Dossier lycée professionnel

Ce dossier est le compte rendu de l'essentiel du Séminaire APMEP 1996 "Moyens et méthodes de l'enseignement des mathématiques, du BEP au BTS ; quels apports pour les autres cycles ?".

Les Statistiques, le calcul des probabilités et l'enseignement professionnel

Jean-Louis Piednoir

La statistique, voire le calcul des probabilités, sont peu enseignés en France dans l'enseignement secondaire et même dans l'enseignement supérieur. Une exception notable : les enseignement à finalité professionnelle, du BEP au BTS. Cela n'est pas étonnant, compte tenu de leur applications dans l'industrie et dans le secteur tertiaire. Les professeurs n'ayant, sauf exception, que peut de contacts avec ces branches des mathématiques, sont souvent réticents devant leur enseignement et n'y voient trop souvent que procédures à faire apprendre par cœur ou occasions à appliquer des techniques mathématiques enseignées par ailleurs. L'objectif de cet exposé est de proposer une approche de la statistique et du calcul des probabilités proche des préoccupations de l'enseignement professionnel.

I - Les a-priori de départ

Tout le monde sait que les élèves de l'enseignement professionnel ne font pas partie de l'élite scolaire, qu'ils ont des lacunes dans la maîtrise des connaissances mathématiques du collège. Pourtant l'expérience montre qu'il est possible de leur enseigner des choses relativement complexes (pas toujours, hélas !), dès lors que leur présentation est adaptée à leurs préoccupations. Les axiomes qui suivent, et qui bien entendu sont discutables, explicitent les a priori qui fondent la suite de l'exposé.

Axiome 1 : Ces élèves ne sont pas plus bêtes que les autres, mais ils apprennent autrement. Ils apprécient ce qui est lié avec l'expérience courante, ils privilégient l'aspect efficace de la démarche mathématique. D'où la nécessité d'une approche interdisciplinaire qui donne un sens aux apprentissages mathématiques.

Axiome 2 : Comme les autres élèves, ils ont droit à des ouvertures culturelles. La statistique le permet ; elle n'est pas seulement un prétexte pour trouver l'équation d'une droite connaissant deux points.

Axiome 3 : L'efficacité de l'enseignement mathématique passe par l'enseignement de notions nouvelles rompant avec la reprise des notions classiques sur lesquelles on est en échec depuis parfois plusieurs années.

Axiome 4 : Les mécanismes non acquis au collège peuvent peu à peu être maîtrisés à condition que cette maîtrise apparaisse comme incontournable pour la compréhension de notions indispensables.

II - Une proposition en calcul des probabilités

1 - Les programmes

Les éléments de calcul des probabilités qui figurent aux programmes des BEP, des baccalauréats professionnels, des BTS, sont justifiés, dans le secteur industriel, par les procédures de contrôle de fabrication qui sont d'usage courant. Certes, ces procédures sont entièrement codifiées, elles font l'objet de normes AFNOR. On peut donc les mettre en route sans rien connaître de leur justification. Mais l'objet de l'enseignement est précisément de former des opérateurs intelligents.

Au niveau du BTS, pour la fiabilité, on applique aussi le calcul des probabilités aux durées de vie des équipements.

2 - Étude des situations aléatoires

La première étape de l'étude du calcul des probabilités consiste à observer des situations aléatoires et à collectionner des mesures qui sont différentes les unes des autres pour des facteurs contrôlés tous identiques. On

peut choisir ses exemples dans des champs très variés. Les cotes de pièces fabriquées en série en font partie. Dans l'industrie, on prélève des échantillons, les mesures sont reportées sur des cartes de contrôle que l'on peut utiliser.

Les sciences du vivant sont, par excellence, le champ de l'aléatoire. Les ouvrages de statistique et probabilités spécialisés regorgent d'exemples.

Dans le second degré, les exemples les plus répandus proviennent des jeux de hasard. Mais on introduit alors une difficulté supplémentaire. Dans ces jeux, on fait souvent une hypothèse d'équiprobabilité (cf. ci-dessous, la modélisation).

On peut également, à l'aide de la touche RANDOM des calculettes ou de l'ordinateur, fabriquer du hasard. Celui-ci a l'avantage d'être connu, la probabilité d'un événement est alors une valeur numérique contrôlée.

A partir de ces observations, on peut mettre en évidence les lois des grands nombres : quand on répète un grand nombre de fois une expérience aléatoire simple, succès/échec, la fréquence relative du nombre de succès se stabilise autour du nombre qui est la probabilité de l'événement "succès". Les simulations sur calculette se prêtent facilement à cette observation.

3 - La modélisation

Le bon langage pour rendre compte des situations aléatoires est le langage mathématique. Le cas le plus simple est celui où l'ensemble des cas possibles est fini. On associe à l'ensemble $\Omega = \{1, 2, \ldots n\}$ les nombres $(p_1, p_2, \ldots p_n)$ où pi est la probabilité de l'issue $\{i\}$. A partir de là, il est facile de calculer la probabilité de E sous-ensemble de Ω à partir des règles régissant les mesures. Pour les enseignements qui nous intéressent, il est inutile de multiplier les situations d'équiprobabilité plus ou moins complexes. Le calcul de la probabilité se réduit alors à un problème de combinatoire. Il faut compter le nombre de cas favorables et le nombre de cas possibles.

L'introduction de la variable aléatoire engendre une difficulté supplémentaire. A la situation (Ω, p) où $p = (p_1, p_2, \ldots p_n)$ on ajoute une application X dans l'ensemble des réels $\mathbb{R}$. On se contentera des cas simples : nombre de succès dans la répétition n fois d'une situation à deux issues succès, échec, par exemple, sans trop insister dans une première approche sur l'indépendance desdites répétitions. On peut affirmer alors que les mathématiciens ont démontré la loi des grands nombres c'est-à-dire que la probabilité pour que la fréquence des succès s'écarte de la probabilité dudit succès tend vers zéro avec le nombre d'essais. Par simulation, il est facile de mettre expérimentalement en évidence le phénomène.

4 - La loi normale

On aimerait bien ne pas parler de probabilité sur un ensemble infini comme IR. Malheureusement, cela est pratiquement impossible. Les applications industrielles font appel à des variables continues : diamètre d'une pièce, durée de vie d'un équipement, etc. La loi normale est omniprésente. La manipulation de la table, quand on donne la moyenne et l'écart-type de cette loi est une vraie difficulté. Il faut pourtant être capable de calculer les probabilités d'un intervalle dans un cas donné : *quelle est la probabilité pour que la dimension d'une pièce donnée soit comprise entre a et b ?*

L'importance de la loi normale peut, par contre, être mise en évidence par expérience. Des fabricants vendent des dispositifs dans lesquels la fameuse "courbe en cloche" apparaît facilement. On peut alors énoncer ce que les mathématiciens appellent le théorème central limite : *quand un phénomène exprimable par un nombre réel est la résultante additive de causes aléatoires nombreuses, de faible ampleur chacune, toutes d'un même ordre de grandeur, alors il suit une loi normale.*

Cependant, il est bon d'ajouter que tout n'est pas gaussien. Les histogrammes rendant compte des temps d'attente d'un taxi à un carrefour donné, les durées de vie d'un équipement, n'ont pas l'allure d'une courbe en cloche.

5 - Sensibilisation au hasard

L'expérience montre, qu'habitués au déterminisme, nos contemporains sont peu sensibles au hasard. A l'aide de jeux, il est possible d'acquérir cette sensibilité. Par exemple, on jette 100 fois une pièce de monnaie honnête. Le résultat 43 piles et 57 faces est-il plaubible ?

La légende raconte que le biologiste Mendel, quand il a énoncé les lois de l'hérédité, aurait rapproché la proportion de pois à peau lisse à la deuxième génération du chiffre simple de sa théorie. Réexaminés a posteriori, ses résultats paraissent trop beaux pour être vrais !

III - La statistique

1 - Observation des mœurs de la tribu des statisticiens

La démarche statistique est omniprésente dans nos sociétés et, pour certains problèmes, fort ancienne. Mais elle est mal connue de nos contemporains, même de ceux qui ont une culture scientifique. En France, elle est fort peu enseignée, pourtant elle est incontournable quand il s'agit de rationaliser des prises de décision. Quelques exemples illustrent cette assertion.

Exemple 1 : Un commerçant doit faire des commandes à ses fournisseurs,

ses ventes fluctuent d'un jour à l'autre. C'est pourtant à partir des observations du passé qu'il déterminera sa commande.

Exemple 2 : Un armurier passe, à la veille de l'ouverture de la chasse, une commande. Il veut contrôler si le lot de cartouches livré est de bonne qualité. Il a une solution pour le savoir : essayer toutes les cartouches...mais il n'a plus rien à vendre ! Il lui reste une solution : juger sur un échantillon extrait du lot.

Exemple 3 : L'économiste étudie les salaires pratiqués dans un pays déterminé. Il a de très nombreuses données. Comment les synthétiser pour pouvoir raisonner sur quelques chiffres simples ?

Exemple 4 : Le contrôleur a mesuré les dimensions d'un lot de pièces prélevé dans une fabrication en série. Lui aussi cherche quelques indications capables de caractériser la série.

Tous ces exemples illustrent la démarche statistique. On a des individus appartenant à une population. Sur chaque individu, on fait une mesure. Les mesures faites sont différentes, d'un individu à l'autre il y a variation. La statistique se nourrit de cette variabilité.

2 - Réactions de la vie courante

Les individus sont donc différents. Des différentes mesures faites sur eux, la statistique prétend dégager une mesure sur la population dont ils font partie. Cette démarche choque toujours. En effet, la manipulation des résultats statistiques est délicate. Par exemple, si une liaison entre deux phénomènes est établie, on glisse inconsciemment à une explication de nature causale qui peut être fausse. La réaction naturelle est alors d'invalider la procédure alors que seule son interprétation est en cause.

Un exemple illustre cela. Dans le cadre du "politiquement-correct", on veut savoir outre-Atlantique si la sélection dans les universités est ou n'est pas sexiste. Les premiers résultats tendraient à le prouver. Dans les chiffres ci-dessous, le numérateur est le nombre de candidats sélectionnés, le dénominateur le nombre total de candidats. Cela est établi pour les hommes et pour les femmes.

	Hommes	Femmes
Total université	534/1198 = 0,446	113/449 = 0,252

Mais on fait maintenant l'analyse par département. L'université comprend les départements A et B.

	Hommes	Femmes
département A	512/825 = 0,621	89/108 = 0,824
département B	22/373 = 0,059	24/341 = 0,070

Chaque département paraît privilégier un recrutement féminin, pourtant l'université paraît privilégier un recrutement masculin. L'explication est simple : les femmes ne se comportent pas comme les hommes et vont majoritairement porter leur candidature au département le plus sélectif. Il existe une variable, le département, importante et non apparente dans le résultat brut.

L'étude comparée de la réussite scolaire des enfants de français et des enfants étrangers conduit à des conclusions analogues. Globalement, les premiers réussissent mieux que les seconds. Comme il est connu que la catégorie socioprofessionnelle (C.S.P.) des parents influe sur la réussite, on compare à même C.S.P.. L'avantage reste aux enfants de Français, même s'il est plus réduit. En terme de remédiation, on pensa alors à proposer des cours de français pour les population allogènes. Le CEFISEM approfondit l'étude et introduisit une nouvelle variable : la taille de la fratrie. A C.S.P. et à taille de fratrie égale, les résultats s'inversent. Les enfants d'étrangers réussissent mieux que leurs camarades français. En terme d'action scolaire, il ne s'agit plus de faire des cours de langue, mais plutôt de proposer des études surveillées. Dans une famille nombreuse, il est plus difficile de faire ses devoirs, d'apprendre ses leçons. Le tableau ci-dessous résume le phénomène. L'indicateur pris est la proportion d'élèves arrivant en Sixième sans redoublement en primaire.

Profession du père	Français	Etrangers
Total	64,1	43,4
Ouvriers non qualifiés	47,2	38,8
Ouvriers non qualifiés famille de 3 enfants et plus	33,2	35,4

La conclusion trop vite tirée par le sceptique est alors la suivante : *« on fait dire ce qu'on veut aux statistiques »* . La version sophistiquée est plus plaisante : *«la statistique est comme le bikini elle montre tout mais elle cache l'essentiel »*. La mise en cause de la méthode est d'autant plus forte que les conclusions heurtent les préjugés. On adopte alors la méthode que l'on peut appeler voltairienne : l'idée a priori vaut mieux que tous les résultats empiriques obtenus par la méthode statistique. Pourquoi voltairienne ? Au XVIII[e] siècle, régnait l'opinion suivante : la population du royaume de France baisse. Les guerres de Louis XIV, les mauvaises récoltes et donc les famines pouvaient étayer cette opinion. Les disciples de Vauban eurent l'idée d'estimer, par une méthode proche des sondages actuels, ladite popu-

lation. Loin de diminuer, cette dernière augmentait. La publication des résultats heurta le sens commun et Voltaire entreprit, par un libelle fort bien troussé, de se moquer de ces statisticiens qui s'étaient forcément trompés car tout le monde savait bien que la population du royaume baissait.

3 - La démarche statistique

Les exemples précédents ont montré quel était l'objectif de la statistique. On définit une population $\mathcal{P}$ composée d'individus et on cherche des caractéristiques de $\mathcal{P}$. Mais il n'est pas possible d'effectuer des mesures directement sur $\mathcal{P}$. On est obligé de passer par les individus. Pour cela, on prend dans $\mathcal{P}$ un échantillon $\mathcal{E}$ (il est possible que $\mathcal{E} = \mathcal{P}$; on a alors un recensement). Sur chaque individu i de $\mathcal{E}$, on effectue une mesure x_i prise dans un ensemble X. Comme il y a variabilité, les x_i ($i \in \mathcal{E}$) ne sont pas tous identiques. Le problème statistique est d'agréger ces n mesures pour trouver une mesure sur la population. On cherche donc une application g de X^n dans un ensemble Y de mesures caractérisant la population et on cherche $y = g(x_1, \ldots x_n)$.

Les statisticiens appellent g un résumé.

Prenons l'exemple des cartouches. Dans le lot fourni, on prend un échantillon de 100. On tire chaque cartouche. On pose $x_i = 1$ si elle est bonne et $x_i = 0$ sinon. On a : $X = \{0 , 1\}$. Le résumé sera la proportion de bonnes cartouches dans l'échantillon : $y = \frac{1}{100}(x_1 + x_2 + \ldots + x_n)$ et $Y = [0 , 1]$. Dans l'enquête sur les salaires, chaque salarié a un salaire x_i, $X = \mathbb{R}^+$, les résumés peuvent être divers selon les besoins de l'étude : salaire moyen $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} x_i$ si N est le nombre de salariés ; salaire médian : $\tilde{y} = \text{Méd}(x_1 \ldots x_N)$ c'est le salaire tel que la moitié gagne moins de y et l'autre moitié plus. On peut aussi caractériser l'inégalité de la distribution mais c'est une autre affaire.

Le travail du statisticien est donc de construire des résumés adaptés aux objectifs du spécialiste concerné : commerçant, biologiste, sociologue, etc. Il faut donc choisir parmi tous les résumés possibles celui qui serait le meilleur ou plus modestement pas trop mauvais. Pour cela, il faut des critères. On peut en observer, qui permettent un premier tri qualitatif :

(i) Le résumé doit être peu sensible à la présence ou à l'absence dans la population d'un individu donné.

(ii) Si on peut comparer plusieurs populations identiques, on s'attend à ce que les variations entre résumés soient de moindre ampleur que les variations entre individus d'un même échantillon. En quelque sorte, on amortit les fluctuations individuelles.

Mais ce tri est insuffisant. Par exemple, il ne permet pas d'éliminer les "stratégies voltairiennes". Pour aller plus loin, pour dégager des critères quantitatifs plus précis, il faut construire un modèle mathématique décrivant des hypothèses supplémentaires que l'on peut faire sur la population..

A ce stade, on distingue, suivant le type de données, de modèle, deux types de statistique. Si on considère les individus de l'échantillon comme "tirés au hasard", la population est caractérisée par une loi de probabilité qu'il s'agit de déterminer. On appelle cette statistique la statistique inductive ou statistique mathématique. Elle n'est enseignée dans les lycées qu'au niveau BTS (cf. ci-dessous). On peut également considérer que telle ou telle structure mathématique est adaptée à la description des données. Si par exemple on suppose que les mesures numériques (prises dans $\mathbb{R}$) peuvent être plongées dans un espace euclidien, alors la moyenne, la variance, le coefficient de corrélation prennent tout leur sens. On appelle cette statistique la statistique descriptive qui se prolonge en analyse des données.

4 - La statistique descriptive

Il n'est pas possible dans un aussi bref exposé de faire le tour des techniques enseignées dans les lycées. On insistera sur quelques ambiguïtés. Les données les plus simples sont relatives à un caractère qualitatif à deux catégories : la cartouche est bonne ou mauvaise, l'élève a parcouru le cycle primaire sans ou avec redoublement. La difficulté commence quand il faut, soit comparer plusieurs caractères quantitatifs, soit vérifier que la population considérée a bien une existence scientifique.

Pour illustrer le propos, reprenons les exemples précédents. Si on étudie la sélection dans les universités américaines, on s'aperçoit que la population de départ est loin d'être homogène. En fait on a affaire à deux sous-populations, les hommes et les femmes qui se comportent de façons très différentes quant à leurs vœux. Il vaut mieux étudier séparément ces deux sous-populations. Mais il n'est pas toujours aisé de les déterminer.

Cet exemple, mais surtout le suivant, relatif à la réussite scolaire, peuvent être interprétés en termes de variable cachée. En première observation on pouvait conclure que l'origine ethnique (français/étrangers) était à l'origine d'une réussite scolaire différenciée. En fait intervenait une troisième variable : la taille de la fratrie. Si on la fait rentrer dans l'analyse, on voit que l'origine ethnique ne joue pas de rôle négatif, elle ne peut être la cause de la réussi-

te scolaire moindre des étrangers. On peut poser comme hypothèse que la taille de la fratrie en est la cause. Des études plus approfondies pourraient peut-être l'invalider.

On voit donc que, même sur des données aussi simples que la proportion on peut faire travailler les élèves, les faire réfléchir, former leur sens critique.

Si on prend un caractère quantitatif comme le salaire on peut se demander quel est le meilleur indicateur de centralité. En d'autres termes, quel est l'ordre de grandeur des salaires versés. Si on fait de la sociologie, il paraît évident que le salaire médian est le bon indicateur : 50% des salariés gagnent moins, 50% plus. Par contre, si on fait de l'économie, pour avoir une idée de la capacité des salariés à dépenser, le salaire moyen paraît plus pertinent. Or, ces deux indicateurs sont différents. La distribution des salaires étant asymétrique, les hauts salaires tirent la moyenne vers le haut. Le salaire moyen est donc plus élevé que le salaire médian.

Les mêmes remarques peuvent être faites sur les mesures de dispersion : écart-type ou intervalles interquartile, interdécile. Tout dépend du type de données et de l'usage qui peut en être fait. Si par exemple il existe un individu dont la valeur prise est nettement plus grande que celle des autres observations, les indicateurs moyenne, écart-type vont être anormalement élevés. Sa présence ou son absence va influer fortement sur le résumé violant une des caractéristiques du bon résumé. Il faut alors trouver d'autres indicateurs. En caricaturant, illustrons le propos par les données suivantes : dix mesures entre 1 et 1,1 et une autre à 10, à t avec $t \to \infty$; que deviennent moyenne et écart-type ?

5 - La statistique inductive

Quand il existe un modèle probabiliste, celui-ci est justifié par des conditions très strictes d'expérience. A priori, on ne peut déterminer entièrement la loi de probabilité. Elle dépend de paramètres inconnus qui caractérisent la population. L'objet de la statistique est alors, au vu de l'expérience, de donner une valeur à ces paramètres. La statistique inductive est alors l'art de mesurer des probabilités.

Illustrons le propos par le problème des cartouches. On reçoit un lot très grand de cartouches. On va tirer au hasard 100 cartouches dans ce lot. Cela veut dire que chaque cartouche a la même probabilité de figurer dans le lot et que les tirages sont indépendants (l'un n'influe pas sur les suivants). Cela est raisonnable si N, le nombre de cartouches du lot est très grand devant 100 (N>>100). La probabilité d'avoir k bonnes cartouches ($0 \leqslant k \leqslant 100$) est alors calculable. Soit p_k cette probabilité. On a : $p_k = C_{100}^{k} p^k (1-p)^{100-k}$ mais p

est inconnu. Soit ω l'échantillon tiré, $k(\omega)$ le nombre de bonnes cartouches dans l'échantillon. Une théorie sophistiquée, mais intuitivement compréhensible nous proposera comme valeur raisonnable de p la quantité : $\hat{p}(\omega) = \frac{k(\omega)}{100}$ appelée estimation de p. La variable $\hat{p}(\cdot)$ qui à tout échantillon ω associe $\hat{p}(\omega)$ est dite estimateur de p. On peut mesurer, dans le cadre de la théorie précitée, sa pertinence.

Dans ce modèle, il est bien sûr possible d'analyser beaucoup de situations plus complexes, de définir pour une estimation un intervalle de confiance, de tester des hypothèses, bref, de prendre des décisions rationnelles.

IV - La statistique et l'enseignement

Avec la statistique et le calcul des probabilités, les deux sont liés quand on parle de statistique inductive, on a des domaines des mathématiques dont les applications sont de plus en plus présentes dans l'industrie et les services. Elles sont donc inévitables dans les enseignements à finalité professionnelle. Leur aspect concret, inévitablement interdisciplinaire, plaît aux élèves. Comme il s'agit de champs nouveaux, ils sont davantage prêts à s'y investir que dans la répétition de choses déjà vues au collège. Des exemples simples et instructifs peuvent être dégagés. Le présent exposé n'avait d'autre ambition que de dégager quelques perspectives.