Une activité pour initier à la statistique inférentielle en classe de seconde

Brigitte Sotura(*)

Après l'introduction des probabilités au collège, les nouveaux programmes du lycée abordent dès la classe de seconde des questions qui relèvent de la statistique inférentielle : *que dire d'une population à partir de données recueillies sur un échantillon de cette population ?*
En particulier la question de *la prise de décision à partir de l'observation d'une fréquence observée dans l'échantillon* est introduite dès la classe de seconde et reprise dans la quasi-totalité des programmes de première et terminale des séries générales et technologiques ; la problématique reste la même sur les trois niveaux et seuls les moyens disponibles pour y répondre diffèrent : la simulation en classe de seconde, la loi binomiale en première, la loi normale en terminale.
Le tableau ci-dessous extrait des programmes officiels, illustre le propos introductif.

Niveau	Contenu	Capacités attendues
2de	Simulation. Intervalle de fluctuation obtenu, de façon approchée, par simulation.	*Utilisation de la simulation pour une prise de décision à partir d'une fréquence observée.* *Estimation d'une proportion.*
1ère	Détermination de l'intervalle de fluctuation à partir de la loi binomiale (avec tableur ou algorithme).	*Utilisation de la loi binomiale pour une prise de décision à partir d'une fréquence observée.*
Tle	Approximation de la loi binomiale par une loi normale. Intervalle de fluctuation asymptotique à 95% $\left[p-1{,}96\sqrt{\frac{p(1-p)}{n}}, p+1{,}96\sqrt{\frac{p(1-p)}{n}}\right]$. Intervalle de confiance $\left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right]$.	*Utilisation de l'intervalle de fluctuation asymptotique pour une prise de décision à partir d'une fréquence observée.* *Estimation d'une proportion inconnue à partir d'une fréquence observée sur un échantillon.*

La classe de seconde est le moment privilégié pour faire comprendre l'enjeu de cet enseignement qui sera décliné sur les trois années de scolarité au lycée pour la plupart des élèves. Le programme de seconde précise que *l'objectif est d'amener les élèves* ***à un questionnement*** *lors des activités de prise de décision à partir d'un*

(*) Irem Paris Diderot

échantillon ou d'estimation d'une proportion. C'est ce point de vue qui a été choisi pour cette activité de la classe de seconde.

La situation choisie

Dans la réserve indienne d'Aamjiwnaang, située au Canada à proximité d'industries chimiques, il est né, entre 1999 et 2003, 132 enfants dont 46 garçons. Cette situation doit- elle être considérée comme anormale et interroger les autorités sanitaires ?
(Le sexe-ratio canadien est d'environ 105 garçons pour 100 filles.)
Cette situation est proposée dans les documents ressources de lycée professionnel (disponible sur le site *eduscol*) et est issue des travaux de l'IREM Paris Nord (statistique et citoyenneté).

Elle est choisie, d'une part, en raison de son caractère réel et de l'actualité de ces questions sanitaires. Elle offre l'occasion de montrer en quoi les mathématiques donnent des outils pour traiter de vraies questions scientifiques. En effet depuis quelques années, un grand nombre de pays industrialisés observent une baisse du rapport de masculinité (rapport du nombre de garçons sur nombre de naissances). Des produits chimiques pourraient en être la cause. C'est ainsi que suite à des études menées sur les effets du bisphénol A sur le sexe-ratio les autorités françaises ont décidé d'en réglementer l'usage.

Cette situation permet facilement de considérer ces 132 naissances comme un échantillon constitué des résultats de 132 répétitions indépendantes d'une même expérience aléatoire, autrement dit comme un tirage de boules avec remise dans une urne.

Les prérequis des élèves

Cette activité en classe de seconde commencée le 30 avril 2012 vient après un enseignement des probabilités de deux semaines en janvier durant lesquelles une expérimentation en réel (lancer de deux pièces) et des simulations d'expériences aléatoires sur tableur ont été réalisées par les élèves.

Les observations faites au cours de ces expérimentations ont conduit à admettre que, *lorsqu'on répète une même expérience aléatoire, la fréquence d'apparition d'un événement qui fluctue d'un échantillon à l'autre, se stabilise autour de la probabilité de cet événement lorsque la taille de l'échantillon augmente*.

À l'occasion de simulations d'expériences aléatoires et des traitements statistiques des séries obtenues, les élèves ont acquis une certaine familiarité avec le tableur et ont utilisé les adressages absolus et relatifs, les fonctions NB.SI, ALEA.ENTRE.BORNES ainsi que l'assistant graphique.

Le déroulement

1ère séance (30 avril) : le problème ci-dessus est posé devant la classe entière.

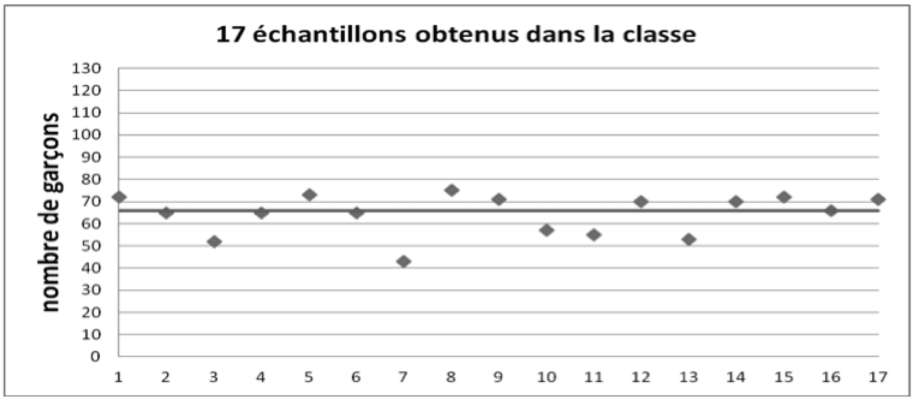
On est amené à préciser qu'il naît en moyenne 105 garçons pour 100 filles au Canada (soit une proportion de garçons de 105/205 soit 0,51) mais on suppose, pour simplifier (au moins dans un premier temps), qu'il naît en moyenne autant de filles que de garçons.
Il s'agit donc de comparer 46/132, soit environ 0,35, à 0,5.
On constate qu'il est né beaucoup moins de garçons que de filles dans cette réserve.

Simulation des naissances à l'aide de bouteille

On considère que la naissance est une expérience aléatoire à deux issues *fille* ou *garçon* ayant chacune la même probabilité de se réaliser. Sous cette hypothèse, on va étudier si le hasard peut suffire à expliquer un tel résultat à savoir 46 garçons sur 132 naissances.
Pour cela on propose aux élèves de simuler cette expérience à l'aide de « machines à fabriquer du hasard » : il s'agit de bouteilles contenant des boules oranges et noires. À chaque retournement on note la couleur obtenue. Si la boule est noire on convient qu'il s'agit d'un garçon sinon d'une fille.
Un consensus s'établit dans la classe pour mettre trois boules oranges et trois boules noires dans chaque bouteille (le nombre total ne semble pas avoir d'importance pour les élèves).

Les bouteilles sont distribuées et chaque binôme simule 132 naissances en effectuant 132 retournements et note le nombre de garçons obtenu.
Le tableau et le graphique ci-dessous présentent les résultats obtenus par les 17 binômes de la classe.

Gr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
G	72	65	52	65	73	65	43	75	71	57	55	70	53	70	72	66	71
F	60	67	80	67	59	67	89	57	61	65	77	62	79	62	60	66	61



Aucun binôme n'a obtenu le résultat 46 sur 132 mais un binôme a obtenu 43 boules noires.

Les élèves repartent de cette séquence avec l'idée que le hasard pourrait produire un résultat comme celui observé dans le village d'Aamjiwnaang. Il reste à étudier si un tel phénomène est fréquent ou ne l'est pas.

Deuxième séance sur ordinateur (en demi-groupe)

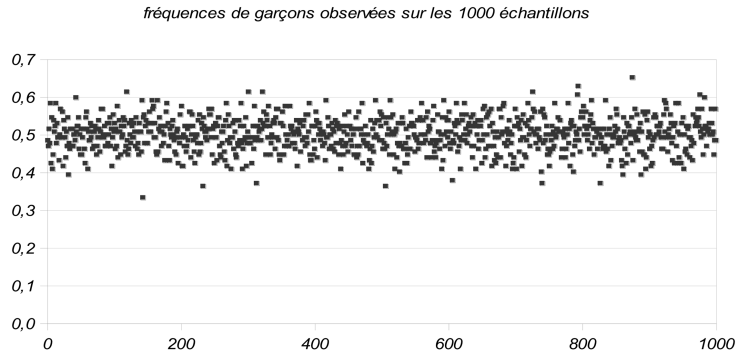
Pour disposer de plus de données, on demande aux élèves de simuler sur tableur un grand nombre d'échantillons. Chaque élève dispose d'un ordinateur et le document suivant leur est distribué en début de séance et ramassé en fin d'heure.

On se propose de simuler sur tableur de nombreux échantillons de 132 naissances afin d'affiner les observations faites ce matin au sujet du nombre de naissances de garçons observées dans le village d'Aamjiwnaang

1) **Simuler 132 naissances** disposées *en ligne* sur le tableur ; faire afficher le nombre de garçons ainsi que la fréquence de garçons obtenue. Relancer le calcul à l'aide de la touche F9.
Entre quelles valeurs semble varier le nombre de garçons sur un échantillon de 132 naissances?

2) À l'aide d'un copier coller **simuler 1000 échantillons de taille 132** et pour chacun d'entre eux, afficher le nombre de garçons et la fréquence de garçons obtenus.
Intéressons nous à la série statistique (brute) dont le caractère est *la fréquence de garçons* dans l'échantillon. Représenter cette série comme ci-dessous (*insérer diagramme* XY) (N'oubliez pas d'activer la touche F9 pour observer les fluctuations). Faites valider par le professeur.

3) À l'aide du graphique reproduit ci-dessous répondre aux questions suivantes :



a) Est ce que le hasard peut donner une fréquence de garçons égale ou inférieure à 46/132 ?

b) Pour 132 naissances, dans quel intervalle se trouve le plus souvent la fréquence de garçons ?

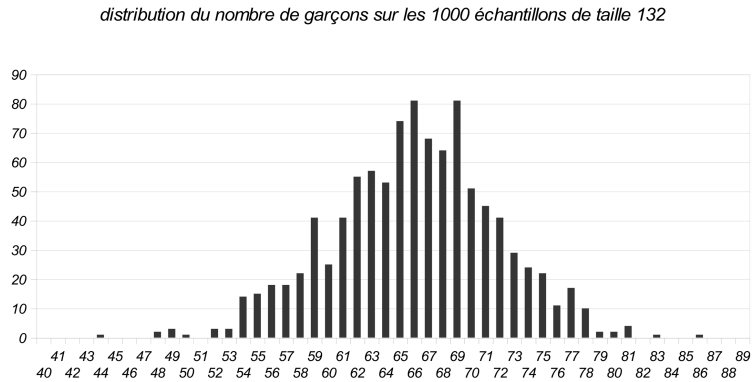
c) Sur les 1000 échantillons combien y en a-t-il pour lesquels cette fréquence est en dehors de l'intervalle [0,4 ; 0,6] ?

d) Si vous étiez responsable des questions sanitaires dans la région, diriez vous que ce qui se passe dans la réserve indienne d'Aamjiwnaang doit être considéré comme le fait du hasard ou diriez vous que la situation doit être considérée comme anormale et qu'une enquête doit être menée pour déterminer ce qui pourrait en être la cause ?

4) **Traitement statistique de la série du nombre de garçons par échantillon**

On se propose de regarder plus précisément la façon dont se distribue le nombre de garçons sur ces 1000 échantillons.

On a réalisé le diagramme en bâtons ci dessous à partir de la simulation de 1000 échantillons de taille 132.



a) Qu'a-t-on indiqué sur l'axe des abscisses ?

b) Sur l'axe des ordonnées ?

c) Un des bâtons correspond à 69 en abscisse et 81 en ordonnée. Interprétez ces données.

d) Commentez ce graphique. Quelle conclusion en tirez-vous par rapport à la question posée au sujet de la ***réserve indienne d'Aamjiwnaang*** ?.

e) Réalisez vous-même ce type de diagramme sur le tableur à partir de la simulation de vos 1000 échantillons. Faites valider par le professeur.

3ème séance : mise en commun en classe entière et conclusion

On a observé que le hasard pouvait produire « 46 garçons » sur 132 naissances ou moins. Pour autant on a observé que cela se réalise très rarement (sur ces 1000 échantillons seuls 18 ont une fréquence en dehors de l'intervalle [0,4 ; 0,6]).

L'étude précédente montre que le hasard peut donner un résultat comme celui qui est observé dans ce village mais il ne le peut que très rarement. Il y a un risque de se tromper en affirmant que ce n'est pas le hasard qui explique à lui seul ce qui se passe dans ce village mais **ce risque d'erreur est jugé suffisamment faible pour qu'on décide de considérer la situation comme anormale**.

Les autorités sanitaires de la région doivent donc s'inquiéter de la situation et des enquêtes doivent être menées pour en chercher les causes car cette étude statistique ne dit rien sur les raisons d'un tel taux de naissance de garçons.

Deux autres situations issues du document ressource du lycée professionnel ont été proposées aux élèves avec cette fois peu de directives.

Quelques observations, en guise de conclusion.

Comparer des proportions ne va pas de soi pour certains élèves en difficulté : par exemple 46/132 est à comparer à 0,5 (ou à 105/205). Cette difficulté apparaîtra encore dans les copies des élèves lors de l'évaluation qui suivra cette séquence.

Cette activité conduit à raisonner tantôt sur le nombre de garçons tantôt sur la fréquence. Il en sera de même en classe de première et terminale où on raisonnera tantôt sur la variable aléatoire X égale au nombre de garçons et tantôt sur la variable aléatoire F égale à la fréquence de garçons.

La simulation d'une naissance et du sexe de l'enfant à l'aide d'un retournement de la bouteille et de la couleur de la boule obtenue prépare la modélisation qui sera faite en classe de première par une épreuve de Bernoulli. Un échantillon de 132 naissances apparaît ainsi dès la classe de seconde comme la répétition 132 fois de la même épreuve. La variable aléatoire X donnant le nombre de garçons sur un échantillon de taille 132 suit une loi binomiale de paramètres $n = 132$ et $p = 0{,}5$.

La réalisation sur tableur du diagramme en bâtons de la distribution du nombre de garçons sur les100 échantillons pose problème à une majorité d'élèves : il leur faut

prendre conscience que la réalisation de ce diagramme nécessite au préalable de déterminer pour chaque nombre de garçons possible (a priori de 0 à 132) le nombre d'échantillons ayant donné ce résultat. Il s'agit de la série statistique (x_i,n_i) où x_i correspond au nombre de garçons et n_i au nombre d'échantillons comportant x_i garçons.

Dans le deuxième graphique les valeurs de la série (c'est-à-dire le nombre de garçons obtenu dans l'échantillon) sont portées en abscisse. Cette représentation anticipe la représentation de la loi binomiale qui sera faite en classe de première.

Enfin une difficulté et non des moindres est la conclusion : pour certains élèves le fait que le hasard puisse produire un résultat égal ou inférieur à 46/132 les conduit à dire qu'il n'y a pas lieu de considérer la situation comme anormale ; ils ne prennent pas en compte le caractère peu probable de cet événement. Il y aura nécessité de déterminer un seuil à partir duquel on considérera que l'événement est trop peu probable. Parvenir à le faire comprendre (sans nécessairement aller jusqu'à le quantifier) c'est préparer ce qui suivra ensuite dans les classes de première et terminale.

Donner prématurément une règle de décision selon que la fréquence observée est ou n'est pas dans l'intervalle $\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]$ risque de priver les élèves de tout questionnement : or comme le dit explicitement le programme de seconde l'objectif est d'amener les élèves à un questionnement. C'est ce point de vue qui a prévalu dans cette activité afin de faciliter la compréhension de la démarche qui sera vue en première à partir de la loi binomiale.

Une autre question aurait pu être soulevée en classe : « anormal » signifie-t-il que la fréquence observée est inférieure à une valeur donnée » « ou en dehors d'un intervalle (approximativement) centrée en p ?

- le premier point de vue correspondant à la question « la fréquence de naissances de garçons est-elle anormalement faible ? » et conduit à la notion d'intervalle de fluctuation unilatéral à un seuil donné.
- le deuxième correspondant à la question « la fréquence de naissances de garçons est elle anormalement éloignée de 0,5 » et conduit à la notion d'intervalle de fluctuation bilatéral à un seuil donné.