Travaux pratiques n°1

La côte de popularité (Hachette-Déclic 2nde - 2000)

On fait deux sondages sur 100 personnes choisies au hasard en France :

- le premier, au mois de septembre, donne 56 % d'opinions favorables pour un personnage politique important ;
- le second, au mois de décembre, lui donne 58 % d'opinions favorables.

Peut-on en conclure de façon certaine que la côte de popularité a augmenté de 2 points de pourcentage entre septembre et décembre ?

D'ailleurs, est-on certain qu'elle a augmenté ?

Pour le savoir, on simule plusieurs fois un sondage à l'aide de la calculatrice.

On suppose qu'il y a 55 % d'avis favorables dans la population totale, et que ce taux reste **parfaitement constant** au cours du temps.

Le programme donne le pourcentage d'opinions favorables d'un échantillon de taille T, avec un pourcentage réel de 55 %.

TI 80, 82, 83	Casio GRAPH
: 0 → N	0 → N ↵
: Input «T=» , T	«T=?» → T ↵
: For (K, 1, T)	For 1 → K To T ↵
: int (rand + 0.55) → A	Int (Ran#+0.55) → A ↵
: A + N → N	A + N → N ↵
: End	Next ↵
: Disp N / T × 100	N / T × 100 ◢

1° Exécuter ce programme 20 fois en choisissant une taille d'échantillon égale à 100 et en notant à chaque fois le pourcentage obtenu.

a) Quelle est l'étendue des valeurs ainsi obtenues ?

b) Combien de résultats sont au-dessus du pourcentage réel ? en dessous du pourcentage ?

c) Combien de résultats diffèrent de plus de 2 points du pourcentage réel ? de 5 points ? Quel est l'écart moyen ?

2° Recommencer avec des échantillons de 500. Commenter.

Prolongement :

Si le pourcentage réel d'opinions favorables a baissé de 2 points (53 % au lieu de 55 %), un sondage de 100 personnes permet-il de le révéler ?

On pourra exécuter le programme après l'avoir modifié, en remplaçant 0.55 par 0.53.

Travaux pratiques n°2

Simulations de sondages (Nathan – Transmath 2nde – 2000)

Les sondages jouent un rôle important dans notre société. Il en est souvent question pour indiquer la cote de popularité d'hommes politiques, mais, en réalité, ils interviennent dans bien d'autres domaines, moins médiatiques, mais importants.

1 De quoi s'agit-il ?

On peut en résumer le principe de la manière suivante : dans une population contenant un « grand nombre » N d'individus, un pourcentage p de ces individus possède une certaine propriété $\mathcal{P}$, par exemple « ont l'intention de voter pour tel candidat ». Pour trouver p, on pourrait, **en théorie**, examiner **tous** les individus et compter ceux qui ont la propriété $\mathcal{P}$. Connaissant le nombre total d'individus, on en déduirait alors immédiatement le pourcentage p cherché.

2 Estimation

Dans la pratique, il est très difficile d'examiner tous les individus : cela nécessiterait trop de temps, trop d'argent... Parfois même, l'examen exhaustif de tous les individus donnerait p, certes, mais la connaissance de p ne servirait alors à rien !

Ainsi, par exemple, imaginons que l'on veuille connaître, dans un lot de 10 000 ampoules, avant de les mettre en vente, le pourcentage p d'ampoules ayant une durée de vie supérieure à 1 500 heures. En allumant sans interruption chacune des 10 000 ampoules, on verrait quelle est la durée de vie de chacune d'elles, et on en déduirait p ; mais alors, toutes les ampoules seraient grillées, et les calculs effectués relatifs à ce lot seraient inutiles !

Dans la pratique, on ne considère donc qu'une partie de la population, contenant un nombre n d'élements, n étant nettement inférieur au nombre total N d'éléments. On dit qu'une telle partie est un **échantillon de taille n**. On peut calculer alors, dans cet échantillon, le pourcentage f d'individus ayant la proprieté $\mathcal{P}$.

Intuitivement, si la taille de l'échantillon n'est pas « trop petite », on peut alors dire que le vrai pourcentage p, relatif à toute la population, devrait être « voisin de f ». On dit que f est une **estimation ponctuelle de p**. On peut concevoir qu'une telle estimation n'est pas très satisfaisante. En effet, si, par « malchance », l'échantillon choisi à des caractéristiques très distinctes de celles de la population, p et f peuvent ne pas être voisins.

La théorie des probabilités et des statistiques permet d'obtenir, grâce à des calculs qui ne sont pas au programme de la classe de Seconde, une estimation plus satisfaisante, faisant intervenir, cette fois, un intervalle de centre f susceptible de contenir p.

En voici des exemples.

3 Fourchettes de sondage

Reprenons les notations précédentes :

- n est la taille de l'échantillon ;
- p désigne la fréquence des individus, dans la population totale, ayant la propriété $\mathcal{P}$;
- f la fréquence des individus de l'échantillon ayant la propriété $\mathcal{P}$.

On suppose que $n > 30$ et que n est « très petit » par rapport à l'effectif total de la population.

On suppose enfin que f est compris entre 0,3 et 0,7.

On peut alors démontrer les propriétés suivantes :

> Il y a environ 95 chances sur 100 pour que l'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$ contienne p.

On dit que l'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$ est la **fourchette de sondage de p** (on dit parfois **« l'intervalle de confiance de p »**) **au niveau de confiance 0,95.** On dit aussi **« avec un risque de 5 % »**.

Dire que p appartient à l'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$, revient à dire que $|f - p| \leqslant \frac{1}{\sqrt{n}}$.

Lorsqu'on estime p par f au niveau de confiance 0,95, on dit que la précision est $\frac{1}{\sqrt{n}}$.

- La fourchette de sondage de p au niveau de confiance 0,90 se définit de manière analogue : c'est l'intervalle $\left[f - \frac{0,82}{\sqrt{n}} ; f + \frac{0,82}{\sqrt{n}}\right]$.

> Il y a environ 90 chances sur 100 pour que l'intervalle $\left[f - \frac{0,82}{\sqrt{n}} ; f + \frac{0,82}{\sqrt{n}}\right]$ contienne p.

• La fourchette de sondage de p au niveau de confiance 0,99 est l'intervalle $\left[f - \frac{1,5}{\sqrt{n}} ; f + \frac{1,5}{\sqrt{n}}\right]$.

> Il y a environ 99 chances sur 100 pour que l'intervalle $\left[f - \frac{1,5}{\sqrt{n}} ; f + \frac{1,5}{\sqrt{n}}\right]$ contienne p.

4 Pour mieux comprendre

1. On suppose que $f = 0,6$ et $n = 100$.

a) Représentez sur un axe la fourchette de sondage de p au niveau de confiance 0,95. (Prendre pour unité 10 cm.) On note I cet intervalle.

b) Sur le même axe, représentez la fourchette de sondage I_1 de p au niveau de confiance 0,90, puis la fourchette de sondage I_2 de p au niveau de confiance 0,99. (Utilisez des couleurs différentes.)

c) Sur le graphique, on peut constater que $I_1 \subset I \subset I_2$. Expliquez pourquoi, sans calculs, il est intuitivement prévisible que l'intervalle I_1 soit contenu dans l'intervalle I, et que I soit contenu dans I_2.

2. On conçoit que, lorsque la taille de l'échantillon augmente, la précision de l'estimation est meilleure. Expliquez pourquoi ceci est en accord avec les propriétés des fourchettes de sondage indiquées dans le paragraphe 3.

5 Application

Un candidat à une élection fait effectuer un sondage. Sur 1 000 personnes interrogées au hasard, 530 déclarent vouloir voter pour lui. Le nombre d'électeurs est égal à 95 438 (donc la taille de l'échantillon est nettement inférieure à l'effectif total). On supposera que les électeurs ne changent pas d'avis le jour du vote. On note p la proportion des votes favorables au candidat.

1. Quel est l'intervalle de confiance de p au niveau 0,95 ?

2. Complétez la phrase suivante :
« Il y a environ 95 chances sur 100 pour que le pourcentage des votes favorables au candidat soit compris entre ... % et ... %. »

6 Une simulation à présent

On suppose que, dans une urne contenant un « grand nombre de boules », il y a 60 % de boules rouges. Donc, ici, $p = 0,6$. On prélève au hasard un échantillon de 100 boules et on note f la fréquence de boules rouges tirées ; f est donc la fréquence des boules rouges dans un échantillon de 100 individus (ici, $n = 100$).

On sait que, dans ces conditions, la fourchette de sondage de p au niveau de confiance 0,95 est, approximativement, l'intervalle $\left[f - \frac{1}{\sqrt{100}} ; f + \frac{1}{\sqrt{100}}\right]$, c'est-à-dire :

$$[f - 0,1 ; f + 0,1].$$

On peut simuler cette situation à l'aide de chiffres au hasard. Voici un programme simulant l'étude d'échantillons de taille 100, et indiquant, pour chacun d'eux, la fréquence correspondante.

Commentaires	TI	Casio (avec « For »)	Casio (sans « For »)
Début du programme	Lbl 0	Lbl 0	Lbl 0
Mise à zéro de la mémoire S comptant le nombre de boules rouges	0 → S	0 → S	0 → S
Boucle : calcul de S, au fur et à mesure, pour les 100 tirages successifs*	For (I, 1, 100) If rand < 0,6 Then S + 1 → S End End	For 1 → I To 100 If Ran# < 0,6 Then S + 1 → S Ifend Next	0 → I Lbl 1 I + 1 → I Ran# < 0,6 ⇒ S + 1 → S I < 100 ⇒ Goto 1
Calcul de la fréquence	S ÷ 100 → F	S ÷ 100 → F	S ÷ 100 → F
Affichage de la fréquence	Disp F Pause	F ◢	F ◢
Retour au début du programme	Goto 0	Goto 0	Goto 0

Pour les TI 89 et 92, « rand » est remplacée par « rand( ) » ; « for(I, 1, 100) » est remplacé par « For I, 1, 100 ».

*** Remarque importante :**
Prélever un échantillon de 100 boules, revient à effectuer 100 tirages successifs **sans remise.** Or la taille n de l'échantillon (ici $n = 100$) est nettement inférieure au nombre total de boules de l'urne. On conçoit donc que le fait d'enlever quelques boules dans l'urne ne modifie pas la proportion p de boules rouges. **On peut donc assimiler un tirage sans remise et un tirage avec remise**. C'est ce qui est fait dans ce programme. (Dans la démonstration des propriétés ci-dessus relatives aux fourchettes de sondage, il s'agit également de tirages avec remise.)

1. Effectuez, par simulation, 50 prélèvements (de 100 boules chacun), et, pour chacun d'eux, indiquez la fourchette de sondage de p au niveau de confiance 0,95.

2. Vous devriez pouvoir constater que 95 % environ des 50 intervalles ainsi obtenus contiennent le nombre $p = 0,6$. Est-ce le cas ?

Remarque :

Dire que p appartient à l'intervalle $[f - 0{,}1 \; ; \; f + 0{,}1]$ équivaut à dire que $|p - f| \leqslant 0{,}1$, c'est-à-dire $|f - p| \leqslant 0{,}1$, c'est-à-dire f appartient à $[p - 0{,}1 \; ; \; p + 0{,}1]$, c'est-à-dire, puisque $p = 0{,}6$: f appartient $[0{,}5 \; ; 0{,}7]$.

La question 2 ci-dessous peut donc également se formuler ainsi :

« 95 % environ des 50 valeurs de f, c'est-à-dire 47 valeurs environ, devraient appartenir à l'intervalle $[0{,}5 \; ; 0{,}7]$ »

COMMENTAIRE : Le tableur Excel peut permettre de visualiser directement sur l'écran les intervalles de confiance, et donc de voir immédiatement ceux qui contiennent le nombre 0,6 et ceux qui ne le contiennent pas.

Voici le type de graphique, relatif à ce programme, que l'on peut obtenir. (Source : « Simulations d'expériences aléatoires », *Irem Paris Nord*.)