ministère
éducation
nationale

ÉduSCOL
[dgesco]

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Mathématiques

Lycée

Ressources pour la classe de seconde

- Probabilités et Statistiques –

Juin 2009

Table des matières

I. Introduction

L'enseignement de la statistique et des probabilités constitue un enjeu essentiel pour la formation du citoyen, lui donnant des outils pour comprendre l'information chiffrée, décider et choisir de façon éclairée et participer au débat public. Il est par ailleurs très utile aux autres disciplines qui s'appuient fréquemment sur des modèles statistiques ou probabilistes. Le rapport de la commission de réflexion sur l'enseignement des mathématiques exprimait en ces termes l'enjeu de cet enseignement : « *l'objectif d'une initiation aux probabilités et à la statistique au niveau collège et lycée est d'enrichir le langage, de repérer des questions de nature statistique, de définir des concepts qui fonderont un mode de pensée pertinent, rassurant, remarquablement efficace* ».

Les programmes du collège ont inscrit l'étude des séries statistiques avec des indicateurs de position et de dispersion. Des notions de probabilité sont abordées en classe de troisième à partir de situations familières permettant, entre autres, de rencontrer des probabilités qui ne soient pas uniquement définies à partir de considérations intuitives de symétrie mais qui prennent appui sur l'observation d'épreuves répétées et la stabilisation des fréquences.

	Classe de sixième	Classe de cinquième	Classe de quatrième	Classe de troisième
Organisation et gestion de données	Organiser des données en choisissant un mode de représentation adapté. Lire et compléter une graduation sur une demi-droite graduée. Lire et interpréter des informations à partir d'une représentation graphique.	Repérage sur une droite graduée et dans le plan. Classes, effectifs, fréquences. Tableaux de données : lectures, interprétation, élaboration, représentations graphiques.	Moyenne pondérée.	Caractéristiques de position : médiane, quartiles. Approche de caractéristiques de dispersion : étendue. Notion de probabilité

En conséquence, il s'agit de veiller à bien inscrire l'enseignement de la classe de seconde en continuité avec celui du collège. Cela vaut autant pour la seconde générale et technologique que pour la voie professionnelle dont les programmes en vigueur à la rentrée 2009 font une part importante aux probabilités et à la statistique[1]. Il conviendra notamment d'éviter des révisions systématiques et de proposer des situations permettant le réinvestissement des notions abordées dans les classes précédentes.

Le travail statistique sur données réelles, brutes ou préalablement traitées avec l'aide incontournable de l'outil informatique, est de nature à favoriser la prise d'initiative et la conduite de raisonnements pour interpréter, analyser ou comparer des séries statistiques sur des sujets en prise avec l'actualité des élèves. Pour reprendre le rapport de la commission de réflexion sur l'enseignement des mathématiques, « *le matériau brut travaillé par la statistique est constitué de données expérimentales, les outils théoriques utilisés sont essentiellement la géométrie et l'algèbre linéaire pour la statistique exploratoire et les probabilités pour la statistique inférentielle et l'outil matériel est l'ordinateur* ».

Ainsi les outils de statistique descriptive travaillés au collège pourront être efficacement sollicités lors de l'exploration de fichiers comportant des données réelles afin d'exhiber des

[1] BO spécial n°2 du 19/02/2009

régularités, des dominantes ou des caractéristiques que la masse de données ou le nombre de variables étudiées ne livrent pas facilement. Cette pratique de fouille de données (statistique exploratoire, data mining) est de nos jours essentielle tant les données disponibles sont nombreuses et constituent des repères indispensables pour les entreprises ou les politiques publiques dans leurs choix stratégiques et décisionnels. Histogrammes, représentations graphiques, calculs de moyennes, de médianes, de quartiles sont autant d'outils que les élèves pourront utiliser pour faire émerger des questions sur les données, dégager des problématiques d'étude ou résumer l'information essentielle, d'autant plus que les outils informatiques accessibles aux élèves permettent de travailler sur des gros fichiers.

Les premiers éléments de probabilité ont été abordés au collège essentiellement dans des situations de jeux (lancers de dés ou de pièces, loteries, tirages dans des urnes). Cela a permis une première approche de quelques lois de probabilité qui seront progressivement décontextualisées au lycée en vue de fournir des modèles pour d'autres champs d'application, tant dans les domaines scientifiques (sciences physiques et chimiques, sciences de la vie et de la terre, sciences de l'ingénieur, etc.) que dans les sciences économiques et sociales. Ainsi les élèves ont été familiarisés, par des lancers de pièces équilibrées, avec la loi uniforme sur {0 ; 1} qui permettra, par exemple, d'aborder de nombreux sujets (sex-ratio, parité). De même le tirage dans une urne composée de boules de deux couleurs différentes, dans des proportions p et $(1 - p)$, fournit une première approche de la loi de Bernoulli, qui s'ouvrira sur des applications aux sondages (estimations à la sortie des urnes les soirs d'élection), au contrôle de qualité des productions industrielles (maîtrise statistique des processus de production) ou aux diverses estimations sur échantillon. Toutes ces questions relèvent de la statistique inférentielle, qui fonde ses résultats sur des considérations probabilistes et permet l'induction à partir de données observées.

Les questions de fluctuation d'échantillonnage constituent un axe important de la formation du futur citoyen, qui aura ainsi été sensibilisé au lycée à la nécessaire prudence à avoir avant d'interpréter une évolution ou d'effectuer des comparaisons. En effet toute évolution de moyenne ou de proportion, toute comparaison d'échantillons doit être nuancée et relativisée au regard des variations liées à la fluctuation d'échantillonnage.

Afin d'entrer vraiment dans une démarche statistique en lien avec les concepts probabilistes, on gagnera à utiliser, comme fil rouge, un fichier de données réelles pour mettre en œuvre ou pour illustrer les différentes notions inscrites au programme[2]. En procédant ainsi, on limite le temps d'appropriation des données et les élèves peuvent plus rapidement se concentrer sur les outils mathématiques, la situation étudiée devenant familière.

Par ailleurs, c'est en ayant recours à des données réelles que l'on développe les capacités d'observation et de raisonnement des élèves : comprendre la nature des données, repérer l'organisation d'un tableau, imaginer et réaliser des représentations ou des calculs adaptés, comprendre un graphique sont autant d'occasions de raisonner et d'exercer l'esprit critique. Dans les exemples développés dans ce document, on a tenu à souligner ces différents aspects de la formation, montrant l'ampleur et l'ambition des raisonnements conduits, ainsi que la place légitime de cet enseignement dans le programme de mathématiques.

---

[2] Voir le dossier proposant une progression et des activités utilisant le fichier des populations des communes françaises : Consulter le tableau récapitualtif
Télécharger l'ensemble des fichiers (25 Mo, fichiers à laisser dans le même dossier pour le fonctionnement des liens)

II. Des statistiques aux probabilités

1. Statistique descriptive, analyse de données

1.1. Résumé des notions abordées au collège

Les notions de moyenne, médiane, étendue, quartiles et écart interquartile ont été développées au collège ainsi que leurs interprétations.

Ces notions pourront être sollicitées en classe de seconde dans plusieurs domaines en lien avec les autres disciplines, par exemple pour étudier des séries de mesures expérimentales en sciences physiques. Il est alors possible de traiter plusieurs questions autour de l'intervalle interquartile, de son amplitude, etc.

1.2. Analyse de données

La classe de seconde est l'occasion d'une part de consolider l'utilisation des fonctions statistiques des calculatrices et d'autre part de traiter, à l'aide d'un tableur, des séries statistiques riches et variées comportant un grand nombre de données brutes en lien avec des situations réelles. À titre d'exemples, on pourra trouver de telles données :

- sur le site de l'INSEE , population des régions, départements et communes :
http://www.insee.fr/fr/ppp/bases-de-donnees/recensement/populations-legales/france-departements.asp

- sur le site de l'INED, espérance de vie à la naissance :
http://www.ined.fr/fichier/t_telechargement/18154/telechargement_fichier_fr_sd2006_t2_fm.xls

- sur le site de météo France ou des sites de particuliers tel :

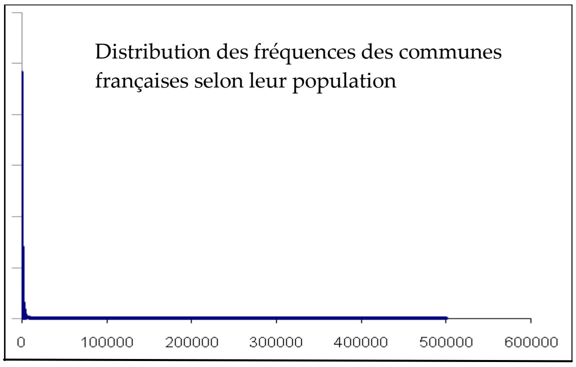
http://www.meteociel.com/climatologie/climato.php

Par ailleurs on pourra exploiter, dans le cadre de travaux interdisciplinaires, des données issues d'autres disciplines (sciences physiques, sciences de la vie et de la terre, sciences de l'ingénieur etc.).

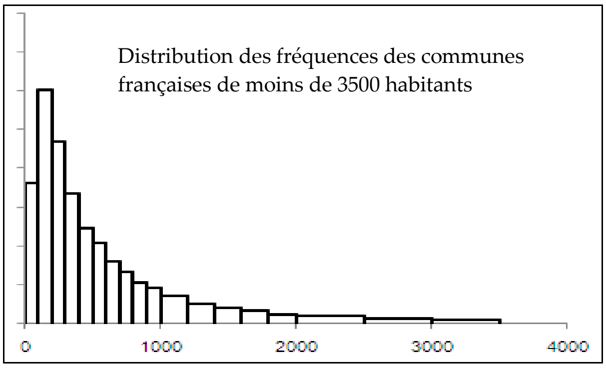
Exemple : fichier des communes françaises
Le nombre important de données peut inciter, dans un premier temps, à effectuer une représentation graphique (graphique 1).

Le résultat obtenu est surprenant et soulève immédiatement la question de la répartition des communes en fonction de leur nombre d'habitants, suggérant de scinder la série, par exemple en isolant les communes de moins de 3500 habitants qui est le seuil retenu dans la loi électorale de 2007. Le graphique 2 donne la répartition de ces communes[3].

[3] Réalisation d'histogrammes à classes d'amplitudes inégales, on pourra consulter :
http://www.ac-grenoble.fr/maths/guppy/pages/fiches/Mediane/Tableur_stats_1_Var.htm



Graphique 1



Graphique 2

Il est clair, en observant le graphique 1 ci-dessus que la moyenne (1760 habitants) n'est pas pertinente pour résumer cette série statistique. On lui préfèrera ici la médiane et les quartiles.

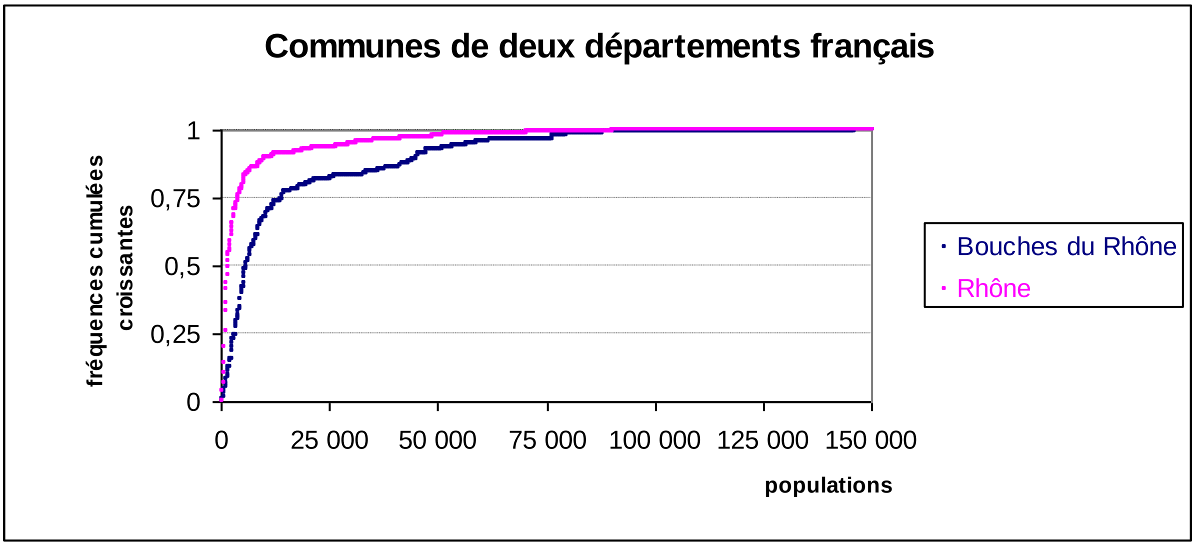
1.3. Fréquences cumulées croissantes

La courbe des fréquences cumulées croissantes[4] permet de représenter la distribution d'une série statistique ainsi que l'illustre l'exemple ci-dessous.

Exemple : comparaison entre deux départements[5]

Le graphique ci-après est réalisable à partir du fichier INSEE des régions, départements et communes de France. La compréhension des deux courbes conduit les élèves à des raisonnements formateurs pour répondre aux questions suivantes :

- comment interpréter l'antécédent[6] de 0,5 (resp 0,25 ; 0,75) par chacune de ces fonctions ?
- l'une des courbes est située en dessous de l'autre ; comment interpréter cette propriété ?



---

[4] Elle permet de retrouver une valeur approchée de la médiane, mais ce n'est pas une méthode à préconiser lorsque l'objectif est uniquement de déterminer une valeur médiane.

[5] Activité à l'adresse suivante : http://www.ac-grenoble.fr/maths/docresseconde/Exemple de progression/activite_communes.doc

[6] Sous l'hypothèse que les points aient été reliés.

2. Probabilité sur un ensemble fini

Le travail sur les probabilités, initié en classe de troisième, est stabilisé et consolidé en classe de seconde avec, en perspective, une démarche de modélisation de phénomènes réels.

2.1. Résumé des notions abordées en troisième

Notion de probabilité, calcul dans des situations familières (lancer de pièces ou de dés, roue de loterie, urnes). Probabilités estimées par des fréquences observées sur de longues séries. Application pour modéliser des situations de la vie courante. Les expériences aléatoires concernent des situations à une ou deux épreuves qui n'excèdent pas 6 éventualités au total.

2.2. Distribution de probabilité sur un ensemble fini, probabilité d'un événement

Il s'agit dans un premier temps de consolider les notions abordées en classe de troisième.

Une distribution de probabilité sur un ensemble Ω est définie par la donnée des probabilités des éléments de Ω. Un événement est défini comme sous-ensemble de Ω. C'est cette définition ensembliste qui permet de calculer la probabilité d'un événement en ajoutant les probabilités des éléments qui le constituent. On consolide à l'occasion la notion d'ensemble ou de sous-ensemble, ce qui permet, entre autres, d'ancrer l'idée que dans un ensemble on ne répète pas les éléments et que leur ordre n'importe pas.

Les distributions de probabilité peuvent être estimées par observation de la stabilisation des fréquences[7] sur de longues séries d'expériences[8] ou bien par des considérations géométriques ou physiques en référence à l'équiprobabilité.

De même qu'au collège les élèves ont utilisé, sans formalisme, quelques éléments de langage sur les probabilités[9] , de même en classe de seconde certaines notations usuelles seront utilisées pour leur commodité et sans donner lieu à un formalisme excessif :
$\overline{A}$, p({1,2,3}), $p(A \cap B)$ ou p(A et B), $p(A \cup B)$ ou p(A ou B), card(A).

2.3. Modélisation, modélisations ?

La simulation a pour préalable de choisir un modèle.

Exemple 1 : somme de deux dés
Cette situation se prête volontiers à la mise en œuvre d'une démarche consistant à proposer un modèle et à le confronter aux données d'expérience. Les résultats compris entre 2 et 12 peuvent conduire certains élèves à faire porter l'équiprobabilité sur l'ensemble des 11 résultats observables. Par quelques expérimentations avec des dés, puis en ayant recours à des simulations, on est conduit à rejeter ce modèle pour proposer de faire porter l'équiprobabilité sur les 36 couples de résultats {(1,1) ;(1,2) …(6,6)}, présentés usuellement sous forme de tableau croisé.

---

[7] On en restera à la perception intuitive de la loi des grands nombres.
[8] On trouvera un exemple d'approche fréquentiste page 13 dans le document ressource des nouveaux programmes de lycée professionnel http://www.ac-grenoble.fr/maths/docresseconde/Proba_stat_LP.doc
[9] Cf. document ressource du collège :
http://www.ac-grenoble.fr/maths/docresseconde/doc_ressource_clg_probabilites.pdf

D'autres exemples[10] pourront être développés, montrant aux élèves les cheminements de pensée entre les modèles retenus et les expériences réelles, et posant la question de l'ensemble sur lequel porte l'équiprobabilité.

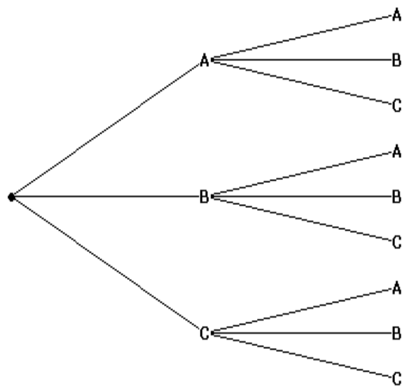
Exemple 2 : le problème des deux personnes qui s'assoient "au hasard "

Un square est équipé de trois bancs à deux places. Deux personnes arrivent successivement et s'installent au hasard. Quelle est la probabilité que ces personnes soient assises côte à côte ?

En faisant l'hypothèse de l'équiprobabilité des issues :

Première modélisation[11] : on numérote les six places 1, 2, 3, 4, 5 et 6, chaque paire représente les deux places occupées. En comptant les paires {1,2}, {1,3} … {5, 6}, on obtient pour cet événement une probabilité de $\frac{3}{15}$ soit $\frac{1}{5}$.

Deuxième modélisation[12] *:* on note les trois bancs A, B et C, les résultats de l'expérience peuvent être codés par des couples, par exemple (B,A) : la première personne s'assoit sur le banc B et la deuxième sur le banc A. On peut aussi construire l'arbre des possibles :



On obtient une probabilité de $\frac{1}{3}$.

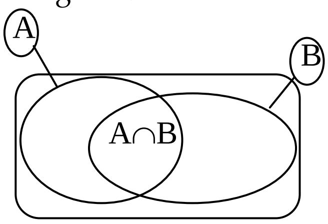
Ces deux modèles donnent des résultats différents, « tirer au hasard » n'étant pas une information suffisante, même en rajoutant une hypothèse d'équiprobabilité. Une description plus précise de l'expérience doit être fournie pour induire un choix de modèle.

3. Calculs de probabilités

3.1. Réunion et intersection de deux événements

Les symboles d'union et intersection sont introduits en liaison avec les conjonctions « ou » et « et », en comparant leur sens mathématique avec leur usage dans la langue commune.

La formule $p(A\cup B) + p(A\cap B) = p(A) + p(B)$ pourra être illustrée par un diagramme de Venn :



[10] Le problème du Duc de Toscane : lancer de trois dés, quelle est la somme la plus probable ? : http://www.ac-grenoble.fr/maths/docresseconde/Le paradoxe du Duc de Toscane.doc

[11] On peut imaginer que le choix de la place s'effectue par tirage sans remise dans une urne avec 6 jetons numérotés de 1 à 6.

[12] Le choix du banc pourrait s'effectuer par tirage avec remise dans une urne contenant 3 jetons A, B et C.

Exemple : avec le fichier des communes françaises
On choisit au hasard une commune dans l'ensemble des communes de France.
Quelle est la probabilité de l'événement A : « cette commune est dans votre région » ?
Quelle est la probabilité de l'événement B : « sa population est inférieure à 1000 habitants » ?
Définir les événements $A \cup B$ et $A \cap B$ puis calculer leur probabilité.
Cet exemple, à partir d'un fichier comportant de très nombreuses données, peut conduire les élèves à pratiquer des instructions logiques dans les conditions appelées par la fonction NB.SI.

3.2. Tableaux croisés

Les tableaux croisés rencontrés dans des résumés d'enquêtes se prêtent à des calculs simples de probabilités d'intersections ou d'union[13].

3.3. Arbres des possibles

Les arbres décrivant de façon exhaustive les issues d'une expérience ont pu être abordés en classe de troisième ; on peut consolider cette pratique pour aider les élèves à se construire des images mentales fiables et être plus assurés dans les modélisations et les calculs.

Ces arbres aident au dénombrement et sont des supports de raisonnement. Il n'est pas toujours nécessaire ni matériellement possible d'en représenter toutes les branches. On peut développer les capacités d'abstraction des élèves en utilisant des pointillés dans leur construction.

Exemple : probabilité d'avoir la même date anniversaire[14]
En regardant les dates anniversaires des élèves dans les classes du lycée, on peut être surpris du nombre de classes dans lesquelles deux élèves fêtent leur anniversaire le même jour.

Ce résultat étonnant peut inciter à faire un calcul de probabilité.

On pourra proposer de commencer par un exemple plus simple :

« Dans un groupe de quatre personnes prises au hasard, quelle est la probabilité qu'au moins deux d'entre elles fêtent leur anniversaire le même mois ? On suppose que, pour chaque personne, tous les mois d'anniversaire sont équiprobables et on les numérote de 1 à 12. ».

On peut reformuler ce problème en assimilant l'expérience à un tirage aléatoire dans une urne : « une urne contient 12 boules numérotées de 1 à 12, on effectue au hasard et avec remise quatre tirages successifs, et l'on note les numéros obtenus, dans l'ordre d'apparition. Chaque numéro tiré correspond au mois d'anniversaire d'une des personnes ».

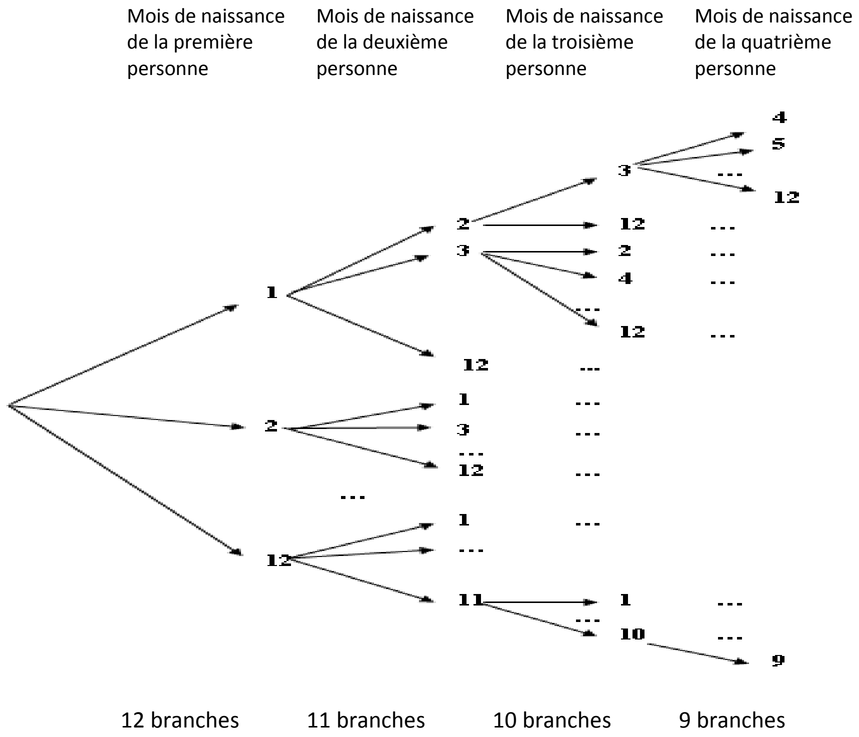
On peut compter le nombre total des issues avec un arbre comportant des pointillés.

Ensuite on peut rechercher le nombre d'éléments de l'événement étudié, montrer la difficulté que l'on rencontre pour décrire et compter directement les cas favorables, puis faire réfléchir à l'intérêt et à l'énoncé de l'événement contraire (négation de "au moins…").

---

[13] On peut utiliser des tableaux croisés disponibles sur le site de l'INSEE, par exemple le diplôme le plus élevé selon l'âge et le sexe : http://www.insee.fr/fr/themes/tableau.asp?reg_id=0&ref_id=NATCCF07235
[14] Cf. feuille de calcul
http://www.ac-grenoble.fr/maths/docresseconde/Anniversaires_probabilit%E9s_simulation.xls

Illustration de l'arbre des possibles de l'événement contraire:



Pour bien s'assurer de la compréhension de l'arbre, on peut suivre un trajet de la racine à une extrémité de branche et demander aux élèves d'interpréter le résultat par une phrase.

On trouve au total $12 \times 11 \times 10 \times 9 = 11880$ issues possibles. Ce qui donne comme probabilité de l'événement contraire : $\frac{12 \times 11 \times 10 \times 9}{12^4}$, soit environ 0,43 ; d'où la probabilité cherchée (environ 0,57). On peut ensuite adapter ces calculs de probabilité pour un groupe de cinq, puis six personnes[15].

Retour au problème initial : le même raisonnement, immédiatement transposé, permet de résoudre la situation des mêmes dates anniversaires et de rechercher la taille du groupe de personnes pour avoir une probabilité supérieure à 0,8 qu'au moins deux d'entre elles fêtent leur anniversaire à la même date.

Utilisation d'un algorithme

Si n désigne le nombre de personnes du groupe, il s'agit de déterminer à partir de quelle valeur de n le nombre $q = \frac{365 \times 364 \times \cdots \times (365-(n-1))}{365^n}$ est inférieur à 0,2, avec $q = 1 - p$.

En remarquant que q peut s'écrire comme une répétition de multiplications : $q = \frac{365-0}{365} \times \frac{365-1}{365} \times \frac{365-2}{365} \times \cdots \times \frac{365-(n-1)}{365}$, on peut élaborer un algorithme de calcul de ce nombre selon la valeur de n (algorithme 1), puis par essais successifs, déterminer la première valeur de n qui répond à la question[16].

[15] Réponses: 0,62 et 0,78.

[16] Pour n = 35, on trouve $p \approx 0,814$, alors que pour n = 34 on a $p \approx 0,795$.

Algorithme 1	Algorithme 2
Variables n, i entiers et q réel Entrées Saisir n Initialisations q prend la valeur 1 Traitement Pour i variant de 1 à $n-1$ q prend la valeur $q\times\frac{365-i}{365}$ Sorties Afficher $1-q$	Variables i entier, p et q réels Entrées Saisir p Initialisations q prend la valeur 1, i prend la valeur 0 Traitement Tant que q est supérieur à $1-p$ q prend la valeur $q\times\frac{365-i}{365}$ i prend la valeur $i+1$ Fin du Tant que Sorties Afficher i

On peut aussi utiliser un algorithme avec boucle et condition d'arrêt (algorithme 2) pour éviter le tâtonnement et répondre rapidement à cette question ou à une question du type : déterminer la plus petite valeur de n à partir de laquelle p est supérieur à 0,99, ou à 0,999, et observer l'évolution de n.

Le calcul précédent pourrait également être réalisé à l'aide d'un tableur.

3.4. Arbres pondérés

Les situations simples à deux épreuves ont pu être travaillées au collège à l'aide de petits arbres pondérés[17]. Il s'agit d'entretenir, sans aucun nouveau développement ni aucune complexification, ce type de présentation et son mode opératoire, comme l'illustre l'exemple ci-dessous. Toute connaissance sur le conditionnement est hors programme.

Exemple :

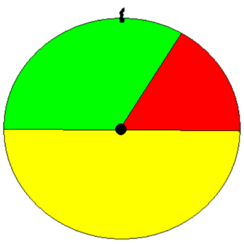
Un forain propose le jeu suivant : "À tous les coups l'on gagne"

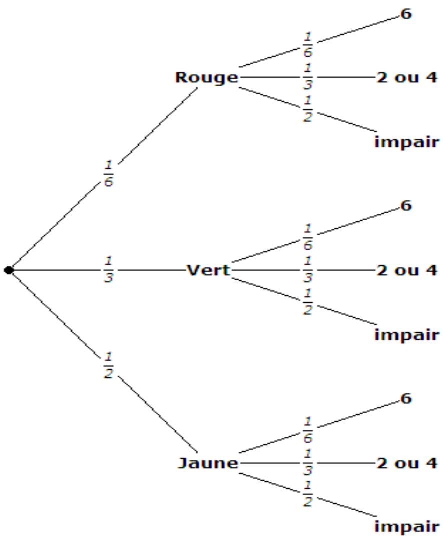
Le joueur fait tourner une roue divisée en secteurs de mesures 60°, 120° et 180° puis il lance un dé équilibré. Il gagne un petit lot si la couleur sortie sur la roue est le vert et si le dé sort un numéro impair. Il gagne un gros lot si la couleur sortie sur la roue est le rouge et si le dé sort un six. Dans les autres cas, il gagne une pacotille.

Quelle est la probabilité de gagner un gros lot ?

Quelles sont les probabilités de gagner un lot (petit ou gros) ?

De gagner une pacotille ?





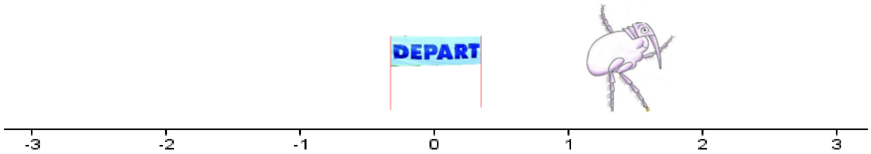
[17] Cf. document ressource du collège : http://www.ac-grenoble.fr/maths/docresseconde/doc_ressource_clg_probabilites.pdf

L’arbre pondéré ci-dessus permet de répondre à la première question où il s’agit d’évaluer la probabilité de l’événement (R,6) : dans un arbre, la probabilité du résultat auquel conduit un chemin est égale au produit des probabilités rencontrées le long du chemin[18]. Cet arbre permet d’autre part de trouver facilement les formules donnant la probabilité d’une réunion ou d’un événement contraire, en leur donnant du sens.

3.5. Exemples d’algorithmes : marche aléatoire et temps moyen

Une marche aléatoire est une trajectoire constituée de pas successifs, aléatoires et indépendants. Les applications des marches aléatoires sont multiples tant en botanique (le botaniste Robert Brown, notait en 1827, le caractère apparemment erratique du déplacement de particules de pollen dans l'eau, appelé par la suite mouvement brownien), en sciences physiques (trajectoires d’une molécule dans un liquide ou un gaz), en économie (variations du cours d’une action en bourse) ou en sciences informatiques (certains moteurs de recherche utilisent des marches aléatoires pour parcourir les pages internet) etc.

Exemple : des sauts de puce[19]



Une puce se déplace sur un axe gradué : à chaque saut elle se déplace d'une unité, de manière aléatoire et équiprobable vers la droite ou vers la gauche. Elle part de l’origine et effectue une marche de 30 sauts.
Proposer un algorithme donnant la position d’arrivée de la puce, c'est-à-dire la position du trentième saut.
Enrichir l’algorithme précédent pour donner la liste des positions d’arrivée de N marches aléatoires.

La fonction alea fournit un nombre aléatoire dans l'intervalle [0;1[; l'entier x désigne la position de la puce :

```
Algorithme 1
        Variables
            x, i entiers
        Initialisations
            x prend la valeur 0
        Traitement
            Pour i variant de 1 à 30
                Si alea < 0,5
                alors x prend la valeur x + 1
                sinon x prend la valeur x – 1
                Fin du Si
            Fin du Pour
        Sorties
            Afficher x
```

[18] On en reste à une perception intuitive, lorsque les épreuves successives sont indépendantes (au sens où la première épreuve n’a aucune influence sur la seconde), comme cela est évoqué dans le document ressource sur les probabilités au collège en page 11 :
http://www.ac-grenoble.fr/maths/docresseconde/doc_ressource_clg_probabilites.pdf
[19] Cf. une proposition de fiche élève :
http://www.ac-grenoble.fr/maths/docresseconde/fiche_eleve_marches_aleatoires.doc

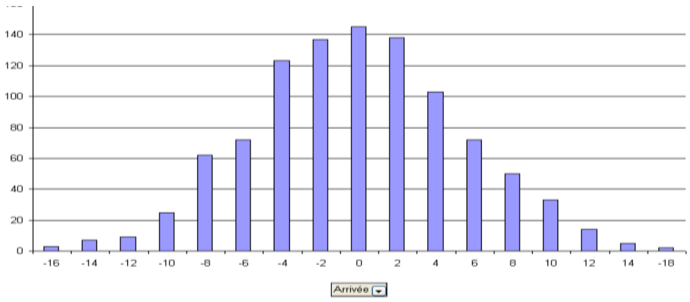
On peut demander aux élèves d'examiner ce que fait l'algorithme suivant :

```
Algorithme 2
        Variables
                N, S, x, i, j entiers
        Entrées
                Saisir N
        Initialisations
                S prend la valeur 0
        Traitement
                Pour j variant de 1 à N
                  x prend la valeur 0
                  Pour i variant de 1 à 30
                      Si alea < 0,5
                      alors x prend la valeur x + 1
                      sinon x prend la valeur x – 1
                      Fin du Si
                  Fin du Pour
                  S prend la valeur S+x
                Fin du Pour
        Sorties
                Afficher S/N
```

Enfin, on peut s'interroger sur la fréquence des différentes positions d'arrivée et pour cela simuler N marches aléatoires de 30 sauts, grâce à l'algorithme suivant :

```
Algorithme 3
        Variables
                N, x, i, j entiers, L liste
        Entrées
                Saisir N
        Initialisations
                Vider la liste L
        Traitement
                Pour j variant de 1 à N
                  x prend la valeur 0
                  Pour i variant de 1 à 30
                      Si alea < 0,5
                      alors x prend la valeur x + 1
                      sinon x prend la valeur x – 1
                      Fin du Si
                  Fin du Pour
                  L(j) prend la valeur x
                Fin du Pour
        Sorties
                Afficher la liste L
```

Pour mieux se représenter les résultats, il peut être utile de représenter les données aléatoires ainsi obtenues, sous la forme du graphique suivant :



Exemple : nombre de lancers pour sortir tous les numéros d'un dé cubique[20]

On souhaite estimer le nombre de lancers nécessaires pour sortir toutes les faces d'un dé cubique.

Comme pour l'exemple précédent, la réponse est aléatoire. On peut, dans un premier temps, écrire un algorithme permettant de simuler une expérience, puis faire tourner plusieurs fois cet algorithme et observer que la moyenne des résultats obtenus se stabilise[21] pour un grand nombre d'expériences. Cette moyenne, elle aussi aléatoire, permet néanmoins de proposer une réponse au problème en termes de probabilités.

```
Algorithme
        Variables
              L liste
              S, n, x entiers
        Initialisations
              Vider la liste L
              S et n prennent la valeur 0
        Traitement
              Tant que[22] n est inférieur à 6
              S prend la valeur S +1
              x prend la valeur d'un nombre aléatoire entier entre 1 et 6
              si x ne figure pas dans la liste L alors
                     n prend la valeur n+1
                     L(n) prend la valeur x
              Fin du Tant que
        Sorties

             Afficher S
```

On peut, pour estimer le nombre moyen de lancers permettant de sortir les six faces du dé, modifier l'algorithme précédent en s'inspirant de l'algorithme 2 ci-avant.

III. Échantillonnage

1. Fluctuation d'échantillonnage

1.1. Notion d'échantillon

Le terme « échantillon » prenant différents sens, il convient de préciser ce qu'il recouvre dans le programme de seconde.

Dans le sens commun des sondages, ce terme s'apparente à un sous-ensemble obtenu par prélèvement aléatoire dans une population. Ainsi parle-t-on usuellement de résultats estimés sur échantillon.

[20] On peut examiner en variante le nombre moyen de tablettes de chocolat qu'il faut acheter pour obtenir la collection complète des six images qui sont jointes. Cf. une expérimentation en classe à l'adresse http ://www.statistix.fr/spip.php ?article15

[21] On en restera à une approche intuitive de la loi des grands nombres.

[22] On peut observer que cette condition n'assure pas que l'algorithme se termine en un nombre fini d'opérations. Il suffirait alors de modifier légèrement le critère de sortie de boucle, en fixant par exemple un nombre maximum pour S au-delà duquel le programme s'arrête.

En statistique, un échantillon de taille n est la liste des n résultats obtenus par n répétitions indépendantes de la même expérience[23]. Par exemple, un échantillon de taille 100 du lancer d'une pièce est la liste des résultats pile ou face obtenus successivement en répétant 100 fois le lancer de la pièce. De même pour un échantillon de taille 100 relatif au lancer d'un dé dont on observe l'apparition ou non de la face 6, ou bien encore pour un échantillon obtenu par tirages successifs avec remise d'une boule dans une urne contenant 2 boules blanches et une boule verte. Ces trois exemples relèvent en fait du même modèle, celui de Bernoulli qui affecte la probabilité p au nombre 1 et la probabilité $(1 - p)$ au nombre 0, seule situation abordée en classe de seconde.

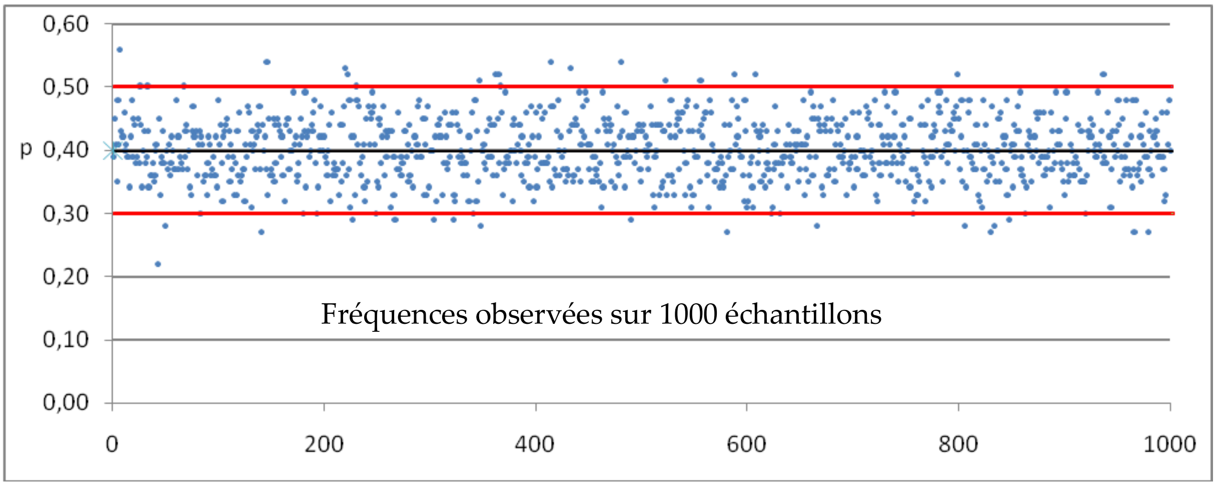
Cette notion d'échantillon fournit un cadre théorique pour démontrer les résultats énoncés ci-dessous sur la fluctuation d'échantillonnage.

En outre, ces résultats théoriques pourront s'appliquer aux sondages destinés à estimer une proportion p (par exemple le pourcentage de votes « oui » lors d'un référendum), en remarquant qu'un tirage sans remise d'un échantillon dans une population suffisamment nombreuse est assimilable à la répétition d'un tirage avec remise dans une urne.

1.2. Intervalle de fluctuation[24]

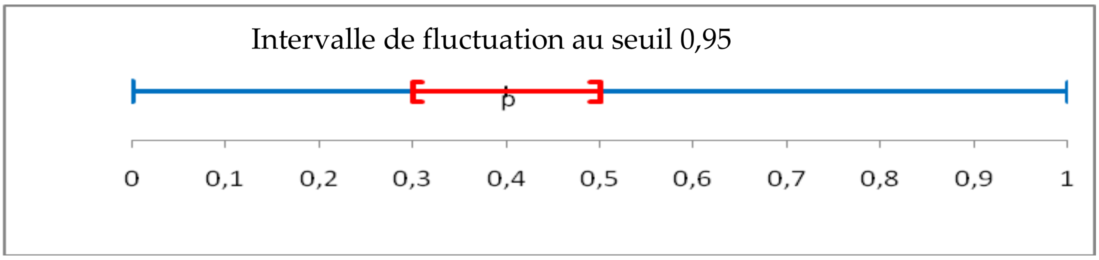
On peut, par expérimentation et simulation, faire observer aux élèves que les échantillons de taille n obtenus à partir d'un modèle de Bernoulli ont, pour environ 95% d'entre eux, des fréquences d'apparition du nombre 1 qui fluctuent[25] dans un intervalle centré en p et d'amplitude $\frac{2}{\sqrt{n}}$.

On a simulé ci-dessous[26], 1000 échantillons de taille 100 d'un modèle de Bernoulli avec $p = 0,4$. Chaque échantillon est représenté par un point dont l'ordonnée est sa fréquence d'apparition du 1. On observe que la plupart des échantillons ont des fréquences d'apparition du 1 dans l'intervalle [0,3 ; 0,5].



[23] C'est-à-dire relative au même modèle.
[24] Voir les simulations avec un tableur :
http ://www.ac-grenoble.fr/maths/docresseconde/intervalles_fluctuation_confiance.xlsx
ou http ://www.ac-grenoble.fr/maths/docresseconde/intervalles_fluctuation_confiance.ods
[25] Pour n assez grand, on observera que $\frac{1}{\sqrt{n}}$ est l'ordre de grandeur de cette fluctuation autour de p au seuil 95%.
[26] Ici la simulation a été effectuée au tableur à l'aide de la formule ENT(ALEA()+0,4)

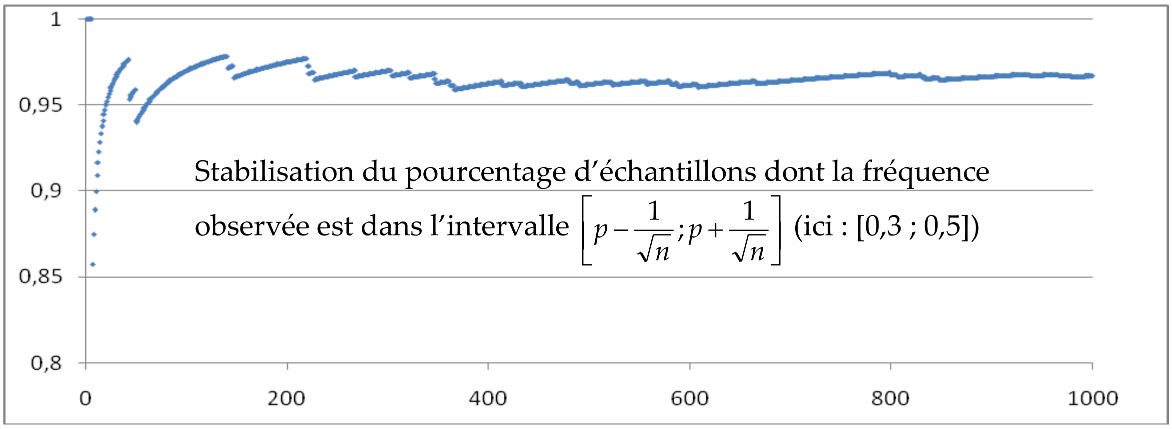
On peut alors définir[27] l'intervalle de fluctuation au seuil de 95%, relatif aux échantillons de taille n, comme l'intervalle centré autour de p, où se situe, avec une probabilité égale à 0,95, la fréquence observée dans un échantillon de taille n.



Dans la pratique, on utilise l'intervalle[28] $\left[p-\frac{1}{\sqrt{n}}\,;p+\frac{1}{\sqrt{n}}\right]$, pour des probabilités p comprises entre 0,2 et 0,8, et des échantillons de taille n supérieure ou égale à 25.

Pour p donné, on peut faire calculer les bornes de cet intervalle pour quelques valeurs de n, et remarquer qu'il faut multiplier la taille de l'échantillon par k^2 pour diviser par k l'amplitude de l'intervalle. On pourra calculer l'amplitude correspondant aux échantillons de taille 1000, taille souvent retenue dans les sondages.

Il est possible de visualiser[29] le pourcentage d'échantillons dont les fréquences d'apparition du 1 sont situées dans l'intervalle $\left[p-\frac{1}{\sqrt{n}}\,;p+\frac{1}{\sqrt{n}}\right]$. Le graphique[30] suivant est obtenu à partir des 1000 échantillons simulés et représentés ci-avant.



[27] Il faudrait en fait considérer le plus petit intervalle où se situent les fréquences observées avec une probabilité au moins égale à 0,95. Mais pour une première approche de cette notion, on s'est limité à l'énoncé ci-dessus.

[28] Il s'agit d'un résultat asymptotique, résultant de la convergence en loi de la variable aléatoire f_n correspondant à la fréquence d'un échantillon de taille n vers la loi normale de moyenne p et d'écart type $\sigma = \sqrt{p(1-p)}$. Ainsi, pour n assez grand, f_n appartient avec une probabilité d'environ 0,95 à l'intervalle $\left[p-\frac{1{,}96\sigma}{\sqrt{n}}, p+\frac{1{,}96\sigma}{\sqrt{n}}\right]$ qui est inclus dans $\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]$ (car $\sigma = \sqrt{p(1-p)} \leq \frac{1}{2}$). On pourra se référer au document d'accompagnement du programme 2001 de la classe de seconde à l'adresse :
http ://www.ac-grenoble.fr/maths/docresseconde/doc_proba_stat_seconde2001.pdf

[29] cf. feuille de calcul : http://www.ac-grenoble.fr/maths/docresseconde/intervalles_fluctuation_confiance.xlsx ou http://www.ac-grenoble.fr/maths/docresseconde/intervalles_fluctuation_confiance.ods

[30] Lecture : parmi les 200 premiers échantillons de taille 100 obtenus, environ 97% ont une fréquence d'apparition du 1 dans l'intervalle [0,3 ; 0,5]

2. Applications de la fluctuation d'échantillonnage

2.1. Prise de décision à partir d'un échantillon

On pourra développer quelques exemples de prises de décision sur échantillon fondées sur la connaissance de l'intervalle de fluctuation.

Exemple : la parité, c'est quoi ?
Deux entreprises A et B recrutent dans un bassin d'emploi où il y a autant de femmes que d'hommes, avec la contrainte du respect de la parité. Dans l'entreprise A, il y a 100 employés dont 43 femmes ; dans l'entreprise B, il y a 2500 employés dont 1150 femmes (soit 46%). Or 46% est plus proche de 50% que 43% : les chiffres parlent d'eux-mêmes, pourrait-on dire, et l'entreprise B respecte mieux la parité que l'entreprise A. Si on admet que la parité, c'est exactement 50% de femmes, il est vrai que B en est plus proche que A. Mais une telle définition, à l'unité près, de la parité n'aurait ici pas de sens.
La parité, cela signifie que l'identité sexuelle n'intervient pas au niveau du recrutement, c'est-à-dire qu'au niveau du caractère homme ou femme, les résultats observés pourraient être obtenus par choix au hasard des individus dans la population. Dans ce cadre, l'entreprise A est assimilable à un échantillon de taille 100 du modèle de Bernoulli (avec $p = 0{,}5$) dont l'intervalle de fluctuation est [0,4 ; 0,6] et l'entreprise B à un échantillon de taille 2500 dont l'intervalle de fluctuation est [0,48 ; 0,52]. La valeur 43% est donc dans l'intervalle de fluctuation, alors que 46% ne l'est pas. Autrement dit, pour l'entreprise B, la proportion de 46% s'observe dans moins de 5% des échantillons obtenus selon le modèle accordant une probabilité égale d'obtenir un homme et une femme. On peut alors rejeter l'hypothèse que cette entreprise respecte la parité.

En résumé, le raisonnement pour apprécier si une fréquence observée f sur un échantillon de taille n est compatible ou non avec un modèle de Bernoulli de probabilité p, est le suivant : on regarde si cette fréquence est dans l'intervalle de fluctuation à 0,95[31] relatif aux échantillons de taille n du modèle, c'est-à-dire si l'écart entre f et p est probable, au sens où le hasard produirait un tel écart dans 95% des échantillons envisageables. Si f est en dehors de l'intervalle de fluctuation, on considère que l'observation n'est pas compatible avec le modèle, en ce sens que dans un tel modèle elle ne s'observerait que dans 5% des échantillons de taille n (avec le risque de prendre la mauvaise décision dans 5% des cas). On s'interrogera alors, par exemple dans le contrôle de qualité industrielle, sur le réglage d'une machine lorsque dans un lot de pièces produites la fréquence de défauts observés est peu probable au regard du modèle indiquant une probabilité p de défauts.

Ce type de raisonnement, qui intègre la connaissance de l'intervalle de fluctuation, est à la base de ce que l'on appelle parfois une « preuve statistique ».

On trouvera de nombreuses activités pour la classe dans le document ressource relatif au nouveau programme de mathématiques de la voie professionnelle, disponible à l'adresse suivante : http://www.ac-grenoble.fr/maths/docresseconde/Proba_stat_LP.doc. On s'y interroge par exemple, à l'aune de la fluctuation d'échantillonnage, sur la fréquence des cas de leucémie chez les garçons de moins de 15 ans dans une petite ville des États-Unis (9 cas sur 5969, soit une fréquence de 0,0015 alors que la proportion constatée sur l'ensemble du territoire est de 0,00052). De même on questionne les 16 naissances masculines sur les 20

[31] Le seuil 0,95 est conventionnel. D'autres seuils sont retenus selon les questions ou domaines étudiés.

naissances d'un petit village de Chine ou encore les 46 naissances de garçons parmi les 132 naissances dans une petite ville du Canada. Dans ces différents exemples, le raisonnement statistique permet de mettre en évidence des observations « rares » au regard de la fluctuation d'échantillonnage, qui ont déclenché des investigations complémentaires, extérieures au champ des mathématiques, pour rechercher et analyser les causes des phénomènes observés.

Par ailleurs, une activité étudie, au regard des probabilités de défections constatées à l'embarquement, le risque de surréservation que prend une compagnie aérienne lorsqu'elle vend plus de billets que de places.

On y aborde aussi un exemple de contestation d'un jugement au Texas au motif que la désignation des jurés était discriminante à l'égard des américains d'origine mexicaine. Cette possibilité de quantifier, en termes de probabilité, si la fréquence observée (ici la proportion de jurés d'origine mexicaine) est rare ou fréquente dans un modèle considéré, fournit au juge des éléments qui, agrégés à d'autres arguments, fondent son jugement.

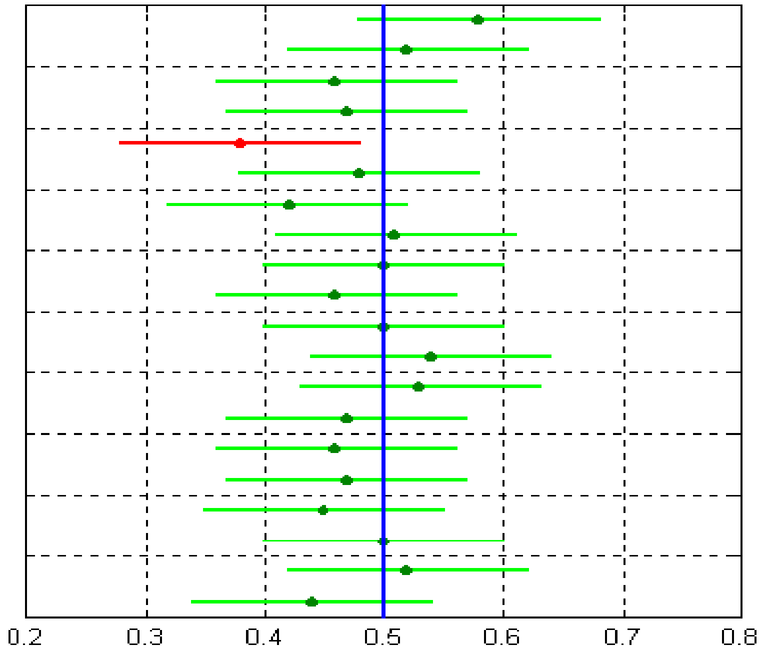
Exemple : élection annulée
Dans un entretien retranscrit dans le livre : "le hasard aujourd'hui ", Jean-Louis Boursin relate ce qui s'est passé lors d'une élection dans la région parisienne :
[… Un candidat aux élections législatives soupçonnait de fraude un certain nombre de bureaux de vote. Il pensait que dans ces bureaux là, il y avait un risque parce qu'il n'avait pas confiance en ceux qui tenaient les bureaux. Il a fait faire des sondages extrêmement précis, il a fait une étude sur les élections précédentes et, muni de ces chiffres et des résultats, centaine par centaine, il est allé au tribunal administratif et a affirmé que le hasard ne pouvait pas produire cela : " voilà une centaine de bulletins qui donnent 98 bulletins à mon adversaire et 2 à moi, cela n'est pas possible. Le calcul des probabilités montre que cela a une chance sur dix ou quinze millions de se produire ". Le tribunal lui a donné raison sur un simple raisonnement probabiliste…].

2.2. Estimation d'une proportion

On se place dans la situation de référence suivante, d'une urne contenant plusieurs centaines de petites billes de couleur blanche ou verte dans une proportion p inconnue de billes vertes. On cherche à estimer p à partir d'un échantillon de taille n (on imagine aisément des applications de cette situation, par exemple à l'estimation du résultat d'un référendum).

On considère alors un échantillon de taille n (par répétition de n tirages aléatoires avec remise dans l'urne) et on calcule la fréquence des billes vertes dans cet échantillon. On dispose ainsi d'un échantillon parmi tous ceux qu'on aurait pu obtenir et, d'après le paragraphe 1.2, on sait qu'environ 95% des fréquences observées sont dans l'intervalle $\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$. Or l'appartenance de f à $\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$ équivaut à celle de p à l'intervalle $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ (appelé aussi fourchette de sondage), ce qui permet de dire que, parmi tous les échantillons de taille n possibles, 95% des intervalles associés $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ contiennent le nombre p.

On a représenté ci-dessous[32], les fourchettes associées à 20 échantillons de taille 100 dans le cas p = 0,5. Un échantillon (en rouge) donne une fourchette ne contenant pas p, les autres (en vert) contiennent p.



Pour exprimer l'idée qu'avant tirage[33] de l'échantillon on avait 95% de chances d'obtenir une fourchette $\left[f-\frac{1}{\sqrt{n}}\,;f+\frac{1}{\sqrt{n}}\right]$ qui contienne p, on dira que la fourchette obtenue, une fois l'échantillon tiré, est un intervalle de confiance au niveau 0,95 de p. On notera que l'intervalle de confiance associé à un échantillon au niveau 0,95 ne dépend que de la taille n de l'échantillon et non pas de la taille de la population[34].

Exemple : estimation par sondage du pourcentage d'artisans parmi les entreprises d'un département.

À partir de données réelles[35] indiquant, pour les entreprises d'un département, si elles ont ou non le statut d'artisan, on peut estimer grâce à un échantillon de taille 1000 la proportion d'artisans. Un tirage aléatoire de 1000 entreprises[36] a donné 36,5% d'artisans. Ainsi,

[32] D'après le document d'accompagnement du programme de seconde publié en 2001 consultable à l'adresse : http://www.ac-grenoble.fr/maths/docresseconde/doc_proba_stat_seconde2001.pdf
Voir aussi des simulations :
http://www.ac-grenoble.fr/maths/docresseconde/intervalles_fluctuation_confiance.xlsx
ou http://www.ac-grenoble.fr/maths/docresseconde/intervalles_fluctuation_confiance.ods

[33] Il est clair qu'une fois l'échantillon tiré, l'intervalle de confiance associé est entièrement fixé et qu'il n'y a alors que deux possibilités : p appartient ou n'appartient pas à l'intervalle. C'est pourquoi il est incorrect de dire que p a une probabilité 0,95 d'appartenir à cet intervalle.

[34] Dans son livre " Les structures du hasard", Jean Louis Boursin explique ceci à l'aide d'une métaphore : étudier un échantillon c'est comme prendre une louche dans une grande marmite pour goûter la soupe, peu importe la taille de la marmite, on aura déjà beaucoup de renseignements sur la soupe à partir de cet échantillon.

[35] D'après l'INSEE, fichier à l'adresse :
http://www.ac-grenoble.fr/maths/docresseconde/Exemple de progression/activites_artisans.doc

[36] On trouvera un exemple d'algorithme qui permet d'extraire un échantillon d'une liste, utilisable aussi sur calculatrice : http://www.ac-grenoble.fr/maths/docresseconde/Annexe_algorithme_echantillons.doc

l'intervalle [33,3% ; 39,7%] est un intervalle de confiance au niveau 0,95 du pourcentage d'artisans dans ce département (la vraie valeur de 37,34% pouvant en l'occurrence être calculée). Cette situation réelle permet aux élèves de pratiquer une estimation par sondage statistique.

IV. Repères pour l'évaluation

La diversité des objectifs visés par l'enseignement des statistiques et des probabilités, tels qu'ils sont précisés dans le programme, invite à proposer des formes d'évaluation variées, prenant davantage en compte l'usage des TIC ou l'expression orale.

Bien sûr, de nombreuses capacités attendues peuvent aisément s'évaluer sous la forme habituelle d'un devoir en temps limité, comme par exemple tout ce qui a trait aux calculs de probabilités, aux représentations graphiques ou aux résumés statistiques.

En revanche, s'agissant de la fluctuation d'échantillonnage, l'objectif est de faire réfléchir les élèves à la conception et à la mise en œuvre d'une simulation et de les sensibiliser aux notions d'intervalle de fluctuation, d'intervalles de confiance et à l'utilisation qui peut en être faite. Aussi, semble-t-il prématuré d'exiger dans des contrôles écrits une autonomie totale des élèves pour conduire les raisonnements qui sont attachés à ces notions : on prendrait en effet le risque de restitutions par cœur pour compenser une assimilation naissante et encore fragile.

C'est pourquoi, l'évaluation des capacités attendues suivantes :

- concevoir, exploiter et mettre en œuvre des simulations de situations concrètes à partir d'un tableur ou d'une calculatrice,
- exploiter et faire une analyse critique d'un résultat d'échantillonnage,
- utiliser un logiciel ou une calculatrice pour étudier une série statistique,

devrait majoritairement être réalisée sous forme de comptes-rendus de travaux pratiques ou de devoirs à la maison. Ces modalités d'évaluations donnent accès à des situations statistiques ou probabilistes appelant l'usage d'un logiciel ou d'une calculatrice, et permettent, en raison de l'interaction possible entre le professeur et les élèves, d'ouvrir le champ de l'évaluation à des situations plus riches et plus ouvertes, qui mobilisent davantage les capacités de recherche, d'expérimentation et d'initiative.

Par ailleurs, concernant tout particulièrement ce chapitre, la place de l'oral gagnerait à être développée, tant cette forme de communication facilite, par le questionnement interactif qu'elle permet, l'explicitation de certains raisonnements statistiques délicats à consigner à l'écrit. Dans ce cadre, on peut envisager de proposer des situations dont l'étude est réalisée en classe et dont le compte-rendu, rédigé à la maison, est suivi d'un exposé en classe ou bien d'échanges avec le professeur permettant d'approfondir certaines argumentations ou démarches imparfaitement restituées à l'écrit afin de les améliorer.

L'expérience acquise lors de l'expérimentation de l'épreuve pratique de mathématiques[37] et les critères d'évaluation qui y ont été explicités constitueront de précieux points d'appui, transférables à ces nouvelles modalités d'évaluation.

---

[37] Cf. le site du groupe des mathématiques de l'inspection générale : http://igmaths.net/

Exemple : discrimination et statistique[38]

À partir de l'activité « contester un jugement » présentée en page 15 dans le document ressource relatif aux nouveaux programmes de lycée professionnel[39], on peut proposer un travail pratique en classe amenant les élèves à reformuler la question sur la discrimination éventuelle des jurés d'origine mexicaine (339 sur 870 jurés) dans un contexte statistique prenant en compte la fluctuation d'échantillonnage. Il s'agit alors de concevoir l'ensemble des jurys désignés à ce jour comme un des échantillons de taille 870 de la loi de Bernoulli de paramètre p, où p désigne la proportion au Texas d'américains d'origine mexicaine ($p \approx 0{,}791$). Cette reformulation étant effectuée, les élèves sont amenés ensuite à simuler des échantillons de taille 870 pour voir si la fréquence observée est dans la zone de fluctuation au seuil 95% ou bien ils peuvent utiliser l'intervalle $\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$. Il leur faut enfin conclure au regard du problème étudié.

Il est clair que ces différentes étapes peuvent donner lieu à des échanges oraux avec le professeur et à des productions écrites partielles qui peuvent être évaluées. L'évaluation peut porter, par exemple, sur la capacité des élèves à tirer profit du questionnement oral du professeur ou de ses indications pour reformuler ce problème de discrimination dans le cadre statistique de l'échantillonnage, ou aussi sur la capacité à mettre en œuvre une simulation sur tableur ou sur calculatrice, ou bien encore sur la capacité à interpréter les résultats de la simulation pour conclure. Tous ces éléments peuvent ensuite être intégrés dans l'évaluation globale de la situation étudiée, soit sous la forme d'une notation intermédiaire, soit par des appréciations qui seront prises en compte lors de la notation finale à la suite de la production écrite. D'une façon générale la notation ne peut être dissociée de l'évaluation des compétences acquises.

Vouloir évaluer sous une forme traditionnelle les capacités mobilisées dans de telles activités amènerait le professeur à proposer des questions intermédiaires qui limiteraient la prise d'initiative et l'intérêt que les élèves peuvent porter à l'étude de situations en lien avec la vie courante, avec des faits de société ou des questions scientifiques.

Signalons, pour terminer, l'idée originale qui a amené les professeurs de mathématiques du groupe « statistique et citoyenneté » de l'IREM de Paris Nord[40] à travailler avec des professeurs de français sur le même thème et à comparer la preuve statistique ainsi apportée aux critères d'un texte argumentatif en français[41].

Une bibliographie et des compléments pour le professeur sont disponibles à la fin du document ressource pour les lycées professionnels.

---

[38] Activité du groupe « statistique et citoyenneté » de l'IREM de Paris Nord présentée sur le site statistix (http://www.statistix.fr/spip.php?article27) et disponible aussi à l'adresse http://www.ac-grenoble.fr/maths/docresseconde/discrimination.doc.

[39] Voir document ressource pour les lycées professionnels : http://www.ac-grenoble.fr/maths/docresseconde/Proba_stat_LP.doc

[40] Les travaux de ce groupe sont disponibles à l'adresse : http://www-irem.univ-paris13.fr/spip/spip.php?rubrique15

[41] Cf. http://www.ac-grenoble.fr/maths/docresseconde/discrimination.doc