Conseil Scientifique des IREM – Université Paris Diderot – 5 décembre 2014

Débat sur le thème :

« Enseignement de la statistique et éducation à la citoyenneté »

Témoignage de Jeanne Fine[1]

Dans le cadre de ce débat, il m'a été demandé de témoigner :
- de mon expérience avec le secondaire, en particulier à propos des sondages,
- de la vision internationale de l'enseignement de la statistique[2] dans le secondaire.

Présentation rapide de mes activités en lien avec le thème :
De 1973 à 1993, j'enseigne la statistique à l'université de lettres et sciences humaines de Toulouse. De 1993 à 2010, professeure à l'IUFM Midi-Pyrénées, j'enseigne la statistique dans le cadre de la préparation aux capes de mathématiques et de sciences économiques, je crée un centre de statistique au sein de l'institut pour une aide à la recherche et pour réaliser des études et des enquêtes, je suis chargée de mission à l'égalité entre les filles et les garçons puis chargée de mission à l'évaluation des formations. De 2011 à 2014, j'ai été corédactrice en chef de la revue *Statistique et Enseignement*[3]. J'ai profité de cette tribune pour exposer mes propositions pour l'enseignement (les « pourcentages », « l'approche sondage ») et développer des thèmes qui m'intéressent tout particulièrement : « interdisciplinarité », « curriculum statistique dans le secondaire en France et à l'étranger », les « MOOC ». Par ailleurs, je poursuis une activité de conseil et de formation en statistique en entreprises ou collectivités territoriales ; cette formation d'adultes m'apporte un autre éclairage sur la formation à la statistique.

---

1. Introduction

Les deux termes du débat sont « éducation à la citoyenneté » et « enseignement de la statistique » ; plus largement, la question est la suivante : «comment assurer un socle commun de connaissances, de compétences et de culture à tous les élèves et fournir des bases solides en mathématiques à ceux qui poursuivront des études scientifiques ? »
Il me semble qu'il existe encore un large champ de manœuvre pour former l'ensemble des élèves de collège aux « mathématiques pour tous » et qu'un plus grand nombre d'entre eux puisse approfondir le formalisme mathématique leur permettant de se spécialiser.
Je développerai ci-après la question des « pourcentages », « l'approche sondage », la comparaison du curriculum statistique français et du curriculum proposé dans le rapport GAISE de l'*American Statistical Association* (U.S.) avant de revenir aux deux termes du débat et de conclure.

2. Les pourcentages

On trouvera en annexe 1 une page sur « les pourcentages », première page distribuée à mes interlocuteurs dans le cadre de mes formations. On définit « fréquence, proportion, probabilité » et leurs propriétés, « taux de comparaison, taux d'évolution » et leurs propriétés avant d'introduire l'écriture du résultat en « pourcentage ». Pour un développement des raisons m'ayant conduite à insister sur ce point, je renvoie le lecteur à mon libre-propos *Quelle est votre définition de « pourcentages » ? Proposition pour l'enseignement* [1].

---

[1] Statisticienne, Toulouse – jeanne.fine@gmail.com
[2] Dans tout le texte l'enseignement de « la statistique » est indissociablement lié à celui des « probabilités »
[3] http://www.statistique-et-enseignement.fr

3. L'approche sondage en statistique

Nous n'avons malheureusement en français qu'un seul mot pour désigner la théorie des sondages aléatoires et les sondages d'opinion. La théorie des sondages (« sampling » en anglais, « muestreo » en espagnol) concerne l'échantillonnage et l'estimation dans les populations finies et aborde le sondage aléatoire simple ou stratifié, proportionnel ou non, avec ou sans remise, sondage par grappes, redressement d'échantillon, la prise en compte du défaut de couverture ou des non-réponses… Le sondage d'opinion (« poll » en anglais, « sondeo » en espagnol) est largement popularisé par les sondeurs qui, grâce à la méthode des quotas, prétendent s'appuyer sur la théorie des sondages pour justifier leur pratique.
Ces deux aspects sont développés dans l'article *Les sondages : délaissés par les statisticiens et malmenés par les politologues* [2], support d'une communication que j'ai faite lors de la journée *Statistique et Sondages* organisée à Toulouse en mars 2007 à l'intention des professeurs de mathématiques du secondaire. Le contenu de cette communication a été retravaillé par Philippe Dutarte et le groupe IREM Paris-Nord pour une utilisation plus directement accessible par les professeurs [3]. On trouvera une présentation moins formelle de la comparaison entre sondage aléatoire et sondage par quotas dans un numéro spécial sur les sondages préélectoraux de la nouvelle revue *Statistique et Société* [4].

Pour revenir à la théorie des sondages, c'est en 1986 que je découvre cette théorie dans le cadre de *Journées d'Études en Statistique* organisées par la *Société Française de Statistique* pour ses adhérents. Cette partie de la statistique n'était quasiment pas enseignée à l'université car considérée comme relevant de la statistique publique (INSEE, INED, …). Cette approche m'a, au contraire, parue lumineuse pour mieux comprendre la statistique et l'enseigner autrement. J'en développe les grandes lignes dans l'article « *Probabilités et statistique inférentielle : approche sondage versus approche modèle* » [5].

La statistique enseignée dans le secondaire ou à l'université en France ne traite quasiment que des variables réelles et repose sur la modélisation probabiliste et l'échantillonnage i.i.d. (variables aléatoires indépendantes et identiquement distribuées) (*approche modèle*) alors que la statistique des sciences humaines et sociales concerne très souvent des variables catégorielles définies sur des populations finies, populations desquelles il est possible d'extraire des échantillons en utilisant des procédures aléatoires (*approche sondage*).

La théorie des sondages modifie profondément les concepts étudiés en probabilités et statistique inférentielle. Dans l'approche sondage, aucune hypothèse n'est faite sur la distribution de probabilité ; l'aléatoire vient de l'échantillonnage (procédure aléatoire de génération de l'échantillon). L'approche sondage permet de sortir du paradigme de l'approche modèle dans des contextes simples, de redonner du sens au vocabulaire de la statistique (population, individus, échantillons indépendants et échantillons appariés, …), de faire la place aux variables catégorielles et à la statistique des sciences humaines et sociales.

Le débat est repris dans le numéro sur le thème de l'interdisciplinarité. Dans le premier article de ce numéro spécial paru en décembre 2012, intitulé *La preuve par les chiffres (evidence based) : de quoi s'agit-il ?* [6], Claudine Schwartz développe, à partir de nombreux exemples, la lente construction d'une preuve scientifique. Dans une deuxième partie de l'article, elle décrit les images et le vocabulaire à employer pour l'approche de la notion de test : elle revient sur l'histoire de la théorie des tests (les *tests de signification* de Fisher et les *tests d'hypothèses* de Neyman et Pearson) et propose d'unifier leur présentation ; par ailleurs, elle critique l'utilisation de l'expression « vraie valeur du paramètre » et propose de « déconstruire la métaphore des urnes », ajoutant que, « aujourd'hui, le cadre scientifique est celui de la manipulation de modèles ». Cf. aussi sur le site *Statistix*[4], ses articles n° 90 et 91. Elle montre bien la nécessité de maîtriser le cadre de la modélisation (même en théorie des sondages, des modèles sur la population peuvent être utilisés) ; mais c'est d'un point de vue pédagogique qu'il me semble préférable de repérer les différents contextes plutôt que de les gommer. Dans ma contribution aux débats [7], je prône la double approche et insiste sur les différents types de données (données de sondages aléatoires, données issues d'expériences aléatoires, données d'observation ne relevant pas des deux premières catégories). C'est le point de vue développé dans le rapport GAISE présenté ci-après.

[4] http://www.statistix.fr

4. Comparaison du curriculum statistique français et du curriculum statistique du rapport GAISE (US)

Le rapport GAISE [8], édité aux États-Unis par l'*American Statistical Association* (*ASA*), présente un cadre pour un curriculum statistique de la maternelle à l'université. Dans le numéro spécial de la revue *Statistique et Enseignement* sur « le curriculum statistique dans le secondaire et comparaisons internationales », paru en mars 2013, je propose [9] une traduction de la présentation du cadre général du rapport GAISE, un résumé de ce cadre (cf. annexe 2) et une comparaison avec le curriculum français dont voici les grandes lignes.

En statistique et probabilités, le curriculum français est assez limité au niveau du collège et, au contraire, très détaillé au niveau du lycée mais sur un nombre restreint de notions. De plus ces notions sont déconnectées entre elles et l'aspect calculatoire l'emporte sur le sens et l'interprétation.

En effet, c'est en troisième qu'il est fait, pour la première fois, référence à la statistique et qu'une initiation aux probabilités est proposée. Avant ce niveau d'enseignement, il est question de traitement de données sans se préoccuper de l'origine ou du recueil des données, il est proposé de travailler la mise en forme de tableaux et de graphiques sans en préciser le contenu : s'agit-il d'effectifs ? de données brutes ? de données agrégées ? On introduit le vocabulaire de la statistique, effectifs, fréquences, catégories, … sans définir ces notions. Aussi, dans les manuels scolaires ou les épreuves d'évaluation, l'utilisation du vocabulaire est parfois faite à contresens et les questions sont souvent mal posées. Les exigences en collège sont très faibles, surtout en regard des objectifs affichés dans l'introduction aux programmes.

Au lycée, dans la série scientifique, l'enseignement des probabilités est assez déconnecté de l'enseignement de la statistique : l'indépendance de deux événements et les exercices s'y référant est déconnectée de la notion d'échantillon (un échantillon de taille *n* est constitué des résultats de *n* répétitions indépendantes de la même expérience). La notion de probabilité conditionnelle est déconnectée de la notion de fréquence conditionnelle, puisque le traitement statistique d'un couple de variables catégorielles n'est pas au programme.

Par manque de temps, l'initiation à la statistique inférentielle est centrée sur l'estimation et le test pour une « proportion »…

L'exigence d'un programme linéaire et cumulatif, reposant sur des définitions mathématiques et admettant le moins possible de résultats non démontrés, aboutit à travailler des notions inutiles pour une formation à la statistique…

L'image de la statistique en fin de curriculum est surtout bien étroite : pas de vue d'ensemble des différents types de données et des différents objectifs de la statistique (en particulier la comparaison de deux groupes ou l'association entre deux variables), pas de formation à la démarche scientifique (formuler la question, recueillir des données, analyser les données, interpréter les résultats).

À l'opposé, le curriculum proposé dans le rapport GAISE est rédigé sans pratiquement aucun formalisme. Le contexte, le recueil et l'interprétation des données ainsi que le repérage du type de variabilité forment le cadre du curriculum. C'est une vision synthétique de la statistique qui est ainsi proposée. Il s'agit ensuite d'approfondir tel ou tel aspect et de revenir en boucle sur les apprentissages.

Le curriculum proposé répond à trois objectifs de niveau d'exigence croissant :
- une initiation aux concepts, idées, terminologie et techniques de base de la statistique,
- une formation du futur citoyen à la littératie statistique (comprendre l'information quantitative, en avoir une lecture critique, savoir argumenter et communiquer à partir de données),
- une initiation à la démarche scientifique.

5. Éducation à la citoyenneté et enseignement de la statistique

Réaliser une enquête avec les élèves

Afin d'introduire les concepts de base de la statistique des sciences humaines et sociales (variables catégorielles, tableaux d'effectifs croisés, liens entre les variables, travail sur les fréquences), réaliser une enquête avec les élèves est une bonne stratégie. En effet, les données disponibles sur les sites institutionnels sont des données « agrégées », ce sont des moyennes ou proportions de différentes catégories de la population. Leur traitement statistique n'est pas toujours bien intéressant d'un point de vue pédagogique. De plus, il n'est pas toujours facile de savoir comment ces « données » ont été produites, sur quelles « conventions » elles ont été construites.

J'ai mené des enquêtes avec des étudiants de Deug Mass (Mathématiques et Sciences Sociales) ; l'expérience avait été très appréciée des étudiants mais très coûteuse en temps de travail pour l'avancée du programme dans le volume horaire imparti. En particulier, la construction du questionnaire est longue est difficile.

Début 2006 à l'IUFM, lorsque des collègues m'ont sollicitée pour un PER (Parcours d'Étude et de Recherche) en statistique dans le cadre d'une recherche collaborative de l'INRP, nous avons repris l'idée de l'enquête. Nous avons tenté de rejoindre le projet international « Census at school », créé en 2000 au Royaume-Uni, dont une version française « Recensement à l'école » est accessible sur le site de la Société statistique du Canada[5]. On trouvera une présentation du projet par Brigitte Chaput sur le portail des IREM [10]. Mais des critiques se sont élevées : nous sommes en France très vigilants sur la protection des données personnelles. Rejoindre le projet sans accompagnement institutionnel et sans mise en garde des professeurs sur les questions sensibles (taille, poids, religion, origine, mode de vie…) est dangereux. Le danger reste entier pour les enquêtes par questionnaire proposées par des professeurs tentés par cette entrée de la statistique. Pourtant, même en 2006, il était possible de choisir, parmi les activités proposées autour du projet « Recensement à l'école », celles qui utilisaient une partie non discutable du questionnaire mais nous avons préféré renoncer à ce projet.

Davantage tourné vers la théorie des sondages, avec des sujets, des données et un glossaire, l'IREM de Dijon et la *Société Française de Statistique* ont organisé le « Challenge Graines de Sondeur »[6] pour les lycéens de l'Académie de Bourgogne à l'occasion du 8ème congrès francophone sur les sondages qui vient de se dérouler (novembre 2014).

La statistique pour le citoyen

A la rentrée 1973, une unité de valeur obligatoire de « mathématiques » en première année de sociologie, psychologie, histoire, géographie est créée. C'est pour mettre en place cet enseignement que je rejoins l'équipe pédagogique. Sans perdre de vue les contenus que nous nous étions fixés, notre objectif était que les étudiants puissent « maîtriser » les informations quantitatives, tableaux et graphiques présentés dans les journaux d'information. C'est ce que l'on appelle aujourd'hui la « littératie statistique ». Tous les étudiants n'y ont pas pris le même intérêt mais beaucoup d'entre eux, et surtout d'entre elles, ont repris confiance dans leur capacité à faire des « mathématiques ». Quant au contenu statistique, il est difficile de faire l'impasse sur les concepts présentés dans l'annexe sur les pourcentages.

Au niveau du lycée, c'est sur le site de statistique de Philippe Dutarte, cité en [3], que l'on trouve des activités très intéressantes et directement utilisables par les professeurs sur le thème *Statistique et citoyenneté* et il propose des exemples plus récents dans les ateliers des Journées de l'APMEP. On a déjà cité aussi le site *Statistix*. Le problème est la maintenance de ces sites et la difficulté pour les nouveaux professeurs de se repérer dans la multitude des travaux en ligne.

Les statistiques sexuées pour l'éducation à l'égalité entre les filles et les garçons à l'école

C'est en janvier 1999, un an avant la signature de la première « convention interministérielle sur l'égalité entre les filles et les garçons, les femmes et les hommes dans le système éducatif », qu'avec une dizaine de formatrices et formateurs de l'IUFM Midi-Pyrénées et de professeurs en postes, nous créons une équipe de recherche et formation sur ce thème. Quatre axes de recherche sont retenus : la représentation des hommes et des femmes dans les livres scolaires et la littérature enfantine, l'orientation dégagée des stéréotypes sexués, en particulier l'orientation scientifique et technique des filles, la construction de l'identité sexuée, l'éducation à la mixité scolaire et l'éducation à la citoyenneté. Des collègues faisaient de même à l'IUFM de Lyon. Il y avait à l'époque très peu d'études sur la mixité à l'école et aucun document pédagogique pour aborder la question.

[5] http://www.censusatschool.ca/fr/
[6] http://sondages2014.sfds.asso.fr/graines-de-sondeur/

Aujourd'hui, les supports de formation existent et sont largement diffusés ; le problème n'est pas dans la production de données statistiques mais dans leur interprétation car on touche à sa propre *représentation du masculin et du féminin*. Par exemple, pour le choix d'un métier, une fois que l'on a fait le constat d'une répartition très fortement sexuée des élèves selon les sections des lycées, la difficulté est de répondre à la question « où est le problème si c'est le choix des garçons d'aller vers les sciences et les techniques et des filles d'aller vers le social et le soin ? »

6 Conclusion

J'ai volontairement élargi le débat proposé vers la question : «comment assurer un socle commun de connaissances, de compétences et de culture à tous les élèves et fournir des bases solides en mathématiques à ceux qui poursuivront des études scientifiques ? ». Car, le problème aujourd'hui, mais ce peut être une chance, est la révolution numérique qui, d'une part, transforme la société et les attentes vis-à-vis de l'école, d'autre part, bouleverse le fond et la forme des enseignements et des apprentissages.

Les attentes vis-à-vis de l'école changent ; on a du mal à trouver une véritable référence aux mathématiques dans le projet de socle commun de connaissances, de compétences et de culture (cf. sur le portail des IREM les commentaires de Michèle Artigue et Jean-Pierre Raoult [11]). Le nombre d'heures de cours de mathématiques en première et terminale S diminue de réforme en réforme pour faire place à des activités en interdisciplinarité ou pour introduire probabilités et statistique, ce qui ébranle les bases d'une solide formation de scientifiques. Aujourd'hui, l'introduction de l'informatique inquiète la communauté mathématique.

Grâce aux MOOC (Massive Online Open Courses), aux environnements numériques de travail dans les universités ou dans les académies, aux sites d'exercices corrigés en libre accès, à la « classe inversée » (les notions et les exercices d'applications directes sont faits à la maison, les problèmes plus complexes encadrés par le professeur dans la classe), il est permis d'espérer proposer une formation à l'école à la fois plus générale (ce qui ne veut pas dire sans contenu et sans concept) et plus approfondie sur certains sujets, que ce soit en mathématiques ou dans les autres disciplines.

Pour finir sur cette autopromotion de mes articles sur l'enseignement de la statistique, je renvoie à l'article intitulé *Statistique, informatique, mathématiques et interdisciplinarité* [12] dans lequel je développe les arguments présentés ici et dont voici les dernières lignes :

> Afin que les mathématiciens aient une vue d'ensemble des sciences mathématiques, c'est dès les premières années de licence qu'une initiation à l'informatique et aux mathématiques appliquées (analyse numérique, probabilités, statistique) ainsi qu'une formation à l'épistémologie et l'histoire des sciences devraient être introduites. Une plus grande unité disciplinaire des mathématiques et un plus grand intérêt de la part des mathématiciens pour les autres disciplines et pour les questions relevant de l'enseignement sont nécessaires pour la formation des futurs citoyens et des futurs mathématiciens.

Références

[1] Fine, J. (2013). Quelle est votre définition de « pourcentages » ? Proposition pour l'enseignement. *Statistique et Enseignement*, Vol. 3, n° 2, 87-97.
En ligne : http://publications-sfds.math.cnrs.fr/ojs/index.php/StatEns/article/view/188/174

[2] Fine, J. (2007). Les sondages : délaissés par les statisticiens et malmenés par les politologues. *Journée Statistique et Sondages*, Toulouse, 13 mars 2007, http://jeannefine.free.fr/Sondages-Toulouse2007/
En ligne : http://jeannefine.free.fr/Sondages-Toulouse2007/documents/doc_JFine.pdf

[3] Dutarte, P. et al (2007) Statistique et citoyenneté. Brochure n° 135 de la commission inter-IREM Lycées Technologiques. Partie III. Sondages 75-102
En ligne : http://dutarte.perso.neuf.fr/statistique/brochure135.pdf

[4] Fine, J. (2013). Pour une plus grande transparence sur la méthodologie des sondages électoraux. *Statistique et Société*, Vol. 1, n° 2, 23-28.
En ligne : http://publications-sfds.fr/index.php/stat_soc/article/view/195

[5] Fine, J. (2010). Probabilités et statistique inférentielle : approche sondage versus approche modèle. *Statistique et Enseignement*, Vol. 1, n° 2, 5-21.
En ligne : http://publications-sfds.math.cnrs.fr/ojs/index.php/StatEns/article/view/14

[6] Schwartz, C. (2012). La preuve par les chiffres (evidence based) : de quoi s'agit-il ?, *Statistique et Enseignement*, Vol. 3, n° 2, 3-21
En ligne : http://publications-sfds.math.cnrs.fr/ojs/index.php/StatEns/article/view/125

[7] Fine, J. (2013). L'enseignement de la statistique en interdisciplinarité. Contribution aux débats, *Statistique et Enseignement*, Vol. 2, n° 2, 77-86.
En ligne : http://publications-sfds.math.cnrs.fr/ojs/index.php/StatEns/article/view/187

[8] Franklin et al (2007). Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report, a pre-K-12 curriculum framework. American Statistical Association (ASA), Alexandria, VA 22314
En ligne: http://www.amstat.org/education/gaise/

[9] Fine, J. (2013). Le rapport GAISE (U.S.) Cadre d'un curriculum statistique de la maternelle à la Terminale *Statistique et Enseignement*, Vol. 3, n°1, 25-54.
En ligne : http://publications-sfds.math.cnrs.fr/ojs/index.php/StatEns/article/view/137

[10] Chaput, B. (2011) Présentation du projet « Census at school – Recensement à l'école »
http://www.univ-irem.fr/spip.php?article498

[11] Commentaires sur le projet de socle commun de compétences, de connaissances et de culture par M. Artigue et J.-P. Raoult. Portail des IREM.
http://www.univ-irem.fr/IMG/pdf/Socle_Commentaires_de_M_Artigue_et_JP_Raoult.pdf

[12] Fine, J. (2012). Statistique, informatique, mathématiques et interdisciplinarité, *Statistique et Enseignement*, Vol. 3, n°2, 33-59
En ligne : http://publications-sfds.math.cnrs.fr/ojs/index.php/StatEns/article/view/127

Les pourcentages

Fréquence, proportion, probabilité

En statistique
Si la population est de taille (ou d'effectif) *N* et *A* un sous-ensemble de taille *n* alors la fréquence (ou proportion) de *A* (relativement à la population) est le rapport *n/N*.
La fréquence d'un ensemble est comprise entre 0 et 1 et vérifie la propriété d'additivité : si *A* et *B* sont deux ensembles sans élément commun, la fréquence de la réunion des deux ensembles est la somme des fréquences.

En probabilité
Les propriétés de la fréquence sont également vérifiées par les probabilités finies.

Taux de comparaison et taux d'évolution

Pour comparer une quantité **positive** x entre deux situations *A* et *B*, on peut utiliser le *taux de comparaison* relativement à *A*, défini par le rapport (on suppose x_A non nul) :

$$\frac{x_B - x_A}{x_A}$$

En particulier, s'il s'agit d'une quantité positive x évoluant dans le temps entre les instants 0 et 1, on peut utiliser le *taux d'évolution* (relativement à l'instant 0), défini par le rapport (on suppose x_0 non nul) :

$$\frac{x_1 - x_0}{x_0}$$

Ces taux peuvent être négatifs (alors compris entre – 1 et 0) et peuvent être supérieurs à 1 (jusqu'à l'infini

Si *t* désigne le taux, alors le nombre *m* défini par *m* = 1 + *t* est appelé *multiplicateur associé au taux t*.

Les taux vérifient certaines propriétés ; en particulier pour les évolutions successives. Si une quantité **évolue** au taux *t* entre les instants 0 et 1 et au taux t' entre les instants 1 et 2, alors le taux d'évolution t'' entre les instants 0 et 2 vérifie : $1 + t'' = (1 + t)(1 + t')$. On n'additionne pas les taux, on multiplie les multiplicateurs associés.

Où sont passés les pourcentages ?

Le pourcentage est une écriture :

$$\frac{1}{4} = 0{,}25 = \frac{25}{100} = 25\ \%$$

Les fréquences, proportions, probabilités, taux de comparaison, taux d'évolution s'écrivent en pourcentage. Il s'agit d'un choix d'unité de mesure :

- les effectifs peuvent être exprimés en milliers ou en millions
- les proportions et les taux peuvent être exprimés en % ou en ‰

Il est conseillé de n'utiliser l'écriture en pourcentage que dans la présentation des tableaux et des résultats et non pour les calculs.

Vigilance dans la lecture et l'utilisation des pourcentages

Pour certains le pourcentage est le numérateur sur 100
Pour d'autres c'est un abus de langage pour fréquence, proportion, probabilité ou taux

Le rapport GAISE (U.S.)

En ligne sur le site de l'ASA (American Statistical Association) http://www.amstat.org/education/gaise/

Description succincte du rapport

Le rapport GAISE donne un cadre conceptuel d'une formation en statistique avec trois niveaux de développement. Ces trois niveaux, A, B et C, peuvent être pensés comme un programme pour le primaire (A), le collège (B) et le lycée (C). Ces niveaux reflètent en fait les étapes de l'apprentissage en statistique et pas seulement l'âge des élèves ; aussi, les lycéens ou étudiants qui n'ont pas eu de formation antérieure auront besoin de passer par les niveaux A et B avant d'aborder le niveau C.

Pour chaque niveau de développement, les élèves rencontreront les quatre étapes de la *démarche statistique* (appelées aussi dans le rapport les quatre composantes du processus de recherche pour la résolution d'un problème statistique) :

- formuler une question,
- collecter des données,
- analyser les données,
- interpréter les résultats.

Une attention particulière est portée à chaque niveau sur l'interprétation des résultats en revenant au *contexte de la question* posée au départ.

L'omniprésence de la variabilité dans les données est un des thèmes repris dans les trois niveaux ; les élèves sont amenés à repérer les sources de variabilité dans leurs données.

Différents types de variabilité sont introduits sur des exemples et repris en détails tout le long du cursus :

- la *variabilité des données de mesures* (plusieurs mesures d'une même grandeur),
- la *variabilité naturelle* d'une grandeur mesurée sur différentes unités statistiques d'une population,
- la *variabilité induite* par différents niveaux de facteurs (dans le cadre, par exemple, de plans d'expériences),
- la *variabilité due au « hasard » dans le cadre d'échantillonnages aléatoires,*
- la *variabilité due au « hasard » lors d'une assignation aléatoire des unités expérimentales à des groupes dans le cadre de plans d'expériences*.

Il y est question aussi :

- de *variabilité intragroupe* et de *variabilité intergroupe*,
- de *co-variabilité*,
- de *variabilité résiduelle* dans le cadre de l'ajustement d'un modèle.

Sont nommées « *données d'observation* » les données qui ne sont pas issues d'une sélection aléatoire d'un échantillon d'unités statistiques ou de l'assignation aléatoire de traitements à des unités expérimentales ; le traitement de ces données diffère de celui de données obtenues selon des procédures aléatoires.

Le rapport insiste sur le fait qu'une formation à la statistique devrait présenter ces différents types de variabilité car, dans la pratique de la statistique, la résolution d'une question statistique et la prise de décision dépendent de la compréhension, de l'explicitation et de la quantification de la variabilité dans les données.

Le cadre conceptuel ne détaille pas les concepts et méthodes permettant d'analyser les données recueillies. C'est dans les pages suivantes du rapport que nous trouvons ceux préconisés à chaque niveau de la formation en statistique. Nous les résumons ici.

Au niveau A, sont présentés, sur de nombreux exemples, les *tableaux d'effectifs* et les *diagrammes en barres* pour une variable catégorielle, le *diagramme en points* (dotplot) et le *diagramme tige et feuilles* (stem and leaf plot) pour une variable numérique (éventuellement pour deux groupes), le *diagramme de dispersion* (scatterplot) pour un couple de variables numériques (appelé aussi graphe plan ou graphe X-Y d'un nuage de points), un *graphe chronologique* (timeplot) pour une variable numérique dépendant du temps, des *tableaux d'effectifs à double entrée*, les indices de centralité (*moyenne, médiane, mode*) et de dispersion (*étendue*) pour une variable numérique, la *modalité modale* pour une variable catégorielle. Il s'agit bien d'un premier niveau, sans doute bien ambitieux pour le primaire !

En ce qui concerne les probabilités au niveau A, il est indiqué que les élèves doivent comprendre que *la probabilité est une mesure des chances que quelque chose se réalise. C'est une mesure du certain ou de l'incertain.* Les évènements pourraient être placés sur un segment gradué de 0 (événement jugé impossible) à 1 (événement jugé certain) ; pour 1/2, il est jugé également probable que l'événement se réalise ou non ; de part et d'autre de 1/2 sont

situés les événements dont la réalisation est jugée moins probable (avant 1/2) ou plus probable (après 1/2) que leur non réalisation.

Les élèves doivent réaliser des expériences pour estimer des probabilités à partir de fréquences calculées sur des données empiriques ; les lancers de pièces ou de dés, ou de roues de loterie, sont les outils utilisés à ce niveau.

Au niveau B, sont présentés, pour les variables numériques, les *distributions d'effectifs* et de *fréquences*, les *histogrammes*, les *quartiles*, *l'écart interquartile* et *l'écart absolu moyen*, le *diagramme en boîte* (boxplot) à partir des cinq indices *min, premier quartile, médiane, troisième quartile, max*.

Des indices sont proposés pour mesurer l'association entre deux variables catégorielles ayant chacune deux modalités. Ces indices sont calculés en fonction des quatre effectifs de la table de contingence.

Ils sont réutilisés pour mesurer l'association entre deux variables numériques : la table de contingence 2 x 2 est construite en répartissant les observations au-dessous et au-dessus de la moyenne de chacune des deux variables. Si on note *a, b, c, d* ces effectifs et *n* l'effectif total, alors l'indice $[(a+d)-(b+c)] / n$ fournit une première indication sur l'association entre les deux variables.

Moins naturel est l'indice $Phi = (ad-bc)/(l_1 l_2 c_1 c_2)^{1/2}$ où l_1, l_2, r_1, r_2 désignent la somme des effectifs des lignes 1 et 2 et des colonnes 1 et 2. Il s'agit d'un indice descriptif (c'est-à-dire ne dépendant pas de l'effectif total *n*) ; il mesure un écart entre le tableau de fréquences observées et le tableau d'indépendance entre les deux variables. On a la relation $Phi^2 = Khi^2/n$ où Khi^2 est la statistique du test d'indépendance des deux variables. On peut vérifier que, si l'on remplace les variables catégorielles par l'indicatrice de la première modalité, alors le coefficient de corrélation linéaire des deux indicatrices est égal à *Phi*.

A partir du graphe plan d'un nuage de points représentant les observations de deux variables numériques, est abordée la modélisation par ajustement linéaire, dans ses aspects descriptif et prédictif. C'est tout d'abord à l'œil qu'il est proposé de tracer une droite passant « au mieux » par le nuage de points puis en utilisant les centres de gravité des deux sous-nuages définis à partir de la médiane de la variable explicative (droite de Mayer).

C'est au niveau B que la fluctuation d'échantillonnage d'une moyenne d'échantillonnage ou d'une fréquence d'échantillonnage est étudiée, par simulation.

Au niveau C, sont introduites les notions de *variance*, *écart-type*, *coefficient de corrélation linéaire* et *droite d'ajustement par les moindres carrés*.

L'étude des résultats d'un plan d'expériences est réalisée à partir de résumés numériques et de graphiques.

La marge d'erreur, à un niveau de confiance donné, associée à l'estimation d'une moyenne (ou d'une proportion) à partir d'un échantillon aléatoire, est étudiée par simulation. Le test d'hypothèse d'égalité d'une moyenne (ou d'une proportion) à une valeur fixée est abordé ainsi que la notion de *p*-valeur. La *p*-valeur est définie comme la probabilité d'observer un résultat comparable ou plus extrême que celui observé quand la valeur de cette moyenne (ou de cette proportion) dans la population est égale à la valeur fixée et elle est approchée par simulation.

En probabilités, les notions *d'espérance mathématique, variance, écart-type d'une variable aléatoire numérique* sont directement obtenues à partir des notions analogues de la statistique descriptive. L'accent est mis sur *l'indépendance en probabilité* et sur l'utilisation des probabilités pour prendre des décisions et tirer des conclusions.

Outre les notions et outils rappelés ci-dessus, les auteurs du rapport insistent sur l'importance de différencier les différentes sources de données :

- données obtenues à partir d'un sondage aléatoire,
- données obtenues à partir d'un plan d'expériences aléatoires,
- données d'observation (celles qui ne viennent pas d'un plan de recueil aléatoire, sondage ou plan d'expériences).

Ils ajoutent que la plupart des questions auxquelles il est possible de répondre à partir d'un recueil, de l'analyse et de l'interprétation des données exigent un plan de recueil aléatoire : sondage aléatoire ou plan d'expériences aléatoire. Ces deux plans de collecte ont des objectifs différents mais tous deux permettent de réduire les biais et d'utiliser l'inférence statistique pour l'estimation de la marge d'erreur et pour l'évaluation de la *p*-valeur d'un test d'hypothèse. Les données d'observation (et études observationnelles) peuvent néanmoins fournir des indications sur les distributions des variables d'intérêt et sur les associations qui peuvent exister entre elles, mais une relation de cause à effet ne peut pas être établie. C'est parfois la seule source disponible de données. Des exemples commentés illustrent cette partie du rapport.