15 Simulation et échantillonnage

Livre p.302.

Objectifs :

- Expliciter la notion de simulation
- Définir un échantillon
- Introduire la notion d'intervalle de fluctuation

Introduction :

Dans le chapitre 8 nous avons appris des outils permettant d'étudier une série statistiques. Dans ce chapitre nous allons apprendre à obtenir des séries statistiques grâce à la simulation d'expériences aléatoires et nous allons aussi nous intéresser à l'échantillonnage et à un de ses problème : comment déterminer si un échantillon est représentatif d'une population ou d'une situation ?

1. Simulation

Pour obtenir des données statistiques sur un phénomène aléatoire (jeu de dés, de cartes, de loterie, n'obtenir que des feux verts lors d'un trajet, réussite au bac [1],...) on peut répéter un grand nombre de fois une expérience et consigner les résultats pour ensuite les traiter. Cette méthode peut être longue et fastidieuse si on veut par exemple s'intéresser à la somme des nombres obtenus sur les faces supérieures de deux dés jetés simultanément [2]. Pour y remédier on peut utiliser une calculatrice, un tableur ou mieux un langage de programmation quelconque sur un ordinateur.
Lorsqu'on veut simuler une expérience où intervient le hasard, on va demander à la machine (calculatrice ou ordinateur) des nombres aléatoires. Ces nombres s'obtiennent avec la fonction *Random* qui nous renvoie un nombre aléatoire compris entre 0 et 1 (mais différent de 1). En le multipliant par 6, on obtient un nombre entre 0 et 6 (6 exclu). En prenant la partie entière, on obtient un entier compris entre 0 et 5. Il reste à ajouter 1 pour obtenir la simulation d'un jet de dé à six faces !
Nous allons résumer ici quelques commandes de la calculatrice et du tableur permettant de simuler des expériences aléatoires.

Voici d'abord quelques séquences de touches pour obtenir certaines fonctions :

	Sur Casio		Sur TI	
Fonction	Pour obtenir	On appuie sur	Pour obtenir	On appuie sur
Random	`RAN#`	`OPTN`, puis `PROB` et `RAN#`	`RAND`	`MATH`, puis `PRB` et `RAND`
Partie entière	`INT`	`OPTN`, puis `NUM` et `INT`	`INT`	`MATH`, puis `NUM` et `INT`

1. Oups ! La réussite au bac n'est pas aléatoire !
2. Ou pire : si on s'intéresse à la réussite au bac

Quelques exemples de simulations :

Expérience à simuler	Commande calculatrice	Formule tableur
Obtenir un nombre aléatoire $x \in [0;1[$	Casio : `RAN#`. TI : `RAND`.	Saisir la formule suivante dans une cellule : `=ALEA()`
Simuler le jet d'un dé à six faces numérotées de 1 à 6	Casio : `INT(6*RAN#)+1`. TI : `INT(6*RAND)+1`.	Saisir la formule `=ENT(6*ALEA())+1`
« Pile ou face » : 0=pile, 1=face	Casio : `INT(2*RAN#)` TI : `INT(2*RAND)`	`=SI(ALEA()<0,5; "face";"pile")`
Simuler une famille de trois enfants (chiffre pair = garçon, chiffre impair = fille)	Casio : `INT(1000*RAN#)` TI : `INT(1000*RAND)`	`=ENT(1000*ALEA())`
Somme de deux dés	Casio : `INT(6*RAN#+1) +INT(6*RAN#+1)` TI : `INT(6*RAND+1) +INT(6*RAND+1)`	`=ENT(6*ALEA()+1) +ENT(6*ALEA()+1)`

2. Échantillonage

Définition 15.1 Lorsqu'on « prélève » n individus d'une population, on dit que ces n individus constituent *un échantillon* de la population. L'entier n est appelé la taille de l'échantillon.

Propriété 15.1 Plus la taille d'un échantillon est grande plus la fréquence d'un caractère dans l'échantillon est proche de celle de la population totale.

Exemple Pour un sondage « à la sortie des urnes », plus on interroge de personnes sur leur vote, plus la fréquence des personnes ayant voté pour le candidat A dans l'échantillon sera proche de la fréquence des personnes ayant choisi ce candidat dans la population totale [3].

Théorème 15.1 (de l'intervalle de fluctuation) On considère une population sur laquelle on étudie la proportion p d'individus ayant un caractère C avec $p \in [0,2;0,8]$ et dans laquelle on prélève un échantillon de taille $n \geq 25$.
Dans ces conditions, dans 95 % des échantillons, la proportion p' des individus de l'échantillon ayant le caractère C est comprise dans l'intervalle $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$.

Conséquence :
On considère une population sur laquelle on étudie la proportion p d'individus ayant un caractère C et dans laquelle on prélève un échantillon de taille $n \geq 25$. Si dans l'échantillon, la proportion p' d'individus ayant le caractère C vérifie $p' \in [0,2;0,8]$ alors dans 95 % des cas on peut dire que $p \in \left[p' - \frac{1}{\sqrt{n}}; p' + \frac{1}{\sqrt{n}}\right]$.

3. Il faut tout de même que les personnes interrogées soient choisies au hasard dans la population totale.