Association Mathématique du Québec

Jeux de données statistiques au collégial

Robert Bilinski,
Collège Montmorency

Résumé

Avec les différentes réformes, tant au collégial qu'à l'université, la finalité des cours de statistiques et de méthodes quantitatives donnés au cégep a changé. En effet, ils sont maintenant, hélas, pour une part non négligeable des étudiants, les derniers cours de ce genre avant une pratique éventuelle.

Ce texte est un plaidoyer pour incorporer des données réelles dans les cours afin d'aborder des situations de base dans la pratique statistique qui ne sont pas couvertes dans les livres et dans la plupart des cours présentement offerts.

1 Introduction

Avec l'apparition des programmes universitaires en sciences humaines « sans mathématiques » de l'UQAM (gestion, sociologie, etc.), nous sommes confrontés à une nouvelle réalité qui influence inévitablement notre planification de certains cours de cégep : nos cours de méthodes quantitatives et de statistiques peuvent être les derniers lieux d'exposition d'un contenu statistique avant son utilisation par les étudiants dans leur « profession ».

De plus, avec une vision « simpliste » de la statistique au secondaire[1] qui couvre sensiblement le même contenu, il est de notre ressort de faire progresser les étudiants dans une utilisation réelle de leurs connaissances, soit de les outiller pour une utilisation « professionnelle » des statistiques avec toutes les embûches que des données réelles peuvent présenter.

Le but de cet article est de montrer qu'il y a une différence significative entre les données réelles et celles que l'on propose dans les cours et que, en conséquence, les étudiants ne sont pas prêts à affronter les difficultés d'une « vraie analyse de données ».

Depuis maintenant une quinzaine d'années, j'enseigne les statistiques au cégep dans de nombreux programmes (comme tout le monde, j'ai aussi dans ma tâche des cours d'algèbre et de calcul...). Il y a quelques années de cela, j'ai vécu des expériences en consultation statistique qui ont changé et qui changent encore ma manière d'enseigner, les connaissances que je transmets et la finalité de l'exercice.

[1] Je crois que cette approche est justifiée et « normale » au secondaire. Par contre, je crois qu'elle est dommageable au cégep.

Je n'ai pas encore atteint l'état d'équilibre de ce processus. Ainsi, je développe encore de nouvelles manières de présenter la matière et je poursuis mon questionnement personnel et professionnel sur le sujet...

Je présenterai ici la raison d'être d'un de ces changements, soit l'utilisation de « vraies données statistiques » dans mes cours. En effet, pour quelqu'un ayant pratiqué la statistique, il est évident qu'il y a une différence entre les données « réelles » et les données « académiques » utilisées dans les cours.

Par la suite, je m'attarderai à ce que j'ai pu observer quant aux conséquences pédagogiques de ces changements dans la transmission de la matière.

La plupart des remarques que je transmets ici ne sont ni « scientifiques », ni « vérifiées », mais, j'ose espérer, vous seront utiles pour en arriver à un questionnement fondamental sur les objectifs des cours de statistiques et de méthodes quantitatives que nous enseignons.

2 Une « vraie » base de données

Le tableau 1, présenté à la page suivante, est une « vraie » base de données.

3 Les différences

Fondamentalement, les données tirées des livres sont « trop propres ». J'illustrerai mon propos en prenant pour exemple la base de données présentée au tableau 1. Cette base de données a été colligée par une doctorante qui a composé son questionnaire, a effectué son sondage et a compilé ses données elles-mêmes, et ce, sans l'aide d'un statisticien. Par la suite, lors de l'étape de l'analyse, elle a fait appel aux services de statisticiens qui « ont été mis devant le fait accompli ». Pour des raisons de confidentialité, je ne donnerai pas ici plus de détails que nécessaire pour illustrer mes propos, en veillant à garder confidentielles les spécificités de la recherche.

La base de données présente un minimum de 6 caractéristiques que l'on retrouve rarement dans les données utilisées généralement dans les livres de statistiques et de méthodes quantitatives au cégep. Nous présenterons ces 6 caractéristiques, ainsi que quelques autres, en les illustrant, lorsque c'est possible, à l'aide des cellules et des groupes de cellules numérotés dans le tableau 1. Par ailleurs, l'ordre dans lequel les différences sont énoncées ne présuppose aucune hiérarchie entre elles.

1) *Modalités ignorées* (voir numéro 1 dans le tableau : colonne B de la base de données)

Dans le questionnaire à 4 choix de réponses, la modalité 1 qui correspond à « pas du tout » n'a été choisie par personne dans l'échantillon et la modalité 2 a été rarement choisie.

Tableau 1 : La base de données.

On peut faire l'analogie avec la création des nombres aléatoires. En effet, lorsqu'on demande à une personne de créer des nombres « aléatoires », ceux-ci ont plutôt tendance à être distribués plus uniformément que les nombres réellement aléatoires. Autrement dit, la chance permet des irrégularités que la création par un humain ne permet pas, comme pour la création de données construites.

2) *La nature de certaines données ne figure pas dans les 4 catégories de base*[2]
(voir le numéro 2 dans le tableau : cellules C1 à C4 de la base de données)

La question est « Combien de fois magasinez-vous par mois ? ». Bien qu'apparemment la variable soit « quantitative discrète », on ne recueille qu'une vague valeur moyenne. En effet, rares sont les personnes qui pourraient donner ce nombre avec précision, et de plus, l'étendre à une valeur « par mois », car personne ne garde normalement un registre de cette information.

En ce sens, la variable est plutôt qualitative que quantitative. L'utilisation des nombres n'est ici que représentative d'un ordre de grandeur. Ce manque de précision explique d'ailleurs le regroupement en classes prédéfinies dans le questionnaire, avant la cueillette des données (procédé généralement à éviter puisqu'il peut mener à des biais et des difficultés d'interprétation dans des séries chronologiques par exemple).

En effet, la question revient à demander si on magasine « un peu » (1 à 4 fois par mois), « moyennement » (5 à 8 fois par mois), « beaucoup » (9 à 12 fois par mois) ou « très » (plus de 12 fois par mois). L'utilisation de réponses chiffrées vise à donner au répondant un repère plus fort et universel que les descriptions verbales habituelles pouvant sembler plus « vagues ». De plus, les mots comme « beaucoup » n'ont pas la même signification pour chaque personne et donc l'utilisation de classes prédéfinies permet une uniformisation des réponses dans ce cas-ci.

3) *Des réponses non analysables* (voir le numéro 3 dans le tableau : colonnes E à I de la base de données)

La question 4 du questionnaire était mal posée. En effet, la partie erronée de la question se lit : « Sur une échelle de 1 à 5, 1 étant le moindre et 5 le plus, comment répartissez-vous leur utilité parmi ces options... ». L'auteure de l'étude fournit ensuite 5 énoncés. Cette question peut avoir deux interprétations contradictoires :

– soit on demande d'ordonner les 5 énoncés du moins important au plus important,
– soit on demande d'attribuer une cote de 1 à 5 à chaque énoncé, indépendamment des autres, 5 pour *important* et 1 pour *pas important*. J'affirme que la question n'est pas analysable. En effet, la probabilité qu'on obtienne par hasard, avec la seconde interprétation, une permutation des entiers de 1 à 5 au moins une fois dans l'échantillon dépasse le seuil de risque d'erreur de 10 % [3], qui est le plus grand risque d'erreur habituellement toléré « en sciences humaines ».

Dans les recherches que j'ai effectuées, je n'ai pas encore trouvé de données non analysables dans

[2]Soit qualitative nominale/ordinale et quantitative discrète/continue

[3]La preuve de ce résultat invoque des arguments combinatoires que le lecteur pourra trouver, ou que je pourrais lui fournir sur demande.

un manuel scolaire de niveau collégial. Par ailleurs, pour briser certains stéréotypes sur les risques d'erreur, je cite ici un extrait de mon livre à paraître *Méthodes quantitatives* :

« Il est à noter que dans des domaines comme la santé, on tolère dans certaines situations bien pointues des risques d'erreurs allant jusqu'à 60 %. Nommons ici une de ces situations exceptionnelles : l'acceptation de médicaments expérimentaux pour des maladies à ce jour mortelles. »

4) *Des données manquantes* (voir le numéro 4 dans le tableau : cellule L24 de la base de données)

Dans les résultats de la question 7, on retrouve la compilation d'une valeur « 0 », alors que cette option n'est pas incluse dans les choix de réponse proposés. La personne interrogée a décidé de ne pas répondre. Il s'agit d'une donnée manquante.

Cette différence entre les données réelles et « académiques » est d'ailleurs la plus marquante. En effet, on en retrouve même dans les enquêtes menées à bien par des statisticiens chevronnés dans des conditions contrôlées et bien financées. Ce dernier point est complètement négligé dans les manuels (à part pour affirmer que « les recensements coûtent plus cher que les sondages »). Pourtant, il revêt une importance non négligeable dans la « vraie vie », car les compagnies et les chercheurs tentent de limiter la taille des échantillons, pour des raisons financières, tout en préservant la qualité des résultats. C'est d'ailleurs un élément qui a influencé et continue d'alimenter la recherche en statistique.

Je note en passant les techniques les plus fréquentes pour faire face à ce problème :

- On les ignore, par exemple pour les calculs de moyenne ;
- On impute des valeurs à partir de différents critères selon le but ;
- On remplace par une mesure de tendance centrale ;
- On utilise des techniques de censure dans l'analyse.

5) *Biais* (ce point n'est pas illustré par les données du tableau 1) :

En faisant une analyse plus poussée, par exemple sur le sens des résultats de la question 1, on remarque rapidement des biais dans la collecte des données qui influencent leur qualité et leur interprétation.

Par exemple, si on regarde la colonne M des données dans le fichier EXCEL du tableau 1, on remarque que l'échantillon est composé majoritairement de jeunes. On ne peut donc qu'avec timidité étendre les résultats d'analyse de cet échantillon à une population d'âge d'or. Des raisonnements similaires s'appliquent aux colonnes A, B, C et N. Dans le tableau suivant, je récapitule les biais dans l'échantillonnage que l'on peut déduire directement des statistiques descriptives sur la composition de l'échantillon :

Colonnes (format EXCEL)	Biais d'échantillonnage
A-B-C	Échantillon très majoritairement féminin
M	Échantillon majoritairement jeune
N	Échantillon exerçant majoritairement la même profession
Conclusion	La population est restreinte

Tableau 2 : Biais.

En analysant les facteurs de l'étude, on remarque qu'il y a un biais de surreprésentation tellement prépondérant qu'on en arrive à la conclusion que la population n'est pas représentative de « la population du Québec » comme le voulait le responsable de l'étude, mais bien plutôt de « la population des jeunes femmes exerçant la profession de désigner et connues de la personne faisant l'étude » (en se basant sur le critère de représentativité).

De plus, on conçoit que l'échantillonnage a été effectué de manière empirique, plus précisément par la technique des quotas, et qu'il n'est pas possible d'effectuer des estimations par intervalles de confiance, car ceux-ci sous-estiment l'erreur commise.

6) *La complexité des données*[4] (ce point n'est pas illustré par les données du tableau 1)

La base de données du tableau 1 contient au total 77 variables dont les données ont été récoltées sur 28 individus. De ces 77 variables, il y en a 13 qui constituent les facteurs permettant de décrire l'échantillon, et les 64 autres représentent les variables d'intérêt.

La complexité de la base de données présentée dans cet article dépasse largement celle des données présentées dans les livres de cégep. De plus, même dans les données fournies sur support informatique pour étude dans les laboratoires, on ne retrouve pas cette complexité. Pourtant, on n'est pas sujet aux limitations du format papier.

7) *Des données extrêmes*[5] (ce point n'est pas illustré par les données du tableau 1)

En biologie moléculaire, par exemple, il est très fréquent de retrouver, dans un même échantillon, 500 cellules ayant une certaine propriété chez un individu et 1 000 000 de cellules avec la même propriété chez un autre individu. Les 2 données se trouvent dans un même échantillon et ont été réellement observées. Aucune d'entre elles ne peut être considérée comme aberrante. Il faut donc trouver un moyen de représenter des données ayant cette structure de manière efficace et compréhensible pour le lecteur.

Qui plus est, il ne sera pas rare pour un étudiant en sciences humaines ou en sciences de rencontrer des données ayant cette caractéristique. Ce problème peut aussi survenir en ressources humaines, en gestion (par exemple, des salaires s'échelonnant de 30 000 $ et de 1 000 000 $ au sein de la même compagnie) ou en finance (par exemple, des rendements allant d'une perte de 10 000 $ à un profit de 500 000 $ pour le même courtier). Pourtant, de telles données ne se retrouvent que rarement dans

[4] La présence de plusieurs facteurs et variables d'intérêt.
[5] Pas aberrantes.

un livre de cégep.

Les étudiants sont démunis par rapport à ces situations lorsqu'ils les rencontrent dans leurs recherches. Comment représenter de telles données sur un même graphique ? Comment faire un tableau approprié ?

Caractéristique	« Vraies données »	Dans beaucoup de livres
Complexité	Multiplicté de facteurs et de variables d'intérêt	1 variable ou 1 variable et 1 facteur
Données manquantes	Fréquentes	Inexistantes
Codage	Modalités chiffrées	Modalités verbales
Natures des variables	4 types de bases + semi-quantitatif, approximatif, etc.	Des variables bien définies dans les 4 types de bases
Distribution des réponses	Variété de distributions : réponses rares, réponses dominantes, réponses sans effectif, etc.	Il est rare de trouver des distributions « anormales ».
Biais	Présence de biais	Absence de biais
Qualité	Pas toujours analysable	Toujours analysables
Base de données	La base de données est la manière d'entreposer les données à des fins d'analyse. C'est la « matière première » d'un consultant en statistique.	Rare

Tableau 3 : Comparaison « vraies données » vs données de manuels.

4 Historique des expériences d'utilisation de vraies données en classe

Comme je l'ai annoncé dans l'introduction, j'ai procédé à des expériences pédagogiques dans mes différents cours à contenu statistique. J'ai progressivement introduit des situations et des données qui ressemblent à celles de « la vraie vie », pour aboutir à des données tirées d'une « vraie expérience de consultation ».

La population étudiante ciblée est assez large, allant des techniques aux sciences, en passant par les sciences humaines.

Programme	Cours	Contenu statistique	Laboratoire
Technologie de l'architecture	201-F33-MO 45 heures au total dont 30 heures de statistiques	Tableaux et graphiques, mesures, contrôle qualité (initiation), loi normale, introduction à l'estimation	Minimum de 6 heures pour le cours, mais il y a aussi 15 heures d'optimisation
Science de la nature	201-ESH-MO 75 heures de statistiques	Tout ce qui peut se donner de statistique au cégep, sauf le contrôle de la qualité	
Techniques en informatique	201-P15-MO 75 heures dont 45 de statistiques	Combinatoire, tableaux et graphiques mesures, lois	Minimum de 16 heures pour le cours, mais il y a aussi 30 heures de maths discrètes
Campeurs du camp de l'AMQ du secondaire	Camp mathématique de l'AMQ du secondaire 3 heures	« Jeu de rôle »de consultation statistique, connaissances du secondaire, soit un peu plus de la moitié d'un cours de Méthodes quantitatives	
Sciences humaines	360-300-RE 60 heures de statistiques	Tableaux et graphiques mesures, régression, khi-deux, loi normale, estimation d'une moyenne et d'une proportion	

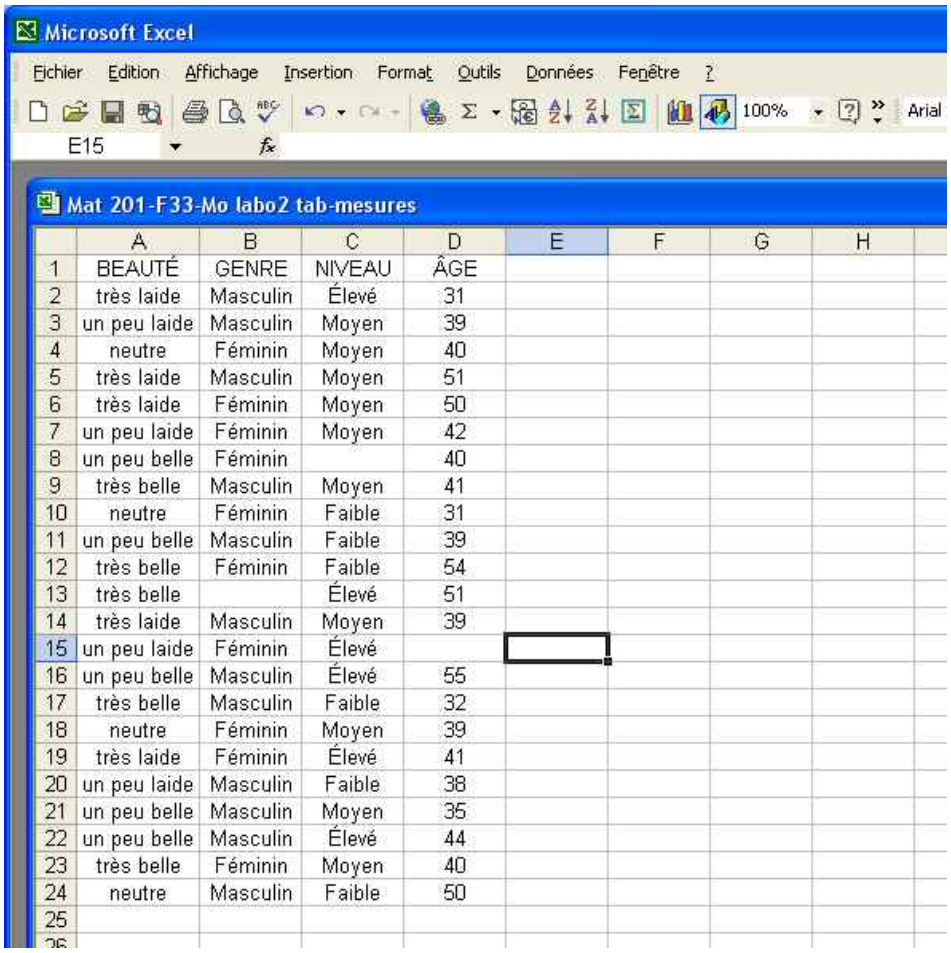
Tableau 4 : Expériences.

Exemple 1 : *Des données manquantes en architecture*

Confrontés à des données manquantes, les étudiants ont réagi fortement.

Après un premier laboratoire informatique sur Excel où les données étaient « académiques » (propres, pas de données manquantes, etc.), j'ai surpris les étudiants en présentant un deuxième laboratoire où il y avait des données manquantes. À ce stade de la session, les laboratoires n'étaient pas encore notés. J'ai présenté les données en faisant un « jeu de rôle » dans lequel j'étais un patron leur demandant d'analyser les données provenant « d'un sondage auprès d'anciens clients » sur leur satisfaction par rapport à nos services, de trouver la signification des données et d'émettre des recommandations pertinentes sur d'éventuels ajustements à apporter aux services offerts par la compagnie.

Les données sont « fictives », mais j'ai inclus le défaut évident des données manquantes. Elles s'inspirent des données du tableau 1, mais en plus simples.



	A	B	C	D	E	F	G	H
1	BEAUTÉ	GENRE	NIVEAU	ÂGE				
2	très laide	Masculin	Élevé	31				
3	un peu laide	Masculin	Moyen	39				
4	neutre	Féminin	Moyen	40				
5	très laide	Masculin	Moyen	51				
6	très laide	Féminin	Moyen	50				
7	un peu laide	Féminin	Moyen	42				
8	un peu belle	Féminin		40				
9	très belle	Masculin	Moyen	41				
10	neutre	Féminin	Faible	31				
11	un peu belle	Masculin	Faible	39				
12	très belle	Féminin	Faible	54				
13	très belle		Élevé	51				
14	très laide	Masculin	Moyen	39				
15	un peu laide	Féminin	Élevé					
16	un peu belle	Masculin	Élevé	55				
17	très belle	Masculin	Faible	32				
18	neutre	Féminin	Moyen	39				
19	très laide	Féminin	Élevé	41				
20	un peu laide	Masculin	Faible	38				
21	un peu belle	Masculin	Moyen	35				
22	un peu belle	Masculin	Élevé	44				
23	très belle	Féminin	Moyen	40				
24	neutre	Masculin	Faible	50				
25								

Tableau 5 : Base de données présentant des données manquantes.

La réaction a été différente dans les deux groupes du programme *Technologie de l'architecture*. Dans le premier, ils se sont mis à répondre immédiatement (« les variables sont… », « la nature est… », etc.) puis j'ai eu la question « qu'est-ce qu'on fait avec les trous ? ». Dans le second groupe, un étudiant s'est écrié « c'est quoi ces données toutes croches ? (sic) ». Il avait remarqué les trous dans les colonnes B, C et D.

Dans les 2 cas, les réactions ont mené à un débat très intéressant et engagé entre les étudiants, où j'ai agi comme modérateur. Tout au long, je jouais le rôle du « patron ignorant en statistique » désirant des résultats, qui ne voulait « surtout pas dépenser une cenne de plus pour interroger plus de monde » et qui refusait absolument de « jeter à la poubelle des données incomplètes ».

Les étudiants ont assez rapidement identifié les stratégies possibles face aux données manquantes (sauf la censure). Ils ont aussi remarqué le coût de la récolte de données.

Exemple 2 : *Des données non analysables dans un cours d'intégration en sciences de la nature*

La première fois que j'ai présenté de vraies données, je me suis permis de le faire avec des finissants

du DEC en sciences de la nature.

Les étudiants avaient un mois pour remettre un travail sur les réponses des questions 1 à 5 de la base de données de la section 2, sachant qu'il y avait une des cinq questions qui n'était pas analysable, qu'il fallait l'identifier et expliquer pourquoi elle ne l'était pas, tout en analysant les 4 autres.

Compte tenu de l'importance du cours dans leur DEC, j'ai procédé à une rencontre à mi-chemin pour m'assurer qu'ils étaient sur le bon chemin et éventuellement redresser le tir. Évidemment, les étudiants s'y sont pris à la dernière minute, et cela en dépit des nombreux rappels en classe. À la fin, ce fut pour eux une entreprise ardue.

Par contre, l'expérience fut enrichissante. Lors des rencontres, j'ai dû procéder à des questions sur le mode socratique pour mettre les étudiants sur la piste. Certains avaient des intuitions « qui tournaient autour du pot », mais je constatais malheureusement qu'ils n'avaient pas le temps de se concentrer sur les données en raison de la « lourde charge » de travail qu'ils avaient dans leurs autres cours.

Exemple 3 : *Utiliser la base de données du tableau 1*

Au camp mathématique de l'ordre secondaire de l'AMQ, je me suis permis de présenter à nouveau les données du tableau 1, mais sous une forme complètement différente. Le groupe était constitué d'une vingtaine d'étudiants s'étant très bien classés au concours du secondaire de l'AMQ, donc « forts » en mathématiques et très attirés par les mathématiques, mais affichant généralement un certain dédain pour la statistique perçue comme une utilisation non réfléchie de formules.

L'atelier a commencé avec une question ouverte « Qu'avez-vous fait en statistiques au secondaire ? ». Les étudiants ont cité le contenu complet et classique des programmes (tableaux, graphiques, mesures de tendances centrales, de dispersion, de position et la régression). La question piège suivait « Était-ce facile ? ».

J'ai alors présenté aux étudiants un jeu de rôle où je jouais le rôle d'un client qui venait chercher conseil auprès de « statisticiens ». J'ai alors projeté la base de données à l'écran. Un deuxième piège les guettait : « Comment analyse-t-on la première question ? ». Un malheureux a mordu et a proposé de calculer la moyenne...

On a alors passé 3 heures à parler de variables, de leur nature, de la meilleure manière de les analyser, des erreurs de questions, des biais présents, de la structure de l'échantillon et des données, etc.

La réaction a été « c'est la première fois qu'on trouve la statistique amusante ! » et « on n'aurait jamais cru qu'il y avait autant de subtilité en statistique ! ». Fort de cette expérience, j'ai essayé de la transposer dans le milieu scolaire avec « ses objectifs à atteindre » et surtout son temps limité.

Exemple 4 : *Une session d'essai en informatique*

Quelques semaines après l'expérience du camp mathématique, j'ai commencé mon cours de statistiques/mathématiques discrètes en informatique, en essayant de répéter l'expérience du camp. Il est à noter que le cours venait de « naître », issu de la réforme du programme, et il nécessitait déjà

une charge de préparation très grande avec, d'un côté un nouveau contenu unique à ce cours et de l'autre, la nécessité de voir le contenu de l'ancien 201-257, mais en seulement 45 heures. Dans ce cours, l'accent doit être mis sur les applications, et il faut aussi l'adapter à la clientèle par la présence de plusieurs laboratoires informatiques.

Je me croyais donc devant un terreau fertile. En effet, le côté « jeu de rôle » de la présentation devait satisfaire les amateurs de jeu vidéo, le côté multimédia de la présentation plairait, étant donné l'intérêt informatique des étudiants, l'utilisation du logiciel EXCEL (c'était d'ailleurs la dernière fois que je m'en servais au profit de SPSS) et le côté appliqué seraient tous des éléments qui donneraient le nouveau ton du cours. L'effet a été très fort dans les deux groupes. Les étudiants réalisaient l'étendue des outils statistiques qui étaient déjà en leur possession. L'approche interactive plaisait.

Exemple 5 : *Le test ultime, les sciences humaines...*

Lors de trois sessions consécutives, j'ai eu à donner des cours de Méthodes quantitatives. Chaque fois, j'ai commencé la session avec le jeu de rôle présenté dans l'exemple 3. L'expérience n'a pas été aussi concluante qu'au camp de l'AMQ. Par contre, j'ai remarqué (de manière très subjective) une hausse d'intérêt chez les étudiants pour la matière.

Je rappelle que les contenus suivants sont déjà vus au secondaire et constituent donc normalement des révisions pour ces étudiants : tableaux, graphiques, échantillonnage, mesures de tendance centrale, mesures de dispersion, mesures de position et régression.

Lors de deux des trois sessions, j'ai remarqué une amélioration de la persévérance. La présente session semble être l'exception, mais j'attribue cela à un facteur externe, le fait que ce soit la dernière promotion du secondaire avant la réforme.

5 Les conséquences pédagogiques

Quand on présente des bases de données avec données manquantes, les étudiants réagissent de différentes manières, allant du jugement négatif à l'égard de cette base de données à l'expression de leur ignorance quand au choix d'une stratégie d'analyse adéquate.

Lorsqu'on présente des bases de données réelles et que l'on force les étudiants à y réfléchir à haute voix, ils s'emballent sur le coup et s'intéressent davantage au cours. Faire durer l'intérêt est un autre problème.

Dans tous les cas, le « statu quo » est rompu. Les cours de cégep prennent leur place comme une suite et une évolution du secondaire. Ce n'est plus une copie et une répétition de ce qui a été vu précédemment.

L'important en présentant des données à des étudiants est de les faire réfléchir au sens de celles-ci. Après tout, ce sont les données qui sont à la source de tous les raisonnements statistiques que les étudiants auront à faire. À mon avis, le fait de revenir aux données au début de chaque raisonnement est un des piliers de la pensée statistique.

En enseignement, la marge est mince entre la simplification et l'infantilisation de la matière. Cette dernière peut engendrer la démotivation. Par contre, il y a déjà un problème en mathématiques au Québec, car il est inconcevable qu'après 11 ans d'étude des mathématiques, certains étudiants émettent des commentaires tels que : « Les mathématiques sont compliquées pour rien ! » ou « En statistique, il faut seulement « plugger » des formules ! (sic) ».

Il en est de même avec la complexification qui peut donner un sentiment d'authenticité ou un sentiment d'incompréhension. Par contre, ce dernier sentiment ne serait-il pas dû malheureusement à des problèmes langagiers, voir même à un illettrisme, comme le dit De Serres (2003) ? Sans faire porter tout le blâme sur ce problème, ne faudrait-il pas le régler avant de modifier et de remodifier les programmes ?

Au cégep, on ne peut pas régler complètement ce problème dans un cours de mathématiques ou de statistiques. À ce niveau, il faut prendre les étudiants avec les capacités langagières qu'ils ont. On ne peut qu'insister lors des travaux sur une rédaction correcte et porter attention aux erreurs de français. De plus, une majorité d'étudiants montrent des signes évidents de fatigue aigüe ou accumulée se traduisant par un manque de concentration, de l'anxiété, de la distraction, etc.

Hormis ces considérations « systémiques », l'utilisation des données réelles mûrit le raisonnement statistique des étudiants. Par exemple, le codage de variables qualitatives par des nombres force les étudiants à aller au-delà des apparences et à se demander si le calcul de la moyenne est faisable sur des données pourtant « numériques en apparence ». L'étudiant ne peut plus appliquer des formules bêtement : il doit « penser ses données » !

Chacune des caractéristiques des données réelles soulevées au point 3 de ce texte amène à un réel questionnement statistique que l'on récapitule dans le tableau suivant :

Caractéristique	Conséquence pédagogique
Complexité	Nécessité de comprendre le but de la question, ne pardonne pas les automatismes (en régression, il faut prendre les bonnes colonnes... dans le bon ordre, contrairement au cas où il n'y en a que 2)
Données manquantes	Questionnement sur la validité de mesurer sans ces valeurs (robustesse des méthodes)
Codage	Nécessité de comprendre la nature de la variable (même si on a une suite de nombres, on ne peut pas toujours faire la moyenne, etc.)
Nature des variables	Questionnement sur le sens que l'on peut donner à une mesure calculée sur des variables ambigües (fiabilité) + Nécessité de récolte de données claires
Distribution des réponses	Questionnement sur la représentativité de la population au sein de notre échantillon. Sur quel groupe peut-on réellement étendre les résultats de notre échantillon ?
Biais	Questionnement sur la fiabilité de nos réponses et de la « définition »de notre population.
Qualité	Questionnement sur le sens et la finalité des questions que l'on pose + Nécessité de bien formuler ses questions
TOTAL	**DÉVELOPPEMENT DE LA PENSÉE STATISTIQUE**

Tableau 6 : Conséquences pédagogiques.

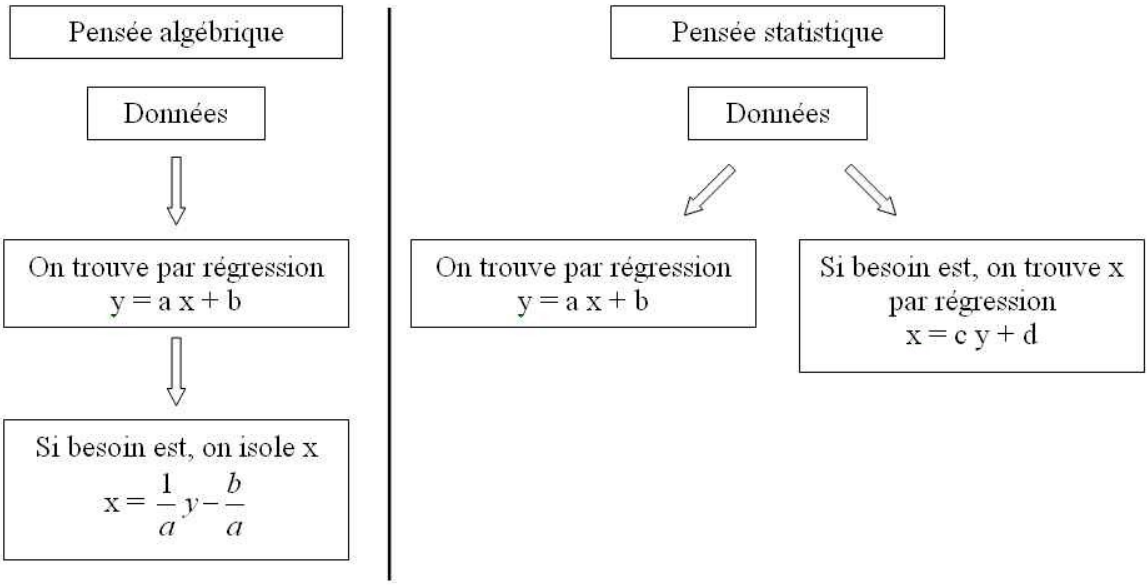
Avec le format « jeu de rôle », sans autres preuves, je pense atteindre plusieurs objectifs pédagogiques intéressants :

- les statistiques ne sont plus une suite d'utilisations irréfléchies de formules,
- les étudiants deviennent des acteurs de leur apprentissage,
- les étudiants perçoivent les aspects authentiques (vraies données, « jeu de rôle »,...),
- la pression sociale forcerait-elle les plus faibles à faire comme les autres, plus forts, « qui se souviennent de leur secondaire » ?

C'est un plus grand risque, car si on joue mal le rôle, on risque d'assister à une démotivation des étudiants. Il faut créer de l'enthousiasme dans le groupe pour qu'il y ait une synergie qui se crée.

6 La pensée statistique

Une petite note finale pour illustrer la différence entre la pensée statistique et la pensée algébrique ; prenons pour cela l'exemple des droites de régression. Un statisticien qui a trouvé à partir des données la droite de régression donnant y en fonction de x, s'il a besoin de x, n'isolera pas x dans l'équation de la droite de régression, mais calculera, à partir des données, une droite qui prédit la valeur de x. Alors qu'un algébriste, ou d'autres mathématiciens, soustraient et divisent dans l'équation bâtie jusqu'à isoler et exprimer y.



Mathématiquement, dans un diagramme de dispersion avec 2 variables (x et y), il y a 2 droites de régression. En effet, dans un cas, on minimise les « moindres carrés verticaux » pour trouver une droite qui prédit le y à partir du x. Dans l'autre, on minimise les « moindres carrés horizontaux » pour trouver une droite qui prédit le x à partir du y. Les 2 droites ne sont concourantes que si la corrélation de Pearson $r = 1$.

Sinon, il est un fait connu que r est la moyenne géométrique des pentes des deux droites. Ainsi, r permet de déterminer la mesure de l'angle entre les deux droites de régression. Si $r = 1$, l'angle entre les deux droites est 0°et si $r = 0$, l'angle est 90°.

En statistique, on fait deux diagrammes différents, les deux faisant intervenir deux variables x et y... l'un donnant y comme une fonction de x et l'autre x comme une fonction de y, qui n'est pas la fonction réciproque de la précédente.

D'ailleurs, en statistique, on ne pense pas en termes de variables dépendantes et indépendantes, mais en terme de facteurs et de variables d'intérêt.

7 Conclusion

Un des buts est de briser les idées reçues et les stéréotypes sur les statistiques et le traitement des données qui sont véhiculés inconsciemment par l'enseignement tel qu'il est proposé. Je suggère de toujours revenir à la base et de comprendre les données.

Un autre but est de transformer les étudiants non pas en statisticiens, mais plutôt en utilisateurs consciencieux des statistiques n'ayant pas de problème à consulter des statisticiens pour les aider à

analyser les données et répondre à des questions.

En somme, le but ultime de la démarche est de bâtir la PENSÉE STATISTIQUE, soit de réfléchir aux données, à leurs limites, à leurs propriétés et à ce qu'elles veulent dire, au lieu de se limiter à une application systématique de recettes stéréotypées sur des données quelles qu'elles soient ou quelle que soit leur structure.

En fin de compte, le changement de cap lors de l'activité est très grand par rapport à l'enseignement magistral « traditionnel ». Par contre, globalement, il reste qu'il y a apprentissage et une majorité d'étudiants ont fait montre d'une plus grande motivation. Il ne faut pas croire à une panacée, puisque cette session-ci semble être un contre-exemple. Les étudiants ont maintenant de plus en plus de problèmes à prendre des notes de cours, à écouter et écrire en même temps et plus fondamentalement, à décerner ce qui est important dans un discours. Serait-ce un effet pervers de l'implantation massive « d'aides pédagogiques » (Powerpoint, feuilles de notes à trous pour aller plus vite, etc.) ? Heureusement, compte tenu du fait que mon « jeu de rôle » ne fait partie que de la première journée de classe, mon but est autre, celui de réveiller la mémoire perdue du secondaire.

Références

[1] Bilinski R. (à paraître), *Méthodes quantitatives*, Montréal, R & R éditions.

[2] De Serres M. (2003), *Intervenir sur les langages en mathématiques et en sciences*, Mont-Royal, Modulo.

Pour aider à une réflexion sur le sujet, le lecteur peut consulter :

[3] Dytham C. (2003), *Choosing and using statistics*, London, Blackwell Publishing.

[4] Hawkins D (2005), *Biomeasurement*, Oxford, Oxford University Press.

[5] Robert C.(2003), *Contes et Décomptes de la statistique*, Paris, Vuibert.

[6] Robert C. (2006), *Pratique de la statistique*, Paris, Vuibert.

[7] Saporta G.(2006), *Probabilités, analyse des données et statistiques*, Paris, Technip.