FOURCHETTES DE SONDAGE EN CLASSE DE SECONDE

Dans le programme de seconde nous avons à traiter un certain nombre de thèmes d'étude. Un des thèmes proposés est le suivant :

Simulations d'un sondage ; à l'issue de nombreuses simulations, pour des échantillons de taille variable, on pourra introduire la notion de fourchette de sondage, sans justification théorique. La notion de niveau de confiance 0,95 de la fourchette peut être introduite en terme de chances (il y a 95 chances sur 100 pour que la fourchette contienne la proportion que l'on cherche à estimer); on pourra utiliser les formules des fourchettes aux niveaux 0,95 ; 0,90 et 0,99 pour une proportion observée voisine de 0,5 afin de voir qu'on perd en précision ce qu'on gagne au niveau de confiance. On incitera les élèves à connaître l'approximation usuelle de la fourchette au niveau de confiance 0,95 ; issue d'un sondage sur n individus (n > 30) dans le cas où la proportion observée f est comprise entre 0,3 et 0,7 ; à savoir :$[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}]$*.

Dans une première partie nous proposerons des exercices que l'on peut effectuer avec les élèves et dans une deuxième partie nous allons montrer pourquoi on utilise cette fourchette (*).
La deuxième partie n'est pas à effectuer avec les élèves.

<u>Première partie</u>

Après avoir effectué quelques exercices concernant les simulations on pourra passer aux exercices suivants :

<u>Exercice n°1</u> :

Les élèves pourront rechercher des sondages qui paraissent dans la presse où il est indiqué le nom de l'institut, la méthode retenue et le mécanisme qui a permis de sélectionner l'échantillon.
Ce sera l'occasion de leur demander :

- de faire des recherches sur les principaux instituts de sondage en France (SOFRES, IPSOS, BVA, IFOP, CSA, BVA). Ce que signifie tous ces sigles, les dates de fondation de ces instituts et ce que font ces instituts à part du sondage d'opinion ;

- pourquoi appelle-t-on les sociétés qui produisent des études d'opinion des instituts ?

On peut également leur faire rechercher la définition de mots qui sont souvent employés dans les enquêtes par sondages : qu'est-ce qu'une enquête, un recensement, un sondage, une base de sondage, un taux de sondage ? Qu'est-ce que par exemple un panel, une étude omnibus ?

<u>Exercice n°2</u> :

Le journal du dimanche publiait le 8 décembre 2002 le sondage suivant réalisé par IFOP.
De manière générale, dites-moi si vous faites tout à fait confiance, plutôt confiance, plutôt pas confiance ou pas confiance du tout aux syndicats ?

	Ensemble de la population (en %)	Ensemble des salariés (en %)	Salariés du privé (en %)	Salariés du public (en %)
Tout à fait confiance	7	7	8	5
Plutôt confiance	49	51	51	51
Total confiance	**56**	**58**	**59**	**56**
Plutôt pas confiance	24	24	20	30
Pas confiance du tout	18	17	20	13
Total pas confiance	**42**	**41**	**40**	**43**
Ne se prononcent pas	2	1	1	1

Sondage réalisé auprès de 953 personnes âgées de 18 ans et plus. La représentativité a été assurée par la méthode des quotas (sexe, âge, profession du chef de famille) après stratification par région et catégorie d'agglomération. Les interviews ont eu lieu par téléphone au domicile des personnes interrogées du 5 au 6 décembre 2002.

1) Quelles explications peut-on donner au pourcentage si peu élevé de "Ne se prononcent pas" ?

2) 8 % des salariés du privé et 5 % des salariés du public ont tout à fait confiance. Comment expliquez-vous que pour l'ensemble des salariés 7 % d'entre eux ont tout à fait confiance ?

3) 17 % de l'ensemble des salariés n'ont pas confiance du tout. Comment expliquez-vous que pour l'ensemble 18 % n'ont pas confiance du tout ?

Pour les deux exercices suivants il s'agit d'appliquer les formules données dans le programme de seconde des différentes fourchettes de sondage aux niveaux de confiance 0,90 ; 0,95 et 0,99. Dans les deux exercices qui suivent, p représente une proportion inconnue dans une population et f une proportion connue, obtenue à partir d'un échantillon aléatoire prélevé dans cette population.

Exercice n°3 :

Déterminer une fourchette de sondage de p lorsque f = 0,6 et n = 1 024 avec un niveau de confiance de 0,90 puis 0,95 et enfin 0,99. Que constate-t-on ?

Exercice n°4 : Dans cet exercice on supposera que $0,3 \leq f \leq 0,7$

1) On désire une fourchette de p au niveau de confiance de 0,99 avec une marge d'erreur de 3 %. Quelle est la taille minimale de l'échantillon nécessaire afin d'obtenir ce résultat ?

2) On souhaite à partir d'un échantillon de taille 225 une fourchette de p avec une marge d'erreur de 5,5 %. Quel est le niveau de confiance du sondage ?

Deuxième partie

Supposons qu'un candidat A se présente à une élection. Il souhaite connaître la proportion p d'électeurs prêts à voter pour lui. Un sondage est effectué auprès de 1 000 personnes (n = 1 000). 60 % des personnes interrogées sont prêtes à voter pour lui. L'institut de sondage affirme que ce résultat est entaché d'une marge d'erreur qui est de 3 % au niveau de confiance de 0,95.

Il est à noter qu'actuellement la loi oblige les journaux publiant des résultats de sondage à indiquer le nom de l'institut l'ayant effectué, la date de l'enquête, le nombre de personnes interrogées, le mécanisme qui a permis de les sélectionner et la méthode retenue.

Mais en aucun cas le journal est obligé d'ajouter la marge d'erreur. Ainsi un sondage sur les Français et la décentralisation, publié dans le journal "Le Télégramme de Brest" du 18/10/02, stipulait que le sondage BVA avait été réalisé auprès de 1 067 personnes, âgées de 18 ans et plus, interrogées du 4 au 7 octobre 2002 en face à face, selon la méthode des quotas.
Il serait de toute manière illusoire d'indiquer cette marge d'erreur avec la méthode des quotas employés en France pour sélectionner les individus. Car pour calculer la marge d'erreur, il faut utiliser un mode d'échantillonnage aléatoire (chaque individu doit avoir une probabilité connue a priori d'appartenir à l'échantillon), ce qui n'est pas le cas avec la méthode des quotas.

Revenons à notre résultat du sondage. La proportion p est inconnue (et à jamais inconnue) mais celle-ci, en supposant qu'elle conserve une valeur constante pendant la durée de réalisation du sondage, est une valeur fixe et par suite appartient ou n'appartient pas à la fourchette construite à partir des observations. Que faut-il alors comprendre ?
Considérons une population dans laquelle il y a une proportion p d'individus qui sont prêts à voter pour le candidat A, on va chercher à déterminer une fourchette de p.

On va prélever dans la population un échantillon aléatoire de taille n. Notons X la variable aléatoire qui, à chaque échantillon de taille n, associe le nombre d'individus de cet échantillon qui se déclarent prêts à voter pour le candidat A.

X est distribuée selon une loi binomiale (ou une loi qui peut être approchée par la loi binomiale[1] de paramètres n et p) d'espérance $E(X) = np$ et de variance $V(X) = np(1-p)$.

Si, de plus, lorsque n est grand et p ni trop proche de 0 ni de 1 (ce qui est souvent le cas car dans une population il est rare qu'un homme politique soit complètement désapprouvé par la population ou qu'il fasse l'unanimité)[2] on peut approcher la loi de probabilité de X par la loi normale de moyenne n p et d'écart type $\sqrt{np(1-p)}$.

Posons $F = \frac{X}{n}$, F est la variable aléatoire qui désigne la proportion d'individus de l'échantillon prêts à voter pour A, la loi de probabilité de F peut être approchée par la loi normale d'espérance p et d'écart type $\sqrt{\frac{p(1-p)}{n}}$ donc la variable $U = \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}}$ est distribuée suivant la loi normale centrée et réduite[3].

[1] Si le prélèvement est effectué successivement et sans remise, alors X suit une loi hypergéométrique qui peut être approchée par une loi binomiale sous l'hypothèse que la taille de n est faible face à la taille N de la population et que N est suffisamment grand (par exemple N > 10 n et N > 100).
[2] On peut se poser la question suivante : comment savoir que p (qui est inconnu) n'est pas trop proche de 0 ou de 1 ? On peut s'en tirer en disant que même lorsqu'on ignore la proportion exacte dans la population on en a malgré tout une vague idée.
[3] Pour simplifier, on négligera la correction de continuité qu'il faudrait appliquer du fait que l'on passe d'une variable aléatoire discrète à une variable aléatoire continue.

Soit $\alpha \in]\,0\,;\,1\,[$ et on note $u_{1-\alpha}$ le nombre tel que $P(U \leq u_{1-\alpha}) = 1 - \alpha$.

Donc $P(-u_{1-\frac{\alpha}{2}} \leq U \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha \Leftrightarrow P\left(F - u_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq p \leq F + u_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$.

On obtient ainsi l'intervalle aléatoire de p symétrique en probabilité au niveau de confiance 1 - α qui est :

$$\left[F - u_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \;;\; F + u_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \right] \subset \left[F - u_{1-\frac{\alpha}{2}}\frac{1}{2\sqrt{n}} \;;\; F + u_{1-\frac{\alpha}{2}}\frac{1}{2\sqrt{n}} \right]$$

car $p(1-p) \leq 0{,}5$ pour $p \in [\,0\,;\,1\,]$

donc $P\left(\left[F - u_{1-\alpha/2}\frac{1}{2\sqrt{n}} \;;\; F - u_{1-\alpha/2}\frac{1}{2\sqrt{n}} \right]\right) \geq 1-\alpha$ (1)

Donc la fourchette de sondage de p avec un niveau de confiance supérieur ou égal à $1 - \alpha$ est :

$$\left[F - u_{1-\alpha/2}\frac{1}{2\sqrt{n}} \;;\; F + u_{1-\alpha/2}\frac{1}{2\sqrt{n}} \right]$$

Ainsi on peut montrer que la fourchette de sondage de p :

- au niveau de confiance supérieur à 0,90 est $\left[F - \frac{1{,}65}{2\sqrt{n}} \;;\; F + \frac{1{,}65}{2\sqrt{n}} \right] \subset \left[F - \frac{0{,}83}{\sqrt{n}} \;;\; F + \frac{0{,}83}{\sqrt{n}} \right]$;
- au niveau de confiance supérieur à 0,95 est $\left[F - \frac{1{,}96}{2\sqrt{n}} \;;\; F + \frac{1{,}96}{2\sqrt{n}} \right] \subset \left[F - \frac{1}{\sqrt{n}} \;;\; F + \frac{1}{\sqrt{n}} \right]$;
- au niveau de confiance supérieur à 0,99 est $\left[F - \frac{2{,}58}{2\sqrt{n}} \;;\; F + \frac{2{,}58}{2\sqrt{n}} \right] \subset \left[F - \frac{1{,}30}{\sqrt{n}} \;;\; F + \frac{1{,}30}{\sqrt{n}} \right]$.

<u>Remarques</u> :

1) Si on prend $1-\alpha = 0{,}95$ on a donc $1-\frac{\alpha}{2} = 0{,}975$ et on en déduit que :
   $u_{1-\frac{\alpha}{2}} = u_{0{,}975}$ soit 1,96 (d'après la table de la fonction de répartition de la loi normale centrée réduite). De la même façon nous déterminons les deux coefficients 1,65 et 2,58.

2) Dans l'énoncé du programme on considère que le niveau de confiance est de 0,95 et non pas supérieur ou égale à 0,95, voir l'inégalité (1).

3) Au niveau de confiance 0,95, la fourchette de sondage est $\left[F - \frac{1}{\sqrt{n}} \;;\; F + \frac{1}{\sqrt{n}} \right]$ (pour $n = 1000$ $\frac{1}{\sqrt{n}} = 0{,}03$ à 10^{-3} près), soit 3% exprimé en pourcentage.

Le résultat du sondage nous indiquait qu'il y avait 60 % des personnes interrogées prêtes à voter pour le candidat A. L'erreur commise est au maximum de 3 % si on estime p par 0,6.
On dit que la marge d'erreur est de 3 %.
Le niveau de confiance 0,95 signifie que si on effectue des paquets de 100 sondages il y aura en moyenne 95 fourchettes qui contiendront la valeur p et 5 en moyenne ne contiendront pas p.

Vous pourrez vous référer au document d'accompagnement de la classe de seconde p 49 pour visualiser ce que signifie au niveau de confiance 0,95.