Sujet 4 : algorithmes de sélection d'un échantillon

L'étude statistique d'une population se base généralement sur l'étude d'un échantillon. Supposons que l'on souhaite estimer l'argent de poche reçu en moyenne chaque mois par les N=20 élèves d'une classe. Pour éviter d'avoir à interroger chacun des 20 élèves, on peut par exemple sélectionner un échantillon de n=4 élèves selon un tirage avec remise.

Num de tirage	Individus de la population																			
1	A	B	C	D	E	**F**	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
2	A	B	**C**	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
3	A	B	C	D	E	**F**	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
4	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	**S**	T

Dans l'exemple ci-dessus, l'individu F est sélectionné 2 fois (tirages 1 et 3), et les individus C et S sont sélectionnés chacun 1 fois. Si l'argent de poche mensuel des individus C, F et S est de 30, 40 et 10 € respectivement, l'argent de poche moyen calculé sur l'échantillon des élèves vaut :

$$(1 * 30 + 2 * 40 + 10)/4 = 30€ .$$

On peut l'utiliser comme **estimation** (cf. glossaire) de l'argent de poche moyen de l'ensemble de la classe. On peut montrer que cette estimation coïncide en **espérance** (cf. glossaire) avec la véritable valeur moyenne de l'argent de poche dans la classe.

Le tirage avec remise est une méthode possible permettant de sélectionner un échantillon dans une population, mais en pratique on utilise plutôt des méthodes de tirage **sans remise** (i.e., un individu ne peut être sélectionné qu'une seule fois). Parmi les méthodes de sélection les plus utilisées, nous donnons deux exemples.

Echantillonnage systématique : on tire d'abord au hasard un individu parmi les N/n=5 premiers. Si l'individu numéroté i est sélectionné, on tire également tous les individus espacés de 5 par rapport à l'individu i, c'est-à-dire les individus i, i+5, i+10 et i+15.

Num de tirage	Individus de la population																			
1	A	B	C	**D**	E															
2						F	G	H	**I**	J										
3											K	L	M	**N**	O					
4																P	Q	R	**S**	T

Dans l'exemple, le 1er tirage ne concerne que les individus de A à E, et on tire l'unité D. On prend ensuite dans l'échantillon les individus I (espacé de 5 par rapport à D), N (espacé de 10) et S (espacé de 15). On peut là aussi utiliser comme estimation l'argent de poche moyen calculé sur l'échantillon.

Echantillonnage simple sans remise : on procède comme pour l'échantillonnage avec remise, mais si un individu est sélectionné lors d'un tirage, il ne peut plus être sélectionné lors des tirages suivants.

Num de tirage	Individus de la population																			
1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	**O**	P	Q	R	S	T
2	A	B	C	D	E	F	G	**H**	I	J	K	L	M	N		P	Q	R	S	T
3	A	B	C	D	E	F	G		**I**	J	K	L	M	N		P	Q	R	S	T
4	A	B	C	D	E	F	G			J	K	L	M	N		P	Q	R	S	**T**

Dans l'exemple, on tire l'individu O lors du 1er tirage : cet individu ne peut donc pas être concerné par les tirages suivants. On sélectionne ensuite successivement

les individus H, I et T dans l'échantillon. On peut là aussi utiliser comme estimation l'argent de poche moyen calculé sur l'échantillon.

Travail à réaliser

L'objectif de ce projet est d'étudier les propriétés de l'un et/ou l'autre de ces deux algorithmes sans remise, et de les comparer à la méthode d'échantillonnage avec remise. Il est demandé d'implémenter un des deux algorithmes sans remise pour une population de taille N quelconque (pour la taille d'échantillon, on pourra se limiter au cas où N est un multiple de n, comme dans les exemples).

En outre, pouvez-vous répondre aux questions suivantes (la liste n'est pas limitative) :

1. Avec l'échantillonnage systématique ou l'échantillonnage simple sans remise, chaque individu a-t-il la même probabilité d'être sélectionné dans l'échantillon?
2. Les trois méthodes décrites ci-dessus présentent-elles des avantages/inconvénients respectifs? Par exemple, si vous voulez réaliser une enquête à la sortie du lycée, laquelle vous semble adaptée?
3. Pouvez-vous trouver des cas où un des trois algorithmes fournit, par rapport aux deux autres, une estimation généralement plus proche de la vraie moyenne?
4. Pouvez-vous trouver des cas où un des algorithmes permet une estimation parfaite de la vraie moyenne (au sens où, quel que soit l'échantillon sélectionné, l'estimation coïncide avec la vraie valeur que l'on veut estimer)?