REPERES - IREM . N° 23 - avril 1996

POURQUOI IL NE FAUT PAS LAISSER DE CÔTÉ LES CHAPITRES DE STATISTIQUES AU COLLÈGE

Jean-Claude GIRARD
Irem de Lyon

L'idée de cet article part d'un constat : les chapitres de statistiques au collège sont souvent négligés, reportés à la fin de l'année ou tout simplement "sautés" sous prétexte que l'on n'a pas le temps de tout faire ! L'étude sérieuse en est alors différée d'année en année jusqu'à ce que l'on considère (en seconde généralement) que tout a été vu avant ! On observe d'ailleurs la même attitude pour l'utilisation de la calculatrice, dont on peut trouver l'idée très intéressante et en reporter pourtant l'utilisation d'année en année par manque de temps ou parce que c'est trop tôt ! A cet égard, les statistiques ont rejoint la géométrie dans l'espace, fréquemment repoussée le plus loin possible dans l'année, rapidement traitée, éventuellement pas traitée du tout suivant le temps disponible.

La première raison de ce choix (parce que c'est un choix !) est que beaucoup de professeurs se sentent moins à l'aise dans ce chapitre, "moins mathématique", que dans les autres, mais cela ne me paraît pas être la raison principale. Tout professeur consciencieux oublierait, en effet, ses états d'âme s'il était convaincu de l'intérêt de cette partie du programme et des difficultés qu'elle présente pour les élèves. Ce n'est malheureusement pas le cas.

Je vois au moins trois intérêts majeurs à développer l'enseignement des statistiques, en tout cas pour qu'il atteigne le niveau que le programme lui assigne :

- au niveau des graphiques, en *liaison avec différentes parties du programme* de mathématiques et pas seulement pour servir d'outil à d'autres matières,
- au niveau des calculs (fréquences,

moyennes, médianes) en liaison avec l'*idée de distribution statistique*,
– au niveau conceptuel en liaison avec l'*idée de hasard et de variabilité des résultats* dans une expérience répétée dans les mêmes conditions (et que l'on qualifiera alors d'aléatoire).

L'étude de ces trois aspects des statistiques peut concourir au développement intellectuel des élèves et en particulier à l'aspect *"formation du citoyen"* confronté de plus en plus aux statistiques (graphiques, pourcentages, moyennes, sondages, etc.). L'objectif visé serait que les élèves se posent eux-mêmes des questions sur ce qu'ils voient ou entendent (chiffres ou graphiques). Cette étude me semble également indispensable en vue de faciliter l'*enseignement des probabilités* en première et terminale, si l'on ne veut pas se contenter de constater à ce moment là "qu'ils ont des difficultés".

I. Les graphiques

C'est la partie des statistiques qui est la moins souvent "oubliée" car elle a des applications dans les autres matières et, de plus, elle fait assez souvent l'objet de questions au brevet des collèges. D'autre part, on saisit l'occasion de la construction des graphiques statistiques (camemberts, barres, histogrammes) pour réinvestir la notion de proportionnalité sous ses différentes formes : pourcentages, échelles, règle de trois. L'hypothèse implicite est que ces graphiques ne posent pas de problèmes (autres que ceux liés à la proportionnalité) aux élèves. Et pourtant, en dehors des difficultés purement statistiques (définition des variables, récolte des données), il reste beaucoup de points d'interrogation.

D'abord sur le sens des graphiques eux-mêmes :

– Quel est l'avantage d'un graphique sur un tableau de valeurs ?
– Le graphique sert-il d'illustration ou permet-il de découvrir une structure des données que le tableau ne mettait pas en évidence ?
– Peut-on repasser du graphique au tableau ?
– Quelle perception de la réalité a-t-on en regardant un graphique ?
– Pourquoi tel graphique plutôt que tel autre ? Dans quels cas, chacun est-il pertinent ?

Ensuite sur d'autres notions qui renvoient à différents domaines mathématiques :

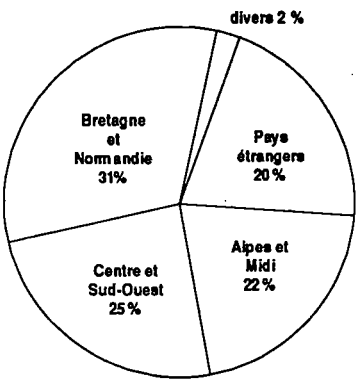
– Les camemberts utilisent la *notion d'angle* et de *mesure d'angle* qui ne sont pas toujours acquises. Comment peut-on prendre en compte cet état de fait ? Que représente le disque complet ? Autrement dit, quel est l'ensemble sur lequel on calcule les *pourcentages* ?
– Les histogrammes et les graphiques en barres ou en bâtons utilisent une *échelle* verticale sur laquelle on porte des effectifs ou des *fréquences*. Sur quel ensemble de référence ces fréquences ont-elles été calculées ?
– Lorsque l'on représente des *variations*, sont-elles calculées de façon absolue ou relativement à une valeur de référence ?

Exemple (extrait d'un livre de CM1 : Objectif Calcul-Editions Hatier)

Le livre pose les questions suivantes :

1) Observe ce graphique.
2) Essaie de le lire.

3) Quels renseignements donne-t-il ?
4) Essaie de traduire ce graphique par un tableau de nombres.

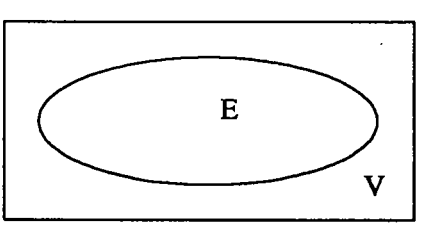


LES VACANCES DES FRANÇAIS

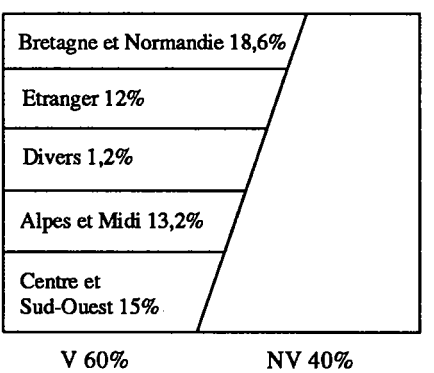
On pourrait aussi demander (en CM2, en 6e ou plus tard !) :

– Sur quoi sont calculés les pourcentages ?
– Est-ce 20% des français qui partent en vacances à l'étranger ou 20% de ceux qui partent en vacances qui vont à l'étranger ?
– Peut-on calculer combien de français partent à l'étranger ? Combien partent en vacances ? etc.

Cela pourrait être l'occasion d'une initiation à la représentation ensembliste :



On peut raisonner sur la population française ou, pour simplifier, sur 100 personnes. Si l'on considère que 60 % des français partent en vacances, les 20 % qui vont à l'étranger représentent en fait 20% de 60% c'est à dire 12% de la population. La question fondamentale est : calcule-t-on les pourcentages sur l'ensemble de la population ou sur l'ensemble des français qui partent en vacances ? Ce genre de questions permet de donner du sens aux pourcentages bien plus que l'entraînement à la virtuosité dans les calculs.



2) Ces questions sont une préparation à l'étude des probabilités car on retrouvera les mêmes problèmes lorsque l'on raisonnera (en première) en termes de probabilités : Un français étant choisi au hasard, quelle est la probabilité qu'il prenne ses vacances à l'étranger si l'on sait qu'il part en vacances (probabilité conditionnelle E sachant V, soit 20%) ou la probabilité pour le même français de partir en vacances à l'étranger (E et V, soit 12%).

D'ailleurs de nombreux problèmes de probabilités sur les ensembles finis se ramènent à des problèmes de fréquences ou de pourcentages.

Exemple (extrait du livre de Terminale ES "Déclic" collection hachette 1994) :

Lors d'un sondage auprès de 24 000 personnes, 14 280 sont parties en vacances et 5 340 sont parties en vacances d'hiver. Calculer la probabilité des événements suivants :
a) "une personne, prise au hasard, est partie en vacances"
b) "une personne, prise au hasard, est partie en vacances d'hiver"
c) "une personne, partie en vacances, est partie en vacances d'hiver".
On peut remarquer que les probabilités présentent les mêmes difficultés que les pourcentages au niveau de l'ensemble de référence.

On peut les ajouter (ou les soustraire) si les calculs ont été faits sur les mêmes ensembles de référence :
$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ et } B)$

On les multiplie si un calcul a été fait sur un premier ensemble et l'autre sur un sous-ensemble de celui-ci:

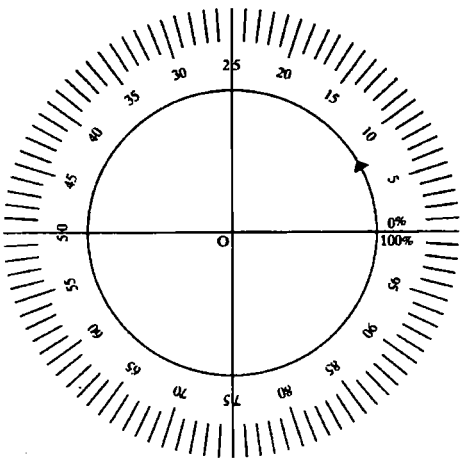
$P(A \text{ et } B) = P(A/B) \times P(B)$

3) Ces questions concourent également à l'apprentissage de la lecture de graphiques. Celle-ci est au moins aussi importante que la construction. A quoi peut-il servir de construire des graphiques si l'on ne sait pas lire les graphiques déjà construits ? Quelle idée un élève se fait-il en regardant un histogramme ou un camembert ? L'a-t-on entraîné à lire un graphique ? A-t-il une perception globale des quantités représentées ou se fait-il une idée des unes par rapport aux autres ? ou par rapport à un tout ? On peut faire le pari que la lecture d'un graphique statistique est du même ordre que la lecture d'une figure de géométrie dans l'espace. Le décryptage n'est pas inné. La première perception est visuelle mais l'interprétation est cognitive, elle demande des connaissances. La lecture de l'expert n'est pas celle de l'élève (1). Il doit donc y avoir apprentissage de la lecture d'un graphique statistique. Les conceptions d'un élève sont souvent dans la comparaison plus grand, plus petit, et ceci sur des grandeurs prises dans l'absolu. L'objectif devrait être de les amener à comparer les valeurs les unes par rapport aux autres ou par rapport à un tout, c'est-à-dire raisonner en valeur relative, en pourcentage, et alors, être capable d'identifier l'ensemble de référence ?

Par conséquent, il pourrait y avoir un grand intérêt à travailler les graphiques statistiques autrement que comme application de la proportionnalité. Ils devraient être un moyen de développer ce concept lui-même, les deux concepts s'éclairant mutuellement.

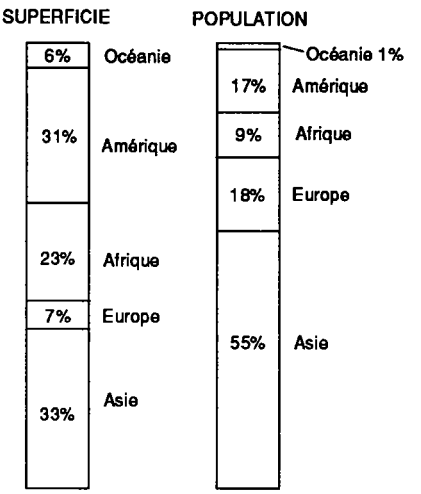
Tout comme la notion d'angle ne saurait être acquise sans en avoir une bonne image mentale, il me semble nécessaire de faire acquérir une image mentale d'un pourcentage. Cela nécessite un apprentissage. Des séquences peuvent être construites (2) à partir de la lecture et de la construction de graphiques statistiques, bien avant de maîtriser parfaitement les calculs de pourcentages, en utilisant par exemple un rapporteur à pourcentages (figure page suivante) (3).

(1) Voir à ce sujet l'article de Jacques Courivaud, "Le traitement graphique des images de géométrie", *Repères-IREM* n°4, Juillet 1991.
(2) Voir, par exemple, Daniel Gros, "*Des graphiques à l'école*", Mémoire CAFIPEMF-IUFM de Lyon, Centre local de St-Etienne, 1994.
(3) Matériel en vente à l'IREM de Lyon.



La perception de la proportionnalité n'est pas la même sur les longueurs que sur les aires (4). Pour ceux qui sont plus sensibles à une "vision" linéaire de la proportionnalité, on peut aussi travailler sur les barres. Exemple : (Objectif Calcul CM1)

– Que représente la longueur de chaque barre ?
– Sur quoi ont été calculés les pourcentages ?
– Peut-on comparer ces différents pourcentages ?
– Quelle idée veut donner ce graphique ?



SUPERFICIE ET POPULATION
DES CONTINENTS

II. Les paramètres et les distributions statistiques

On étudie classiquement au collège les effectifs et les fréquences d'apparition d'une modalité d'un caractère qualitatif ou d'une valeur d'un caractère quantitatif ainsi que la moyenne pour ce dernier cas. La médiane est au programme de troisième mais son étude est souvent esquivée, pour plusieurs raisons. Tout d'abord sa connaissance n'est pas exigible des élèves et, donc, elle ne peut figurer dans les sujets du brevet des collèges. D'autre part, encore une fois, beaucoup de professeurs ne voient pas l'utilité de ce concept.

On peut se demander, en effet, pourquoi calculer certains paramètres d'une série statistique ? Si l'on s'en tient au calcul de la moyenne, par exemple, il faut reconnaître que cela n'a pas beaucoup de sens, et même dans certains cas, pas du tout.

(4) Et encore moins sur les graphiques en perspective, qui sont la plupart du temps faux du point de vue mathématique. On lira avec profit l'article de Gérard Pornin "Des impôts à l'ellipse" dans *Des chiffres et des lettres au collège*, Bulletin Inter-Irem Premier Cycle 1991-1992, dans lequel on présente une activité statistique dont les objectifs sont géométriques (Théorème de Thalès, trigonométrie, cercle circonscrit, angles, symétries, tracés).

18 et -	19 à 23	24	25	26-27	28-29	30 et +
4,9%	24%	14,8%	14,6%	25,1%	12,9%	3,7%

Tableau 1

L'objectif est de donner une idée d'une série statistique par une valeur numérique ou de comparer deux séries statistiques. On ne peut le faire avec les seules moyennes. Que peut-on dire, par exemple, d'un endroit où la température annuelle moyenne est de 20° ? La sensation ne sera pas exactement la même si les températures sont situées toute l'année entre 18° et 22° ou si elles évoluent entre –40° l'hiver et +30° l'été.

La moyenne ne prend son sens que si elle est associée à une mesure de la dispersion des valeurs de la série. Par exemple l'écart type qui prend en compte les écarts (par rapport à la moyenne) de chacune des valeurs de la série. L'inconvénient de ce paramètre est qu'il n'a pas de représentation concrète simple, qu'il est long à calculer, qu'il ne figure qu'au programme de seconde et, de plus, étant lié à la moyenne il est sensible, comme cette dernière, à des valeurs anormalement grandes. Il existe heureusement d'autres paramètres de dispersion. Lorsque l'on porte dans un bulletin scolaire la moyenne, la note la plus basse et la note la plus haute, on caractérise la distribution des notes par un paramètre de tendance centrale, sa moyenne, et par un paramètre de dispersion, son étendue, c'est à dire l'écart entre le minimum et le maximum de la série. Ce dernier paramètre est simple à comprendre, d'un calcul aisé et peut être présenté au collège ! L'inconvénient, cette fois, est qu'il est assez frustre et encore plus sensible aux valeurs extrêmes.

Une alternative, réalisable en collège, est de caractériser une série statistique par sa médiane, pour la tendance centrale, et par l'intervalle interquartile pour la dispersion. Derrière ce vocabulaire un peu barbare se cache en réalité une notion assez simple, de calcul relativement aisé et qui présente, de plus, le double avantage de réinvestir la médiane et de se prêter à une représentation graphique. La conjonction de ces paramètres permet alors d'analyser une série statistique ainsi que de comparer des séries statistiques d'un double point de vue (tendance centrale et dispersion) en donnant du sens à ces deux concepts.

Exemple : D'après les statistiques de l'éducation nationale (5), le nombre d'élèves par classe de sixième (établissements publics de France métropolitaine en 1989-1990), se répartit comme le montre le tableau 1.

La moyenne (même source) s'élève à 24,6 élèves par classe.

On peut se livrer dans un premier temps à des calculs classiques sur les pourcentages et les moyennes.

1) Combien de classes de sixième avec 24 élèves, 25 élèves etc. ? (Il faut donc transformer les pourcentages en effectifs en supposant que le nombre total de classes de sixième est, par exemple, de 30.000.)

(5) *Repères et références statistiques sur les enseignements et la formation 1991-1992*, Ministère de l'éducation nationale, Direction de l'évaluation et de la prospective, 1993.

Numéro d'ordre	1	2	...	15000	15001	...	29999	30000
Valeur de l'observation	16	16	...	25	25	...	32	32

Tableau 2

Numéro d'ordre	1	2	...	7500	7501	...	15000
Valeur de l'observation	16	16	...	23	23	...	25

Tableau 3

Numéro d'ordre	15001	15002	...	22500	22501	...	30000
Valeur de l'observation	25	25	...	27	27	...	32

Tableau 4

2) Comment calculer la moyenne lorsque les données sont regroupées en classe et, qui plus est, que les classes extrêmes ne sont pas bornées ?

(On prend comme valeur de chaque classe, le centre de classe en faisant l'hypothèse, par exemple qu'il n'y a pas de classe d'effectif inférieur à 16 ni supérieur à 32 ce qui donne comme valeurs des centres de classe : 17 ; 21 ; 24; 25 ; 26,5 ; 28,5 ; 31 et comme moyenne 24,55.)

On peut remarquer que ceci n'est qu'une valeur approchée puisque l'on a perdu des informations en regroupant les données alors que l'on peut penser que la valeur du ministère a été calculée à partir des données brutes et qu'elle est exacte.

3) Comme on l'a déjà fait remarquer, la moyenne ne nous donne pas de renseignements sur les variations des effectifs dans les classes. On peut passer alors à l'analyse de la série par les paramètres proposés au début de ce paragraphe.

Si l'on considère que la série comporte 30 000 classes de sixième, alors la médiane est l'effectif de la 15 000^{e} classe de la série ordonnée (plus exactement, la moyenne entre la 15 000^{e} et la 15 001^{e} !

Il convient de ne pas confondre (erreur fréquente chez les élèves) le rang des observations (dans un classement dans l'ordre croissant par exemple) et la valeur de ces observations (tableau 2).

La série est alors partagée en deux sous-séries d'effectifs 15 000 que l'on peut de nouveau partager en deux par leurs médianes respectives (tableaux 3 et 4).

La médiane de la première série, 23, correspond au premier quartile Q1 de la série d'origine.

La médiane de la deuxième série, 27, correspond au troisième quartile Q3 de la série d'origine.

Ces deux valeurs combinées à la médiane Me de la série d'origine, partagent cette série en quatre parties de même effectif (tableau 5).

Numéro d'ordre	1	7500	15000	22500	30000
Valeur de l'observation	16	23	25	27	32
% des valeurs	25%	25%	25%	25%	

Tableau 5

Nombre d'élèves	18 et -	19	20	21	22	23	24
Fréquence	4,9%	4,8%	4,8%	4,8%	4,8%	4,8%	14,8%
Fréquence cumulée croissante	4,9%	9,7%	14,5%	19,3%	24,1%	28,9%	43,7%

Tableau 6

Nombre d'élèves	25	26	27	28	29	30 et +
Fréquence	1469%	12,55%	12,55%	6,45%	6,45%	3,7%
Fréquence cumulée croissante	58,3%	70,85%	83,4%	89,85%	96,3%	100%

Tableau 7

On peut trouver la valeur de la médiane et ainsi que celles des quartiles en construisant le tableau des fréquences cumulées. On fait pour cela l'hypothèse que dans les classes comportant plusieurs effectifs possibles d'élèves, la répartition est uniforme entre les différents effectifs (tableaux 6 et 7).

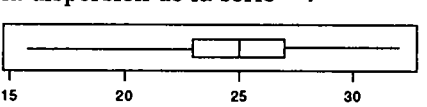
Les 50% sont atteints pour une valeur de la classe 25 élèves donc la médiane Me est 25.

Les 25% sont atteints pour une valeur de la classe 23 élèves donc le premier quartile Q1 est égal à 23.

Les 75% sont atteints pour une valeur de la classe 27 élèves donc le troisième quartile Q3 est égal à 27.

La différence entre Q3 et Q1 est l'écart interquartile et l'intervalle [Q1 ; Q3] est l'intervalle interquartile. Il contient 50% des valeurs de la série, c'est donc une mesure de la dispersion de la série.

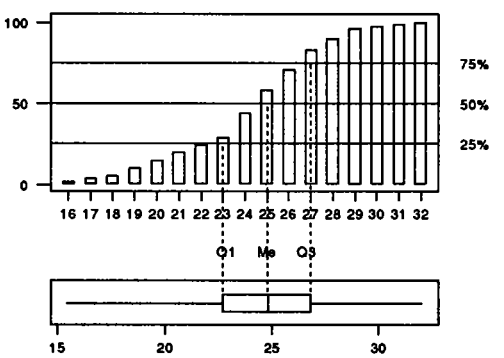
On peut représenter la série par un graphique, très utilisé dans les pays anglo-saxons mais encore peu répandu en France, (bien que figurant dans les potentialités de certains calculatrices commes les TI 80 ou 82 et HP 38G), et appelé box-plot (6), ou graphique en boîte ou encore boîte à moustaches, qui fait intervenir la médiane et les quartiles Q1 et Q3 et donne donc une illustration à la fois de la médiane, c'est-à-dire la tendance centrale, et de l'intervalle interquartile, c'est-à-dire de la dispersion de la série (7).



(6) John W. Tukey, *Exploratory Data Analysis*, Addison Wesley, Reading MA, 1977.

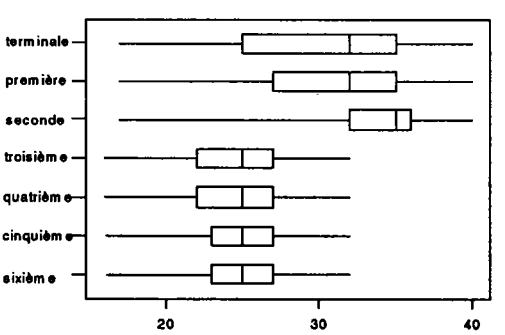
(7) Pour plus d'explications sur le calcul de la médiane et la construction des graphiques en boîte, voir par exemple, Jean Claude Girard, "La médiane, pour quoi faire ? Un exemple d'utilisation : les boîtes de dispersion", *Actes du premier colloque de la commission Inter Irem, Enseignement de la statistique et des probabilités*, Toulouse 14 et 15 mai 1993.

Le passage des fréquences cumulées au graphique en boîte peut se faire sur le graphique suivant :



Ce graphique illustre en particulier ce qu'on peut lire dans les données initiales (mais il faut avoir l'idée d'aller le chercher) que plus de 50% des classes de sixième ont entre 23 et 27 élèves.

Ce genre de graphique prend tout son intérêt lorsque l'on veut comparer plusieurs séries statistiques. Par exemple si l'on veut analyser les effectifs des différentes classes du lycée et du collège. On peut alors refaire le même travail pour chaque classe à partir des chiffres du ministère (même source) puis représenter côte à côte les sept graphiques en boîte.



On constate alors que les classes de collège sont assez semblables de par leurs médianes (25) et de par leurs dispersions (même intervalle interquartile pour les 6^e-5^e, même intervalle pour les 4^e-3^e, peu de différence entre les deux catégories) par contre l'effectif médian est très supérieur en lycée et spécialement en seconde, les classes de ce niveau présentent, de plus, les effectifs les plus élevés et ce de façon homogène alors que les premières et les terminales présentent plus de variations autour de la valeur centrale (effet des différentes séries du baccalauréat qui provoquent de petits effectifs dans les lycées de taille moyenne).

Pour conclure, ce genre de travail peut donner du sens aux concepts de tendance centrale et de dispersion ainsi qu'à l'idée de comparaison de séries statistiques autrement que par le calcul des moyennes, ce qui n'a souvent pas de sens. En cela, c'est une bonne préparation au travail sur l'écart type qui sera vu en seconde et une approche des distributions de probabilités qui seront vues plus tard.

III. Le hasard et la variabilité

L'étude des statistiques en collège devrait être une préparation à l'étude des probabilités au lycée. Si l'on veut que cette louable intention soit suivie d'effet il faudrait que soient abordés au moins deux aspects qui constituent le cœur des problèmes où l'on fait intervenir des modèles probabilistes :

– la notion de hasard
– la notion de variabilité des résultats de certaines expériences que l'on qualifie justement d'aléatoires, c'est à dire dont on ne peut prévoir ni calculer le résultat.

1) La définition d'épreuve aléatoire devait faire l'objet de travaux pratiques en première (8) mais la version définitive du programme n'en fait plus état. Faut-il en conclure que c'est inutile ou que l'on suppose que cela a été fait avant ? La première hypothèse est à rejeter de façon évidente. La plupart des difficultés rencontrées en probabilités proviennent du passage de la réalité de l'expérience à la modélisation dans laquelle on effectuera les calculs. La première condition pour trouver le bon modèle est de bien définir l'épreuve aléatoire et par conséquent d'avoir une bonne représentation de ce qu'est une telle épreuve: "L'objectif est d'entraîner les élèves à décrire quelques épreuves aléatoires simples... Il est important que les élèves puissent se familiariser avec les probabilités pendant une durée suffisante. L'étude de ce chapitre ne doit pas être bloquée en fin d'année." (9)

Il n'est pas évident de faire prendre conscience de la variabilité des résultats dans la répétition de certaines expériences que l'on qualifie alors d'aléatoires. Pléonasme peut-être mais comment les élèves ne seraient-ils pas surpris que, dans les mêmes conditions, une même expérience ne donne pas toujours le même résultat. La physique (déterministe) a dû les convaincre que si les conditions initiales sont données, alors les résultats peuvent être calculés aux erreurs de mesure près ! Il n'est pourtant pas difficile de trouver des contre-exemples sans revenir une fois de plus au lancer d'un dé !

– des graines de qualité semblable plantées en grande quantité dans un même champ produisent des plants de tailles différentes. On peut modéliser cette situation par une expérience aléatoire.

– des frères et sœurs ont le même patrimoine génétique et pourtant il existe de nombreuses différences entre eux. Là encore, les lois de l'hérédité font intervenir le "hasard" comme "explication".

– de la même façon, on observera des variations entre des échantillons issus d'une même population car le hasard ne les aura pas constitués rigoureusement identiques. Par exemple, lorsque l'on étudie la variation de l'opinion par deux sondages successifs, on peut s'attendre à des résultats différents même s'il n'y a pas eu de modification au niveau de la population. C'est le rôle de la statistique inférentielle de faire la part de variation qui revient au hasard et celle qui traduit un réel changement de l'opinion. Le calcul des probabilités permet de calculer la probabilité de l'écart observé dans l'hypothèse où les deux échantillons seraient issus d'une même population, c'est-à-dire si l'opinion n'avait pas évolué. Si cette probabilité est trop petite (inférieure à 5% par exemple), le hasard, d'où découle la variabilité à laquelle on peut s'attendre dans la répétition d'une telle expérience ne permet pas d'expliquer raisonnablement la différence observée et alors on refuse l'hypothèse d'une opinion stable. On parlera alors de différence significative.

Il me semble que l'on peut faire en collège un travail d'approche de cette notion de variabilité c'est-à-dire des variations des résultats dans un même épreuve aléatoire. Pour cela on échappe à la manipulation de chiffres, ce qui peut paraître fastidieux mais qui me semble

(8) Projet de programme de première.
(9) *Ibid.*

269,7	263,4	268,8	272,9	266,4	262,2	268,7	262,3
263,6	260,7	260,3	264,5	255,8	271	261,2	261,2
264,4	265	263,4	266,2	267,1	264,4	263,1	262,1
259,7	267	267,6	265,9	265,5	269,8	264,6	261,4
262,4	265,6	264,1	265,3	264,5	266,1	258,7	264,8

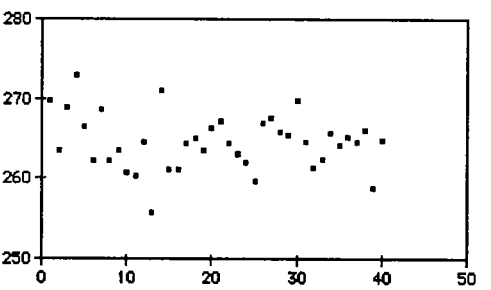
Tableau 8

indispensable au moins une fois dans une scolarité si l'on veut mettre le doigt sur cette idée de variabilité. Il faudrait évidemment trouver des données, réelles si possible, et qui aient un sens pour les élèves. On peut en recueillir par exemple à l'occasion d'une visite dans une usine. Pour aller plus vite, on peut prendre un exemple dans un livre, mais les élèves vont-ils comprendre que sur une machine réglée de la même façon, avec la même matière première et à des instants très rapprochés (production en continu) les résultats obtenus puissent être différents et surtout que l'on ne puisse pas prévoir le suivant ?

Exemple : Les données suivantes (10) représentent le poids en grammes d'un joint d'étanchéité utilisé dans l'industrie automobile et obtenu d'une production continue. Chaque valeur correspond à une production de 30 secondes. La variation dans l'écoulement du caoutchouc provenant de l'extrudeuse affecte directement les dimensions du joint. Quarante données ont été obtenues sur une période de production d'environ 30 minutes. Elles représentent un échantillon de la production(tableau 8).

(10) Les données de l'exemple sont extraites de *Maîtrise statistique des procédés*, Gérald Baillargeon, Les éditions SMG, Trois Rivières, Québec, 1992.

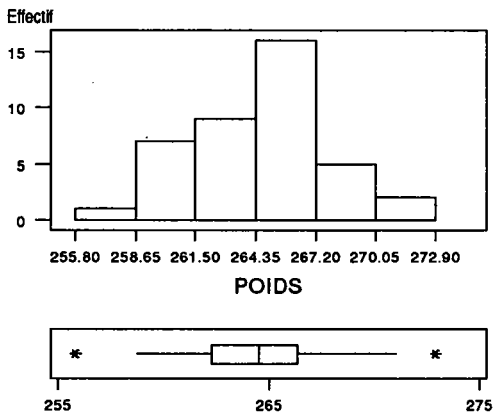
Les données sont dans l'ordre où elles ont été obtenues et peuvent être représentées dans cet ordre chronologique sur le graphique suivant :



Les valeurs semblent arriver au hasard, on ne peut prédire la suivante. Que peut-on faire ou dire dans ces conditions ? Pour dépasser les remarques naïves ou évidentes des élèves ("ça varie", "c'est pas très précis", "la machine n'est pas bonne") et être efficace (mettre en place un contrôle de qualité, pouvoir dire quand il est nécessaire de régler la machine, savoir si un outil est adapté ou non à la production etc.), on peut se placer dans un cadre statistique, c'est à dire dans un modèle expliquant en partie par le hasard, la variabilité des résultats.

Une étude possible au collège pourrait commencer par une représentation graphi-

que. On pensera bien sûr à l'histogramme puisque que les mesures individuelles sont variées (36 valeurs différentes sur 40).



Explications : On a fait figurer en dessous de l'histogramme le graphique en boîte présenté au paragraphe précédent avec une variante qui consiste à détacher des "moustaches" les valeurs trop éloignées et que l'on qualifiera d'aberrantes (11). La règle choisie (Règle de Tukey (12)) est de qualifier d'aberrante une valeur supérieure à Q3 + 1,5 × I ou inférieure à Q1–1,5 × I où I est l'écart interquartile c'est-à-dire Q3 – Q1.

Mais les règles de construction de l'histogramme sont difficiles (le choix du nombre d'intervalles change la forme du graphique, problème des intervalles semi-ouverts, etc.) et, de plus, il peut ne pas avoir de sens pour les élèves. On pourrait procéder, en guise de préalable, à la construction d'un graphique plus simple qui a l'avantage de donner à peu près la même représentation de la série et cela en ne perdant aucune information. Il s'agit du graphique en tige et feuilles (13) (ou *stem and leaf*).

```
255   8
256
257
258   7
259   7
260   37
261   224
262   1234
263   1446
264   1445568
265   03569
266   124
267   016
268   78
269   78
270
271   0
272   9
```

Explications :

La valeur 255,8 est représentée par
255 | 8
tige | feuille
La ligne 261 224 représente les valeurs 261,2 261,2 et 261,4 .

On peut, avec la même règle que pour le graphique en boîte, faire figurer les valeurs très éloignées. On les repère par LO et HI. Il conviendrait de voir si elle n'ont pas fait l'objet d'une erreur de mesure ou de transcription.

(11) Car ces valeurs ont une très faible probabilité d'apparaître. On peut calculer cette probabilité pour certains modèles. Par exemple, pour la loi normale, elle est d'environ 3,5.

(12) *Op. cit.*

(13) John W. Tukey, *Ibid.*

De ces graphiques, on peut faire ressortir quelques observations.

```
LO    2558,

258   7
259   7
260   37
261   224
262   1234
263   1446
264   1445568
265   03569
266   124
267   016
268   78
269   78
270
271   0

HI    2729,
```

La distribution des valeurs n'est pas quelconque, encore moins uniforme. On a beaucoup de "chances" de se trouver proche de la valeur médiane qui est 264,5. On remarque que 31 valeurs (soit plus de 75%) se trouvent entre 261,2 et 267,6. On retrouve un peu d'ordre dans notre hasard.

Cette observation est à mettre en relation avec ce qui se fera en seconde quand on pourra faire usage de l'écart type : 95% des valeurs à moins de deux écarts types de la moyenne. Malheureusement ce dernier résultat n'est valable que les pour les distributions normales.

Prolongements possibles (compréhensibles au collège) :

– Que peut-on conclure de l'observations des ces 40 valeurs si l'on suppose la machine réglée convenablement au début de la production ?
– Que penser, si dans la suite de la production, une heure plus tard par exemple, on mesure une valeur de 268,5 ?
– Que penser, si dans la suite de la production, on mesure une valeur de 255 ?
– A partir de quelle(s) valeur (s) doit-on suspecter un dérèglage de la machine ?
– Que penser de la machine utilisée si on doit avoir impérativement un poids compris entre 264 et 266 pour que la production soit acceptable ?

Remarques

Il s'agit seulement de faire prendre conscience de quelques faits :

– dans de nombreuses expériences, répétées pourtant dans les mêmes conditions, les résultats présentent une certaine variabilité,
– les mathématiques prennent en compte ce genre de situations en fournissant des modèles faisant intervenir le hasard; on peut alors retrouver une certaine stabilité au cœur de ces variations et faire des prévisions,
– une valeur éloignée de la moyenne ou de la médiane n'est pas impossible mais doit attirer notre attention,
– dans le cas étudié, la variabilité des résultats est liée à la précision de la machine c'est-à-dire à sa capacité à produire des pièces dont la mesure est plus ou moins proche de la valeur de réglage.

En conclusion

Ce genre de travail devrait permettre :

– d'analyser une série statistique de façon critique,
– d'être confronté à une épreuve aléatoire concrète,
– d'appréhender les effets du hasard,

– de retrouver des régularités au sein des résultats aléatoires,
– de sensibiliser les élèves à une application des statistiques : le contrôle de qualité (14),
– et surtout de montrer que l'on ne fait pas de statistiques uniquement pour obtenir un beau graphique ou une moyenne avec quatre décimales.

* * *

L'objectif de cet article était d'illustrer ce que l'étude des statistiques au collège (et même après) pouvait apporter, en dehors de ce contexte, à la formation générale ainsi qu'aux autres domaines mathématiques, à la préparation à l'étude des probabilités et à la formation du citoyen en particulier concernant la familiarisation à l'idée de variabilité qui relativise l'importance quasi mythique donnée à la moyenne et montre la nécessité de prendre en compte la dispersion d'une série statistique.

S'il présente beaucoup d'intérêt, cet enseignement présente aussi de nombreuses difficultés. La réflexion doit être poursuivie dans différentes directions, par exemple sur la pertinence de l'introduction des probabilités (au moins des expériences aléatoires) au collège, sur l'apprentissage des pourcentages et sur celui de la lecture de graphiques (Quelles sont les conceptions spontanées des élèves devant un graphique ? Comment les aider à construire de bonnes images mentales ?). Cela ne se fera qu'en réfléchissant à des exemples concrets et intéressants qui donnent, en prime, du sens aux concepts statistiques étudiés à ce niveau de scolarité (15).

(14) Dans la réalité, les contrôles sont effectués à partir de la moyenne des valeurs d'un échantillon dont l'effectif est souvent égal à 5. Voir, par exemple, *Maîtrise statistique des procédés*, Gérald Baillargeon, Les éditions SMG, Trois Rivières, Québec, 1992.

(15) Voir par exemple, *L'empereur et la girafe. Leçons élémentaires de statistiques*, Claudine Robert, Diderot éditeur, Paris, 1995.