mathbb{I}_{[J=i]} = \sum_{i=1}^{n} h(S_n - X_i - X_i^*)\, \mathbb{I}_{[J=i]}$.

(ii) En déduire que $E(h(T_n)) = \sum_{i=1}^{n} P(J = i) E(h(S_n - X_i + X_i^*))$.

d. Pour $i \in \{1, \ldots, n\}$, montrer que pour tout réel $s, E(h(s + X_i^*)) = \frac{1}{\mu_i} E(X_i h(s + X_i))$.

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

e. En déduire que $E(h(T_n)) = E(S_n h(S_n)) / E(S_n)$.

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-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

$$\theta = \left(\sum_{i=1}^{n} y_i \right) / \left(\sum_{i=1}^{n} x_i \right)$$

pour déterminer à quel point elle a mal facturé ses clients.

9. Soit m un entier fixé tel que $1 \leqslant m \leqslant n$. On note $\mathcal{P}_m$ l'ensemble des parties $A \subset \{1, \ldots, n\}$ 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 $A \in \mathcal{P}_m, \mathbb{P}([R = A]) = \frac{1}{\binom{n}{m}}$. 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 $\{1, \ldots, n\}$, puis un deuxième élément s_2 uniformément dans $\{1, \ldots, n\} \setminus \{s_1\}$, etc... puis un m-ème élément s_m uniformément dans $\{1, \ldots, n\} \setminus \{s_1, \ldots, s_{m-1}\}$. On note $S = (s_1, \ldots, s_m)$, 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 $\mathbb{P}([R = A]) = \frac{m!(n-m)!}{n!}$. En déduire que l'ensemble R a été choisi uniformément dans $\mathcal{P}_m$.

b. On rappelle que `rd.randint(1,n+1)` renvoie un entier au hasard entre 1 et n. Compléter la fonction `selection(V)`, 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 `len(V)` renvoie le nombre d'éléments du vecteur V .

```
def selection(V) :
  n = len(V)
  a = rd.randint(1,n+1)
  x = .....
  W = del(......)
  return [x,W]
```

c. 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 $\{1, \dots, n\}$:

```
def choix(m,n) :
  V = range(1,n+1)
  R = []
  for i in range(m) :
    .....
  return R
```

10. Pour une partie $A \in \mathcal{P}_m$, on définit $\bar{x}_A = \frac{1}{m} \sum_{i \in A} x_i$, $\bar{y}_A = \frac{1}{m} \sum_{i \in A} y_i$, et aussi $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

La compagnie décide d'utiliser $\theta_R = \bar{y}_R / \bar{x}_R$ comme estimateur de θ [1].

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

(i) Montrer que $E(X) = \binom{n}{m}^{-1} \sum_{A \in \mathcal{P}_m} \bar{x}_A$.

(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 que

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

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

(v) Exprimer θ en fonction de $E(X)$ et $E(Y)$.

b. Montrer que $E(\theta_R) = E\left(\frac{Y}{X}\right)$.

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

1. Remarque pour les carrés : *L'estimation* est le sujet du dernier chapitre de notre cours mais les questions qui suivent vous sont accessibles. Un estimateur est une variable aléatoire dont on observe une réalisation. Cette réalisation nous donne une estimation du paramètre θ (c'est-à-dire une valeur approchée). On essaye de comprendre dans la fin du sujet quelles sont les qualités de ce processus d'approximation.

(i) Montrer que $E(1/X) \geqslant 1/E(X)$.

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

(iii) Conclure que $E(1/X) = 1/E(X)$ si et seulement si $x_i = \bar{x}$ pour tout i.

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

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

11. Ce problème peut être résolu en choisissant les m clients non de manière uniforme comme dans la question **9**, mais de manière biaisée par la taille. Par analogie avec la construction de T_n dans la question **8**, on commence par choisir une variable aléatoire J à valeurs dans $\{1, 2, \ldots, n\}$, dont la loi est donnée par $\mathbb{P}([J = i]) = x_i / \sum_{r=1}^{n} x_r$. 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 $A \in \mathcal{P}_m$, et tout $i \in A$, on a

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

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) Montrer que

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

(ii) En déduire que

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

12. Une fois choisi le groupe de clients R (par la procédure de la question **11**), on définit $\hat{\theta}_R = \bar{y}_R / \bar{x}_R$.

a. Montrer que

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

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