Statistiques : deuxième partie

1 Simulation, échantillonage

1.1 Simulation d'un tirage dans une urne

Dans une urne, on a placé 10 boules indiscernables au toucher : 7 sont rouges et 3 sont vertes.
On tire au hasard une boule de l'urne, on note sa couleur puis on la remet dans l'urne.
Le but de cette activité est de réaliser une simulation de 100 tirages, puis de 500 tirages à l'aide d'un tableur.

1.1.1 Simulation de 100 tirages

Pour réaliser cette simulation à l'aide d'un tableur, numérotons les boules rouges de 1 à 7 et les boules vertes de 8 à 10.
Ouvrir un tableur et compléter la ligne 1 afin d'obtenir l'écran suivant :

	A	B	C	D
1	Tirages		Fréquence de "rouge"	Fréquence de "vert"
2				

Tapez dans la cellule A2 l'instruction : =ALEA.ENTRE.BORNES(1 ;10) [1]
Recopiez la cellule A2 vers le bas jusqu'à la cellule A101.
Vous obtenez 100 résultats qui correspondent aux 100 tirages.
Saissisez maintenant les instructions suivantes :
- dans C2 : =NB.SI(A2 :A101 ;"<=7")/100 [2]
- dans D2 : =NB.SI(A2 :A101 ;">=8")/100

Pour obtenir une nouvelle simulation, appuyez sur la touche **F9**. Après plusieurs simulations, que remarque-t-on concernant les fréquences d'apparition du "rouge" et du "vert" ?

1.1.2 Simulation de 500 tirages

En vous inspirant de ce qui précède, réalisez plusieurs simulations de 500 tirages.
Que remarque-t-on concernant les fréquences d'apparition du "rouge" et du "vert" ?

▶ Lorsque l'on répète plusieurs fois la même expérience aléatoire, **la distribution des fréquences varie**. On parle alors de **fluctuation d'échantillonage**.

1. Cette instruction donne un nombre aléatoire entier compris entre 1 et 10.
2. Cette instuction vous affiche la fréquence de "rouge"

1.2 Simulation du lancer d'un dé à 4 faces

On souhaite simuler 10000 fois le lancer d'un dé tétraédrique, dont les faces sont numérotées de 1 à 4.
Quelle formule faut-il entrer en B2 pour obtenir un nombre aléatoire compris entre 1 et 4 ? Recopier cette formule jusqu'à la cellule B10001.

Quelle formule faut-il entrer en C2 pour obtenir, après recopie de celle-ci jusqu'à la cellule C10001, la fréquence d'apparition du 4 ?

	A	B	C
1	n° du lancer	Résultat du lancer	Fréquence d'apparition du 4
2			

Sélectionnez la plage de cellules C2 :C10001, puis insérez le nuage de points correspondant.
Que remarque-t-on ?

▶ **Si l'on étudie des échantillons de tailles de plus en plus grandes, l'amplitude de la fluctuation diminue et les fréquences observées se stabilisent.**
Cela permet par exemple d'estimer une proportion inconnue p dans une population en étudiant des échantillons de taille suffisante.

2 Intervalle de fluctuation, intervalle de confiance

2.1 Intervalle de fluctuation

Au sein d'une population, on connaît la **proportion** p des individus ayant un caractère donné.
Parmi les échantillons de taille n extraits de cette population, la fréquence d'apparition f du caractère varie avec l'échantillon prélevé.

Si $0{,}2 \leq p \leq 0{,}8$ et si $n \geq 25$, on admet que la fréquence d'apparition f observée appartient à **l'intervalle** $[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}]$ avec une probabilité d'au moins 0,95.

▶ Cet intervalle est appelé **intervalle de fluctuation au seuil de 95%**.

L'intervalle de fluctuation permet d'étudier un échantillon donné.
Si la fréquence observée appartient à cet intervalle, on ne rejette pas cet échantillon, sinon on le rejette.
Attention, cela signifie que dans environ 5% des cas, la décision prise risque d'être incorrecte.

Exemple
Un sac de graines de fleurs doit continir 23% de graines de tulipes. On choisit un échantillon de 100 graines et on compte 17% de graines de tulipes.
Doit-on en conclure que la proportion dans le sac n'est pas bonne ?
L'intervalle de fluctuation est $[0,23 - \frac{1}{\sqrt{100}}; 0,23 + \frac{1}{\sqrt{100}}]$, soit [0,13 ;0,33].
On ne peut donc pas conclure que la proportion annoncée ne soit pas bonne.

2.2 Intervalle de confiance

Au sein d'une population, on désire estimer la **proportion inconnue** p des individus ayant un caractère donné.
On étudie un échantillon de taille n, $n \geq 25$, extrait de cette population. Si la fréquence d'apparition f du caractère étudié est comprise entre 0,2 et 0,8 alors on peut estimer que la proportion p du caractère dans la population totale appartient à **l'intervalle** $[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}]$ avec une probabilité d'au moins 0,95.

▶ Cet intervalle est appelé **intervalle de confiance au seuil de 95%**.

Attention, dans environ 5% des cas, l'intervalle de confiance ne contient peut-être pas p.

Exemple
Lors d'une élection présidentielle, un sondage réalisé sur un échantillon de 1000 personnes, donnait à un candidat A 24% des intentions de votes et 20% à un candidat B.
Finalement, le vote a donné les résultats suivants : 21,18% pour le candidat A et 21,86% pour le candidat B.

Le sondage réalisé rendait-il plausible ces résultats ?

L'intervalle de confiance pour le candidat A est $[0,24 - \frac{1}{\sqrt{1000}}; 0,24 + \frac{1}{\sqrt{1000}}]$, soit [0,208 ; 0,0,272].

L'intervalle de confiance pour le candidat B est $[0,20 - \frac{1}{\sqrt{1000}}; 0,20 + \frac{1}{\sqrt{1000}}]$, soit [0,168 ; 0,232].

Les scores effectifs des deux candidats appartiennent bien à ces intevalles, ces résultats étaient envisageables.