Activité échantillonnage, intervalle de fluctuation, prise de décision *(à partir d'un même thème)*

Les trois activités qui suivent s'inspirent du document [« ressources pour la classe de première générale et technologique, statistiques et probabilités »] page 32.

Ces activités poursuivent les objectifs suivants :

- Montrer qu'un même thème peut être exploité de plusieurs points de vue, en seconde et en première : échantillonnage, simulation d'un sondage, mise en évidence de l'intervalle de fluctuation, prise de décision.
- Donner des exemples d'exercices permettant de traiter à la fois le programme de probabilités et statistiques, l'entraînement à l'algorithmique et l'utilisation d'un tableur.

Activité 1 : échantillonnage

Niveau :

Classe de seconde.

Extrait du programme officiel :

Échantillonnage Notion d'échantillon. Intervalle de fluctuation d'une fréquence au seuil de 95%*. Réalisation d'une simulation.	• Concevoir, mettre en œuvre et exploiter des simulations de situations concrètes à l'aide du tableur ou d'une calculatrice. • Exploiter et faire une analyse critique d'un résultat d'échantillonnage.	Un échantillon de taille n est constitué des résultats de n répétitions indépendantes de la même expérience. À l'occasion de la mise en place d'une simulation, on peut : • utiliser les fonctions logiques d'un tableur ou d'une calculatrice, ◇ mettre en place des instructions conditionnelles dans un algorithme. L'objectif est d'amener les élèves à un questionnement lors des activités suivantes : • l'estimation d'une proportion inconnue à partir d'un échantillon ; • la prise de décision à partir d'un échantillon.

Type d'activité :

Recherche en salle informatique (1 heure) et bilan fait en devoir à la maison.

Pré requis :

Sur le plan statistique, connaître les notions de base de statistiques (effectif et fréquence) et avoir vu sur quelques exemples préalables la notion de sondage et d'échantillon.
Sur le plan informatique, l'activité s'adresse à des élèves ayant déjà un peu manipulé un tableur mais qui ne sont pas forcément familiers avec les outils de simulation sur un tableur (c'est pourquoi l'activité est très guidée).

Motivation de l'activité, objectifs et compétences travaillées :

- Faire découvrir sur un exemple la notion de fluctuation d'échantillonnage.
- Initier les élèves à la simulation d'un sondage avec un tableur.
- Utiliser un graphique fait à l'aide d'un tableur pour interpréter des résultats et conjecturer la notion d'intervalle de fluctuation.
- Faciliter la prise de note des élèves en activité informatique.

Compte rendu d'expérience :

Cette activité a été testée durant l'année 2013-2014. Le professeur a gardé et analysé les quatorze bilans rendus par un groupe d'élèves. On trouvera ce bilan à la suite de l'énoncé de l'activité.

Texte de l'activité élève

On désire simuler un sondage effectué dans une population d'électeurs. On admet qu'on connaît la proportion p d'électeurs votant pour le candidat A (il n'y a que deux candidats A et B ; tous les électeurs votent soit pour A, soit pour B).
On prend pour hypothèse que $p = 0,52$.
On sonde 1000 personnes.

1. Simulation d'un sondage

- On décide arbitrairement qu'un électeur ayant voté pour le candidat A est repéré par un nombre décimal appartenant à l'intervalle $[1;1,52[$ et qu'un candidat ayant voté pour le candidat B est repéré par un nombre décimal appartenant à l'intervalle $[0,52;1[$.

  Quelle formule faut-il entrer dans la cellule A1 pour simuler le choix de vote d'un électeur tiré au hasard dans la population ? ____
- Simuler 1000 choix de vote d'électeurs tirés au hasard dans la population.
- Entrer dans la cellule A1001 une formule permettant de compter le nombre d'électeurs ayant voté pour le candidat A.
- Entrer dans la cellule A1002 une formule permettant de calculer la fréquence d'électeurs ayant voté pour le candidat A.
- Relancer une dizaine de fois la simulation, noter les résultats affichés dans la cellule A1002 et vérifier qu'ils sont cohérents : ____

  ____
- Déterminer le plus petit intervalle auquel appartiennent les dix résultats obtenus : _______________

2. Simulation de 100 sondages

- Sélectionner les 1002 premières cellules de la colonne A et recopier-les 100 fois vers la droite (jusqu'à la colonne CV).
- Sélectionner les 100 données de la ligne 1002 et afficher le nuage de point correspondant ; en cliquant sur l'axe des ordonnées, choisir pour minimum $0,44$, pour maximum $0,6$ et pour graduation $0,01$.
- Donner un intervalle centré sur $0,52$, le plus petit possible, qui contienne au moins 95% des résultats : ____
- Compter le nombre de valeurs extérieures à l'intervalle $[0,49;0,55]$: ____________________
- Relancer la simulation une dizaine de fois en notant à chaque fois le nombre de valeurs extérieures à l'intervalle $[0,49;0,55]$: ____

  ____

  ____

3. Travail écrit à rendre

En une dizaine de ligne environ, faire un bilan de l'aspect statistique de cette activité. Ce bilan ne devra pas décrire les aspects techniques liés à l'utilisation du tableur.

Activité 1 compte rendu d'expérience

Cette activité a été expérimentée dans une classe de seconde de niveau moyen au mois de mai 2014.
Les copies d'un groupe de 14 élèves ont été gardées et analysées.

Deux élèves ne font que décrire les opérations à effectuer et en quelque sorte se contentent de réécrire l'énoncé.

Parmi les 12 autres :

Seuls 2 élèves parlent d'échantillons de taille 1000. Les autres se contentent de parler d'un « sondage de 1000 personnes » qui est répété un certain nombre de fois. Il n'est pas très étonnant que le mot échantillon n'apparaisse pas dans les copies puisqu'il n'est pas présent dans l'énoncé.

Un premier groupe de 6 élèves parle d'intervalle de fluctuation mais en a une idée approximative. Ils font référence à un intervalle de fluctuation avec des formulations peu claires comme par exemple
« *Nous avons répété l'opération 100 fois afin d'établir un intervalle de fluctuation. Nous avons conclu que la fluctuation est proche de 0.52* »
« Nous avons cherché un intervalle qui contient 95% des résultats » sans que ce qui précède permette de déterminer de quels résultats il s'agit.
« *L'intervalle des 10 résultats se rapproche de l'alea* »

Un deuxième groupe de 6 élèves a compris que la fréquence n'était pas la même dans chaque échantillon, que les valeurs des différentes fréquences restaient proches de 052 et qu'il existait un intervalle centré sur 0,52 contenant au moins 95% des résultats.
Dans ces copies, on peut voir dans la rédaction que l'élève a compris la différence entre ce qui se passe dans la population globale (proportion de 0.52 connue) et les différentes fréquences calculées dans les échantillons alors que ce n'est pas clair dans les copies du premier groupe.
Dans ce groupe on notera que 2 élèves s'interrogent sur le nombre de valeurs extérieures à l'intervalle [0,49 ; 0,55]
Ils constatent que *lors de la succession de plusieurs simulations, les valeurs extérieures sont parfois plus de 5.*
Ils expliquent alors ce phénomène par *l'imprécision des fréquences beaucoup trop arrondies*
Autrement dit, le fait que des valeurs soient extérieures à l'intervalle [0,49 ; 0,55], n'est pas expliqué avec un raisonnement probabiliste ;
A ce niveau, et après ce travail exploratoire, il restera donc au professeur à expliquer que **l'intervalle de fluctuation au seuil de 95% relatif aux échantillons de taille n,** est l'intervalle centré autour de p où se situe avec une probabilité de 0.95 la fréquence observée dans un échantillon de taille n.

Activité 2 : Découverte de l'intervalle de fluctuation $\left[p-\frac{1}{\sqrt{n}}\,;p+\frac{1}{\sqrt{n}}\right]$ en classe de seconde

Niveau :

Classe de seconde

Extrait du programme officiel :

Échantillonnage Notion d'échantillon. Intervalle de fluctuation d'une fréquence au seuil de 95%*. Réalisation d'une simulation.	• Concevoir, mettre en œuvre et exploiter des simulations de situations concrètes à l'aide du tableur ou d'une calculatrice. • Exploiter et faire une analyse critique d'un résultat d'échantillonnage.	Un échantillon de taille *n* est constitué des résultats de *n* répétitions indépendantes de la même expérience. À l'occasion de la mise en place d'une simulation, on peut : • utiliser les fonctions logiques d'un tableur ou d'une calculatrice, ◇ mettre en place des instructions conditionnelles dans un algorithme. L'objectif est d'amener les élèves à un questionnement lors des activités suivantes : • l'estimation d'une proportion inconnue à partir d'un échantillon ; • la prise de décision à partir d'un échantillon.

Type d'activité :

Recherche en salle informatique d'une durée de deux fois une heure.

Pré-requis :

- Avoir une expérience des simulations sur tableur.
- Avoir traité le chapitre « statistiques descriptives ».

Motivation de l'activité :

Il s'agit de donner un sens à la notion d'intervalle de fluctuation et plus particulièrement à l'influence de l'effectif sur l'amplitude de cet intervalle.

Compétences travaillées :

- Utilisation des différentes fonctionnalités du tableur : formules, références absolues, représentations graphiques.

- Observation, description et interprétation de différents résultats.

Texte de l'activité élève

Dans une population donnée, la proportion p d'électeurs votant pour un candidat A est égale à 0,6.
On sait que lorsqu'on réalise un sondage, la proportion de votants pour le candidat A trouvée dans l'échantillon est « généralement voisine » de p. L'objectif de l'activité est de donner un sens plus concret à l'expression « généralement voisine » et de mettre en évidence l'influence du nombre n de personnes sondées, c'est-à-dire l'effectif de l'échantillon.
Dans le fichier EXCEL fourni, il ne faut pas modifier les cellules en couleur.

Partie 1

Ouvrir le fichier EXCEL « Approche…. ». Sur la première feuille de calcul, « Simulation », on a réalisé la simulation d'un sondage d'effectif n=400 dans une population où la proportion de votants pour le candidat A est p=0,6.

	A	B	C	D
1	Electeur n°	Simulation N°1		
2	1	1		
3	2	1		
4	3	1		
5	4	0		

1) Dans la cellule B2 on a mis la formule suivante : « =SI(ALEA()<0,6 ;1 ;0) ». Expliquer le fonctionnement de cette formule.
2) La formule placée en B2 a été recopiée 400 fois (jusqu'en B401) pour simuler les votes de 400 électeurs de la population P.
   Dans la cellule B402 écrire la formule qui donne le résultat de cette simulation, c'est-à-**dire la fréquence des votants pour le candidat A** dans l'échantillon simulé.
3) Appuyer plusieurs fois sur la touche F9 (fonction RECALCUL qui demande au logiciel de recalculer toutes les formules du fichier ouvert). Vérifier que le résultat affiché dans la cellule B402 est bien cohérent.
4)
   a) Sélectionner les cellules B1 à B402 et les recopier vers la droite jusqu'à la colonne AO. On obtient ainsi 40 simulations, les résultats de ces simulations étant dans la ligne 402.
   b) Ces résultats sont recopiés automatiquement sur la feuille « Fluctuation » à la ligne 3. Dans les lignes 2 et 4 de cette nouvelle feuille on a mis les résultats de simulations de sondage dans la même population mais avec des effectifs n=100 (pour la ligne 2 et n=1600 pour la ligne 4).
   Sélectionner les 4 premières lignes de cette feuille et représenter ces séries par des nuages de points. (On pourra modifier l'axe des ordonnées pour mieux visualiser).
   Comparer les dispersions de ces trois nuages de points.
5) Pour chacune des 3 séries, on décide d'enlever la plus petite et la plus grande des valeurs obtenues. Les valeurs extrêmes restantes, notées a et b, ont été placées dans les cellules B11 à C13. L'intervalle [a, b] est un intervalle de fluctuation au seuil de 95% (obtenu expérimentalement) de la fréquence de votants pour A.
   a) D'où vient la valeur 95% ?
   b) Calculer l'amplitude de cet intervalle pour les trois valeurs de n.
   c) Comparer les résultats obtenus.

Partie 2

Pour les valeurs de *n*, multiples de 100 inférieures ou égales à 2000, on a réalisé à chaque fois 1000 simulations de sondages de *n* personnes dans une population où la proportion de votants pour A est égale à p=0,6.
Pour chacune de ces valeurs de *n*, on a donc obtenu une série de 1000 fréquences. Certaines caractéristiques de ces séries sont données dans le tableau sur la feuille de calcul « 1000 Simulations_n=100 à 2000 » dont le début a été reproduit ci-contre.

n=	100	200	300	400
Min	0,44	0,495	0,503333	0,52
a	0,5	0,53	0,546667	0,55
q1	0,57	0,575	0,583333	0,5825
q3	0,63	0,625	0,62	0,6175
b	0,69	0,67	0,653333	0,6475
Max	0,75	0,715	0,693333	0,6725
Moyenne	0,59679	0,59978	0,601153	0,598975

- Min et Max sont à chaque fois la plus petite et la plus grande des 1000 fréquences obtenues.
- q1 et q3 sont à chaque fois le premier et troisième quartile de la série.
- En enlevant les 25 plus petites et les 25 plus grandes valeurs de chacune des séries, on définit a et b comme la plus petite et la plus grande valeur restante de la série.
- Moyenne est la moyenne des 1000 fréquences obtenues.

Ainsi, par exemple, la colonne n=400 résume les résultats de 1000 simulations (dans la première partie de l'activité 40 simulations avaient été faites).
La plus petite des fréquences observées au cours de ces 1000 simulations a été 0,52, la plus grande 0,6725.
Parmi ces simulations, 25 ont donné une fréquence inférieure à 0,55 et 25 une fréquence supérieure à 0,6475.
Un quart des fréquences observées étaient inférieures à 0,5825 et un quart supérieures à 0,6175.
Enfin la moyenne des fréquences observées étaient 0,598975.

1) Pour chaque valeur de n, quel est le pourcentage des fréquences calculées qui est dans l'intervalle [a, b]
2) Que représente le nombre (b-a) pour un intervalle [a, b]. Compléter la ligne 12.
3) Dans la suite de cette activité, on recherche si une fonction de la variable n permet d'approcher les valeurs (b-a). Si c'est le cas, que peut-on dire des variations de cette fonction ?
4) On décide de comparer les valeurs (b-a) avec celles des fonctions $n \mapsto \frac{K}{n}$ et $n \mapsto \frac{L}{\sqrt{n}}$ où K et L sont des entiers.
   a) Compléter la ligne 13 en recopiant la formule « =$C18/B11 » vers la droite. Expliquer la signification de cette formule et plus particulièrement le symbole « $ ».
   b) Compléter la ligne 14 pour qu'elle calcule les valeurs de la fonction $n \mapsto \frac{L}{\sqrt{n}}$.
5) Sélectionner les 4 lignes du tableau (11 à 14) et représenter les 3 courbes correspondantes.
6) Modifier les valeurs de K et L (entiers) pour essayer de trouver une courbe approchant au mieux celle correspondant à (b-a).
7) Conclusion : Compléter les phrases suivantes :
   a) Lorsqu'on effectue des simulations de sondages d'effectif n, l'amplitude de l'intervalle de fluctuation au seuil de 95% vaut environ…………
   b) En admettant que cet intervalle a pour centre $p = 0,6$ on obtient ainsi : […………… ;…………..].

Activité 3 : Intervalle de fluctuation et prise de décision avec la loi binomiale

Niveau :

Classe de première ES, S, STI2D, STL ou STMG.

Extrait du programme officiel (commun aux classes de première ES et S) :

Échantillonnage Utilisation de la loi binomiale pour une prise de décision à partir d'une fréquence.	• Exploiter l'intervalle de fluctuation à un seuil donné, déterminé à l'aide de la loi binomiale, pour rejeter ou non une hypothèse sur une proportion.	L'objectif est d'amener les élèves à expérimenter la notion de « différence significative » par rapport à une valeur attendue et à remarquer que, pour une taille de l'échantillon importante, on conforte les résultats vus en classe de seconde. ◇ L'intervalle de fluctuation peut être déterminé à l'aide d'un tableur ou d'un algorithme. Le vocabulaire des tests (test d'hypothèse, hypothèse nulle, risque de première espèce) est hors programme.

Type d'activité :

Il s'agit d'une activité guidée de synthèse à réaliser en classe, utilisant la calculatrice et faisant intervenir de l'algorithmique.

Pré-requis :

- Avoir traité le chapitre sur la loi binomiale.
- Avoir calculé des probabilités dans le cadre de la loi binomiale avec la calculatrice.
- Avoir introduit l'intervalle de fluctuation au seuil de 95% lié à la loi binomiale.
- Avoir utilisé la calculatrice pour programmer des algorithmes.

Motivation de l'activité, objectifs et compétences travaillées :

Cette activité est volontairement détaillée car elle intervient souvent durant le 3^e^ trimestre, donc à un moment où le temps commence à manquer.
Elle fait utiliser les fonctions de calcul des probabilités de la calculatrice.
Elle propose un travail d'interprétation, de modification et de programmation d'un algorithme sur la calculatrice.

Compte-rendu d'expérience :

L'activité a été testée dans une classe de première S de niveau moyen pendant l'année scolaire 2013/2014. Elle n'a pas posé de difficulté particulière.

Remarque :

La situation étudiée est tirée du document « Ressources pour la classe de première générale et technologique, Statistiques et probabilités » page 42, avec une mise en application différente de celle proposée dans ce document.

Texte de l'activité élève

Problème N°1

M. Léon, chef de gouvernement d'un pays lointain, a été élu avec 52% des voix des électeurs.
Un an plus tard, il affirme que sa cote de popularité est toujours égale à 52%.
Pour vérifier son affirmation, on décide de réaliser un sondage en interrogeant 500 électeurs au hasard (la population est suffisamment grande pour considérer qu'il s'agit de tirages avec remise).
On souhaite savoir, au seuil de 95%, quels sont les résultats de ce sondage qui contredisent l'affirmation de M. Léon.

1) On considère l'expérience aléatoire consistant à interroger 500 électeurs au hasard dans une population composée de 52% de personnes votant pour M. Léon (et donc de 48% ne votant pas pour lui).
Soit X la variable aléatoire prenant comme valeurs le nombre de personnes votant pour M. Léon.
Quelle est la loi de probabilité de X ?

____

2) A l'aide du tableau de valeur de la calculatrice, trouver le plus petit entier naturel k tel que $P(X \leq k) \leq 0,025$ et en déduire la borne de gauche de l'intervalle de fluctuation au seuil de 95% lié à la loi binomiale.

____

3) On considère l'algorithme ci-contre, écrit en langage naturel :

```
Entrer une valeur pour N
Entrer une valeur pour P
A prend la valeur 0
Tant que P(X ≤ A) ≤ 0,025 répéter
        A prend la valeur A+1
Fin du Tant que
Afficher A/N
```

Que réalise cet algorithme ?

____

____

____

4) Compléter l'algorithme précédent pour qu'il calcule aussi la borne de droite de l'intervalle de fluctuation au seuil de 95% lié à la loi binomiale.

```
Entrer une valeur pour N

Entrer une valeur pour P

________________________________________
________________________________________
________________________________________
________________________________________
________________________________________
________________________________________
________________________________________
________________________________________
________________________________________
```

5) Programmer l'algorithme précédent sur la calculatrice et exécuter le programme pour trouver l'intervalle de fluctuation au seuil de 95% lié à la loi binomiale.
   Conclure sur le problème posé.

Problème N°2

Dans le monde, la proportion de gauchers est de 12%.
Soit n le nombre d'élèves de votre classe.

1) Déterminer, en utilisant la loi binomiale, l'intervalle de fluctuation au seuil de 95% de la fréquence des gauchers sur un échantillon de taille n.
2) Votre classe est-elle représentative du nombre de gauchers dans le monde ?