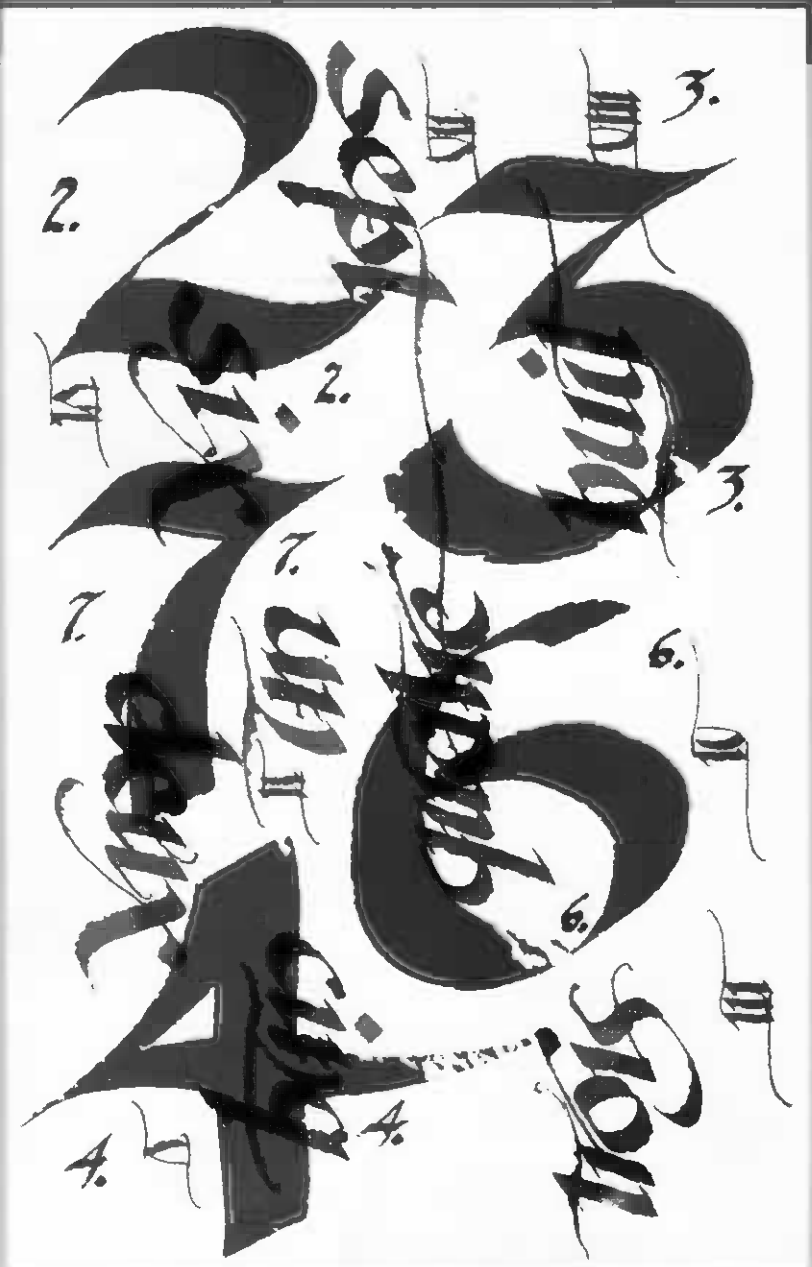
BRIGITTE ESCOFIER
JÉRÔME PAGÈS

DIDACT STATISTIQUE

Initiation aux Traitements statistiques

Méthodes, méthodologie



les PUR
Presses Universitaires Rennes

Brigitte ESCOFIER, Jérôme PAGÈS

Initiation aux traitements statistiques

méthodes, méthodologie

Presses Universitaires de Rennes

Collection Didact mathématiques

A paraître :

Thierry FOUCART, *L'analyse des données, méthodes et études de cas*, septembre 1997.

Le fichier des données utilisé dans cet ouvrage est disponible aux Presses Universitaires de Rennes (adresse ci-dessous) sur simple demande accompagnée d'une enveloppe affranchie portant l'adresse du demandeur.
(fichier ASCII sur disquette)

UHB Rennes 2 - Campus de La Harpe
2, rue du Doyen Denis-Leroy
35044 Rennes

Dépôt légal : 1er semestre 1997
ISBN : 2-86847-231-1

Avant-propos

Objectif général

Ce livre est organisé autour du traitement statistique d'un fichier de données réel, d'une taille assez grande mais tout à fait courante : les notes obtenues au bac et pendant leur année scolaire par environ 1000 élèves de classe terminale scientifique, dite alors terminale C.

Au travers du traitement de ce fichier, l'objectif est de présenter une **culture de base** de statistique opérationnelle, culture dans laquelle on peut distinguer deux composantes :

- un **ensemble d'outils statistiques,** mettant en œuvre des notions simples mais la plupart du temps suffisantes ;
- une **méthodologie**, terme à la fois vague et évocateur qui inclut la façon d'organiser les questions et les traitements statistiques et dont une forme concrète est un guide d'étude d'un fichier de données numériques.

A qui s'adresse ce livre ?

Le public visé est assez large : le contenu du livre est, pour l'essentiel, **accessible sans formation scientifique particulière**, par exemple à des bacheliers littéraires. Dans cet esprit, nous avons restreint autant que faire se peut le recours au formalisme mathématique et largement commenté les inévitables « formules ».

Ce livre doit rendre des services :

- aux **étudiants** en sciences humaines, sciences expérimentales, agronomie, gestion, commerce, etc. Cet ouvrage ne correspond pas au programme d'un cursus particulier, mais il aborde les questions de base qui constituent le cœur de la plupart des programmes de statistique qui s'adressent à des non-mathématiciens. L'étudiant y trouvera une présentation intuitive mais rigoureuse de ces questions, la plupart du temps absente des présentations usuelles plus formalisées.
- aux **enseignants** ayant en charge un cours de statistique ; les présentations et les exemples originaux de ce livre ont été mis au point pour « faire passer » bon nombre de notions réputées difficiles. En particulier, cet ouvrage devrait intéresser les professeurs de mathématiques de l'enseignement secondaire confrontés à l'introduction de la statistique dans les programmes.
- aux **praticiens** qui veulent s'initier à la pratique de la statistique parce qu'ils doivent manier des fichiers de données et/ou interpréter (et critiquer) des résultats statistiques. Ils trouveront un exemple de fichier réel relativement complexe ; ce fichier n'a pas été taillé spécialement pour mettre en valeur les méthodes et son traitement fait apparaître des problèmes dont la résolution nécessite une attitude pragmatique – rarement décrite dans les livres – pour raisonner des choix quelquefois empiriques.

Démarche

A chaque étape, nous introduisons d'abord une question ou un problème puis différents moyens pour y répondre. On illustre ainsi de façon naturelle **quand et comment employer**

les outils statistiques de base. Lorsque plusieurs outils sont utilisables, nous les appliquons systématiquement au fichier des notes : la confrontation de leurs résultats permet de montrer, dans un contexte réel bien précis, leurs intérêts respectifs et leurs limites.

Tous les outils statistiques existants ne sont pas présentés. Nous avons choisi ceux qui nous paraissent devoir être le plus fréquemment employés dans l'étude courante de données ; ils sont pour la plupart assez simples. La présentation de ces outils n'est pas toujours classique : elle a été en partie repensée, à partir de notre expérience d'enseignement et de traitement des données, **en se référant toujours aux questions** auxquelles ils permettent de répondre.

Au-delà des outils, ce livre décrit une méthodologie pour l'étude statistique d'un fichier de données numériques. Son ambition est d'aider un novice en la matière à aborder ces fichiers et même de servir de guide à un utilisateur plus expérimenté. En particulier, de ce point de vue, l'accent est mis sur les premières étapes de l'étude d'un fichier, étapes trop souvent ignorées ou éludées dans les ouvrages classiques : la vérification des données, l'étude et la prise en compte des données manquantes (que beaucoup de logiciels traitent avec des méthodes souvent non précisées et qui peuvent se révéler inadaptées), la recherche et la prise en compte d'éléments remarquables ou "outliers".

Insistons sur l'état d'esprit qui doit accompagner cette méthodologie : avoir l'esprit toujours éveillé, critique et interrogateur devant les données, **ne jamais conclure à partir des seuls chiffres**, mais en conjuguant résultats statistiques et réflexions basées sur des connaissances externes aux données. Il n'est pas inutile de rappeler que la réflexion est au moins aussi importante que le calcul d'indices ou la production de graphiques.

Trois parties

Dans la première partie, les techniques statistiques sont introduites non pas dans l'ordre classiquement rencontré dans les ouvrages, mais au moment où elles sont utiles dans l'étude du fichier des notes. Ce faisant, le pari est de **faire coïncider un traitement de données réaliste et une présentation progressive des techniques** : à quelques exceptions près, qui n'altèrent en rien l'esprit du livre, cette démarche s'est avérée possible.

La seconde partie est consacrée à la mise en évidence d'**éléments remarquables** (valeurs, individus, variables), composante essentielle de la description d'un fichier. En pratique, cette mise en évidence intervient très tôt dans une étude et d'ailleurs nous en mentionnons plusieurs aspects dans la première partie. Mais sa présentation systématique ne pouvait intervenir qu'après la première partie.

La troisième partie comporte **11 fiches techniques**, chacune consacrée à l'un des thèmes essentiels de l'analyse statistique d'un fichier. Ces thèmes sont bien sûr abordés dans la première partie ; mais la présentation des fiches, outre son caractère autonome adapté à une consultation ponctuelle, comprend des aspects techniques dont la présence tout au long de la première partie en aurait rompu le fil directeur. En particulier, ces fiches sont l'occasion d'établir quelques ponts entre la démarche descriptive et la démarche inférentielle classique.

Enfin, il nous est agréable de remercier ici Louise-Marie Dousselin, Jean-Pierre Escofier, Yvette Grelet, Marie-Odile Lebeaux et Annie Morin qui ont accepté de relire tout ou partie du manuscrit.

Sommaire

Partie 1 Traitement d'un fichier de notes

Chapitre 7. Etude de variables quantitatives : répartition des notes

Chapitre 8. Liaison entre deux variables quantitatives : les notes sont-elles liées entre elles ?

Chapitre 9. Synthèse d'un ensemble de variables quantitatives

Chapitre 10. Caractérisation d'une sous-population ; élèves avec données manquantes

Chapitre 11. Comparaison entre plusieurs sous-populations : les élèves d'un même lycée

Partie 2. Eléments remarquables et éléments aberrants

Partie 3. Fiches techniques

Numérotations et renvois

Les parties 1 et 2 se composent de chapitres numérotés de 1 à 14. Les numéros des paragraphes (§), tableaux et figures d'un chapitre commencent par le numéro du chapitre. Exemple : **Fig. 3.2.** désigne la deuxième figure du chapitre 3.
Les fiches sont numérotées de 1 à 11. Les numéros des paragraphes (§), tableaux et figures d'une fiche ne commencent pas par le numéro de la fiche. Exemple : cf. **Tab. 3.** renvoie à la troisième figure de la fiche où se trouve ce renvoi.
Un renvoi à une autre fiche que celle dans laquelle il se trouve est précédé du numéro de fiche. Exemple : cf. Fiche 3 § 3.4 renvoie au paragraphe 3.4 de la fiche 3.

Partie 1

Traitement d'un fichier de notes

Description des données étudiées

On dispose, pour chacun des élèves ayant passé les épreuves écrites du bac C en 1989 dans un même centre d'examen, des notes qu'il a obtenues au bac, des notes trimestrielles telles qu'elles figurent sur son livret scolaire et du numéro du lycée auquel il appartient.

Individus, population, sous-population

Le fichier fourni comporte initialement 993 enregistrements, chacun correspondant à un élève. Un élève est aussi appelé *individu*. Ce terme statistique désigne ici un élève, mais il peut s'agir de bien autre chose. Par exemple, des vins notés sur différents critères par des dégustateurs, et plus généralement des éléments quelconques dont on connaît une ou plusieurs caractéristiques.

L'ensemble des 993 élèves, c'est-à-dire l'ensemble des individus étudiés, s'appelle en statistique la *population*. Lorsque l'on ne considère pas tous les individus mais seulement certains d'entre eux, on parle de *sous-population*. Par exemple, on peut restreindre une partie de l'étude à la sous-population des élèves ayant fréquenté un lycée public.

Enfin, chaque élève est repéré par son numéro d'ordre dans le fichier.

Variable quantitative

Pour chaque élève, on dispose de 20 notes : les notes, au bac et aux trois trimestres, dans les 5 matières de l'écrit du bac (mathématiques, physique, sciences naturelles, histoire-géographie et philosophie). Ces notes sont des valeurs numériques comprises entre 0 et 20. Chaque note (par exemple, la note de mathématiques au bac) est appelée *variable numérique* ou *variable quantitative*. C'est une variable, car sa valeur varie d'un individu à l'autre ; elle est dite numérique car cette valeur est un nombre. On dit aussi variable quantitative (ce sont souvent des quantités qui sont exprimées), terme utilisé le plus souvent dans la suite.

Codage, variable qualitative, modalité

On connaît aussi le lycée auquel appartient chaque individu. Cette appartenance est codée de la façon suivante : à chacun des 22 lycées est affecté un numéro compris entre 1 et 22, le numéro 0 correspondant aux candidats libres. L'appartenance à un lycée est, comme la note, une variable : pour chaque élève, elle prend une valeur qui est le lycée auquel l'élève appartient. Ici cette valeur est codée par un nombre mais elle pourrait être codée tout à fait différemment : par le nom de chaque lycée en entier, par une lettre, par d'autres nombres puisque les nombres ne sont dans ce cas qu'une simple codification. L'appartenance au lycée présente, dans le fichier, une apparence numérique (cela est pratique pour la saisie

informatique des données), mais cette apparence est trompeuse. Ces codifications ne se manient pas comme des nombres : faire la somme ou la moyenne des nombres codifiant le lycée n'a strictement aucun sens. La variable *lycée*, qui prend des valeurs non numériques, est appelée variable *qualitative* (ou variable *nominale*). Ses valeurs sont appelées *modalités* de la variable qualitative. Ici, la variable qualitative *lycée* possède 23 modalités (22 lycées et *candidat libre*).

Groupes de variables

L'ensemble des 20 notes est très structuré. Il y a 5 matières et 4 notes dans chaque matière ; ces 4 notes se correspondent puisqu'il s'agit, pour chaque matière, des notes obtenues au bac et aux 3 trimestres. On peut donc diviser l'ensemble des notes, soit en 5 groupes (les matières) soit en 4 groupes (les dates). Un autre point de vue assez logique consiste à séparer les 15 notes trimestrielles et les 5 notes du bac : on obtient alors 2 groupes de variables d'effectifs différents. Cette situation, dans laquelle les variables se regroupent d'une ou de plusieurs façons, est très fréquente. Elle influe forcément sur les objectifs de l'étude et donc sur les analyses statistiques à envisager ; aussi est-il toujours intéressant d'expliciter l'existence de groupes de variables, ce qui permet d'orienter les traitements à réaliser.

Tableau de données

L'ensemble de ces données peut être présenté comme un grand tableau rectangulaire ayant en lignes les 993 élèves et en colonnes leurs 21 caractéristiques (le numéro de lycée + 20 notes). A titre d'exemple, examinons les 2 premiers élèves du fichier (cf. **Tab. 1.1**). La première ligne concerne le premier élève. Le premier nombre est le numéro de code de son lycée qui est 11. Les 4 nombres suivants sont ses notes de mathématiques aux trois trimestres, puis au bac. Cet élève a donc eu, en mathématiques, 5.5 au premier trimestre, 11.25 et 13.50 aux 2 trimestres suivants et 14 au bac. Sa note suivante (12.50) est celle de physique au premier trimestre. Nous n'avons pas indiqué toutes les notes : des tirets remplacent les 12 notes suivantes. Les 3 dernières sont les notes en philosophie des deuxième et troisième trimestres et du bac. De même, le deuxième élève appartient au lycée codé 18, il a 11.5 en mathématiques au premier trimestre, etc.

11	5.50	11.25	13.50	14.00	12.50	----	9.50	10.50	9.00
18	11.50	12.25	14.75	15.00	16.25	----	8.50	7.00	8.00

Tableau 1.1. *Les deux premières lignes du tableau des données.*

Objectifs de l'étude

2.1 De l'intérêt de préciser des objectifs

Avant d'analyser des données, il est indispensable de préciser un certain nombre d'objectifs de l'étude.

Au cours de l'étude elle-même, ces objectifs se modifient souvent : certains résultats obtenus conduisent à examiner d'un œil nouveau l'ensemble du problème, remettent en question des a priori, invitent à fouiller certains aspects susceptibles de révéler d'intéressantes tendances ou de curieuses anomalies, etc.

Mais ceci n'exclut pas une réflexion préalable, bien au contraire. En effet, toutes les décisions que l'on doit prendre dans la conduite de l'étude n'ont de sens que si l'on a pu préciser au départ un certain nombre d'objectifs. D'autre part, des modifications et/ou des enrichissements de la problématique ne peuvent surgir que si un cadre a déjà été défini. Enfin, beaucoup d'analyses sont possibles sur des données comme celles que nous étudions et il n'est pas concevable de les réaliser toutes : préciser des objectifs guide au moins les premiers choix.

Il est commode de préciser ces objectifs sous forme de questions auxquelles on tente de répondre.

2.2 Quelques questions préalables (non indépendantes)

Dans le cas de notre étude, ces questions ont été formulées à l'issue d'une discussion entre le statisticien et le demandeur. Elles se placent à différents niveaux et parfois se recoupent entre elles. Au départ, il n'est pas forcément nécessaire de structurer les objectifs : il faut seulement réfléchir au problème, en vue de disposer de quelques guides susceptibles d'orienter les traitements statistiques.

- Existe-t-il des matières pour lesquelles on observe fréquemment de très mauvaises notes ? Plus généralement, comment se répartissent les notes dans chacune des matières ? Cette répartition varie-t-elle d'une matière à l'autre ? Varie-t-elle entre les notes trimestrielles et celles du bac ?

- Y a-t-il une relation entre les notes trimestrielles et celles du bac ? Peut-on « prédire » avec précision la note du bac connaissant celles des 3 trimestres ? Les élèves sont-ils, pendant l'année, sous-notés ou surnotés par rapport au bac ? Les réponses aux questions précédentes varient-elles selon la matière ? Dépendent-elles du lycée ? Dépendent-elles des élèves en plus du lycée et de la matière ? Certains élèves obtiennent-ils des notes meilleures

(ou plus mauvaises) au bac que pendant l'année dans toutes les matières ? Si oui qui sont-ils ?

• Observe-t-on une évolution des notes entre les trois trimestres ? (hypothèse du demandeur de l'étude : une augmentation)

• Existe-t-il des différences entre les lycées pour certaines notes, certaines matières ? Certains lycées ont-ils des résultats meilleurs que d'autres ? Certains lycées sous-notent-ils (resp. surnotent-ils) les élèves ? (ceci est important pour les dossiers d'admission des élèves, par exemple dans les classes préparatoires aux grandes écoles)

• Peut-on observer des relations entre les notes des différentes matières ? Par exemple, les élèves très bons en mathématiques sont-ils souvent bons aussi en philosophie ou au contraire très médiocres ?

Plus généralement, on cherche à décrire l'ensemble des informations contenues dans le fichier, c'est-à-dire dans les $993 \times 21 = 20853$ données. L'examen complet de ces données est impensable et de plus, de même que l'arbre cache la forêt, l'examen de chaque nombre ne permet pas d'appréhender d'éventuelles grandes tendances dans le fichier. La description passe donc par des synthèses à l'aide d'indices ou de graphiques qui donnent une vision schématique et simplifiée des informations.

Premières vérifications des données

3.1 Principe presque absolu : les fichiers ne sont jamais propres

L'expérience montre qu'il faut toujours mettre en doute la validité des données recueillies. Généralement, un fichier de données est obtenu en plusieurs étapes : chacune engendre un risque d'erreur.

On connaît un certain nombre d'erreurs types qui apparaissent fréquemment au moment de la saisie ou du remplissage des bordereaux. Il vaut mieux les corriger avant de faire un traitement quelconque, pour ne pas donner des conclusions hasardeuses ou être obligé de tout reprendre à zéro après un long travail.

Des erreurs au moment de la préparation du recueil des données sont aussi très fréquentes. Elles ne sont pas du tout du même ordre. Par exemple, des questions ont été mal formulées et personne n'y a répondu, ou bien deux populations complètement hétérogènes ont été mélangées et il est indispensable de les séparer pour que l'étude ait un sens, etc.

Il est donc nécessaire de faire tout un travail de vérification, de correction, de nettoyage, de scission de fichiers, de recodage, etc. Ce travail est long ; il intervient bien entendu dès le départ où l'on doit absolument détecter et corriger les erreurs les plus visibles et les plus grossières. Il ne peut y avoir aucune dérogation à ce principe. Mais ce travail se poursuit tout le long de l'étude. En effet, des erreurs plus subtiles se détectent au fur et à mesure que l'on utilise des instruments plus sophistiqués (l'un des buts de l'étude des "outliers" du chapitre 12 est la détection d'erreurs). Dans un autre ordre d'idée, on peut être conduit, au vu de résultats, à considérer séparément telle ou telle sous-population pour l'étudier de très près. Maints exemples démontrent qu'un retour à un fichier de données modifié, suggéré par les premiers résultats de l'étude, est utile sinon indispensable.

Les paragraphes suivants proposent un itinéraire de vérification systématique des données. Il faut commencer par ce qui est le plus simple. Aussi, dans un premier temps, cherche-t-on uniquement les erreurs faciles à repérer, donc assez grossières. Pour faciliter leur dépistage, nous en donnons de nombreux exemples parmi les plus courants.

3.2 Mauvaise lecture du fichier

Lorsque l'enregistrement des données dans le fichier informatique ne correspond pas à la description que l'on en a faite, on peut aboutir à un grand nombre de valeurs fausses et même quelquefois à un fichier entièrement faux.

Ce type d'erreur est dramatique. On n'insiste jamais assez là-dessus. Il n'est pas rare de lire des conclusions complètement fausses obtenues à partir d'un fichier mal lu. Pour comprendre le problème, regardons le cas de notre fichier de notes.

Le fichier qui nous a été fourni a été saisi avec un éditeur de texte sous la forme illustrée **Fig. 3.1**.

- Chaque ligne correspond à un élève.
- Le numéro du lycée est codé sur les 2 premiers caractères de la ligne et cadré à droite.
- Ensuite, chaque groupe de 4 chiffres correspond à une note (le point décimal est omis). Les 4 premiers caractères correspondent à la note de mathématiques au premier trimestre ; les 4 suivants à la note de mathématiques au deuxième trimestre, puis au troisième et enfin au bac. Les notes des autres matières apparaissent de la même façon.

Ainsi, le premier élève appartient au lycée 11, a obtenu 5.5 au premier trimestre en mathématiques, 11.25 au second, etc. C'est ce qu'on appelle un format fixe : on peut tracer des lignes verticales qui séparent chacune des 21 variables.

```
11055011251350140012500-------095010500900
18115012251475150016250-------085007000800
```

Figure 3.1. *Début du fichier de notes tel qu'il se présente en format fixe. Seuls le début et la fin de ligne sont indiqués ; les tirets remplacent les autres valeurs. Les lignes verticales matérialisent la grille de lecture (elles n'apparaissent pas dans le fichier).*

Quand on lit un tel fichier en format fixe, il faut pouvoir repérer pour chaque individu les valeurs de chaque variable. Il faut donc indiquer le nombre total de variables et préciser que la première variable (lycée) est codée sur 2 caractères et que les 20 autres sont chacune codées sur 4 caractères et exprimées en centièmes de points. Ces informations sont appelées *format des données* ou *grille* (ou *masque*) *de lecture*.

Pour illustrer la gravité d'une erreur de description, supposons que l'on ait commis l'erreur d'indiquer que le lycée est codé sur les 3 premiers caractères (au lieu de 2) et que l'on ait pas commis d'erreur pour la suite en indiquant que les notes sont enregistrées sur les caractères suivants par paquets de 4 caractères. Ceci revient à décaler les lignes verticales séparant les valeurs des variables (cf. **Fig. 3.2**).

```
11055011251350140012500-------095010500900
18115012251475150016250-------085007000800
```

Figure 3.2. *Même fichier qu'en 3.1. avec une mauvaise grille de lecture.*

On lira alors que l'individu numéro 1 appartient au lycée 110 et qu'il a obtenu 55.01 en mathématiques au bac, 12.51 en mathématiques au premier trimestre, etc. Un simple décalage aboutit donc à considérer un tableau de données entièrement faux !

Toutes les variations sur ce thème du décalage sont possibles et malheureusement souvent vraisemblables si la présentation du fichier est un peu compliquée. On peut être confronté à une erreur systématique sur toutes les variables (notre exemple) ou seulement sur une ou plusieurs d'entre elles. Attention : un décalage ne conduit pas forcément à des valeurs hors plage facilement détectables.

Fichier avec séparateurs

Une autre façon de présenter un fichier de données consiste à mettre entre chaque variable un caractère spécial qui les sépare (dans ce cas, il n'est pas nécessaire que les variables

soient cadrées). Ce caractère, appelé séparateur, est souvent un espace (ou blanc), souvent aussi un nombre indéterminé d'espaces ; ce peut être aussi une virgule, un point virgule ou tout autre caractère qui ne peut se confondre avec les valeurs des variables.

En utilisant des séparateurs, on évite les erreurs de décalage du type de celle rencontrées avec les fichiers en format fixe. Mais il reste bien d'autres causes d'erreurs systématiques. Citons, par expérience, une erreur sur le nombre ou l'ordre des variables (par exemple, on oublie la présence du numéro de lycée ou bien on l'indique comme apparaissant en dernier et non en premier ; avec des séparateurs, le premier élève se retrouve alors avec, en mathématiques, 11 au premier trimestre, 5.5 au second, etc.).

Fichiers provenant de tableurs

Les fichiers qui proviennent de tableurs ou d'un logiciel de saisie n'échappent pas à ce dernier type d'erreurs, que ce soit au niveau de la saisie ou du transfert des données d'un logiciel à l'autre.

Conclusion : Il est indispensable de vérifier que le fichier a été lu correctement. Pour cela, il est impératif de visualiser les valeurs des variables, telles qu'elles ont été lues, pour quelques individus (en pratique, on vérifie les deux premiers et le dernier). Un rapide coup d'œil suffit pour constater des anomalies ; il ne reste alors qu'à corriger la description du format de lecture ou de la liste de variables.

3.3 Quelques erreurs classiques

Dans la saisie des données, on retrouve souvent le même type d'erreurs. Nous donnons ici quelques exemples d'erreurs classiques : les avoir examinés une fois aide par la suite à les repérer (cf. **Tab. 3.3**).

On peut avoir un chiffre ou un nombre erroné ; cela affecte un seul nombre.

On trouve aussi assez fréquemment une inversion entre deux chiffres ou deux nombres ; dans le second cas, deux nombres sont alors erronés.

Plus subtile, mais assez courante, est une omission (ou une répétition) aussitôt compensée, ce qui conduit au nombre exact de valeurs en fin de ligne. Dans ce cas, tous les nombres à partir de l'omission ou de la répétition sont faux (sauf éventuellement le dernier pour une omission). Une omission (ou une répétition) non compensée aboutit à un décalage dans la lecture de toutes les valeurs suivantes qui deviennent fausses.

Erreur	Exemple	Conséquence
Ligne exacte	1 5.50 11.25 13.50	-
Erreur sur un chiffre	1 5.80 11.25 13.50	*Un nombre faux*
Inversion de 2 valeurs	1 11.25 5.50 13.50	*Deux nombres faux*
Omission compensée	1 11.25 13.50 13.50	*Deux nombres faux*
Omission	1 11.25 13.50	*Décalage*
Répétition d'une valeur	1 5.50 5.50 11.25 13.50	*Décalage*

Tableau 3.3. *Quelques erreurs classiques.*

Les omissions ou répétitions peuvent éventuellement être repérées sur un fichier en format fixe de dimension raisonnable en le faisant défiler à l'écran : même à allure rapide, le non-

alignement de certains nombres est immédiatement détecté. Mais les autres erreurs sont difficiles à repérer directement et nécessitent une recherche systématique.

3.4 Détection de valeurs aberrantes

Valeurs hors plage

Dans le cas de notre étude, on sait que toutes les notes sont comprises entre 0 et 20. Les valeurs hors de ces limites sont de façon certaine des erreurs. Si l'on dispose d'un logiciel qui permet d'introduire les bornes de variation des variables (0 et 20 pour toutes les notes ; 0 et 22 pour le lycée) et de mettre en évidence toutes les valeurs hors plage, le travail est grandement facilité.

Maxima et minima

En l'absence d'un tel outil, le plus simple, pour vérifier que l'on reste dans les intervalles de variation des variables, est de calculer les valeurs minimum et maximum de chaque variable et de les contrôler une par une.

Appliqué au fichier des notes, ce principe a permis de détecter une seule valeur hors plage : un 22 en sciences naturelles au bac.

Notons que l'on ne connaît pas toujours exactement les intervalles de variation, mais que l'on a très souvent une idée, même approximative, de l'ordre de grandeur des valeurs. On sait par exemple qu'un être humain n'a vraisemblablement pas une taille de 70 mètres. Là encore, un simple regard sur les maxima et les minima permet de détecter ces anomalies.

3.5 Bilan des valeurs aberrantes

Que faire si on repère des valeurs aberrantes ? Tout dépend de l'étendue des dégâts ! Si la moitié du fichier de données est visiblement incorrecte, la démarche à suivre ne sera certainement pas la même que si une seule erreur a été détectée. Nous avons vu que la première situation n'est pas du tout irréaliste avec une erreur générale de lecture.

Il est donc utile de connaître l'importance du phénomène. S'il s'agit de valeurs "hors plage", il suffit de les compter. Si les bornes exactes ne sont pas connues mais que des valeurs apparaissent visiblement "hors norme", on peut avoir une idée approximative du nombre de ces valeurs en travaillant avec des bornes « naturelles » (un maximum de deux mètres pour un être humain par exemple).

Notre fichier, heureusement, ne comporte qu'une seule valeur "hors plage" : un 22 en sciences naturelles au bac. Ceci permet d'illustrer dans le paragraphe suivant la démarche suivie dans le cas où le nombre de valeurs aberrantes ou hors norme est très faible.

3.6 Cerner le problème de chaque valeur aberrante

Il faut d'abord bien cerner le problème : une valeur "hors plage" est certainement fausse, mais est peut-être l'indication de toute une série de valeurs fausses.

En effet, si une valeur est fausse, il est possible qu'un chiffre ait été saisi pour un autre. Dans ce cas, une seule valeur est fausse. Mais il est au moins aussi probable que l'erreur touche plusieurs valeurs. Le cas le plus fréquent est celui de l'omission d'un chiffre (ou d'un nombre) ou sa répétition erronée. Ces deux erreurs aboutissent à un décalage de toutes les valeurs suivantes qui sont donc mal lues et de ce fait fausses. Il n'est pas certain que ce décalage aboutisse immédiatement à une valeur hors norme. Et, comme ce décalage peut

affecter non seulement les valeurs de cet individu mais aussi celles des individus suivants dans le fichier, il est prudent de visualiser tout le contexte de ce 22 : la ligne entière concernant l'individu, mais aussi par exemple les 2 lignes précédentes et les 2 suivantes (cf. **Tab. 3.4**). Une autre erreur classique faite au moment de la saisie, l'inversion de deux nombres successifs, est aussi à envisager.

mathématiques				physique				sc. naturelles				histoire-géo.				philosophie			
bac	T1	T2	T3	bac	T1	T2	T3	bac	T1	T2	T3	bac	T1	T2	T3	bac	T1	T2	T3
17	16.5	15.0	10.5	14	12.0	14.0	11.5	11	12.0	14.5	14.0	9	8.5	10.0	10.5	6	10.0	11.0	10.0
8	15.1	13.3	15.8	11	12.5	10.8	12.8	13	13.0	15.5	12.0	9	12.0	6.9	13.0	4	8.5	9.7	9.5
7	14.0	11.6	12.7	13	13.0	14.5	11.5	**22**	11.3	12.8	12.0	8	11.6	11.8	15.0	4	7.0	7.4	abs
18	15.2	13.5	15.6	13	12.5	13.5	15.0	14	14.0	15.5	11.0	15	11.8	12.3	15.0	11	9.0	9.4	9.0
15	9.5	8.5	8.5	9	8.7	11.2	11.6	9	10.3	11.8	12.0	12	13.0	12.0	12.6	8	9.0	8.0	9.5

Tableau 3.4. *Contexte de la valeur aberrante 22.*

Le fichier est très structuré : avant même de l'étudier, on dispose de tant de connaissances sur lui que l'on peut faire beaucoup de vérifications. Les valeurs encadrant le 22 ne paraissent pas à première vue étonnantes. Non seulement elles sont comprises entre 0 et 20, mais les notes d'une même matière, qui doivent être vraisemblablement liées, ne présentent pas d'incohérences. Aucun phénomène général, par exemple un décalage, n'est visible.

On ne met donc pas en doute les autres valeurs, mais il reste à agir pour celle-ci.

3.7 Corriger, estimer une valeur, supprimer un individu

Trois différentes actions possibles pour cette valeur erronée (corriger, estimer, ou supprimer) sont discutées ci-dessous. On peut aussi la considérer comme une donnée manquante, puisque la véritable valeur est inconnue, et faire des calculs en tenant compte de cette donnée manquante. Les problèmes liés à cette solution assez peu recommandée sont discutés en 4.6.

Corriger

Si ayant repéré une erreur sur un élève déterminé on peut revenir aux données initiales non erronées, il suffit de corriger l'erreur. Ce serait le cas si l'erreur provenait de la saisie et que l'on disposait des bordereaux. N'en disposant pas, il ne reste que les deux autres possibilités.

Supprimer

On peut supprimer complètement de l'étude l'individu qui présente la valeur erronée. Vu les objectifs généraux de description des données, supprimer un individu sur un effectif qui atteint presque mille ne porte pas à conséquence et ne risque pas de modifier notablement les conclusions. D'autre part, on ne s'intéresse pas particulièrement à cet élève anonyme. C'est donc une solution raisonnable.

Estimer

Cependant, si on le supprime complètement, on perd forcément une petite partie de l'information disponible : toutes les autres valeurs connues pour cet individu. D'autre part, les données sont ici très structurées : les valeurs des 4 notes d'une même matière sont la plupart du temps relativement proches. Il est donc facile de proposer une valeur qui est soit

(par chance) exactement la bonne valeur, soit une valeur assez proche. C'est peut-être une meilleure solution que de supprimer totalement l'individu.

Il reste à proposer une valeur raisonnable. Regardons les notes de l'individu qui bénéficie d'un 22 en sciences naturelles au bac. Ses notes en sciences naturelles indiquées aux trois trimestres sont respectivement 11.3, 12.8 et 12.0. Toutes ces notes sont proches de 12. En outre, pour chacun des 4 autres élèves, les 4 notes de sciences naturelles sont assez proches entre elles, ce qui n'est pas pour surprendre. Il est donc vraisemblable que la note erronée se situe aux environs de 12. Un 22 au lieu d'un 12 advient facilement : à la saisie, on répète le 2 au lieu de mettre un 1. Il est donc très probable que le 22 soit un 12 déguisé par une erreur de saisie sur le premier chiffre. Pour conforter cette hypothèse, on peut vérifier que cet élève n'a pas d'écart systématique entre ses notes trimestrielles et ses notes au bac. En mathématiques, sa note au bac est inférieure à ses notes trimestrielles, mais en physique les notes restent très stables. On décide donc de remplacer le 22 par un 12. Le 12 est une "*valeur estimée*" à l'aide de l'ensemble des informations que l'on peut avoir, la valeur exacte étant inconnue.

Insistons sur le fait que la valeur 12 n'est pas obligatoirement la valeur exacte, mais que cela n'a pas une importance fondamentale : elle ne va pas décider de l'attribution du bac pour un élève, mais sera fondue dans les statistiques générales. Autrement dit, l'estimation ici n'a pas d'intérêt en elle-même : elle ne sert qu'à utiliser plus facilement l'information contenue dans les autres données.

La méthode d'estimation que nous avons utilisée est très simple et basée sur le bon sens et la connaissance de la structure des données ; il en existe de beaucoup plus sophistiquées, mais il est ici inutile d'employer des techniques lourdes qui compliquent énormément le travail sans apporter de résultats plus fiables. Quand le choix est possible, on gagne toujours à favoriser la simplicité. **Bien comprendre ce que l'on fait est nécessaire pour pouvoir le critiquer** : l'utilisation d'une formule magique, même très jolie, est dangereuse car on ne garde plus la maîtrise de son action.

Remarques. Quand on remplace ainsi une valeur inconnue par une valeur estimée, il est raisonnable de le garder en mémoire. En effet, il faut pouvoir la remettre en question si l'individu concerné se révèle par la suite un peu bizarre et que cela puisse infléchir les conclusions. Cette remarque vaut surtout lorsque le nombre de valeurs estimées est relativement important.

Pour la petite histoire, nous avons pu disposer plus tard des bordereaux de saisie : le 22 provenait bien d'une erreur dans la saisie d'un 12.

3.8 Que faire s'il y a beaucoup de valeurs aberrantes ?

Si les valeurs hors plage (ou hors norme) sont nombreuses, il est utile de savoir comment elles se répartissent et en particulier si certaines variables ou certains individus sont particulièrement touchés. Cette démarche est détaillée dans le chapitre 4 concernant l'étude de données manquantes qui, à ce niveau, est tout à fait analogue dans les deux cas. Nous donnons seulement ici quelques conseils.

Erreur générale ?

Si le fichier contient un grand nombre de valeurs hors plage ou aberrantes, il reste peut-être encore une erreur générale de lecture (mauvaise description du fichier, décalage) non corrigée. Il faut le vérifier soigneusement.

Bilan sur les variables

Si ces erreurs concernent plutôt certaines variables, il faut concentrer son attention sur ces variables : vérifier leur format de lecture, penser à un éventuel décalage, etc.

Si l'explication n'est pas là, il faut peut-être remettre en question les bornes de variation de ces variables. Si par exemple dans le fichier de notes on avait indiqué à un logiciel que les notes devaient être strictement supérieures à 0 et strictement inférieures à 20, tous les 0 et les 20 apparaîtraient comme valeurs aberrantes.

La population étudiée n'est-elle pas différente de celle à laquelle on s'attendait ? (les individus de plus de 2 mètres ne sont pas si rares parmi les basketteurs...)

Il peut s'agir de situations plus complexes et beaucoup plus délicates à corriger. Par exemple, une erreur au moment même du recueil des données : une mesure très souvent entachée d'erreurs ou une variable mal définie. Une telle constatation peut conduire à supprimer de l'étude une variable dont trop de valeurs sont douteuses ou erronées. Ce problème ne se pose pas pour notre fichier de notes.

Bilan sur les individus

Lorsque des individus ont beaucoup de valeurs erronées, il faut, comme dans le cas du paragraphe précédent, étudier leurs contextes pour voir s'il ne s'agit pas d'un décalage. Lorsqu'une sous-population importante d'individus a beaucoup de valeurs hors plage, c'est une partie du fichier qui est à remettre en question. Il est raisonnable de visualiser la partie du fichier concernant les premiers individus touchés. S'il ne s'agit pas d'un simple décalage qui se propage, on peut penser à une incohérence entre des sous-ensembles de données : par exemple, les individus présentant des erreurs ne proviennent-ils pas d'un même lycée où l'on aurait rempli les bordereaux différemment ?

Si cela est possible, la correction s'impose, sinon il ne reste guère que la solution de supprimer de l'étude les individus erronés. Mais attention : toute suppression modifie le champ de l'étude et les conclusions ne concernent que les individus réellement étudiés ; ce problème apparaît aussi à propos des données manquantes (cf. Ch. 4).

Erreurs résiduelles nombreuses

Si, après des corrections simples ou des suppressions ponctuelles, le fichier contient encore beaucoup d'erreurs, le recueil des données est à mettre sérieusement en doute. D'autant plus que certaines erreurs ne sont pas détectables ainsi et que, si les valeurs aberrantes sont nombreuses, les erreurs risquent de l'être beaucoup plus. Aucune conclusion fiable ne peut être obtenue à partir de données douteuses. Les précautions les plus grandes doivent être prises au moment de l'élaboration et du recueil des données : on ne le redit jamais assez !

3.9 Grosses erreurs et petites erreurs

Lien entre l'importance de l'erreur et celle de sa correction

Il est assez facile, dans ce fichier, de repérer des erreurs qui aboutissent à des valeurs aberrantes comme un 22, mais il est difficile, voire impossible, de repérer par exemple un 12 au lieu d'un 11. Ceci n'est pas trop dramatique car les perturbations apportées par les premières sont beaucoup plus importantes que celles apportées par les secondes.

En effet, quand on décrit statistiquement un fichier, on donne des éléments du comportement de l'ensemble de la population ou d'une sous-population. Pour les notes, on calcule notamment le maximum, le minimum et la moyenne. On voit bien qu'un 11 au lieu

d'un 12 ne modifie rien aux valeurs des deux extrema tandis qu'un 22 amènerait à conclure que la meilleure note, par exemple en sciences naturelles, dépasse toutes les espérances, mais cela ne serait guère sérieux...

La moyenne, elle, est toujours perturbée par une valeur fausse, mais une valeur « très » aberrante comme un 912 l'augmenterait de presque un point, alors qu'un 11 (au lieu d'un 12) ne la diminue que d'un peu plus d'un millième de point, ce qui n'apparaîtrait sans doute même pas dans un résultat arrondi.

Les grosses erreurs cachent les petites

Il existe des aberrations plus subtiles que celles que nous avons citées. Par exemple, un individu qui obtient 20 en mathématiques au bac après avoir eu toute l'année 0 peut paraître assez étonnant pour que l'on ait un doute et qu'il paraisse nécessaire de revenir aux données pour les vérifier.

Cependant, il est indispensable de commencer par corriger les erreurs grossières comme les valeurs hors plage, car leur présence gêne la détection d'erreurs plus fines.

3.10 Quelques autres types d'erreurs

On peut aussi rechercher systématiquement quelques types d'erreurs assez fréquents.

Dans notre fichier de notes, on a remarqué (après bien des études !) que deux lignes successives sont absolument identiques. Il est hautement improbable que deux élèves aient les mêmes valeurs pour les 21 variables et, par ailleurs, on conçoit fort bien une erreur de répétition au moment de la saisie : aussi avons-nous conclu à la répétition d'un individu lors de la saisie et supprimé l'une des deux lignes. Le nombre d'individus de notre fichier n'est donc plus que de 992.

Au lieu de répéter une ligne, on peut en omettre une. Cette erreur est plus difficile à détecter sauf si l'on connaît le nombre exact d'individus. Si celui-ci est connu, il est indispensable de vérifier qu'il coïncide exactement avec le nombre d'individus du fichier. Cette vérification permet aussi parfois de découvrir une ligne vide en début ou en fin de fichier, ce qui peut échapper à un examen du fichier à l'écran ou sur papier.

3.11 Bilan-résumé sur la recherche d'erreurs

Détecter
Examiner le minimum et le maximum de chaque variable. Vérifier le nombre de lignes. Si le format est fixe, vérifier les alignements des nombres (notamment celui du dernier).
Analyser
Si peu d'erreurs : examiner le contexte de chacune. Si beaucoup d'erreurs : dénombrer les erreurs par individu et par variable.
Décider
Corriger, s'il est possible de retrouver la donnée originale. Estimer par une valeur plausible compte tenu du contexte. Abandonner : des individus, des variables, l'étude.

Données manquantes

Attention danger !

Certains logiciels « tiennent compte » des valeurs manquantes, sans toujours préciser comment. La lecture de ce chapitre convaincra le lecteur que chaque cas mérite réflexion avant de choisir la solution la moins mauvaise et qu'un traitement standard, et par surcroît non explicité, risque de fausser complètement les résultats.

4.1 Remarques préliminaires sur le codage des données manquantes

Notre fichier de notes contient beaucoup de données manquantes. Notre interlocuteur ne nous avait pas prévenus, sans doute l'ignorait-il. L'opérateur qui a saisi les données, ne voyant rien pour certaines notes sur les bordereaux, a mis des espaces dans les positions correspondantes. Ce n'est pas forcément la meilleure solution : en format fixe et avec certains logiciels, les blancs se confondent avec des zéros (c'était notre cas).

En nous lançant dans l'étude sans nous préoccuper de ce problème, nous aurions eu alors l'attention attirée par le nombre important de zéros. Ceci à condition d'avoir l'esprit critique toujours en éveil, attitude indispensable dans l'étude d'un fichier de données. Si notre esprit critique s'était endormi, les conclusions auraient été bien faussées : en prenant, par exemple, les 58 données manquantes en histoire-géographie au troisième trimestre pour des zéros, on aurait soupçonné à tort une grande sévérité des professeurs ou la nullité d'un pourcentage important d'élèves.

Dans un fichier où les séparateurs de nombres sont des espaces, les conséquences d'un tel codage sont bien pires puisque toutes les valeurs sont décalées. On repère vite un problème lorsque l'on connaît le nombre exact d'élèves : il en manque quelques-uns. La fin du fichier peut aussi être anormale avec un nombre de notes insuffisant pour le dernier élève. Mais si, par malheur, on ne connaît pas le nombre exact d'élèves et que les décalages successifs se compensent, aucune anomalie à la lecture n'attire l'attention. La recherche des valeurs hors plage peut ne rien détecter non plus puisque toutes les notes ont le même intervalle de variation ; seuls les codes des lycées 21 et 22, lus comme des notes à la suite d'un décalage, peuvent conduire à une valeur aberrante.

Dans notre cas, l'erreur non détectée lors de la lecture ou des premières vérifications a encore quelque chance d'être soupçonnée dans les résultats grâce à une connaissance préalable très importante sur ce fichier. On sait, par exemple, que les notes d'une même matière sont liées, même si l'on ne sait pas précisément mesurer ce lien. Un décalage a toute chance de briser cette structure et des résultats trop inattendus doivent tout de suite faire soupçonner le pire.

Ces remarques pessimistes sont faites pour convaincre de la nécessité de penser systématiquement au problème des données manquantes. Si on en a la maîtrise, il faut y penser dès la saisie des données en choisissant un codage particulier. Certains logiciels imposent un codage : un signe non numérique ou une valeur numérique très grande (ou très petite) qui ne risque pas d'apparaître dans les données. Les plus puissants permettent de différencier plusieurs types de données manquantes : par exemple, dans une enquête de suivi médical, un décès n'est pas du tout équivalent à un oubli de relevé et ces deux types de données manquantes ne seront sans doute pas traitées de la même façon. Si on n'a pas la maîtrise de la saisie, il est indispensable de connaître leur codage. Comme déjà indiqué, si on ne le fait pas, des anomalies dans les résultats doivent faire soupçonner des erreurs dont les données manquantes sont une des causes possibles. Mais il vaut mieux s'intéresser à ce problème épineux dès le départ plutôt que de risquer de donner des conclusions fausses ou d'être amené à recommencer tout un travail.

4.2 Bilan des données manquantes

Le fichier comporte des données manquantes. Y en a-t-il beaucoup ? La démarche à suivre n'est pas la même selon qu'il y en a très peu ou beaucoup. La première étape de l'étude des données manquantes est analogue à celle des données aberrantes : un bilan global.

Effectif total des données manquantes

Dans notre fichier, 811 valeurs manquantes sont répertoriées : cela représente 3.89 % des valeurs. Ce n'est pas catastrophique (un grand pourcentage de données manquantes peut faire douter de la fiabilité des résultats statistiques et peut même conduire à renoncer à une étude) mais c'est un phénomène très important.

Cette importance insoupçonnée mérite des explications. Que signifient toutes ces valeurs manquantes ? Combien d'élèves sont touchés ? Concernent-elles certaines notes particulièrement ? Peut-on les expliquer ? Autant de questions qui n'étaient pas apparues dans les premiers objectifs, mais que ce résultat inattendu amène à se poser et qui montrent qu'un bilan des données manquantes fait partie tout à fait normalement de la description des données.

D'autre part, les données manquantes posent toujours problème. Que va-t-on faire avec elles ? Comme pour les données erronées (qui sont en quelque sorte manquantes), on peut "estimer" la donnée manquante, supprimer l'individu concerné, supprimer la variable concernée ou bien encore faire chaque calcul avec toutes les valeurs disponibles. Le choix n'est pas toujours facile. Pour prendre des décisions raisonnables, il est nécessaire de savoir si elles concernent une population particulière, une variable particulière, etc.

Dans tous les cas il est utile d'étudier la répartition des données manquantes dans le fichier, d'une part à travers les variables et d'autre part à travers les individus.

Répartition des données manquantes par variable

Les effectifs des valeurs manquantes par note montrent que toutes sauf la variable *lycée* sont concernées (cf. **Tab. 4.1**). Les notes les moins touchées sont celles de mathématiques et de physique au bac avec 17 valeurs manquantes ; ensuite ce sont les trois autres notes du bac avec 32 valeurs manquantes. Pour les notes trimestrielles, le nombre de valeurs manquantes varie entre 36 (mathématiques $2^{ème}$ trimestre) et 58 (histoire-géographie $3^{ème}$ trimestre). Les matières les plus touchées sont l'histoire-géographie et la philosophie.

	maths	physique	sciences nat.	histoire-géo.	philosophie
trimestre 1	39	39	46	51	49
trimestre 2	36	38	48	53	48
trimestre 3	39	40	46	58	51
bac	17	17	32	32	32

Tableau 4.1. *Effectif des valeurs manquantes par note pour l'ensemble des 992 élèves. Exemple : 38 élèves n'ont pas de notes en physique au deuxième trimestre.*

Pour aucune note, l'effectif atteint ne justifie une méfiance particulière. Ce serait d'ailleurs étonnant pour un fichier comme celui-ci dans lequel aucune variable ne risque de poser un problème spécifique de relevé. Il n'est donc pas question ici de supprimer des variables.

Répartition du nombre de données manquantes par individu

83 élèves sur les 992 ont des valeurs manquantes, soit 8.4%. En les classant par ordre décroissant de nombre de valeurs manquantes, on obtient le résultat suivant : 17 individus ont 20 valeurs manquantes (toutes les notes), 6 en ont 18, 13 en ont 15, un seul en a 14, 6 en ont 12, un seul en a 9 et un seul 7, 2 en ont 6, un en a 5 et un aussi 4 ; tandis que 4 en ont 3 et 3 en ont 2 ; enfin 27 ont une seule valeur manquante.

Diagramme en bâtons

Ces chiffres donnent l'information complète sur la répartition du nombre de données absentes par individu mais il est difficile d'en avoir une idée synthétique et la lecture en est rébarbative. Le *diagramme en bâtons* (ou *diagramme en barres)* représente graphiquement cette répartition (cf. **Fig. 4.1**) : au-dessus de chacun des nombres possibles de valeurs manquantes (sauf 0), figure un « bâton » dont la longueur est proportionnelle au nombre d'individus concernés. La somme des longueurs de tous les bâtons représente le nombre total d'individus ayant des valeurs manquantes. On peut considérer que les bâtons représentent les pourcentages d'individus pour chaque valeur puisque ces pourcentages sont proportionnels aux effectifs ; la somme des longueurs des bâtons est alors égale à 1.

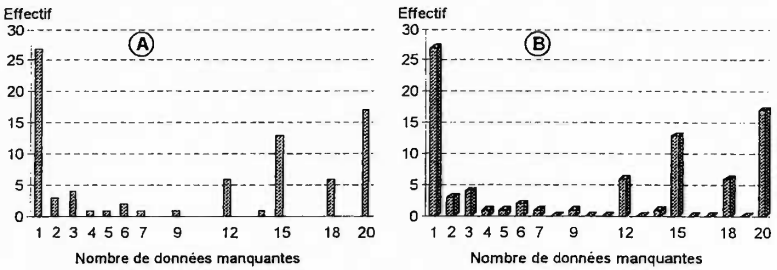


Figure 4.1. *Variable* nombre de données manquantes par élève *: diagramme en bâtons. a) diagramme plat ; b) diagramme en relief, appelé aussi diagramme 3-D, ce qui prête à confusion car une seule dimension (ici le nombre de données manquantes) est étudiée. La mise en relief apporte une touche « sophistiquée » mais n'aide aucunement la lecture.*

On dispose ainsi d'une représentation de la répartition du nombre de données manquantes par individu. Un simple coup d'œil au graphique permet de voir qu'elle est curieuse : des

effectifs relativement nombreux en 12, 15, 18 et 20, encadrent des "trous" pour 19, 17, 16, etc. Cette coïncidence du nombre de données manquantes entre plusieurs individus n'est sûrement pas fortuite ; elle incite à penser qu'il existe des groupes d'individus ayant leurs données manquantes réparties de la même façon dans les différentes matières. Si cela se vérifie, les individus d'un même groupe devront être traités de la même façon.

Histogramme

Ce type de graphique est souvent, abusivement, appelé *histogramme*. Nous indiquons en 7.2 la définition d'un véritable histogramme et sa relation avec le diagramme ci-dessus.

4.3 Etude des données manquantes à travers les individus

Il y a trop de valeurs manquantes pour les étudier une par une (comme nous avons procédé pour la valeur hors norme 22). Aucune variable ne présente d'anomalie particulière. Par contre, la répartition du nombre de données manquantes par individu suggère de considérer les groupes d'élèves qui en ont le même nombre. En particulier, il importe de savoir si pour les élèves d'un même groupe les données manquantes concernent les mêmes notes. Pour cela, nous examinons, pour chaque groupe d'élèves, la répartition des données manquantes dans les différentes matières (cf. **Tab. 4.2**).

nombre	effectif	Maths				Physique				Sc. nat.				Hist.-géo.				Philosophie			
de dm	élèves	b	1	2	3	b	1	2	3	b	1	2	3	b	1	2	3	b	1	2	3
20	17																				
18	6	*				*															
15	13	*				*				*				*				*			
14	1	*	*	*		*	*	*													
12	6	*	*	*	*	*	*	*	*												
9	1	*	*	*	*	*	*	*	*		*	*	*								
7	1	*		*	*	*	*		*	*		*	*	*			*	*			*
6	1	*	*	*	*	*	*	*	*		*	*	*						*	*	*
6	1	*	*	*		*	*	*	*	*				*	*		*	*	*		*
1 à 5	36	*	-	-	-	*	-	-	-	*	-	-	-	*	-	-	-	*	-	-	-

Tableau 4.2. *Répartition des données manquantes (dm).*
Une ligne représente le groupe d'élèves ayant un nombre donné de données manquantes. Au croisement d'un groupe et d'une note, on trouve une étoile (note toujours présente), un espace (note toujours absente) ou un tiret (situation variable d'un individu à l'autre).

- Tout d'abord, 17 élèves ont 20 valeurs manquantes, c'est-à-dire qu'ils n'ont aucune note. Ils forment un groupe parfaitement homogène.
- Ensuite, les 6 élèves qui ont 18 valeurs manquantes n'ont que 2 notes : les notes au bac de mathématiques et de physique. Vraisemblablement ces élèves étaient titulaires d'un autre bac et n'avaient à subir que ces deux épreuves.
- Les 13 élèves qui ont 15 valeurs manquantes n'ont de notes qu'aux 5 épreuves du bac. Comme les 2 premiers groupes, ils forment une population homogène.
- Ensuite, les choses sont moins simples. L'élève qui a 14 données manquantes n'a de notes au bac qu'en mathématiques et en physique, mais il a aussi des notes dans ces matières aux deux premiers trimestres. Les 6 suivants (12 valeurs manquantes) sont

semblables entre eux et analogues au précédent, sauf qu'ils ont des notes pour les trois trimestres. Le suivant (9 données absentes) a, en plus, des notes trimestrielles en sciences naturelles. La caractéristique commune de ces 8 individus (1+6+1) est qu'ils n'ont, au bac, des notes qu'en mathématiques et en physique.

- Le suivant (7 données absentes) a toutes ses notes au bac, mais il lui manque des notes aux deux premiers trimestres.
- Les 2 élèves qui ont 6 données absentes ne se ressemblent pas : l'un n'a de notes au bac qu'en mathématiques et physique alors que l'autre a des notes dans les 5 matières du bac.
- Pour les 36 suivants (de 1 à 5 données manquantes), les notes absentes sont des notes trimestrielles réparties dans les cinq matières et les trois trimestres.

4.4 Répartition des 83 élèves incomplets en 4 groupes homogènes

Parmi les 83 élèves ayant des données manquantes, on distingue quatre sous-populations très typées (cf. **Tab. 4.3**).

1) Ceux qui n'ont aucune note (17 élèves). Ces élèves ont été répertoriés, mais pour une raison inconnue ils n'ont de notes ni au bac ni pendant l'année.

2) Ceux qui, au bac, n'ont de notes qu'en mathématiques et en physique (15 élèves). Il est logique de regrouper ensemble, bien que leur nombre de données manquantes varie, les 6 individus qui n'ont que ces 2 notes et les 9 autres qui ont, en plus, des notes trimestrielles plus ou moins nombreuses. Ces élèves avaient, vraisemblablement, déjà un autre bac que le bac C et sont dispensés des autres matières. Ils forment une sous-population différente de la population générale. Il est clair que ce regroupement se fonde non seulement sur la répartition des données manquantes mais aussi sur la connaissance que l'on a du problème ; une automatisation de cette démarche serait délicate.

3) Ceux qui n'ont comme notes que celles du bac (13 élèves).

4) Ceux (38 élèves) qui ont quelques notes trimestrielles manquantes (entre une et sept), correspondant sans doute à des absences.

Ces quelques chiffres suggèrent donc des explications assez diverses. Ils permettent déjà de décrire en partie le phénomène des données manquantes. Bien entendu, cette description n'est pas complète et la détection de ces 4 groupes suggère d'autres questions. En examinant ces groupes un par un pour étudier les mesures à prendre pour chacun d'entre eux, nous les comprendrons mieux.

Effectif	Cumul	Caractéristique du groupe
909	909	aucune donnée manquante (dont 907 élèves inscrits dans un lycée)
38	947	toutes les notes du bac ; quelques notes trimestrielles manquantes
13	960	toutes les notes du bac ; pas de notes trimestrielles
15	975	au bac, notes seulement en maths et en physique
17	992	aucune note

Tableau 4.3. *Les différents groupes d'élèves selon leurs données manquantes.*

4.5 Groupe des fantômes

17 élèves ont 20 valeurs manquantes. Pour une raison inconnue, ces élèves sont répertoriés dans le fichier mais n'ont absolument aucune note. Ce sont des individus fantômes. Comme la statistique se limite à l'analyse de données bien réelles, il est de notre devoir de supprimer ces ectoplasmes et de considérer que notre fichier ne contient que 975 individus (992–17).

Ces individus fantômes ne sont pas si rares qu'on pourrait le penser : une erreur de manipulation assez fréquente aboutit à la création de ligne blanche, généralement en début de fichier, et transforme ainsi un utilisateur un peu maladroit en médium involontaire. Mais ici, ce ne sont pas des lignes blanches. On connaît le lycée de chacun.

- 8 d'entre eux sont des candidats libres ;
- 4 appartiennent au lycée 14, 2 au lycée 4 et 3 sont uniques dans leur lycée.

Les candidats libres forment une population différente de la population générale et leur grand nombre peut s'expliquer : une inscription, pas de scolarité et finalement le renoncement à l'examen.

Par contre, trouver 4 élèves d'un même lycée paraît étrange. Relever ainsi des anomalies est souvent un intérêt majeur des études statistiques. Celle-ci est tout à fait marginale dans cette étude et nous ne nous y appesantissons pas. Il faudrait d'ailleurs revenir au « terrain » (concrètement se renseigner auprès des administrations des lycées) pour chercher les explications, ce qui n'est pas dans nos moyens ici.

Nouveau bilan

Ces individus éliminés du fichier, il reste 471 valeurs manquantes. Le pourcentage de valeurs manquantes est encore important, il dépasse 2%. Regardons les nouvelles répartitions suivant les variables (cf. **Tab. 4.4**). Les notes de mathématiques et de physique au bac ont des valeurs pour tous les individus, les autres matières du bac ont 15 valeurs manquantes. Pour les notes trimestrielles, les chiffres sont variables. Le plus faible se situe en mathématiques au deuxième trimestre où 19 notes manquent ; le plus élevé est 41 en histoire-géographie au troisième trimestre.

	maths	physique	sciences nat.	histoire-géo.	philosophie
trimestre 1	22	22	29	34	32
trimestre 2	19	21	31	36	31
trimestre 3	22	23	29	41	34
bac	0	0	15	15	15

Tableau 4.4. *Effectif des valeurs manquantes par note pour les 975 élèves ayant au moins une note. Exemple : 21 élèves n'ont pas de notes en physique au deuxième trimestre.*

4.6 Groupe des 15 élèves avec bac incomplet

Le second groupe détecté est celui des élèves qui n'ont, au bac, que les notes de mathématiques et de physique et qui ont vraisemblablement déjà un bac D. Rappelons que 6 d'entre eux n'ont pas d'autres notes et que les 9 autres ont un nombre variable de notes trimestrielles. Discutons des solutions possibles pour une valeur manquante : estimer cette valeur, calculer avec toutes les valeurs connues, supprimer les individus.

Estimer

Il n'est pas du tout envisageable d'estimer les valeurs manquantes comme nous l'avons fait pour la donnée erronée. En effet :

- il y a trop de données manquantes par individu ;
- nous n'avons pas assez d'éléments pour affecter des valeurs "raisonnables". Quelle note "raisonnable" mettrait-on en philosophie à un élève qui n'a pas passé cette épreuve et sur lequel on ne sait rien d'autre que ses notes en mathématiques et en physique au bac ?

Il ne reste que deux solutions : les supprimer du fichier ou les garder en faisant les calculs et les descriptions sur des populations différentes suivant les variables (par exemple calculer la moyenne de mathématiques sur 975 individus et celle de philosophie sur 960). Discutons ces deux solutions par rapport aux objectifs de l'étude.

Calculer avec toutes les valeurs connues

Ceci présente peu d'intérêt. Les 6 premiers individus (qui n'ont que deux notes) sont visiblement totalement inutiles pour comparer les notes de l'année à celles du bac. L'ensemble des 15 individus est totalement inutile pour comparer les répartitions des notes des différentes matières au bac (sauf en mathématiques et physique).

Ils sont non seulement inutiles mais dangereux dès que l'on veut comparer des matières. En effet, supposons (ce qui est faux) ces 15 élèves excellents et ayant des résultats en mathématiques au bac bien meilleurs que l'ensemble de la population. Ils vont faire monter la moyenne de mathématiques ! Ce ne serait pas juste de comparer cette moyenne avec celle de sciences naturelles dans laquelle ces brillants élèves n'interviennent pas. Cette situation imaginaire n'est là que pour illustrer le fait que les comparaisons entre valeurs calculées sur deux populations qui ne se recouvrent pas entièrement risquent d'être scabreuses. En réalité ici, le risque n'existe guère pour les moyennes. qui ne peuvent être notablement modifiées par 15 individus sur près de 1000. Mais si l'on s'intéresse par exemple à des pourcentages, les variations peuvent être très importantes ; supposons que les 15 élèves soient les meilleurs et les seuls à avoir une note supérieure à 18 : avec eux le pourcentage de notes supérieures à 18 serait de 1.5% (=15/975) et sans eux de 0%. Serait-il juste de dire que l'épreuve de mathématiques est plus facile que celle de sciences naturelles car 1.5% des élèves au lieu de 0% atteignent les sommets, alors que les 15 brillants élèves en mathématiques auraient pu être aussi brillants en sciences naturelles ?

Pour illustrer encore cette difficulté, prenons un autre exemple, extrême et tout à fait irréaliste. Supposons que l'on étudie à la fois l'âge et le nombre d'années de travail d'un ensemble d'ouvriers. Par malheur, les moins de 30 ans ont tous des valeurs manquantes pour le nombre d'années de travail et, réciproquement, les plus âgés ont indiqué leur temps de travail mais pas leur âge. Si l'on calcule des moyennes sur les valeurs connues pour chacune des deux variables, on conclura que la population étudiée a, par exemple, une moyenne d'âge de 25 ans et un temps de travail moyen de 20 ans.

Il est donc dangereux de conserver des individus ayant des valeurs manquantes, même si on se limite à décrire la population variable par variable. En effet cela implique que, selon la variable, on n'étudie pas la même population : l'analyse est alors inextricable.

D'autre part, garder des valeurs manquantes limite les possibilités de synthèse car certaines techniques statistiques exigent des tableaux sans aucune donnée manquante (cas de l'Analyse en Composantes Principales, décrite au chapitre 9).

Supprimer les individus

Considérant que les 15 élèves apportent beaucoup plus de problèmes que d'intérêt pour les objectifs principaux de l'étude, qui sont surtout des comparaisons, nous décidons de les supprimer aussi.

Mais attention : si nous supprimons ces individus, la répartition des notes en mathématiques et en physique au bac ne va plus être la même. **Le champ de l'étude change !** Ce n'est plus l'ensemble des candidats que nous considérons, mais l'ensemble des candidats qui ont passé toutes les épreuves du bac. Nos conclusions générales ne concerneront donc pas ces 15 individus. Cette restriction n'est qu'apparente : ces élèves sont tout à fait à part et il paraît naturel de limiter l'essentiel de l'étude aux élèves "normaux".

Cependant, la mise en évidence de cette sous-population amène à se poser des questions sur ces élèves. Qui sont-ils ? Proviennent-ils de lycées particuliers ? Ont-ils des résultats en mathématiques et en physique bien meilleurs que les autres élèves ?

On ne peut se permettre, dans l'analyse complète d'un fichier de données, d'éliminer une sous-population sans l'étudier. Ici, le petit groupe éliminé est très marginal et les questions qui le concernent sont certainement moins fondamentales que celles posées dans les objectifs principaux. Dans une étude rapide des données et en se limitant aux objectifs précisés au départ, il n'est pas absolument nécessaire de s'y attarder beaucoup. Mais attention, ce n'est pas toujours le cas ; quelquefois l'étude de ces sous-populations "différentes" est particulièrement éclairante pour l'ensemble des données. De toute façon, si on en a les moyens, il faut tenter de répondre aux nouvelles questions qui se posent.

Ces questions se réfèrent à deux niveaux : dans le premier, le plus simple, on veut décrire ce groupe d'élèves ; dans le second, on veut les situer dans la population générale.

La description de ce groupe d'élèves est faite au chapitre suivant. Situer nos 15 héros dans la population totale est un problème général et complexe de comparaison entre populations. Nous lui consacrons le chapitre 10.

4.7 Groupe des 13 élèves qui n'ont de notes qu'au bac seulement

Nous avons étudié en détail les deux premiers groupes d'élèves ayant des données manquantes pour introduire des outils et une démarche. Pour le troisième, les 13 individus qui n'ont de notes qu'au bac, nous irons plus rapidement en ne donnant que les conclusions.

Description

Curieusement, il n'y a pas que des candidats libres parmi eux, bien loin de là : 10 sont inscrits dans 8 différents lycées. Pour quelles raisons n'ont-ils pas de notes de scolarité ? Nous l'ignorons.

Suppression

Un des objectifs principaux de l'étude étant la comparaison des notes annuelles avec celles du bac, comparaison pour laquelle ces élèves n'apportent que des perturbations, nous décidons de les supprimer en changeant de nouveau le champ de l'étude. Il ne reste que 947 individus : ceux qui ont toutes leurs notes au bac et la plupart des notes trimestrielles. Nous essaierons (cf. 10.7) cependant de situer les notes de ces 13 élèves dans l'ensemble, pour voir s'ils diffèrent des autres et si leur suppression modifie la répartition générale des notes du bac.

4.8 Groupe des élèves avec quelques valeurs manquantes

Il reste à prendre des décisions pour les 38 élèves auxquels il manque de 1 à 7 notes trimestrielles. L'explication la plus vraisemblable d'une de ces valeurs manquantes est l'absence momentanée de l'élève, à moins qu'il n'y ait eu quelque oubli dans un report de notes.

Supprimer ; estimer ; calculer avec des valeurs manquantes

Jusqu'ici la solution très simple de la suppression s'imposait. Pour ce dernier groupe d'individus, le problème est différent. D'une part, l'information qu'ils apportent (surtout pour ceux qui n'ont qu'une valeur manquante) est presque aussi importante que celle d'un élève normal ; d'autre part, la population de ces élèves plus ou moins souvent absents est sans doute un peu particulière mais beaucoup moins que celle des élèves qui ne passent que mathématiques et physique au bac et il est vraisemblable que les objectifs de l'étude (comparaisons entre les lycées, comparaison entre les notes trimestrielles et celles du bac) les concernent aussi. Enfin, il est envisageable de remplir une unique valeur trimestrielle manquante par une valeur "raisonnable" : la moyenne des deux autres notes trimestrielles.

Faut-il conserver l'information qu'ils apportent en remplissant les valeurs manquantes par des valeurs raisonnables ? Ou faut-il les supprimer de l'étude ? Ou faut-il les conserver en effectuant les calculs sur les seules valeurs connues ? Discutons ces trois solutions.

Calculer avec toutes les valeurs connues

Nous avons déjà parlé du danger de confronter des calculs réalisés sur des populations différentes. De plus, un fichier "incomplet" limite singulièrement les techniques applicables. Nous ne choisirons cette solution qu'en dernier recours.

Supprimer

La suppression présente l'avantage de la simplicité et du fait de travailler sur un fichier entièrement sûr. On restreint alors le champ de l'étude aux élèves qui ont toutes leurs notes, ce qui n'a rien d'absurde en soi.

Pour l'étude générale des notes sans tenir compte du lycée, la suppression de 4% (=38/947) de la population risque peu de modifier les tendances générales des répartitions, qui seules nous intéressent. Sauf si ces 4% représentent un groupe très différent du reste de la population. Or, en regardant la répartition de ces 38 élèves dans les lycées, on s'aperçoit qu'elle est très irrégulière. Certains lycées ne sont absolument pas touchés alors que d'autres le sont beaucoup, le maximum étant détenu par le lycée 13 dans lequel 12 élèves sur 59 ont des notes manquantes soit 21% (2 élèves de ce lycée, n'ayant aucune note trimestrielle, ont déjà été écartés). La suppression pure et simple des 38 élèves risque donc de biaiser les résultats en ce qui concerne la comparaison des établissements entre eux. Si on choisit cette solution, attention surtout aux conclusions concernant le lycée 13.

Estimer

Pour les élèves qui ont été absents, la valeur n'est pas inconnue mais n'existe pas : il ne s'agit donc pas vraiment d'une estimation au sens propre du terme puisque la valeur manquante n'est pas inconnue, mais n'existe pas pour une raison précise (l'absence de l'élève). C'est simplement un moyen pratique d'obtenir un fichier "complet" en gardant le maximum d'informations.

L'*avantage* est de perturber le moins possible la répartition des variables sans données manquantes pour l'ensemble des 947 élèves (variable *lycée* et les 5 notes du bac) ; pour les autres notes, on travaille avec toutes les valeurs connues plus quelques "estimées".

L'*inconvénient* est d'introduire dans le tableau un biais qu'on ne maîtrise pas : en estimant une note à partir des 2 autres notes trimestrielles, les liaisons entre les 3 notes sont systématiquement renforcées. Mais ce n'est pas dramatique, car ce n'est pas le lien entre ces notes qui nous intéresse le plus, mais plutôt leur lien avec la note du bac.

Si l'on veut procéder à cette opération, une nouvelle surprise nous attend : les notes manquantes (pour les élèves qui en ont au moins 2) ne sont pas réparties dans toutes les matières, mais très souvent groupées. Pour l'un, les 2 valeurs manquantes sont en histoire-géographie ; pour 2 autres, les 3 valeurs manquantes sont encore en histoire-géographie ; pour un quatrième, 2 des 4 valeurs manquantes sont (toujours) en histoire-géographie ; la situation ne s'améliore pas pour 5, 6 ou 7 valeurs manquantes : 4 doubles et un triple dans la même matière.

Pour deux valeurs manquantes, il est possible de répéter la seule connue : mais l'incertitude liée à l'estimation prend des proportions inquiétantes pour ces élèves. Quand les trois trimestres de la même matière manquent, il n'y a guère de solution raisonnable pour remplacer ces valeurs.

On peut penser à une solution mixte dans laquelle les élèves ayant plusieurs valeurs manquantes sont supprimés, et ceux qui n'ont qu'une donnée manquante font l'objet d'une estimation.

4.9 Conclusion

La suppression d'individus et l'estimation (pour les individus pas trop incomplets) sont deux solutions acceptables, mais aucune d'entre elles n'est idéale. Nous préférons ici cependant la première. En effet, on est plus maître de la situation en limitant le champ de l'étude qu'en travaillant avec un fichier modifié. On vérifiera cependant ici que la suppression de ces individus ne modifie pas notablement certains résultats simples comme les moyennes ; sinon, nous serions amenés à nuancer les conclusions.

Pour finir, on a donc supprimé tous les individus qui posaient problème. C'est la solution la plus facile : il était possible de le faire d'emblée, mais il est préférable de dresser d'abord un bilan systématique de la situation. On a ainsi mis en évidence l'importance du phénomène et les différents types d'élèves qui ont des valeurs manquantes. Après avoir étudié la population générale, nous examinerons si les élèves éliminés ne forment pas une sous-population très particulière, c'est-à-dire s'ils n'ont pas systématiquement de meilleures ou de moins bonnes notes que les autres. La réponse à cette question est intéressante en elle-même lorsque l'on veut décrire les données de façon complète.

Avant de répondre aux questions posées dans les objectifs en nous restreignant aux 909 élèves sans aucune donnée manquante, nous étudions, dans le chapitre suivant, le premier groupe d'élèves éliminé.

4.10 Bilan-résumé sur les données manquantes

Déterminer :
le nombre total de données manquantes ;
le nombre par individu et par variable (diagrammes en bâtons) ;
la répartition des données manquantes pour les individus qui en ont le même nombre.
Analyser
Si peu de données manquantes : examiner chacune d'entre elles.
Si beaucoup de données manquantes : expliquer par individu (ou groupe d'individus) et/ou par variable.
Décider
Remplacer, s'il est possible de retrouver la donnée originale.
Faire chaque calcul avec le maximum de données disponibles
Estimer par une valeur plausible compte tenu du contexte.
Abandonner : des individus, des variables, l'étude.

Description d'un petit tableau de données : les 15 élèves avec bac incomplet

L'étude du groupe de 15 élèves pour lesquels on ne connaît que trois variables (le lycée et les notes au bac en mathématiques et en physique) permet d'introduire quelques outils très simples de statistique descriptive. L'objectif est ici d'étudier la répartition de variables qualitatives et quantitatives sur une *population d'effectif faible*. Le mot « faible » est important : entre une population de 15 individus et une population d'un millier d'individus, les questions que l'on se pose et les techniques que l'on met en œuvre sont différentes.

Un tableau de 15 individus et de 3 variables peut être examiné directement (cf. **Tab. 5.1**).

lycée	0	22	0	17	3	9	0	0	1	0	0	0	3	0	17	G
maths	9	15	5	16	16	18	8	7	13	8	12	13	8	10	15	11.53
physique	7	11	2	19	12	16	6	4	12	1	9	10	9	10	9	9.13

Tableau 5.1. *Notes au bac des 15 individus avec bac incomplet ; individus rangés dans l'ordre du fichier.*
Un élève est représenté par une colonne qui contient son numéro de lycée (0=candidat libre) et ses notes au bac, en mathématiques et en physique. G : moyenne des 15 élèves.

5.1 Présentation ordonnée d'un tableau

On constate que la lecture de ce tableau, pourtant très petit, n'est pas si facile. On simplifie cette lecture en triant les lignes ou les colonnes suivant un critère, par exemple les colonnes selon la variable *lycée* (cf. **Tab. 5.2**).

lycée	0	0	0	0	0	0	0	0	1	3	3	9	17	17	22	G
maths	9	5	8	7	8	12	13	10	13	16	8	18	16	15	15	11.53
physique	7	2	6	4	1	9	10	10	12	12	9	16	19	9	11	9.13

Tableau 5.2. *Notes au bac des 15 individus avec bac incomplet ; individus rangés suivant la variable* lycée.

Il devient évident sur ce tableau réordonné que les candidats libres sont très nombreux, que deux lycées ont deux élèves dans ce groupe et que trois autres en ont un seul : par un simple tri, la répartition de la variable qualitative lycée est directement lisible. Le phénomène le plus marqué est bien sûr le grand nombre de candidats libres, dont le statut est très différent de celui des autres candidats. Ce tableau montre aussi que leurs résultats sont généralement moins bons que ceux des autres candidats, ce qui est plus difficile à discerner sur le tableau non trié.

On peut aussi trier les colonnes suivant une variable quantitative, par exemple la note de mathématiques (cf. **Tab. 5.3**).

lycée	0	0	0	3	0	0	0	0	1	0	17	22	17	3	9	G
math	5	7	8	8	8	9	10	12	13	13	15	15	16	16	18	11.53
physique	2	4	6	9	1	7	10	9	12	10	9	11	19	12	16	9.13

Tableau 5.3. *Notes au bac des 15 individus avec bac incomplet ; individus rangés suivant la note de mathématiques.*

Comme pour la variable *lycée*, la répartition des valeurs des notes de mathématiques se lit directement. On pouvait penser que les élèves qui ne passent que deux épreuves ont de bonnes notes à ces deux épreuves ; vraisemblablement ces élèves possèdent déjà un bac D : ils ont donc déjà fait leurs preuves et n'ont que deux matières à travailler. Les faits statistiques contredisent cette hypothèse : leurs notes ne sont pas systématiquement élevées (en mathématiques, 6 d'entre elles sont inférieures à 10).

5.2 Représentation axiale d'une variable quantitative

Principe

Un graphique rend la lecture des données encore plus simple et plus rapide que le tableau réordonné (cf. **Fig. 5.1**). Sur un axe gradué de 0 à 20, chaque individu est représenté par un point situé à la position correspondant à sa note. Ainsi, le premier élève est situé à la graduation 5 de l'axe, le second à la graduation 7, etc. On dit que la coordonnée sur l'axe du premier élève est 5.

Figure 5.1. *Note en mathématiques des 15 individus avec bac incomplet ; représentation axiale en distinguant les candidats libres (disques pleins) et les lycéens (cercles).*

Si l'on veut repérer précisément chaque élève, on peut mettre son numéro ou son nom. Ici, les élèves nous sont inconnus : on se contente de points qui montrent simplement la répartition des notes. Les élèves qui ont la même note sont situés exactement au même point ; c'est le cas pour les trois qui ont 8 et les paires qui ont 13, 15 et 16. Il y a plusieurs manières de représenter ces individus confondus. Dans la figure **5.1**, les points de même coordonnée sont superposés. Cette représentation est proche du diagramme en bâtons car il y a peu de valeurs (les notes sont des nombres entiers), mais ici la position des « bâtons » respecte les distances entre les différentes notes.

Représentation de la moyenne

On a figuré aussi le point G, dit centre de gravité, moyenne des notes des 15 élèves. Les élèves situés à gauche de G ont une valeur inférieure à la moyenne ; ceux situés à droite ont une valeur supérieure à cette moyenne. Il sera intéressant de comparer ces résultats à ceux de l'ensemble des élèves quand nous les aurons étudiés.

Comparaison graphique de sous-populations

Nous avons constaté sur le tableau **5.2.** que les candidats libres avaient généralement des notes inférieures aux autres. Une lecture graphique est plus rapide et plus synthétique ; pour cela, dans la représentation axiale, on a distingué par un sigle différent les candidats libres et les lycéens (cf. **Fig. 5.1**). Un simple coup d'œil permet alors de constater que les candidats libres ont des résultats plus faibles que les autres : les élèves des lycées ont tous, sauf un, une note supérieure ou égale à celles des candidats libres. Le graphique ne donne pas de mesure chiffrée mais suffit dans un cas aussi flagrant. Il reste maintenant à interpréter ces faits statistiques : une première hypothèse est que l'encadrement scolaire est efficace, une autre est que les candidats libres sont a priori des élèves à problèmes, etc. On peut aussi s'interroger à propos du lycéen qui n'a que la note 8.

Cette technique simple permet de comparer rapidement les tendances de deux ou plusieurs sous-populations d'effectif faible. Si le nombre de sous-populations est supérieur à 4 ou 5, le graphique peut ne pas être lisible. On peut alors faire plusieurs graphiques en étudiant à chaque fois une sous-population par rapport à l'ensemble.

5.3 Représentation graphique de deux variables quantitatives sur un plan

Principe

On peut faire pour la physique un graphique analogue à celui fait pour les mathématiques. Il est possible aussi de représenter conjointement les deux notes (cf. **Fig. 5.2**).

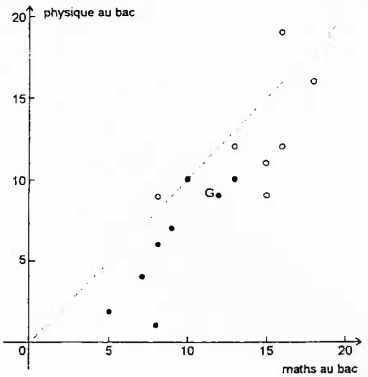


Figure 5.2. *Notes au bac des 15 individus avec bac incomplet ; représentation graphique plane en distinguant les candidats libres (disques pleins) et les lycéens (cercles).*

Sur ce graphique, l'axe horizontal correspond à la note de mathématiques et l'axe vertical à celle de physique. Les deux axes sont gradués de 0 à 20, valeurs extrêmes possibles de ces deux notes. Chaque élève est représenté par un point dont l'abscisse (ou coordonnée sur l'axe horizontal) correspond à sa note de mathématiques et dont l'ordonnée (ou coordonnée sur l'axe vertical) correspond à celle de physique.

Nuage de points ; projections de nuage

L'ensemble des 15 élèves forme ce qu'on a l'habitude d'appeler un *nuage de points*. En le *projetant sur l'axe horizontal*, c'est-à-dire en regardant uniquement les coordonnées des points sur l'axe horizontal, on retrouve la représentation des élèves définie par les notes de

mathématiques (c'est aussi un nuage de points mais sur un axe et non plus dans un plan) ; en le projetant sur l'axe vertical, on obtient le nuage défini par les notes de physique.

Ce nuage visualise la répartition des 15 élèves dans chacune des matières mais aussi suivant les deux notes simultanément. Par exemple, deux points en haut à droite s'écartent des autres : ces deux élèves sont à la fois très bons en physique et en mathématiques.

Représentation de la moyenne

On a figuré le point G, dont les coordonnées sont les moyennes en mathématiques et en physique (11.53 et 9.13). Les élèves situés à gauche de G ont une note de mathématiques inférieure à la moyenne et inversement ceux qui sont à droite ont une note supérieure à cette moyenne. De même, les élèves situés au-dessus (resp. au-dessous) de G ont une note de physique supérieure (resp. inférieure) à la moyenne.

Première bissectrice

Sur le graphique, on a tracé aussi une droite en pointillé appelée "première bissectrice" : c'est la bissectrice de l'angle formé par les demi-axes correspondant aux valeurs positives. Tous les points de cette droite ont mêmes abscisse et ordonnée. Un seul point est situé sur cette droite : l'unique élève ayant la même note en mathématiques et en physique (10). La plupart des points, G compris, sont situés en dessous de cette droite : les notes de mathématiques sont généralement supérieures à celles de physique. Les deux points situés au-dessus constituent deux exceptions.

Etude graphique de la liaison entre deux variables quantitatives

Le nuage présente une forme allongée : les points à droite (note en mathématiques élevée) sont aussi vers le haut (note en physique élevée) et inversement les points à gauche sont en même temps en bas (mauvaises notes dans les deux matières). Les points proches de G pour un axe le sont pour l'autre (notes moyennes dans les deux matières). Deux élèves très bons dans les deux matières se distinguent en haut du graphique ; l'ensemble constitue un nuage de forme allongée dans une direction parallèle à celle de la première bissectrice.

On « voit » donc que les deux variables *mathématiques* et *physique* sont liées entre elles : si l'on connaît la note de mathématiques d'un élève, on peut prévoir que la note de physique sera vraisemblablement un peu plus faible, mais du même ordre de grandeur. La conclusion est par nature nuancée, puisque la statistique décrit des phénomènes réels qui, comme chacun sait, sont rarement tout noir ou tout blanc : on ne peut s'attendre à trouver la formule magique qui donnerait la note en physique d'un élève dont on connaît la note en mathématiques.

La liaison entre deux variables peut donc être étudiée à partir d'un graphique. Si, comme ici, le nuage est allongé dans la direction de la première bissectrice, à une valeur forte de l'une des variables correspond une valeur forte de l'autre : la liaison est dite *positive*. Si la direction d'allongement est perpendiculaire à la première bissectrice, à une valeur forte de l'une des variables correspond une valeur faible de l'autre : la liaison est dite négative (cf. Ch. 8 et Fiche 5).

Les autres petits groupes éliminés de l'étude générale peuvent être examinés de la même manière. Ceci ne conduit pas à introduire de nouvelles techniques ; aussi passons-nous à l'étude du tableau concernant la population des 909 élèves « complets ». Du fait de sa taille, l'étude de ce tableau nécessite une autre démarche et d'autres techniques.

Etude d'une variable qualitative : répartition des élèves dans les lycées

Les moyens de description de la répartition des différentes modalités d'une variable qualitative comme le lycée sont simples et limités.

6.1 Tri à plat

Pour étudier la répartition dans les lycées des 909 élèves ayant toutes leurs notes, il suffit de compter le nombre d'élèves dans chacun d'entre eux. On obtient les *effectifs des modalités* de la variable qualitative *lycée* qui contiennent toute l'information sur la *distribution de cette variable qualitative* (cf. **Tab. 6.1**). L'ensemble de ces effectifs s'appelle le *tri à plat* de la variable *lycée*.

lycée	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	total
effectif	2	27	19	77	65	25	82	11	65	28	38	51	41	47	53	53	17	47	20	71	19	28	23	909
effectif	13	28	20	80	68	26	87	13	67	31	39	52	41	61	56	53	20	51	21	75	20	29	24	975

Tableau 6.1. *Distribution de la variable* lycée *pour les 909 élèves ayant toutes leurs notes et les 975 ayant au moins une note ; tri à plat dans l'ordre des numéros des lycées (0 : candidat libre).*

Ce tableau comporte aussi, à titre indicatif, la distribution des 975 élèves ayant au moins une note. L'objet n'est pas ici de comparer ces deux distributions. Dans le cadre des objectifs généraux, qui incluent la comparaison entre lycées, il faut connaître le fichier sur lequel on travaille. Les deux descriptions ont un intérêt a priori ; mais, ici, seule nous intéresse la répartition des 909 élèves auxquels nous consacrons l'essentiel de l'étude.

6.2 Diagramme en bâtons trié par effectif décroissant

Il est classique et souvent intéressant de visualiser l'ensemble de cette distribution par un diagramme en bâtons (cf. **Fig. 6.1.a**).

Ce diagramme est construit de la même manière que celui de la figure 4.1. décrivant la répartition des données manquantes sur l'ensemble des élèves : les longueurs des bâtons sont proportionnelles au nombre d'élèves de chaque lycée. Ces derniers sont repérés par leur numéro.

Contrairement aux nombres de données manquantes, les numéros des lycées ne traduisent aucun ordre : ils sont arbitraires. Dans ce cas, plutôt que de respecter l'ordre des lycées, il est plus clair de faire le diagramme en ordonnant les lycées par effectif décroissant (cf. **Fig. 6.1.b**).

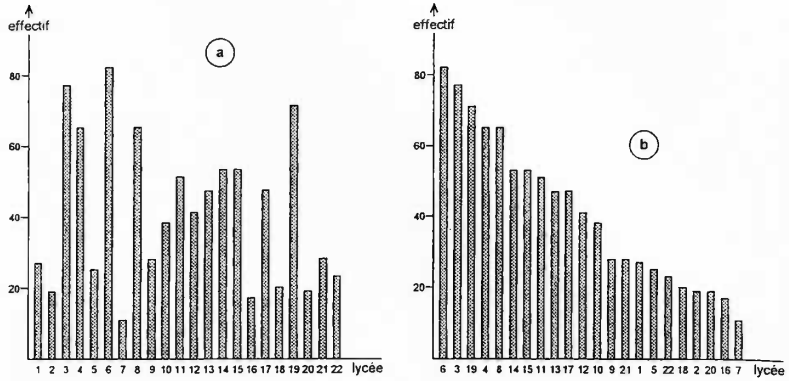


Figure 6.1. *Distribution de la variable* lycée *: diagrammes en bâtons.*
a) lycées rangés par numéro ; b) lycées rangés par effectif décroissant.

On peut remarquer assez vite sur le diagramme trié que les lycées (on excepte les deux candidats libres) peuvent être regroupés, du fait de l'irrégularité de cette répartition, en trois sous-populations correspondant aux effectifs forts, moyens et faibles. Cette structure, qui correspond vraisemblablement au nombre de classes de l'établissement, est difficilement visible avec l'ordre initial des lycées.

6.3 Le regroupement, moyen efficace de description des données

Il est souvent pratique de faire ainsi des regroupements pour décrire les données. Le graphique en suggère un. Pour donner rapidement en quelques chiffres la répartition dans les lycées, on pourra dire, par exemple, que 10 "petits" lycées ont entre 11 et 28 élèves de terminale C, que 7 "moyens" ont entre 38 et 53 élèves et que 5 "grands" ont entre 65 et 82 élèves. Un nouveau problème peut être intéressant : comparer les résultats de ces 3 types de lycées (en fait, les analyses ultérieures ne font apparaître aucune différence entre ces trois types de lycée : cf. **Fig. 9.4** et **Fig. 11.4**).

6.4 Diagramme circulaire

Pour représenter la distribution d'une variable qualitative, on utilise souvent des *diagrammes circulaires* (ou *diagramme en secteurs* ou *camemberts*). Dans ces diagrammes, la surface figurant une modalité est proportionnelle à sa fréquence. Ici encore, s'il n'existe aucun ordre a priori sur les modalités, on gagne en lisibilité en ordonnant les modalités par effectif croissant (cf. **Fig. 6.2**).

Sans doute, dans un but esthétique, certains de ces diagrammes sont représentés comme un disque épaissi en vue perspective. Le malheur est que, très souvent, le graphique ne respecte pas la proportionnalité dans la représentation des surfaces et qu'il donne une représentation faussée de l'importance des différentes modalités (les modalités situées devant apparaissent proportionnellement plus grandes qu'elles ne le sont réellement, à la fois du fait de la perspective et de l'ajout d'un bord qui induit un renforcement visuel).

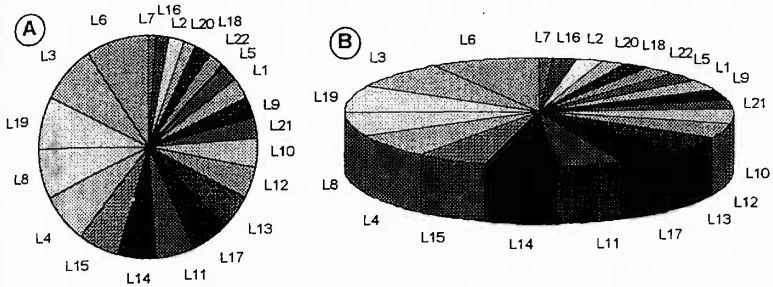


Figure 6.2. *Distribution de la variable lycée : diagrammes circulaires. 907 élèves dans 22 lycées rangés par effectif croissant.*
a) diagramme plat ; b) diagramme en relief, appelé aussi diagramme 3-D quoique une seule dimension (ici l'effectif par lycée) soit étudiée.

Etude de variables quantitatives ; répartition des notes

L'un des objectifs précisés au début de l'étude est de décrire la répartition des notes, pour chaque matière et chaque date. Un autre concerne la comparaison des notes entre elles pour répondre à des questions du type : les notes de mathématiques sont-elles meilleures ou moins bonnes au bac que pendant l'année ?

Le chapitre 5 montre comment décrire des répartitions (ou distributions) de variables quantitatives lorsque l'effectif de la population est faible. Pour près de mille individus, le problème est complètement différent : en particulier des indices synthétiques deviennent utiles et même indispensables si l'on veut effectuer des comparaisons entre variables.

Variable discrète et variable continue

Les notes du bac et les notes trimestrielles sont de natures un peu différentes. En effet, les premières ne prennent que des valeurs entières (contrainte imposée aux correcteurs du bac) alors que les secondes sont données avec deux décimales. Les premières ont seulement 21 valeurs possibles, les secondes en ont a priori beaucoup plus (2001 très exactement). Avec un certain arbitraire, on distingue quelquefois les variables quantitatives avec peu de valeurs possibles de celles qui en ont beaucoup en appelant les premières *variables discrètes* (ou discontinues) et les secondes *variables continues* (terme qui évoque, sans se confondre avec elle, la continuité au sens mathématique).

Ici, le nombre de valeurs possibles pour les notes du bac est tel que l'on peut la traiter aussi bien comme discrète que comme continue.

7.1 Variable discrète ; diagramme en bâtons

Pour disposer d'une information complète sur la distribution des notes au bac, qui ne prennent que 21 valeurs entières au maximum, on peut représenter les effectifs des élèves pour chaque note par un diagramme en bâtons. Il est clair qu'ici l'ordre des notes est primordial et doit être conservé (cf. **Fig. 7.1**), à la différence de la figure 6.1 dans laquelle il convient de classer les lycées par effectif décroissant.

Forme d'une distribution

Un coup d'œil sur le graphique permet d'appréhender la *forme de la distribution*. Par exemple, pour les mathématiques, on voit des effectifs très nombreux pour toutes les notes comprises entre 11 et 17, des effectifs plus faibles pour les deux notes supérieures et les notes inférieures, et de rares très mauvaises notes. Il n'y a pas de "trous" dans la distribution (c'est-à-dire des intervalles contenant très peu ou même pas du tout de notes) qui indiqueraient l'existence de plusieurs sous-populations (comparer avec la distribution des données manquantes, fig. 4.1).

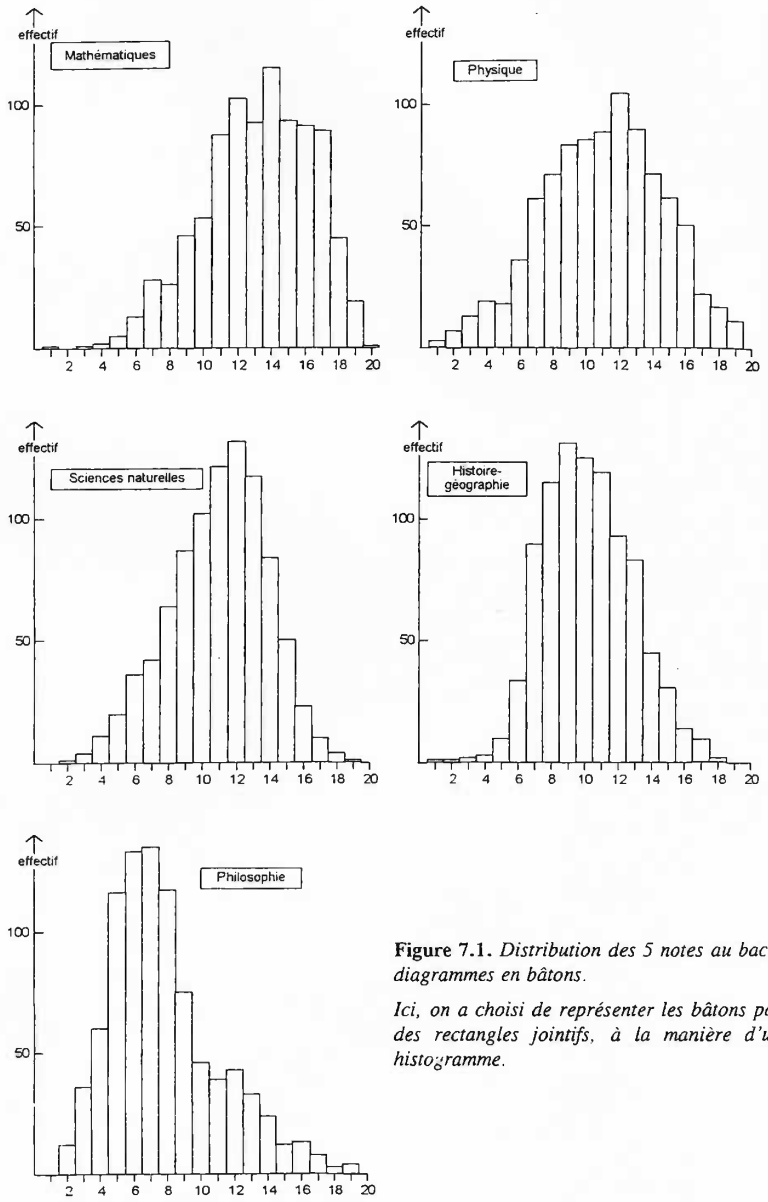


Figure 7.1. *Distribution des 5 notes au bac : diagrammes en bâtons.*

Ici, on a choisi de représenter les bâtons par des rectangles jointifs, à la manière d'un histogramme.

Mode d'une distribution, distribution unimodale

La note la plus fréquemment donnée en mathématiques est 14 : 115 élèves en bénéficient. Cette valeur la plus fréquente est appelée *mode* de la distribution. Ici, ce "pic" n'est pas unique (il existe un autre « pic » à 12 et, plus discrets, deux autres en 7 et 1) : la distribution est dite *multimodale*. Si l'on ne considère que les deux plus saillants (12 et 14), on dira que la distribution est *bimodale*. Lorsqu'il y a un seul « pic », comme en sciences naturelles et en histoire-géographie, elle est dite *unimodale*.

7.2 Variable continue ou discrète ; histogramme

Diagramme en bâtons

Les notes trimestrielles présentent beaucoup de valeurs possibles. Pour de telles variables, il n'est généralement pas possible et/ou intéressant de représenter la distribution par un diagramme en bâtons : ces bâtons risquent d'être trop nombreux, trop serrés et souvent de longueur 1.

Dans ces données, les notes trimestrielles présentent beaucoup de valeurs identiques et les diagrammes en bâtons sont utilisables (cf. **Fig. 7.2**).

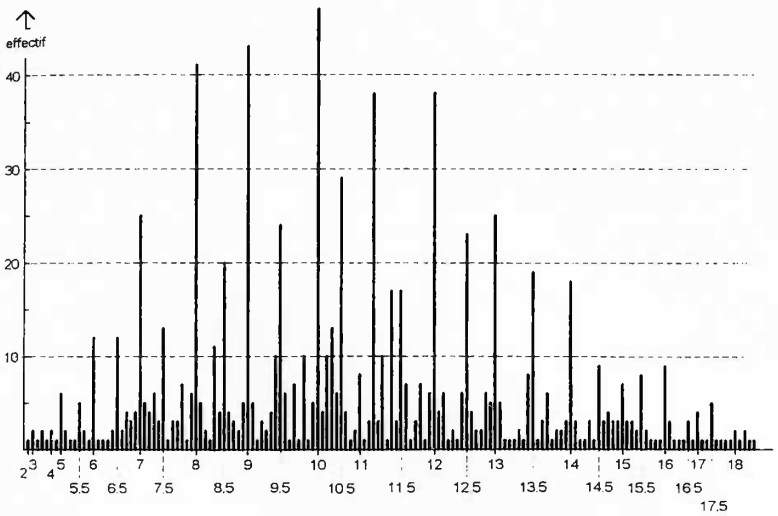


Figure 7.2. *Distribution de la note en maths au premier trimestre : diagramme en bâtons. Les bâtons sont régulièrement espacés ; l'écart entre deux notes est donc proportionnel au nombre de valeurs utilisées au moins une fois entre ces deux notes ; ainsi 2 et 3 sont contiguës alors que 9 et 10 sont séparées par 11 valeurs.*

Ce diagramme est très irrégulier, en ce sens que des effectifs très élevés côtoient des effectifs très faibles ; en particulier, les notes entières et les notes « et demi » (7.5, 8.5, etc.) sont beaucoup plus fréquentes que les autres. Cette structure multimodale découle du fait que les notes dont on calcule la moyenne sont en général entières et que les moyennes

elles-mêmes sont souvent arrondies. Elle révèle plus la façon dont les valeurs ont été calculées que ce qu'elles sont censées mesurer c'est-à-dire le niveau dans une matière. Ainsi, ce qui nous intéresse n'est pas de remarquer que 10 est une valeur plus fréquente que 10.1 : pour juger du niveau d'un élève, ces deux notes sont équivalentes. D'où l'idée de regrouper les valeurs voisines et de construire un histogramme.

Histogramme avec intervalles égaux

Le graphique **7.3.a** représente un histogramme des notes de mathématiques au premier trimestre. On peut le voir comme une généralisation du diagramme en bâtons. Au lieu de représenter l'effectif des élèves qui ont une note précise (10 par exemple), on représente l'effectif de ceux qui ont une note comprise entre deux valeurs (par exemple, de 9.50 non compris à 10.50 compris). Au lieu de bâtons, on figure des rectangles basés sur l'intervalle qu'ils caractérisent. On a choisi ici des intervalles de longueur égale à 1 point et centrés sur une valeur entière (sauf aux extrémités). L'effectif d'élèves dont les notes sont dans chaque intervalle est représenté par la surface du rectangle. Les intervalles étant égaux, surfaces et hauteurs sont proportionnelles et les hauteurs des rectangles représentent aussi les effectifs. Le cas des intervalles égaux est le plus fréquent : l'histogramme peut se lire alors comme un diagramme en bâtons, d'où la confusion fréquente entre ces deux termes (confusion accentuée par le fait que certains logiciels ne font que des histogrammes avec intervalles égaux et ne représentent pas les rectangles mais seulement leur hauteur par des bâtons).

Le commentaire sur la forme de la distribution est tout à fait analogue à celui de la note de mathématiques au bac (distribution unimodale, absence de trous, etc.).

Regroupement en classes

Le tracé de l'histogramme a nécessité un *regroupement en classes* de la population. On a divisé l'intervalle de variation en sous-intervalles disjoints puis on a regroupé, pour les compter, toutes les notes appartenant à chaque sous-intervalle. Bien évidemment le graphique dépend du choix de ces intervalles. La figure **7.3** compare plusieurs histogrammes de la variable *note en mathématiques au 1er trimestre*.

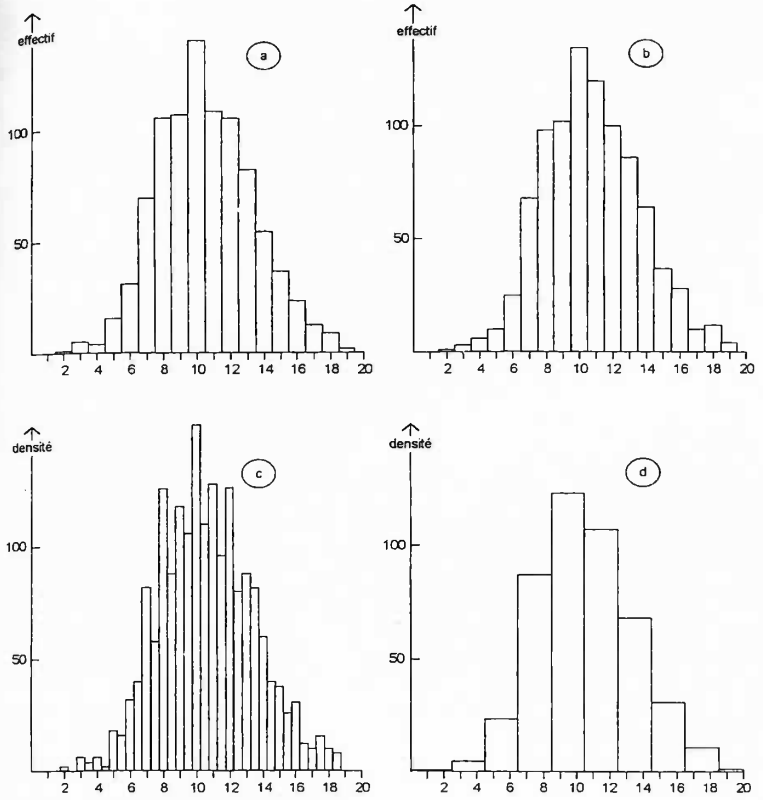
Choix des bornes

Pour subdiviser l'échelle des notes en une vingtaine d'intervalles (on dit aussi classes) de même longueur, il est logique de centrer les classes sur les notes entières. En effet, le centre d'une classe représente en quelque sorte la classe : or les notes entières sont plus simples (10 « parle » plus que 10.23) et en outre correspondent aux effectifs les plus élevés.

Se pose encore le choix des bornes : est-ce la borne inférieure ou la borne supérieure qui doit être incluse dans l'intervalle ? En général ce choix n'a pas d'importance, mais ce n'est pas le cas ici où ces bornes présentent des effectifs importants. On ne peut donner de règles générales, hormis d'utiliser la même définition pour toutes les classes. Les figures **7.3.a** et **7.3.b** illustrent l'influence de ce choix sur l'allure de l'histogramme.

Choix du nombre de classes

Si l'on s'intéresse de très près à toutes les irrégularités d'une distribution, il faut des intervalles petits. Si au contraire on veut examiner la forme générale en gommant les irrégularités, il faut des intervalles assez grands. Le premier choix (20 classes) est assez naturel pour comparer les histogrammes des notes trimestrielles et les diagrammes en bâtons des notes du bac. On peut aussi choisir des intervalles de 2 ou 3 points, l'intérêt de l'intervalle de 3 points étant de centrer sur 10 (i.e. le milieu de l'échelle de 0 à 20) la classe du milieu (cf. **Fig. 7.3.e**).



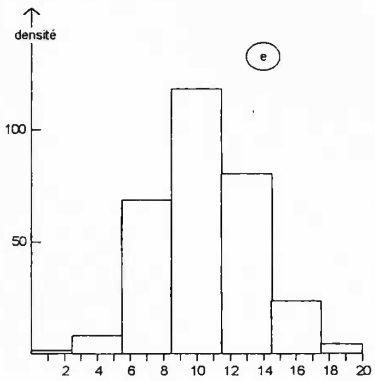


Figure 7.3. *Distribution de la note en maths au premier trimestre : histogrammes.*
a) 20 intervalles, borne supérieure incluse ; ex : 9.50 < note ≤ 10.50 ;
b) 20 intervalles, borne supérieure exclue ; ex : 9.50 ≤ note < 10.50
c) 40 intervalles, borne supérieure incluse ; ex : 9.75 < note ≤ 10.25 ;
d) 10 intervalles, borne supérieure incluse ; ex : 8.50 < note ≤ 10.50 ;
e) 7 intervalles, borne supérieure incluse ; ex : 8.50 < note ≤ 11.50 ;

Remarque : opter pour de grands intervalles rend pratiquement indifférent le choix de la borne (inférieure ou supérieure) incluse dans l'intervalle, l'effectif sur les bornes devenant petit en regard de celui des classes.

Densité d'une distribution

La hauteur des rectangles est le quotient entre l'effectif des individus et la longueur de l'intervalle : c'est la *densité* (i.e. le nombre d'individus pour un intervalle de longueur 1 ; de même qu'une densité de population exprime un nombre d'individus par unité de surface) de la distribution sur l'intervalle. La forme d'un histogramme, donnée par le haut des rectangles, visualise donc la densité moyenne et non les effectifs sur les intervalles. C'est pourquoi les différents histogrammes de la figure **7.3.** sont directement comparables entre eux et même superposables (par exemple, dans l'histogramme en 40 classes d'amplitude .5 point, les effectifs des classes ont été divisés par .5 pour obtenir les hauteurs des rectangles).

Histogramme avec intervalles inégaux

Dans l'histogramme en 7 classes (cf. **Fig. 7.3.e**), choisir des intervalles de 3 points, chacun centré sur une note entière, pose un problème pour les valeurs extrêmes. Les deux classes extrêmes ont une longueur de seulement 2.5 points ; pour ces deux classes, la hauteur du rectangle, c'est-à-dire la densité, est obtenue en divisant l'effectif de la classe par 2.5 et non par 3.

En pratique, il est fréquent de construire comme ici des histogrammes en classes égales, exceptées les classes extrêmes.

Forme d'une distribution, histogramme de la loi normale, symétrie d'une distribution

Avec un nombre d'intervalles inférieur ou égal à 21 l'histogramme est « régulier ». Il a une allure tout à fait classique. En effet, la forme de la courbe constituée par le haut des rectangles (en gras sur le graphique) ressemble à celle d'une « cloche ». Notons simplement ici les principales caractéristiques de cette forme de distribution.

- La moyenne est près du sommet de la cloche. Or, un rectangle très haut signifie une densité importante : un fort pourcentage d'individus est donc proche de la moyenne.
- Les hauteurs des rectangles diminuent assez régulièrement lorsque l'on s'éloigne du sommet : la densité est d'autant plus faible que l'on s'éloigne de la moyenne.
- Les hauteurs des rectangles situés à la même distance de la moyenne, à droite et à gauche, sont presque égales : la distribution est presque *symétrique*.

Cette forme en cloche est très courante et s'apparente au modèle de distribution appelé *distribution* (ou *loi*) *normale* (ou de Gauss ou de Laplace Gauss). Cette loi, qui présente des propriétés théoriques très intéressantes (cf. Fiche 9), est utilisée abondamment.

Nous ne donnons pas les histogrammes de toutes les notes car ils sont tous du même type.

Histogramme de variable discrète

Pour les variables discrètes comme les notes du bac, on peut aussi tracer des histogrammes. On peut choisir des intervalles qui encadrent chaque valeur entière (de .5 à 1.5, etc.), auquel cas on ne change rien par rapport au diagramme en bâtons : les rectangles ont la même hauteur que les bâtons. On peut aussi regrouper plusieurs valeurs discrètes en prenant par exemple des intervalles de 2 ou 3 points.

Conclusion. Les diagrammes en bâtons et les histogrammes décrivent les répartitions de manière très complète, ce qui permet l'étude détaillée de chaque variable. Ils permettent aussi, par un simple coup d'œil, de voir la forme générale d'une répartition, d'en vérifier la régularité et de repérer des anomalies. Mais c'est un instrument assez lourd. Pour avoir une vue globale de la situation et surtout pour effectuer des comparaisons entre répartitions, on a besoin de synthétiser l'information qu'ils contiennent par un petit nombre d'indices.

7.3 Moyennes des notes

Plusieurs indices résument les divers aspects d'une distribution d'une variable quantitative. Commençons par le plus simple et surtout le plus connu : la moyenne.

Moyenne d'une variable quantitative

La moyenne d'une matière se calcule en faisant la somme des 909 valeurs de cette matière et en la divisant par 909. On peut classer les moyennes de toutes les notes par matière et par date (cf. **Tab. 7.1**).

	maths	physique	sciences nat.	histoire-géo.	philosophie
trimestre 1	10.64	10.93	10.85	11.04	8.99
trimestre 2	11.16	11.12	11.13	10.94	9.01
trimestre 3	11.05	11.45	11.08	11.06	9.26
bac	13.21	11.00	10.97	10.21	7.84

Tableau 7.1. *Moyenne de chacune des 20 variables (calcul sur 909 élèves).*

Moyenne, indicateur de tendance centrale

La moyenne est intéressante en elle-même car elle permet de situer globalement les résultats de la population : il est clair que si dans chaque matière au bac la moyenne générale était égale à 2, peu d'élèves obtiendraient leur diplôme et l'on devrait conclure que le niveau de l'épreuve n'est guère adapté à la population ou que cette dernière est mal préparée à l'épreuve. Ce n'est pas le cas : les moyennes sont comprises entre 7.84 et 13.21 . On peut les comparer à la valeur 10, référence habituelle pour un examen, qui porte le même nom de moyenne mais qui recouvre une notion tout à fait différente : c'est la moyenne des deux notes extrêmes 0 et 20 ; elle constitue la barre au-dessous de laquelle on considère qu'un élève n'a pas acquis assez de connaissances et de réflexion. En mathématiques au bac, avec une moyenne de 13.21, les résultats des élèves sont en moyenne bons ; en philosophie au bac, avec une moyenne de 7.84, les résultats sont en moyenne mauvais.

La moyenne d'une variable est l'indice le plus utilisé pour commencer à décrire la répartition de ses valeurs. Il est simple, bien entendu très insuffisant (on ne peut résumer 909 valeurs par une seule), mais permet déjà de situer globalement l'ensemble des notes. On dit que c'est un *indicateur de tendance centrale*.

Moyenne, élément de comparaison entre distributions de variables

On utilise aussi la moyenne pour comparer les répartitions de plusieurs variables. C'est le moyen le plus simple et peut-être le plus informatif pour comparer les différentes notes. Ce qui ne veut pas dire qu'il sera inutile d'aller plus loin.

Nous n'allons pas comparer 2 à 2 les 20 notes (190 comparaisons). Compte tenu de la structure en groupes des variables, nous effectuerons d'abord une comparaison entre les différentes matières en nous restreignant aux 5 notes du bac, puis une comparaison à

l'intérieur de chaque matière, pour voir d'une part si les moyennes de l'année sont stables ou si elles évoluent, et d'autre part si la moyenne du bac diffère de celles de l'année.

Comparaison entre les matières

Pour comparer entre elles les 5 matières au bac, les chiffres parlent d'eux-mêmes : les mathématiques se détachent en tête avec 13.21, la philosophie est en queue avec 7.84 et le peloton des trois autres matières est entre 10.21 et 10.99.

La différence entre mathématiques et philosophie atteint presque 5.5 points, ce qui, pour une échelle de 0 à 20, est énorme. Chacun peut émettre ses commentaires personnels.

Evolution de chaque matière : représentation graphique de séries chronologiques.

Pour chaque matière, les 4 moyennes sont des valeurs successives dans le temps d'une même variable : c'est une *série chronologique*. Certaines questions posées initialement (cf. 2.2) se réfèrent aux évolutions de ces séries. Un graphique facilite la description et la comparaison entre plusieurs séries chronologiques (cf. **Fig. 7.4**).

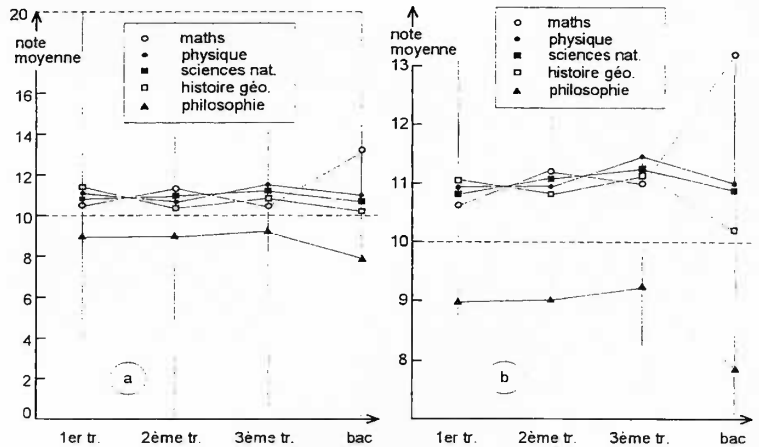


Figure 7.4. *Evolution des moyennes pour les 5 matières.*
Les points correspondant à une même matière ont été reliés. L'échelle des ordonnées est approximativement cadrée sur les notes extrêmes possibles (a) ou observées (b). Par rapport à leur position exacte, les points sont écartés (au minimum) pour éviter les superpositions ; ces écarts sont importants en (a).

Dans un paysage plutôt plat (pour toutes les matières, l'évolution des 3 notes trimestrielles ne dépasse guère le demi-point) on remarque d'abord la note de mathématiques au bac qui dépasse de plus de 2 points les notes trimestrielles, puis à l'opposé, celle de philosophie au bac qui chute de plus d'un point. Les descriptions statistiques sont très souvent de ce type : on s'intéresse aux *tendances générales* mais aussi, voire au moins autant, aux *accidents* ou *anomalies* par rapport aux tendances générales. En mathématiques, les élèves sont donc en moyenne sous-notés pendant l'année ; en philosophie, c'est l'inverse ; dans les autres matières les résultats de l'année reflètent en moyenne ceux du bac.

Ayant relevé par une étude statistique une anomalie comme la croissance importante de la note en mathématiques (par rapport aux autres matières), il faut l'analyser. Ici on peut seulement émettre quelques hypothèses : par exemple, que l'épreuve était moins difficile que celles des années précédentes sur lesquelles se fondaient les enseignants pour préparer les élèves.

7.4 Quelques notations utiles

Bien que nous tentions dans cet ouvrage de limiter au maximum l'emploi de formules, elles sont quelquefois bien pratiques et plus parlantes qu'un discours. Pour les comprendre, il est nécessaire de connaître quelques notations classiques. Ce paragraphe n'est utile qu'à ceux qui les ignorent.

Notation : variable x

Pour parler d'une variable quelconque, on la désigne souvent par la lettre x. Dans ce paragraphe, par exemple, on note x la note de mathématiques au bac.

Notation symbolique x_i

On manie souvent des sommes de très nombreux termes. Il est pratique de les écrire avec une formule générale. Pour cela, il est nécessaire de disposer de notations symboliques. On numérote les élèves de 1 à 909 dans l'ordre du fichier. Pour désigner l'un quelconque des 909 individus, on parle de l'individu *numéro i*, ou d'une manière plus condensée de l'individu i.

Par exemple, on note par x_i (qui se lit x indice i) la valeur de la note de mathématiques au bac de l'individu i. Ainsi x_1 désigne la note de mathématiques du premier élève et vaut 14, x_2 désigne celle du second et x_{909} celle du dernier.

On peut maintenant écrire une formule générale pour calculer la moyenne. Comme il serait un peu long d'écrire une formule comportant 909 termes, on se contente de préciser les premiers et les derniers, mettant entre eux des points de suspension qui indiquent la présence d'autres termes.

$$\text{moyenne de } x = \frac{1}{909}(x_1 + x_2 + x_3 + \quad \ldots \quad + x_{908} + x_{909})$$

Notation Σ

Cette formule est un peu lourde et il est plus pratique d'utiliser la notation condensée avec le signe Σ (qui se lit « somme » ou « sigma »). Pour présenter cette notation, donnons-la d'abord pour la somme des 5 premiers termes :

$$x_1 + x_2 + x_3 + x_4 + x_5 = \sum_{i=1,5} x_i = \sum_{i=1}^{i=5} x_i$$

L'expression condensée à droite du signe égal se lit "somme pour i variant de 1 à 5 des x indice i". Le signe Σ indique que l'on effectue une somme. L'indice i apparaît à la fois sous le signe Σ et dans le terme x_i. Sous le signe Σ, il permet de préciser le nombre de termes de la somme : $i = 1,5$ signifie que cet indice va prendre les valeurs comprises entre 1 et 5. Il y a 5 valeurs (1,2,3,4,5) donc 5 termes dans la somme.

A droite du signe Σ on précise les termes qui sont sommés. Ici ce sont des x_i, donc pour nous les notes de mathématiques au bac d'individus numérotés par i, c'est-à-dire les 5 premiers élèves.

Avec le signe Σ le nombre de termes sommés n'alourdit pas la formule. La moyenne des notes de mathématiques au bac des 909 élèves s'écrit :

$$\frac{x_1 + x_2 + x_3 + ... + x_{908} + x_{909}}{909} = \frac{1}{909} \sum_{i=1,909} x_i$$

Notation $\bar{x}$ de la moyenne

La moyenne de la variable x est classiquement notée par $\bar{x}$: dans les formules ce signe se lit "x barre".

Notation I du nombre d'individus

Dans le fichier, le nombre total d'individus, noté I, vaut 909. La notation I permet d'écrire la formule générale de la moyenne pour une variable x définie sur un ensemble de I individus :

$$\bar{x} = \frac{1}{I} \sum_{i=1,I} x_i$$

7.5 Maximum, minimum, étendue

La moyenne est pratique et facile à manier mais forcément sommaire pour résumer 909 nombres. Elle ne permet pas de répondre aux questions du type : dans quelles matières y a-t-il de très mauvaises notes (ou de très bonnes notes) ?

Pour cela, on calcule la plus faible et la plus forte note de chaque matière, le *maximum* et le *minimum* des 20 variables quantitatives (cf. **Tab. 7.2**). Notons que, contrairement à la moyenne, ces valeurs se lisent directement sur le diagramme en bâtons et, à très peu près, sur l'histogramme.

	maths	physique	sciences nat.	histoire-géo.	philosophie
trimestre 1	2.0 -- 18.7	2.25 -- 19.5	5.0 -- 17.50	5.0 -- 16.6	3.0 -- 17.5
trimestre 2	1.6 -- 19.0	2.40 -- 18.7	2.7 -- 18.50	3.1 -- 16.4	4.0 -- 16.0
trimestre 3	2.0 -- 19.5	2.00 -- 19.6	2.5 -- 18.25	3.0 -- 17.5	1.0 -- 18.0
bac	1 -- 20	1 -- 19	2 -- 19	1 -- 18	2 -- 19

Tableau 7.2. *Minimum et maximum de chacune des 20 notes.*

Les minima varient de 1 à 5 dans toutes les matières : il en existe de très mauvais mais aucun élève n'est considéré comme totalement nul. Les maxima vont de 16 (philosophie deuxième trimestre) à 20 (mathématiques au bac) : il y a des très bons partout et un parfait en mathématiques. Les minima ont plutôt tendance à diminuer avec le temps et au bac, tandis que les maxima ont plutôt tendance à augmenter.

Aucune note n'utilise l'échelle complète de 0 à 20. La note de mathématiques au bac a *l'étendue* (différence entre le maximum et le minimum) la plus grande avec 19. Quelle que soit la matière, l'étendue des notes est plus importante au bac que pendant l'année. L'étendue est l'indice le plus simple (et le plus ancien) pour mesurer la *dispersion des valeurs d'une variable quantitative*.

Il est évident qu'un seul élève complètement nul peut modifier les minima, comme un seul élève génial peut modifier les maxima, et donc l'étendue. Ces indicateurs sont donc assez instables. L'étendue ne permet pas de donner de réponse fiable aux questions : les notes de mathématiques sont-elles groupées autour de la moyenne ou au contraire très dispersées ? Le sont-elles plus ou moins que celles de philosophie ? Pour répondre à ces questions, il est nécessaire d'utiliser d'autres indices de dispersion.

7.6 Dispersion autour de la moyenne : écart absolu moyen, écart-type

Ecart à la moyenne, écart absolu, écart absolu moyen

Comment mesurer par un seul indice la dispersion des notes autour de la moyenne ? Il est facile de voir si la note d'un seul élève est proche ou loin de la moyenne : on calcule la différence entre cette note et la moyenne, appelée écart à la moyenne. L'idée la plus naturelle pour avoir un indice global de dispersion est de calculer la moyenne de ces écarts. Mais attention : lorsque la note est supérieure à la moyenne l'écart est positif et lorsque la note est inférieure à la moyenne l'écart est négatif. Dans le calcul de la moyenne de ces nombres, les écarts positifs et négatifs s'équilibrent et l'on obtient zéro, ce qui n'est guère intéressant. Pour mesurer la dispersion moyenne, on peut considérer la valeur absolue de chaque écart (ou *écart absolu*) et calculer ensuite la moyenne de ces écarts absolus. Cette moyenne s'appelle *écart absolu moyen*.

Pour les mathématiques au bac l'écart absolu moyen vaut 2.61.

Cet indice qui se conçoit facilement n'est pas couramment utilisé. On utilise plutôt un indice assez proche, qui jouit de beaucoup de bonnes propriétés (cf. Fiche 3) : l'écart-type.

Variance et écart-type

Au lieu de prendre la valeur absolue des écarts à la moyenne, on en prend le carré et on calcule la moyenne de ces carrés. Cette moyenne des carrés des écarts s'appelle *variance* de la variable x. Pour les mathématiques au bac :

$$\text{variance de } \textit{maths au bac} = \frac{1}{909}\sum_{i=1.909}(x_i - 13.21)^2 = 10.18$$

Le carré résout le problème de l'annulation des écarts positifs et négatifs comme le fait la valeur absolue pour l'écart absolu moyen. Mais en travaillant sur des carrés, l'échelle de mesure n'est pas du tout la même et n'est pas très parlante. Pour les notes où l'unité est le point, l'unité de la variance est le "point carré" (de même l'unité de la variance d'une longueur mesurée en centimètres est le centimètre carré). Pour se ramener à une valeur de même échelle que x, on prend la racine carrée de la variance, appelée écart-type et notée s ; la variance, carré de l'écart-type, est souvent notée s^2.

$$\text{écart-type de } \textit{maths au bac} = \sqrt{10.18} = 3.19$$

Cet exemple montre que l'écart-type et l'écart absolu moyen peuvent différer sensiblement. La fiche 3 contient des compléments sur ces différences et l'intérêt de choisir l'écart-type, de conception plus compliquée que l'écart absolu moyen. Notons simplement ici que l'écart-type est toujours supérieur ou égal à l'écart absolu moyen (l'égalité n'étant atteinte que dans le cas d'une population concentrée en 2 points symétriques par rapport à la moyenne) et que l'écart-type s'éloigne d'autant plus de l'écart moyen que les écarts à la moyenne sont eux-mêmes hétérogènes.

L'écart-type doit être calculé pour chaque note (cf. **Tab. 7.3**). En pratique, on l'interprète presque comme un écart moyen bien que ce soit inexact.

	maths	physique	sciences nat.	histoire-géo.	philosophie
trimestre 1	2.85	2.97	2.41	2.13	2.00
trimestre 2	2.77	2.79	2.43	2.17	2.12
trimestre 3	2.98	3.16	2.89	2.39	2.33
bac	3.19	3.61	2.88	2.65	3.29

Tableau 7.3. *Ecart-type de chacune des 20 notes.*

Les écarts-types des notes sont de l'ordre de 2 à 3 points. Ceci est à comparer à leurs valeurs extrêmes possibles 0 et 10 : 0 si tous les élèves avaient la même note ; 10 si les élèves se séparaient en 2 groupes égaux (les nuls avec 0 et les parfaits avec 20).

Ecart-type, moyen de comparaison des dispersions des distributions

Un des intérêts fondamentaux des indices numériques est de faciliter des comparaisons. Dans notre étude, la comparaison entre les différents écarts-types est plus significative et fructueuse que l'interprétation de la valeur elle-même de tel ou tel écart-type. A ce niveau, le choix entre écart-type et écart moyen n'est pas primordial : les différences notables de dispersion entre les notes sont traduites par chacun des deux indices. Ici, un phénomène est très net : dans chaque matière, les notes du bac ont un écart-type plus grand que les notes de l'année. La dispersion des notes est donc plus importante au bac que pendant l'année.

Plusieurs hypothèses peuvent expliquer la plus grande dispersion des notes du bac.

- Une raison de nature statistique : les notes trimestrielles sont souvent les moyennes de plusieurs devoirs. **Calculer des moyennes a toujours tendance à réduire la dispersion.** On peut le montrer théoriquement, mais on comprend bien la nature du phénomène : sur plusieurs notes les très faibles ont des chances de se rattraper un peu et les excellents peuvent avoir des défaillances (cf. Fiche 11).
- Une raison venant de la connaissance des données. Dans chaque classe les notations sont toujours influencées par le niveau général de la classe. Par exemple, les élèves brillants regroupés dans une bonne classe sont désavantagés.

Dans le même ordre d'idée, à matière égale, les écarts-types sont plus grands au $3^{ème}$ trimestre qu'aux deux premiers (les différences sont faibles mais systématiques) : en relation avec la première hypothèse précédente, on peut imaginer que le nombre de devoirs dont on a fait la moyenne est en général plus faible au $3^{ème}$ trimestre.

Globalement, si l'on excepte la note de philosophie au bac, les notes des matières scientifiques sont plus dispersées que celles des matières littéraires.

Les notions de moyenne et d'écart-type permettent d'introduire celles, très pratiques, de variable centrée et de variable centrée-réduite que nous n'utilisons pas immédiatement mais dont nous donnons quelques propriétés.

Variable centrée

Dans l'étude de la dispersion de la variable x autour de sa moyenne, ce sont les écarts à la moyenne qui sont considérés. Pour les mathématiques au bac, par exemple, on considérera non plus la note d'un élève mais la différence entre sa note et la moyenne 13.21. Quand on

calcule cette différence pour tous les élèves, on obtient une nouvelle variable : l'écart à la moyenne. Cette nouvelle variable s'appelle *variable centrée*. Pour la variable centrée associée à la note de mathématiques au bac, l'élève qui a 12 aura −1.21 et celui qui a 15 aura 1.79. L'étude directe de cette nouvelle variable permet tout de suite de *situer un individu dans la population*. Celui qui a 12 peut croire a priori qu'il a une bonne note (elle est supérieure à 10), mais son écart est négatif : il est donc au-dessous de la moyenne de l'ensemble des élèves, et même à plus d'un point.

Remarques

- La moyenne d'une variable centrée est égale à 0.
- Si un jury compatissant ajoutait 2 points à tous les candidats, la variable centrée serait inchangée. En effet, l'élève qui a 12 aurait 14, mais la moyenne de l'ensemble des élèves aurait aussi augmenté de 2 points ; son écart à la moyenne serait toujours de −1.21 points. La variable centrée est insensible à l'addition ou la soustraction d'un nombre constant.

Variable centrée-réduite

Il est utile de diviser la variable centrée par son écart-type. La nouvelle variable est dite *centrée-réduite* ou plus simplement *réduite* : son écart-type vaut 1.

En divisant la variable centrée (ou écart à la moyenne) par son écart-type, une note située à un écart-type de la moyenne sera ramenée à 1, une autre située à 2 écarts-types sera ramenée à 2 : l'échelle de référence, ou unité de mesure, d'une variable centrée-réduite est l'écart-type.

Par exemple, la valeur de la variable centrée-réduite pour l'élève qui a 12 en mathématiques au bac vaut :

$$\frac{12 - 13.21}{3.19} = -0.38$$

L'examen de cette valeur permet de situer l'élève dans la population : sa note (12) est inférieure à la moyenne générale (valeur centrée négative), mais assez proche de cette moyenne générale compte tenu de la dispersion des notes (écart à la moyenne générale nettement inférieur à la moitié d'un écart-type).

Les valeurs des variables centrées-réduites sont complètement indépendantes des unités de départ. Une mesure exprimée en mètres ou en centimètres donne exactement la même variable centrée-réduite. On peut ainsi faire des comparaisons entre variables de nature différente : si un enfant est à +3 écarts-types de la moyenne pour sa taille et +1 écart-type pour son poids, on sait qu'il est plus remarquable par sa taille que par son poids.

L'examen des variables centrées-réduites est très pratique pour détecter des valeurs « anormalement » grandes ou « anormalement » petites (cf. Ch. 12).

7.7 Boîte de dispersion, médiane, quartile, outlier

Outre la moyenne et l'écart-type, il existe d'autres indices un peu moins courants. Nous les définissons et examinons leur intérêt dans le cas de nos données. Cette présentation est organisée à partir d'un graphique classique, la *boîte de dispersion* ou *boîte à moustaches*, qui représente la *médiane*, les *quartiles* et les *outliers*.

Boîte de dispersion

Ce graphique, doté aussi du joli nom de boîte à moustaches, est précieux pour comparer entre elles des distributions, par exemple les 4 notes de mathématiques et de philosophie (cf. **Fig. 7.5**). La boîte est le rectangle central, les moustaches sont les 2 traits qui partent au-dessus et au-dessous.

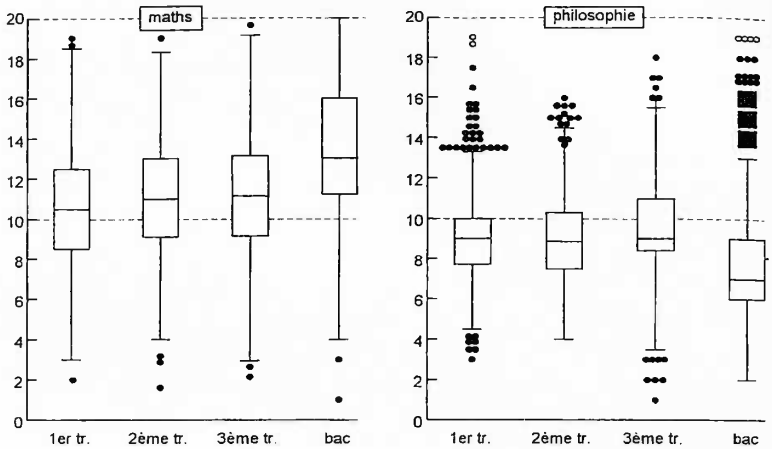


Figure 7.5. *Distributions des 4 notes en maths et en philosophie : boîtes de dispersion. Disque plein (resp. cercle) : individu dont la distance à la boîte est supérieure à 1,5 fois (resp. 2 fois) la hauteur de la boîte ; carré plein : superposition de plus de 10 disques.*

Médiane

La barre dans la boîte indique la valeur de la *médiane*, valeur divisant la population en deux parties égales. Si l'on range les 909 notes par ordre croissant, c'est donc la $455^{ème}$ (avec 908 élèves, effectif pair, la médiane est la moyenne entre les $453^{ème}$ et $454^{ème}$). Pour les mathématiques au premier trimestre, la médiane est 10.5 : la moitié des élèves a donc une note inférieure ou égale à 10.5 et l'autre moitié une note supérieure ou égale à 10.5. C'est un résumé succinct de la distribution du même type que la moyenne (les deux sont des *indicateurs de tendance centrale*) et souvent intéressant. Elle est moins utilisée, l'une des raisons étant qu'il est plus compliqué de trier des valeurs par ordre croissant que de calculer leur moyenne.

Les médianes montrent la même évolution que les moyennes, avec une hausse de 2 points au bac.

Lorsque la distribution est symétrique (exactement la même répartition de chaque côté de la médiane), médiane et moyenne sont confondues. Elles sont ici très proches pour toutes les notes car, comme les histogrammes le montrent, les distributions sont presque symétriques.

Pour les mathématiques au premier trimestre, la médiane est un peu moins élevée que la moyenne (10.5 au lieu de 10.64). En effet, l'histogramme montre que les répartitions des notes extrêmes sont un peu différentes (cf. **Fig. 7.3**) : les notes très élevées sont plus

nombreuses que les notes très faibles, ce qui augmente la moyenne par rapport à la médiane.

Quartiles, intervalle interquartile

Les deux extrémités de la boîte représentent les quartiles ; calculés sur le même principe que la médiane, ils délimitent le premier quart et le dernier quart des notes (supposées rangées par ordre croissant).

En mathématiques au premier trimestre, 226 notes sont inférieures ou égales à 8.5 et 226 notes sont supérieures ou égales à 12.5. La boîte recouvre donc la moitié "centrale" de la population. L'étendue de cette moitié de population (*intervalle interquartile*) est de 4 points : ces notes sont donc assez resserrées.

Pour les trois trimestres, la boîte est assez stable. Au bac, non seulement elle se décale vers les notes supérieures mais elle s'élargit, montrant une dispersion plus grande de la partie centrale de la population. L'écart interquartile, qui est aussi un indicateur de dispersion, montre ici la même tendance que l'écart-type.

Moustaches, outlier, individu hors norme

Il reste quand même la moitié de la population en dehors de l'intervalle délimité par la boîte. Comment est-elle représentée sur ces graphiques ?

Commentons d'abord la note en mathématiques du bac, qui illustre bien notre propos. La boîte va de 11 à 16, soit un écart de 5 points. Le quart supérieur part donc de 16 et va jusqu'à la note maximum observée qui est 20. La "moustache" en s'arrêtant à 20 met en évidence ce maximum. Le quart inférieur s'étend entre 11 et la note minimum observée qui est 1. Ce 1 est à 10 points de la boîte, donc de la partie centrale des notes. C'est très loin ! Si loin qu'il est considéré comme extraordinaire ou "hors norme" et qu'on lui donne souvent le nom d'*outlier* (terme anglais pratiquement passé dans la langue française ; on rencontre les équivalents français *élément aberrant* ou *élément suspect* dont aucun n'a réussi à s'imposer). Un élément très loin des autres est curieux et, par là même, intéressant.

Ce sujet est si important que la seconde partie de cet ouvrage est consacrée à la recherche et l'étude des outliers. Disons simplement ici qu'ils peuvent provenir d'erreurs : si nous n'avions pas corrigé le 22 en sciences naturelles, son statut d'outlier aurait attiré l'attention et aurait conduit à des vérifications. Même s'ils ne proviennent pas d'erreurs, ils sont quelquefois si différents des autres qu'il est alors raisonnable de les écarter lors d'une étude statistique.

Les outliers n'ont pas de définition précise ; ce sont simplement des éléments loin des autres qu'il est utile de repérer d'une façon ou d'une autre. Le critère choisi ici est la distance à la boîte. Mais comment dire qu'une distance est grande ? Il ne peut y avoir de valeur fixe : il faut tenir compte de la dispersion générale de la population. Classiquement, on choisit comme échelle de dispersion la longueur de la boîte (5 points pour le bac) et déclare outlier tout individu situé à plus d'une fois et demie cette longueur. Ce qui, toujours pour le bac, donne ce statut aux notes inférieures à 3.5 et aux notes supérieures à 23.5. Deux notes faibles, 1 et 3, répondent à cette définition. Elles sont mises en évidence sur le graphique car la moustache s'arrête avant eux. Les extrémités des moustaches indiquent le maximum et le minimum de la population hors outliers.

Le critère « d'une fois et demi » découle d'un certain arbitraire, même si l'ordre de grandeur est raisonnable : si on le modifie un tant soit peu, le statut du 3 risque de changer.

On sépare d'ailleurs quelquefois les "fortement outliers" (à plus de deux fois la longueur de la boîte) des autres en les représentant avec un autre symbole. Cette instabilité n'est pas très grave. **Le statut d'outlier ne doit jamais impliquer une attitude systématique, seulement un examen plus attentif de ces individus** (attention : pour certains le terme *outlier* est employé dans le sens plus restrictif de *données devant être écartées*).

Le 1 se repère aussi sur le diagramme en bâtons car non seulement il est éloigné de la majorité de la population mais, en plus, il est séparé des autres par un « *trou dans la distribution* » (personne n'a 2). La présence d'un « trou » séparant une valeur extrême des autres, surtout lorsqu'il est très important (ce qui n'est pas le cas ici), accentue le caractère particulier de la valeur.

Un autre critère simple et très efficace pour détecter des outliers est l'écart à la moyenne mesuré en nombre d'écarts-types, autrement dit la valeur centrée-réduite. On peut déclarer outliers les éléments situés, par exemple, à plus de 3 écarts-types de la moyenne. Ceci permet de traiter de façon homogène l'ensemble des variables. Dans le cas des mathématiques au bac, ces limites sont 3.6 et 22.80 et diffèrent peu des bornes proposées plus haut.

Pour les autres notes, il existe aussi des outliers repérés par un critère ou un autre. Mais, dans ces données, aucun ne révèle un problème particulier : ce sont simplement des notes particulièrement extrêmes que l'on n'a aucune raison d'écarter.

Cas de la philosophie

En comparaison avec celles des mathématiques, ces distributions sont caractérisées par des boîtes de taille restreinte et un grand nombre d'outliers, c'est-à-dire à la fois par une partie centrale très concentrée et la présence de queues de distribution (individus loin de la partie centrale). Ceci peut bien sûr être constaté sur les histogrammes (cf. **Fig. 7.6**). Des indicateurs pour détecter et quantifier la présence de queue(s) de distribution sont présentés chapitre 14.

Médiane et ex æquo

La médiane de la note au bac en maths (13) ne sépare pas réellement la population en deux parties égales. En effet, il y a 364 notes strictement inférieures à 13 et 453 strictement supérieures à 13. Pourtant la médiane est bien de 13, mais 92 notes sont égales à 13. Il y a donc beaucoup d'ex æquo. Pour séparer la population en deux parties égales, il faut considérer 89 notes comme inférieures et 2 comme supérieures. Ce n'est pas très logique, mais ne serait pas gênant si la médiane représentait bien la tendance générale. Un des critères de bonne représentation d'un indicateur de tendance centrale est la *stabilité de cet indicateur pour de petites perturbations des données*. Or, si les correcteurs avaient été un plus généreux avec seulement 3 élèves (sur près de 1000), la médiane aurait augmenté d'un point, ce qui est important. La médiane est donc peu adaptée aux cas de nombreux ex æquo. Le problème ne peut qu'empirer pour les quartiles. La représentation par boîte est donc bien moins adaptée pour la note du bac, variable discrète, que pour les notes de l'année, variables continues.

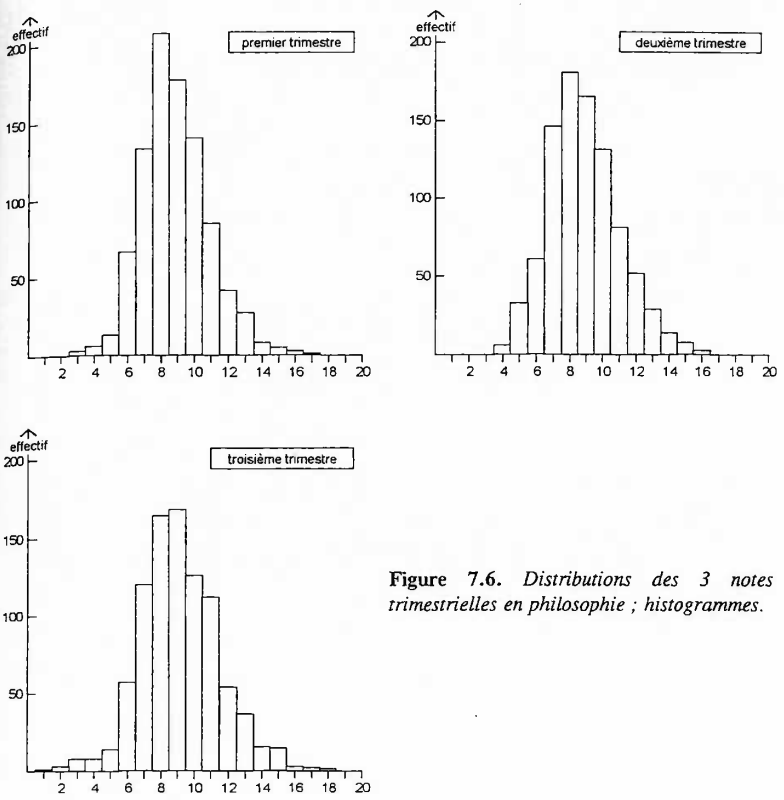


Figure 7.6. *Distributions des 3 notes trimestrielles en philosophie ; histogrammes.*

7.8 Pourcentages par rapport à des valeurs de référence

Pour les notes, 10 est une valeur de référence qui s'impose puisque c'est le minimum exigé pour le diplôme. Le pourcentage de notes inférieures à 10 est un indicateur au moins aussi parlant que la moyenne ou la médiane pour traduire l'adaptation du niveau de l'épreuve.

Restreignons-nous aux notes du bac. En mathématiques, la grande majorité des élèves (87%) atteint au moins 10. En physique, en sciences naturelles et en histoire-géographie, le pourcentage reste assez élevé (65%, 71% et 57%) : nettement plus de la moitié des élèves ont au moins la moyenne. Ce pourcentage est beaucoup plus faible en philosophie (24%).

Le 10 est une valeur de référence « naturelle », mais on peut en choisir d'autres. Si l'on veut examiner les pourcentages de très bonnes notes et de très mauvaises notes, on peut prendre aussi, par exemple, 6 et 14 comme autres valeurs de référence (cf. **Tab. 7.4**).

	$x<6$ mauvais	$6\leq x<10$ faible	$10\leq x<14$ moyen-bon	$14\leq x$ très bon	total
mathématiques	1%	12%	37%	50%	100%
physique	7%	28%	40%	25%	100%
sciences nat.	4%	25%	52%	19%	100%
histoire-géo.	2%	41%	46%	11%	100%
philosophie	25%	51%	17%	7%	100%

Tableau 7.4. *Répartition en 4 classes des notes des 5 matières du bac.*

Commentons les quelques éléments les plus frappants de ce tableau.

Les mauvaises notes sont nettement plus fréquentes en philosophie (25%) que dans les autres matières (au maximum 7%). Elles sont très rares en mathématiques et en histoire-géographie (1% et 2%) et assez rares en sciences naturelles (4%). A l'opposé, les très bonnes notes sont très fréquentes en mathématiques (la moitié des élèves) et rares en philosophie (7%).

Globalement, les notes de mathématiques sont les plus décalées vers le haut et celles de philosophie les plus décalées vers le bas. Ce qui se traduit bien sur les moyennes.

Les très mauvaises notes sont, à l'exception de la philosophie, beaucoup moins nombreuses que les très bonnes. Le record est détenu par les mathématiques avec 50% de très bons et 1% de très mauvais.

L'histoire-géographie suscite moins de notes extrêmes que les autres matières. Ceci était suggéré par les écarts-types (cf. **Tab. 7.3**), légèrement plus faibles en histoire-géographie, mais la différence est bien plus visible ici.

Tableau et regroupement en classes

Ce tableau se déduit facilement des valeurs permettant de tracer l'histogramme. On a simplement regroupé la population en classes plus vastes : d'une vingtaine de classes utilisées pour tracer les histogrammes, on est passé à 4 classes seulement. Il contient donc moins d'information notamment sur la forme de la distribution.

Mais c'est un résumé synthétique assez complet et facile à interpréter. Le problème est quelquefois de *choisir de bonnes valeurs de référence*. Si, comme ici, on étudie simultanément plusieurs variables s'exprimant dans les mêmes unités, les bornes doivent être communes. Quand pour une variable aucun ensemble de valeurs ne s'impose a priori, l'examen de l'histogramme assez détaillé peut en suggérer : si il y a des trous dans la distribution, il est judicieux de choisir des bornes dans ces trous. Si rien ne s'impose, ni a priori ni au vu de l'histogramme, il est recommandé de choisir des intervalles donnant des effectifs analogues, car c'est ainsi que l'information conservée est la plus importante. Notons que si l'on prend deux ou quatre classes et que l'on suit strictement cette règle d'équilibre, on obtient les valeurs données par la médiane et les quartiles. Mais pour un bon résumé, un choix adapté des valeurs est souvent beaucoup plus pertinent. Le choix de ces valeurs peut d'ailleurs être l'aboutissement d'une étude statistique.

7.9 Que choisir ?

Les indices (moyenne, écart-type, maximum, minimum, médiane, quartile), les graphiques (histogramme, boîte de dispersion) et les tableaux de pourcentages visent tous à appréhender

une distribution. Est-il nécessaire de les mettre en œuvre tous pour chaque variable de chaque étude ? Si non, comment choisir entre eux ? Et nous n'avons donné que les méthodes les plus simples et les plus courantes : le nombre d'indices possibles est très grand.

Tout dépend de la nature des données, des questions posées, du temps que l'on veut consacrer, des outils disponibles et des interlocuteurs auxquels on s'adresse.

Il est souvent utile, dans la phase d'analyse, de regarder les données sous plusieurs aspects afin notamment de repérer les éléments les plus saillants (comme les outliers) en utilisant des outils divers. Dans la rédaction finale, on ne conserve que les tendances générales et les résultats intéressants ou inattendus, en choisissant les outils et les graphiques qui les mettent le mieux en évidence.

Ainsi, pour répondre à la première question posée dans les objectifs, "Comment se répartissent les notes, dans les différentes matières et pendant l'année ?", l'étude détaillée exposée ci-dessus peut donner lieu à une synthèse dont nous proposons plus loin un exemple. Les distributions des notes étant assez régulières, les moyennes et les écarts-types (bien mieux adaptés, pour les notes entières du bac, que les médianes et quartiles) les résument bien et il n'y a pas d'anomalie particulière. Mais on ne le sait qu'après une étude détaillée qui ne transparaît pas forcément dans la synthèse ! On peut compléter ces indices par les pourcentages de notes inférieures à 10, qui montrent des nuances, et par le pourcentage très élevé de notes de philosophie inférieures à 6.

Mais avant de faire cette synthèse, il faut apprécier dans quelle mesure l'élimination des élèves ayant des données manquantes influence les résultats.

7.10 Influence des groupes d'élèves ayant des données manquantes

Sans parler des fantômes, trois groupes d'élèves ont été successivement écartés ; nous pouvons maintenant examiner si les distributions des notes telles que nous les résumons varient sensiblement selon que l'on considère ou non ces groupes d'élèves.

Nous nous limitons à examiner si des modifications notables des moyennes et des écarts-types apparaissent lorsque l'on ajoute les trois groupes. On se contente de regarder les notes du bac. Ce sont les plus significatives et de plus, pour les autres notes, les quelques données manquantes éparpillées continueraient à causer quelques soucis pour ces calculs. Il sera toujours temps de s'intéresser de plus près à une note annuelle si l'on remarque des variations notables pour les notes du bac.

Les moyennes des 5 notes du bac ont été calculées (cf. **Tab. 7.5**) en ajoutant successivement le groupe des 38 élèves ayant quelques données manquantes, celui des 13 sans notes trimestrielles et enfin les 15 qui ont un bac incomplet (cf. **Tab. 4.3** pour la définition des groupes d'élèves).

On ne constate pas de modifications notables. Toutes les moyennes baissent un peu lorsqu'un groupe est ajouté, ce qui montre que les groupes d'élèves présentant des données manquantes ont des moyennes plus faibles que la moyenne générale dans toutes les matières. Mais ces variations, qui ne sont que de quelques centièmes de point, ne remettent absolument pas en cause des commentaires basés sur les 909 individus entièrement complets. Par contre, cette tendance faible mais systématique reflète certainement un phénomène réel.

Population	Maths	Physique	Sciences nat.	Histoire-géo.	Philosophie
1 (909)	13.21	11.00	10.97	10.21	7.84
2 (947)	13.13	10.91	10.93	10.15	7.81
3 (960)	13.11	10.89	10.93	10.13	7.78
4 (975)	13.08	10.87	-	-	-

Tableau 7.5. *Moyennes des notes du bac selon la population étudiée. 1 : les 909 élèves sans données manquantes ; 2 : 1 + les 38 élèves ayant quelques données manquantes ; 3 : 2 + les 13 élèves n'ayant pas de notes trimestrielles ; 4 : 3 + les 15 élèves n'ayant que 2 notes au bac.*

Il n'y a pas de modifications importantes non plus pour les écarts-types (cf. **Tab. 7.6**). Ils ne varient que de quelques centièmes de point. L'écart-type augmente généralement avec le nombre d'élèves sauf en philosophie où il diminue.

Population	Maths	Physique	Sciences nat.	Histoire-géo.	Philosophie
1 (909)	3.19	3.61	2.88	2.65	3.29
2 (947)	3.21	3.63	2.93	2.68	3.26
3 (960)	3.23	3.65	2.95	2.69	3.27
4 (975)	3.24	3.67	-	-	-

Tableau 7.6. *Ecarts-types des notes du bac selon la population étudiée* (*cf.* **Tab. 7.5**).

7.11 Exemple de synthèse

Préambule

L'étude se restreint aux 909 candidats sans note manquante. Trois groupes d'élèves ont été écartés : les 15 qui, au bac, n'ont de notes qu'en mathématiques et en physique, les 13 qui n'ont aucune note trimestrielle et les 38 auxquels manquent quelques notes trimestrielles.

Moyennes (cf. Tab 7.1 et Fig. 7.4).

Durant l'année, dans chaque matière, les moyennes sont assez stables. Au bac, la note de mathématiques dépasse de plus de 2 points les notes trimestrielles alors que celle de philosophie chute de plus d'un point.

Au bac, la différence entre les moyennes des 5 matières est accentuée : les mathématiques avec 13.21 et la philosophie avec 7.84 encadrent les trois autres matières situées entre 10 et 11.

Dispersion (cf. Tab 7.2, 7.3 et 7.4)

La note 0 n'est jamais utilisée ; le 20 n'apparaît qu'en mathématiques au bac (une fois).

La dispersion des notes est un peu plus importante au bac que pendant l'année. C'est en histoire-géographie qu'elle est la plus faible (écart-type de 2) et en physique qu'elle est la plus élevée (écart-type de 3.6).

Hormis la philosophie au bac, les matières littéraires présentent des dispersions moindres que les matières scientifiques.

Au bac, les pourcentages d'élèves ayant une note au moins égale à la moyenne sont très variables suivant les matières : 87% en mathématiques, 65% en physique, 71% en sciences

naturelles contre 57 % en histoire-géographie et seulement 24% en philosophie. Comme pour les moyennes, les mathématiques arrivent en tête et la philosophie en dernier. De ce point de vue, les résultats dans les matières scientifiques sont bien meilleurs que dans les matières littéraires.

Le pourcentage de notes strictement inférieures à 6 confirme ce fait pour la philosophie où il atteint presque 25%, mais pas pour l'histoire-géographie où il n'est que de 1.7%, valeur analogue et même plus faible que celles de certaines matières scientifiques.

Liaison entre deux variables quantitatives : les notes sont-elles liées entre elles ?

8.1 Le problème

L'un des objectifs de l'étude est la comparaison entre les notes de l'année et celles du bac. L'idéal serait de pouvoir comparer globalement les 3 notes de l'année à celle du bac mais ce problème est assez complexe et nous commençons, pour ne manier que deux variables à la fois, par comparer l'une des notes trimestrielles à celle du bac.

Qu'est-ce qu'une liaison entre deux variables ?

On sait déjà, par expérience, qu'il existe un lien entre la note du bac et les notes trimestrielles : la note du bac d'un élève qui a de très bonnes notes trimestrielles est en général meilleure que celle d'un élève qui a eu de très mauvaises notes trimestrielles. Mais cette affirmation est un peu vague et l'on veut préciser cette liaison. Il existe plusieurs manières d'aborder le problème de l'étude de la liaison entre deux variables quantitatives. Pour montrer et comparer leur intérêt, on va confronter sur le même exemple : l'utilisation d'un graphique, la construction d'un tableau croisé à partir de valeurs de référence, le calcul d'un indice (coefficient de corrélation), l'étude de la distribution de la différence entre deux notes et enfin la construction d'un modèle. L'exemple détaillé est celui du couple des notes de mathématiques, au bac et au troisième trimestre. La population de référence est celle des 909 élèves ayant toutes leurs notes.

8.2 Etude graphique de la liaison entre deux variables quantitatives

La méthode de base consiste à construire le nuage de points défini par les deux variables, comme nous l'avons fait pour le petit tableau des 15 élèves (cf. **Fig. 5.2**). On obtient ainsi un nuage de points dans lequel chacun des 909 élèves est représenté par un point dont la coordonnée sur l'axe horizontal est la note de mathématiques au troisième trimestre tandis que la coordonnée sur l'axe vertical est la note de mathématiques du bac (cf. **Fig. 8.1**).

Le nombre d'individus à représenter est très grand. Dans certaines zones les points sont très rapprochés ; certains devraient même se superposer, lorsqu'ils correspondent exactement au même couple de notes ou si ces 2 couples de notes sont trop proches pour se différencier sur le graphique. Ici, les points devant se superposer ont été légèrement écartés de leur position réelle.

On remarque une structure assez particulière : les points se répartissent sur des traits parallèles horizontaux qui correspondent aux valeurs entières, seules notes possibles pour le bac. On voit ainsi très clairement la dispersion des notes du 3ème trimestre pour chaque valeur de la note au bac.

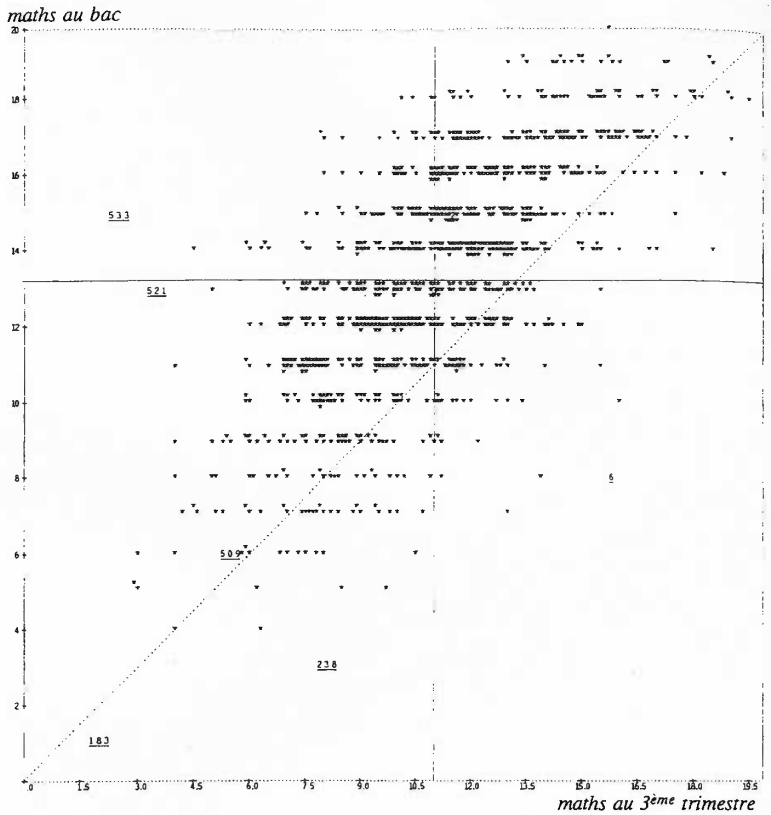


Figure 8.1. *Nuage des 909 élèves défini par les notes en maths du 3ème trimestre et du bac. La bissectrice (en pointillé) correspond aux élèves ayant la même valeur pour les 2 notes. Les lignes continues correspondent aux valeurs moyennes. Quelques élèves remarquables sont repérés par leur numéro d'ordre dans le fichier.*

Forme générale du nuage

Le nuage s'allonge depuis le bas à gauche jusqu'en haut à droite. Une forte note au troisième trimestre implique généralement une forte note au bac et réciproquement, ce qui n'étonne pas vraiment. La liaison entre les deux variables est dite positive. Ce n'est qu'une tendance générale qui ne se vérifie pas pour chaque élève ; pour un 13 au bac par exemple, les notes du 3ème trimestre varient de 3.5 à 16.

Beaucoup de points sont situés au-dessus de la première bissectrice : beaucoup d'élèves ont obtenu des notes meilleures au bac qu'au troisième trimestre. Ceci était déjà apparu lors des comparaisons entre les moyennes et entre les distributions des deux variables.

Distribution d'un couple de variables

Le nuage représente la répartition de la population pour les deux variables simultanément. C'est la *distribution conjointe* des variables ou la distribution du couple de variables. Elle montre par exemple, en haut et à droite du graphique, les élèves qui ont obtenu une bonne note en mathématiques à la fois au 3ème trimestre et le jour du bac.

Couple de variables normales

Schématiquement, si l'on excepte quelques points situés à la périphérie, le contour du nuage présente grossièrement l'allure d'une ellipse. La partie centrale du nuage est plus dense que la partie extérieure. Le nuage qui représente la distribution des 909 élèves pour le couple de variables étudié évoque une *distribution normale bidimensionnelle* (cf. Fiche 9 § 4.4). Nous avons déjà cité (cf. 7.2), pour une variable, le *modèle normal* de distribution associé à une courbe de densité en forme de "cloche". Ici le modèle concerne un couple de variables mais il n'est pas sans lien avec le précédent : chacune des deux variables d'un couple de variables (théoriques) qui a une distribution normale (bidimensionnelle) a elle-même une distribution normale (unidimensionnelle). Mais cela ne suffit pas pour que **le couple** soit distribué « normalement ». Il faut aussi la propriété suivante : sur des ellipses concentriques, la densité du nuage est constante, cette densité décroissant au fur et à mesure que l'on s'éloigne du centre du nuage.

Outlier bidimensionnel

Regardons maintenant les points situés à la périphérie du nuage.

En bas à gauche, l'élève n°183, très mauvais, se détache des autres. Sa note 1 au bac a déjà été remarquée dans l'étude séparée de la distribution du bac et a même été qualifiée d'outlier (cf. 7.7). Pour le troisième trimestre, il est aussi le plus mauvais, mais il se détache moins des autres. Cet élève se remarque sur le graphique du fait de son éloignement par rapport aux autres points ; cet éloignement s'effectue dans la direction générale du nuage car ses deux notes sont tout à fait cohérentes dans leur médiocrité.

Plus étonnants sont les élèves qui s'éloignent « transversalement » de la partie centrale du nuage. Deux élèves, soulignés, sont remarquables au premier coup d'œil.

- L'un (533), en haut et à gauche, obtint 15 au bac après un 2.5 au troisième trimestre. Cela paraît si étonnant que l'on peut se demander s'il ne s'agit pas d'une erreur. Sinon, voilà de quoi donner de l'espoir aux plus mauvais élèves.
- L'autre (6), situé en bas à droite, n'obtint que 8 au bac après un 15.8 au 3ème trimestre. Est-ce une erreur ? Un accident le jour du bac ? Un élève qui provient d'un lycée qui surnote systématiquement ? De toute façon, cet élève suscite des questions.

Ces deux élèves peuvent être qualifiés d'*outliers bidimensionnels* : ils ne se détachent pas du tout des autres pour chacune des notes considérée séparément, mais par une « incohérence » entre leurs deux notes. Le repérage d'outliers bidimensionnels présente le même intérêt que celui des outliers unidimensionnels. Il permet de détecter des erreurs ou des individus très particuliers. Le chapitre 13 contient des compléments sur ce sujet et propose des techniques de recherche systématique d'outliers bidimensionnels.

Conclusion sur l'usage du graphique

Le graphique contient toute l'information sur la liaison entre les deux variables. Son examen est souvent suffisant. Il permet de visualiser la forme générale de la distribution conjointe de deux variables et donc les grandes tendances de leur liaison. Il permet aussi de

repérer des outliers, éléments particuliers qui s'écartent des autres. Cette description qualitative sera complétée par le coefficient de corrélation (cf. 8.4).

8.3 Tableau croisé à partir de valeurs de références

Dans l'étude d'un couple de variables, il est souvent intéressant de construire un tableau croisé (cf. **Tab. 8.1**).

Pour cela, lorsque les variables sont quantitatives, ce qui est le cas ici, le domaine de variation de chacune doit au préalable être divisé en classes. Le choix du nombre de classes et des intervalles correspondants découle toujours d'un certain arbitraire (ce problème a déjà été rencontré en 7.8). Ici, les deux variables étant de même nature, nous avons fixé le même nombre de classes (5) et les mêmes intervalles pour les deux variables. Chaque variable a donc 5 niveaux (comme dans le système de notation littérale où les élèves sont notés par A, B, C, D ou E). Les notes du bac étant entières, on a choisi des intervalles qui ne s'arrêtent pas à ces valeurs entières mais qui les encadrent largement. Nous avons choisi des intervalles de longueur 4, sauf le premier qui est un peu plus grand (les très faibles notes sont rares) et le dernier qui est un peu plus petit. En définitive, les 5 niveaux correspondent aux notes de la façon suivante :

très mauvais : $x \leq 4.5$ (soit pour les notes -entières- du bac : 0, 1, 2, 3 ou 4)
mauvais : $4.5 < x \leq 8.5$
moyen : $8.5 < x \leq 12.5$
bon : $12.5 < x \leq 16.5$
excellent : $16.5 < x$

On peut ainsi construire un tableau (cf. **Tab. 8.1**) dont chaque ligne correspond à un intervalle de la note du bac et chaque colonne à un intervalle de la note du troisième trimestre. Par exemple, au croisement de la ligne *bac bon* et de la colonne *3ème trimestre mauvais*, 33 est le nombre d'élèves qui ont eu à la fois, en mathématiques :

- une bonne note au bac, comprise entre 13 et 16 (inclus) ;
- une mauvaise note au troisième trimestre, comprise entre 4.6 (inclus) et 8.6 (exclu).

		Troisième trimestre					
		très mauvais	mauvais	moyen	bon	excellent	total
B	excellent	0	3	40	86	25	154
A	bon	3	33	221	125	9	391
C	moyen	2	105	155	26	0	288
	mauvais	6	44	19	3	0	72
	très mauvais	2	2	0	0	0	4
	total	13	187	435	240	34	909

Tableau 8.1. *Notes de mathématiques au bac (en lignes) et du troisième trimestre (en colonnes) : tableau croisé des variables divisées en classes.*

Tableau croisé, de contingence, de fréquence

Ce type de tableau est communément appelé tableau croisé (il croise deux variables), tableau de contingence ou même tableau de fréquence(s). Son emploi est systématique pour étudier la liaison entre deux variables qualitatives. Il est utilisé ici pour des variables

quantitatives codées en classes (donc transformées en variables qualitatives). Dans un tel cas, on l'appelle quelquefois tableau de corrélation.

Tableau et histogramme à deux dimensions

Pour synthétiser la distribution d'une seule variable, on la découpe en classes et on trace un histogramme. Pour construire le tableau croisé, on découpe de la même façon chacune des deux variables. Des graphiques de visualisation d'un tableau croisé, dits *histogrammes à deux dimensions* ou *stéréogrammes*, sont possibles, mais il faut bien avouer que leurs qualités sont bien souvent plus esthétiques qu'efficaces (cf. **Fig. 8.2**).

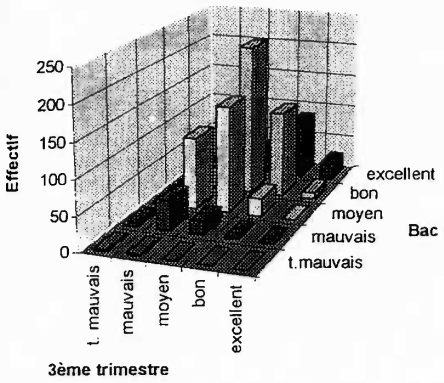


Figure 8.2. *Notes de mathématiques au bac et au troisième trimestre : histogramme à deux dimensions (ou stéréogramme).*

Tableau et graphique

Le tableau **8.1** est une synthèse du graphique **8.1**. Nous avons tracé sur ce graphique la "grille" correspondant aux limites des classes (cf. **Fig. 8.5**). On peut voir, par exemple, dans le carreau en bas à gauche les 2 points définis par les 2 toujours très mauvais élèves et dans le carreau à côté les 2 élèves qui n'étaient que mauvais au troisième trimestre et deviennent très mauvais au bac.

Outliers bidimensionnels

Certaines cases du tableau contiennent des nombres très faibles : ainsi, au croisement de *3ème trimestre très mauvais* et de *bac bon*, on ne trouve que 3 élèves. Si on se reporte à l'étude graphique, on voit que l'un de ces trois élèves est l'outlier déjà repéré, les deux autres étant un peu moins éloignés du nuage. Le tableau met donc moins en évidence les outliers que le graphique puisque l'information est condensée (tous les points d'un carreau sont confondus dans le tableau). Mais comme l'information est plus synthétique, il peut servir de guide dans l'étude de ce dernier. On pourrait examiner plus attentivement les croisements rares ou curieux.

Marges du tableau croisé

La colonne *total* contient la somme des effectifs de chaque ligne. On y trouve donc la distribution de la variable *note du bac*, codée en cinq classes. C'est la *distribution marginale* de cette variable appelée aussi plus simplement *marge*.

De même la ligne *total* contient la somme des effectifs de chaque colonne : c'est la distribution marginale de la variable *note au troisième trimestre*.

Le commentaire d'un tableau croisé commence souvent par celui de ses deux marges. Ici, on peut noter qu'au bac le nombre de *moyens*, *bons* et *excellents* est très important. Au troisième trimestre il y a un décalage : ce sont les *mauvais*, *moyens* et *bons* qui dominent.

Etude du tableau croisé à travers ses colonnes ; distribution conditionnelle

Dans la lecture d'un tableau croisé on privilégie souvent, suivant les objectifs, soit la structure en lignes soit la structure en colonnes. On peut commencer ici la lecture du tableau à travers ses colonnes, ce qui revient à regarder ce que donnent au bac les élèves de chacun des niveaux du troisième trimestre (c'est la *distribution conditionnelle* des notes du bac, le niveau du troisième trimestre étant fixé).

- Les 3 zéros en bas de la dernière colonne indiquent qu'avec un troisième trimestre excellent on obtient un bon ou un excellent résultat au bac.
- L'avant-dernière colonne concerne les bons : ils restent généralement bons (125), beaucoup (86) deviennent excellents, quelques-uns chutent pour devenir moyens et 3 isolés deviennent même mauvais.
- Les élèves moyens au 3ème trimestre ont plus tendance que les bons à améliorer leur note : non seulement ils peuvent grimper de 2 échelons mais ils sont nettement plus nombreux à augmenter leur score qu'à rester stables.
- Le phénomène est plus net pour les mauvais et encore plus pour les très mauvais. Il semble, au-delà des effets de bornes (avec une telle grille, les plus mauvais ne peuvent empirer et les meilleurs ne peuvent s'améliorer), que l'amélioration est d'autant plus marquée que l'on part de bas.
- La première colonne, qui concerne les très mauvais au troisième trimestre, est beaucoup plus éparpillée que la dernière : on observe plus fréquemment des très mauvais qui progressent que des excellents qui chutent.

Ces tendances se lisent directement sur le graphique. Mais cela est plus facile après les avoir repérées sur le tableau croisé où elles sont chiffrées. Le chiffrage facilite aussi la communication des résultats et les comparaisons.

On a souvent avantage à commenter des pourcentages, ici calculés sur chaque colonne, plus parlants et surtout plus comparables entre eux que les chiffres bruts (cf. **Tab. 8.2**).

		Troisième trimestre				
		très mauvais	mauvais	moyen	bon	excellent
B	excellent	0%	1.6%	9.2%	35.8%	73.5%
A	bon	23.1%	17.6%	50.8%	52.1%	26.5%
C	moyen	15.4%	56.1%	35.6%	10.8%	0%
	mauvais	46.2%	23.5%	4.4%	1.3%	0%
	très mauvais	15.4%	1.1%	0%	0%	0%
	total	100%	100%	100%	100%	100%
	effectif	13	187	435	240	34

Tableau 8.2. *Tableau 8.1 transformé en "pourcentages-colonnes" : une colonne donne la répartition de la note au bac pour un niveau donné de la note du troisième trimestre.*

Ce tableau permet de préciser légèrement le commentaire ci-dessus en disant par exemple qu'au bac, parmi les 435 élèves moyens au troisième trimestre, plus de la moitié deviennent bons, 35.6% restent moyens, 9.2% sont excellents et 4.4% ont obtenu de mauvais résultats. On peut ainsi plus facilement faire des comparaisons, en disant par exemple que le passage au niveau immédiatement supérieur concerne 35.8% des bons, 50.8% des moyens et 56.1% des mauvais.

Etude du tableau croisé à travers ses lignes

Le tableau **8.1** peut aussi se lire à travers ses lignes. C'est un point de vue complémentaire du précédent. On regarde alors, pour chaque niveau du bac, où se situaient les élèves au troisième trimestre (distribution conditionnelle des notes du troisième trimestre, le niveau du bac étant fixé). On procède exactement de la même façon que pour les colonnes en calculant cette fois les pourcentages sur chaque ligne (cf. **Tab. 8.3**). Nous ne commentons pas ce tableau, laissant au lecteur, s'il le désire, le soin de le faire. Notons simplement un fait rassurant : la dernière ligne se terminant par 3 zéros, une très mauvaise note de mathématiques au bac ne suit jamais une note correcte au troisième trimestre. Dans les commentaires, il ne faut cependant pas oublier les effectifs totaux donnés par la marge. Des pourcentages calculés sur 4 élèves n'ont pas la même signification que ceux calculés sur 391 et ne sont commentés qu'à titre indicatif voire anecdotique.

		Troisième trimestre						
		t. mauvais	mauvais	moyen	bon	excellent	total	effectif
B	excellent	0%	1.9%	26.0%	55.8%	16.2%	100%	154
A	bon	0.8%	8.4%	56.5%	32.0%	2.3%	100%	391
C	moyen	0.7%	36.5%	53.8%	9.0%	0%	100%	288
	mauvais	8.3%	61.1%	26.4%	4.2%	0%	100%	72
	très mauvais	50.0%	50.0%	0%	0%	0%	100%	4

Tableau 8.3. *Tableau 8.1 transformé en "pourcentages-lignes" : une ligne donne la répartition de la note du troisième trimestre pour un niveau donné de la note au bac.*

Conclusion sur l'usage du tableau croisé

Fondamentalement, le tableau croisé est conçu pour étudier la liaison entre deux variables qualitatives. Il peut néanmoins être utile, même avec des variables quantitatives, en découpant en classes les intervalles de variation.

Il permet alors de mettre en évidence et de quantifier des phénomènes divers, ce qui la plupart du temps contrebalance largement la perte d'information due au regroupement en classes. Utilisé conjointement au graphique, il complète l'étude détaillée de la liaison entre deux variables. Les tableaux des pourcentages en lignes ou en colonnes facilitent le commentaire.

8.4 Coefficient de corrélation

Dans les deux paragraphes précédents, nous n'avons étudié qu'un couple de variables. Pour les étudier tous, le travail est important. De plus, si l'on veut comparer les liaisons entre couples de variables, ni le tableau croisé ni a fortiori le graphique ne sont très pratiques. Pour cela, il est nécessaire de disposer d'un indice numérique qui mesure la liaison. Un seul nombre résume forcément très mal une situation complexe, mais fournit très rapidement une

idée générale et est extrêmement utile pour comparer plusieurs situations. Cette idée vaut aussi pour les indices synthétiques que sont la moyenne et l'écart-type.

Dans l'étude de la liaison entre deux variables quantitatives x et y, la mesure utilisée habituellement est le coefficient de corrélation linéaire, appelé souvent simplement coefficient de corrélation ou même corrélation, bien que l'adjectif linéaire soit important. On l'obtient directement sur toutes les calculettes ayant des fonctions statistiques (cf. Fiche 5 pour des compléments sur cet indice).

Le coefficient de corrélation entre deux variables x et y est souvent noté $r(x,y)$. C'est un nombre compris entre -1 et $+1$. Il est calculé à partir des données centrées-réduites et est donc indépendant des moyennes et des écarts-types de x et de y. Soit, en notant x_i et y_i les valeurs centrées-réduites pour l'individu i :

$$r(x,y) = \frac{1}{I}\sum_i x_i y_i$$

Dans notre étude, il est intéressant de calculer les coefficients de corrélation, pour chaque matière, entre la note du bac et les notes trimestrielles (cf. **Tab. 8.4**).

	1er trimestre	2ème trimestre	3ème trimestre	3 trimestres
Mathématiques	0.579	0.663	0.678	0.698
Physique	0.627	0.711	0.721	0.755
Sciences naturelles	0.423	0.425	0.429	0.508
Histoire-géographie	0.422	0.451	0.464	0.516
Philosophie	0.419	0.434	0.474	0.508
Bac continu	0.754	0.813	0.834	0.841

Tableau 8.4. *Coefficients de corrélation, pour chaque matière, entre la note du bac et les notes trimestrielles ; .579 = r(maths bac, maths 1er trimestre).*
3 trimestres : note moyenne des 3 trimestres ; bac continu : moyenne pondérée, avec les coefficients du bac, des 5 matières.

Que peut-on dire de ces nombres ?

Interprétation du signe du coefficient de corrélation

Tout d'abord ces corrélations sont toutes positives : cela signifie que, dans chaque matière, les élèves qui ont une note élevée au bac ont souvent aussi une note trimestrielle élevée, et ceux qui ont une note faible au bac ont souvent aussi une note trimestrielle faible (cf. Fiche 5). Ce n'est pas systématique pour tous les élèves, mais c'est une tendance générale. Ceci n'est pas pour étonner. Un coefficient négatif, montrant par exemple que les meilleurs élèves au troisième trimestre obtiennent plutôt les plus mauvaises notes au bac, aurait été pour le moins surprenant. Les 24 nuages de points définis par les 24 couples de variables du tableau **8.4** sont donc (plus ou moins) allongés depuis le bas à gauche jusqu'en haut à droite.

Interprétation de la valeur absolue du coefficient de corrélation

Plus la corrélation s'éloigne de 0, autrement dit plus elle est grande en valeur absolue, plus le nuage est allongé. Le cas (théorique) extrême est celui où la corrélation vaut ± 1 ; tous les points se situent alors sur une droite. Si c'était le cas pour le couple *maths bac* et *maths*

troisième trimestre, il existerait alors deux nombres a et b tels que pour tous les élèves, on ait :

$$Maths\ bac = a\ (Maths\ 3^{ème}\ trim.) + b$$

Par exemple, si a valait .8 et si b valait 2, il suffirait de multiplier la note trimestrielle par .8 et d'ajouter 2 pour obtenir celle du bac. Ceci impliquerait notamment des classements des élèves identiques pour les deux notes. Une telle relation entre deux variables s'appelle une *liaison linéaire*.

Ici, bien évidemment, aucun coefficient n'atteint 1 ; mais doit-on les considérer comme grands (en valeur absolue) ou petits ? Autrement dit, les notes du bac sont-elles très liées aux notes trimestrielles ou non ? Ou encore, les nuages sont-ils très allongés ou non ?

Il est toujours difficile de répondre à ce type de question : à partir de quelle valeur va-t-on considérer qu'une corrélation est grande ? Si les corrélations valaient .96 ou .98, on n'hésiterait pas à affirmer que les variables sont très liées ; mais ici elles varient entre .419 et .841. En fait, tout dépend de la référence implicite que l'on prend : si l'on s'attend à des notes très liées (par exemple parce que cela a été observé dans des études similaires), on dira que ces coefficients sont faibles ; à l'opposé, si l'on s'attend à des notes indépendantes (et donc a fortiori non corrélées), on dira que ces coefficients sont élevés.

Pour avoir une idée de l'allongement du nuage, donc de l'intensité de la liaison linéaire traduite par un coefficient de corrélation, le mieux est d'en regarder un ! Par exemple le nuage défini par le croisement *Maths bac* × *Maths 3ème trim.* (cf. **Fig. 8.1**) correspond à une corrélation de .678 : l'association entre notes de même niveau est nette mais le nuage ne peut vraiment pas être assimilé à une droite.

Alors qu'il est difficile d'affirmer qu'une corrélation est grande ou petite, par contre, il est facile de dire si certaines sont plus grandes que d'autres ! Nous commentons ci-après les différences entre les coefficients de corrélation du tableau **8.4** :

- Dans chaque matière, la corrélation avec la note au bac croît légèrement du premier au troisième trimestre : les notes de fin d'année sont plus liées à celles du bac que celles du début d'année. Pour certaines matières les différences entre coefficients sont très faibles, mais comme elles vont toujours dans le même sens et que cette évolution est interprétable, il ne s'agit certainement pas d'un hasard. Là encore, pour donner des conclusions sur les données, on ne se contente pas de l'observation statistique d'une seule différence (qui ne permettrait sans doute pas de conclure) ; on analyse l'ensemble des données et l'on mobilise les connaissances externes aux données. On peut remarquer aussi que l'augmentation de la corrélation est nettement plus marquée en mathématiques et en physique que dans les 3 autres matières. C'est donc dans les deux matières les plus scientifiques que l'évolution vers la note finale est la plus nette.

- C'est en physique que la liaison entre les notes trimestrielles et celles du bac est la plus forte, elle atteint .721 au dernier trimestre. Les mathématiques suivent de peu. Les 3 autres matières, avec leurs corrélations qui tournent autour de .45, réservent beaucoup plus de surprises au bac ! Nous laissons à chacun la liberté de trouver des explications à ce phénomène marqué.

- Pour chaque matière, la note moyenne des 3 trimestres est plus liée à la note du bac que celle de n'importe quel trimestre. Remarquons au passage que $r(Y,X+Z)$ n'est pas la moyenne entre $r(X,Y)$ et $r(X,Z)$ et peut même ne pas être compris entre ces deux valeurs.

Ici, ce fait est à relier à l'intuition selon laquelle une moyenne de notes représente mieux le niveau d'un élève qu'une note isolée.

- Ce dernier phénomène s'observe aussi lorsque l'on considère les moyennes, pondérées par les coefficients du bac, des notes trimestrielles. Ces notes préfigurent ce que pourrait être un bac attribué non pas à partir d'un examen mais d'un contrôle continu. La moyenne effectuée sur les 3 trimestres est liée assez étroitement à la note du bac (.841). Des résultats de ce type peuvent alimenter le débat sur la suppression du bac en tant qu'examen.

Limites du coefficient de corrélation

Attention : s'il y avait par exemple, entre la note du bac et celle du troisième trimestre, une relation du type :

$$Maths\ bac = a\ (Maths\ 3^{ème}\ trim.)^2 + b$$

la liaison entre les deux notes serait très forte puisque l'on pourrait calculer l'une connaissant l'autre. Mais cette liaison n'est pas linéaire : la note de mathématiques est élevée au carré et le nuage ne serait pas sur une droite mais sur une portion de parabole. Dans certains de ces cas, le coefficient de corrélation peut valoir 0 (cela dépend des valeurs de a et de b, sachant que les valeurs observées varient entre 0 et 20) puisqu'il mesure la ressemblance entre des données et une situation de liaison linéaire.

Cette limitation ne concerne pratiquement pas notre exemple, dans lequel tous les nuages ont des formes apparentées à des ellipses. Une mesure de linéarité est tout à fait adaptée : autrement dit, il ne s'impose pas de chercher à mesurer la ressemblance entre nos données et une situation de liaison plus complexe qu'une droite (plus généralement, on montre que si deux variables sont distribuées selon une loi normale bidimensionnelle leur éventuelle liaison ne peut être que linéaire).

Dans notre cas, les effectifs des individus sont très importants et la corrélation est une mesure assez fiable de l'allongement du nuage ; mais lorsque l'effectif est petit, on peut craindre qu'un ou deux individus assez différents des autres n'infléchissent grandement les résultats. On risque alors, en commentant le coefficient de corrélation sans consulter le graphique, de considérer comme un phénomène général ce qui n'est le fait que de quelques cas particuliers.

Ici, nous sommes donc dans un cas où, vu la forme particulièrement régulière des nuages (proche du modèle *normal*) et vu l'effectif important, le coefficient de corrélation résume assez bien la liaison entre deux variables.

8.5 Distribution de la différence entre deux variables

Il existe bien d'autres moyens pour comparer deux variables. Toujours pour le même couple *Maths bac* et *Maths 3ème trim.*, nous avons construit une nouvelle variable en calculant leur différence. Cette nouvelle variable a un sens concret : c'est l'amélioration (ou la détérioration) du résultat de chaque élève entre le troisième trimestre et l'épreuve finale.

Insistons sur le fait que nous exploitons ici une particularité des données. Si les deux variables à comparer étaient de natures différentes avec des unités sans commune mesure (comme un poids et une taille, une vitesse et un prix, etc.), cela n'aurait guère de sens de considérer leur différence.

Pour décrire la distribution de cette nouvelle variable, nous utilisons les moyens les plus appropriés ici parmi ceux présentés dans le chapitre 7, en particulier l'histogramme (cf. **Fig. 8.3.a** et **b**).

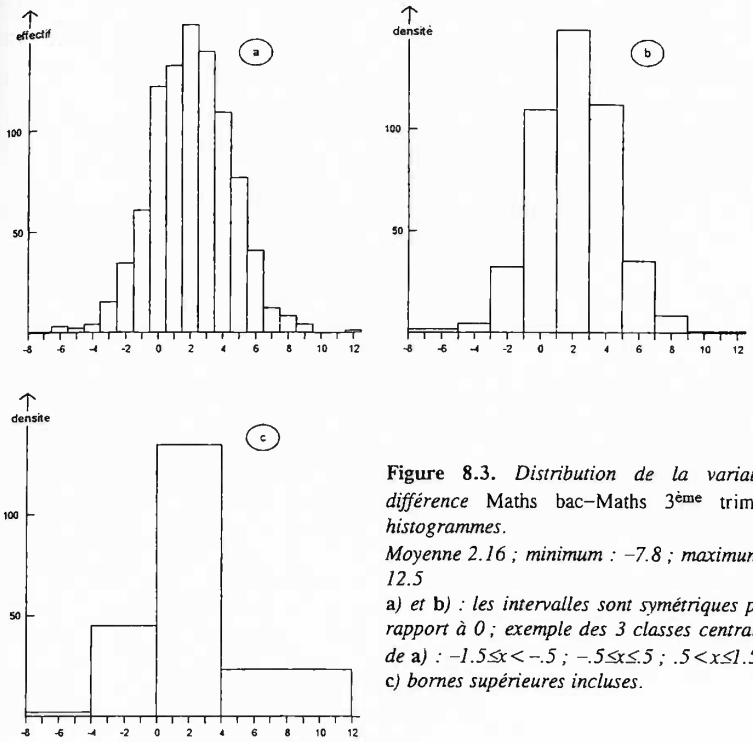


Figure 8.3. *Distribution de la variable différence* Maths bac–Maths 3ème trim. *: histogrammes.*
Moyenne 2.16 ; minimum : –7.8 ; maximum : 12.5
a) *et* **b)** *: les intervalles sont symétriques par rapport à 0 ; exemple des 3 classes centrales de* **a)** *: $-1.5 \leq x < -.5$; $-.5 \leq x \leq .5$; $.5 < x \leq 1.5$;*
c) *bornes supérieures incluses.*

A) Les valeurs extrêmes sont étonnamment élevées : 12.5 et –7.8. La première montre qu'il ne faut pas se décourager devant des résultats trimestriels médiocres : tout est encore possible. La seconde montre que des accidents sont toujours à craindre. A moins qu'il ne s'agisse d'erreurs de saisie !
Mais l'étude de la différence ne permet pas de situer ces élèves. On peut le faire en retournant au graphique 8.1 dans lequel la différence entre les deux notes d'un élève est proportionnelle à sa distance à la première bissectrice (le vérifier par exemple en identifiant les élèves ayant une différence de 1 ou 2 points). Le 12.5 correspond à l'outlier déjà remarqué (élève 533) qui transforme son 2.5 en 15 et le –7.8 à l'autre outlier (6).

B) La moyenne, égale à la différence entre les moyennes des deux notes, est de 2.16.

C) Les histogrammes présentent encore une allure pas très éloignée de celle de la distribution normale, avec beaucoup de valeurs resserrées autour de la moyenne. Pour synthétiser cette distribution, nous avons choisi 3 valeurs de référence : –4, 0 et 4. La

progression des élèves étant générale, pour situer la partie centrale de la distribution, au lieu de prendre un intervalle médian symétrique autour de zéro, nous avons choisi l'intervalle de 0 non compris à 4 compris (cf. **Fig. 8.3.c**) : 59% des élèves ont une amélioration comprise entre 0 et 4 points. Par ailleurs, 20 % montrent une amélioration de plus de 4 points et 20% chutent de moins de 4 points.

8.6 Que choisir ?

Résumons les avantages et les inconvénients de chaque outil.

- Le graphique permet de se faire une idée globale de la liaison et de son type (linéaire ou non) ; il est efficace pour détecter des outliers bidimensionnels.
- Le coefficient de corrélation est très synthétique mais inadapté si la liaison n'est pas linéaire. Il est très utile pour comparer des liaisons linéaires entre elles.
- Le tableau croisé a les défauts de ses qualités : plus complet que le coefficient de corrélation, il est aussi plus complexe à utiliser. Il permet de repérer des anomalies ou des tendances différentes suivant le niveau de chaque note. Il est moins précis que le graphique mais ses effectifs et ses pourcentages sont souvent des indicateurs évocateurs.
- L'étude de la différence entre deux variables n'est possible que pour des variables exactement de même nature. Elle est très parlante dans notre cas, plus synthétique mais moins riche que le tableau croisé. On perd la vision des tendances différentes suivant les niveaux des notes : par exemple, on ne voit pas que les très mauvais au troisième trimestre sont beaucoup plus dispersés au bac que les excellents.

Le choix d'une méthodologie dépend donc de la nature des liaisons, du nombre de variables étudiées et des objectifs. Et, comme nous l'avons déjà dit lors de l'étude d'une seule variable, les outils utilisés dans la phase d'analyse sont beaucoup plus nombreux que ceux qui apparaissent dans la synthèse finale.

Dans notre cas, les couples étudiés sont nombreux et il est intéressant de comparer leurs liaisons. Une méthodologie possible consiste à regarder d'abord rapidement les graphiques. Comme la forme elliptique des nuages permet de considérer que le coefficient de corrélation est un bon indice, ce sont eux que nous retiendrons essentiellement. Mais attention : les coefficients de corrélation ne dépendent ni des moyennes ni des écarts-types. La comparaison est incomplète si on n'évoque pas ces derniers.

Pour pousser plus loin l'étude de certains couples de variables, on peut regarder la distribution de leur différence et présenter quelques tableaux croisés typiques ou remarquables. Si on s'intéresse particulièrement aux outliers, on peut les chercher sur les graphiques ou bien utiliser des techniques spécifiques (cf. Ch. 13).

8.7 Régression

Bien que la problématique associée à notre fichier n'implique pas de modélisation, il est utile, dans une présentation de l'étude de la liaison entre variables quantitatives, d'évoquer la régression. Une bibliothèque entière pourrait être consacrée aux méthodes de régression ; nous donnons ici quelques points de repère sur ces méthodes, conservant comme exemple d'illustration la liaison entre les notes en mathématiques, au $3^{ème}$ trimestre et au bac.

Si le coefficient de corrélation entre ces deux variables était égal à 1, on aurait une relation du type:

$$Maths\ bac = a \times Maths\ 3^{ème}\ trim. + b$$

Ce n'est pas le cas dans nos données, mais on peut quand même chercher une relation de ce type, « compatible autant que faire se peut » avec les données. Géométriquement, cela revient à chercher une droite qui constitue une image stylisée du nuage de points de la figure **8.1** : on dit que l'on procède à un *ajustement* du nuage par une droite.

Dans quels buts ? On emploie les méthodes de régression principalement dans deux perspectives :

- prévision (ou estimation) ;
- mise en évidence d'un modèle explicatif.

Régression : prévision, estimation

Un élève peut vouloir utiliser une telle formule pour tenter, connaissant sa note au troisième trimestre, de prédire, au moins de façon approchée, la note qu'il obtiendra au bac. On utilise ici le terme *prédire* car à l'issue d'un décalage temporel la note du bac finira par être connue. Dans d'autres circonstances, par exemple lorsque l'on cherche à approcher une donnée manquante, on utilise plutôt le terme *estimer*.

A la limite, si l'approximation est très satisfaisante, on pourrait supprimer l'épreuve du bac et utiliser pour les prochaines années la formule obtenue pour l'année étudiée.

Régression et liaison entre variables

Si les notes du bac et celles du troisième trimestre ne sont pas liées, il ne faut pas espérer pouvoir prédire l'une à partir de l'autre ! Avant de mettre en œuvre une technique de régression, il est indispensable d'examiner le graphique croisant les deux variables.

Régression : la formule

La plupart des calculettes fournissent directement les coefficients a et b « optimaux compte tenu des données ». Ici, on obtient :

$$a = .726\ ;\ b = 5.19$$

Ce qui donne comme formule d'estimation de la note du bac en fonction de celle du troisième trimestre (en mathématiques) :

$$Maths\ bac\ estimé = .726 \times (Maths\ 3^{ème}\ trim.) + 5.19$$

Droite de régression et erreur

Cette formule ne donne pas la note exacte du bac pour tout le monde ! Sur la figure **8.5**, on a tracé la droite (D1) d'équation $y = .726\ x + 5.19$, dite *droite de régression*. Pour les points situés sur cette droite, la formule est exacte. Pour tous les autres, les plus nombreux, elle ne l'est pas. Plus le point est éloigné verticalement de la droite, plus la différence entre la vraie note au bac et la *valeur estimée* est importante. Cette « distance verticale » du point à la droite représente l'*erreur d'estimation*. Le graphique montre que ces erreurs peuvent être importantes. C'est le cas particulièrement pour les deux outliers (533 et 6) identifiés au graphique **8.1**.

Pour l'élève 533 (15 au bac, 2.5 au troisième trimestre), on a :

$$Maths\ bac\ estimé = .726 \times 2.5 + 5.19 = 7.005$$

D'où :

$$erreur = note\ exacte - note\ estimée = 15 - 7.005 = 7.995$$

Pour l'élève 6 (8 au bac, 15.8 au troisième trimestre), on a :

$$erreur = 8 - 16.66 = -8.66$$

Les erreurs ne sont pas de même signe : la première valeur est sous-estimée tandis que la seconde est surestimée.

Ces deux exemples, même s'ils sont extrêmes, montrent bien que l'utilisation d'une telle formule n'est pas sans risque.

Formules de la régression

Comment calcule-t-on les coefficients a et b de la droite de régression ? Dans une optique de prévision (ou d'estimation) telle que nous l'avons présentée, il faut approcher le plus possible les notes du bac, donc rendre aussi faibles que possible les erreurs d'estimation. Pour rendre ces erreurs faibles sur l'ensemble des individus, la solution de la régression est de calculer les valeurs de a et b qui minimisent la somme des carrés des erreurs. L'élévation au carré évite que les différences de signes contraires ne s'annulent ; par rapport à l'emploi des valeurs absolues, elle conduit à des formules beaucoup plus simples. Un certain nombre de techniques statistiques cherchent ainsi à minimiser des sommes de carrés ; le principe commun à ces méthodes est appelé *méthode des moindres carrés*.

On peut expliciter les coefficients a et b fournis par la méthode des moindres carrés. Avec les notations suivantes :

y, x : variables étudiées au travers de la relation : $y = ax + b$;

$\bar{y}, \bar{x}$: moyennes de y et de x ; s_y, s_x : écarts-types de y et de x ;

$r(x,y)$: coefficient de corrélation entre x et y ;

on montre que :

$$a = r(x,y)\frac{s_y}{s_x}$$

Dans l'exemple où y=*Maths bac* et x=*Maths 3ème trim.* (*cf.* **Tab. 7.3** et **8.4**) :

$$a = 0.678\frac{3.19}{2.98} = 0.726$$

De même, on montre que :

$$b = \bar{y} - a\bar{x}$$

Cette relation exprime que la droite passe par le point moyen (de coordonnées $\bar{x}$ et $\bar{y}$) du nuage. Dans l'exemple (cf. **Tab 7.1**) :

$$b = 13.21 - (0.726 \times 11.05) = 5.19$$

Dans la régression les deux variables ne jouent pas un rôle symétrique

Il est plus naturel d'essayer d'estimer la note du bac à partir de celle du troisième trimestre que l'inverse. Ne serait-ce qu'à cause de la succession temporelle. Néanmoins, bien que le terme de prédiction s'applique alors difficilement, on peut techniquement appliquer exactement la même méthode en inversant les rôles des deux notes ; le but, un peu artificiel, est alors d'estimer la note du $3^{ème}$ trimestre à partir de celle du bac. Tous calculs faits (avec les relations ci-dessus explicitant a et b), la formule obtenue est la suivante :

Maths 3ème trim. estimé = .635 *Maths bac* + 2.66

Ce résultat peut troubler. En effet, si l'on estime y en fonction de x par : $y=ax+b$, on peut s'attendre à estimer x en fonction de y par : $x=-(b/a)+(1/a)y$. En fait, il résulte de l'expression de a en fonction de $\{s_y, s_x, r(x,y)\}$ que cette intuition n'est vraie que si $r(x,y)=\pm 1$, c'est-à-dire si x et y sont reliés exactement par $y=ax+b$. Or, en pratique, cette relation n'est jamais exactement vérifiée : $ax+b$ ne permet que d'approcher y ; la véritable relation entre les deux variables est $y=ax+b+e$, relation dans laquelle e représente l'écart au modèle (e est aussi appelé *erreur* ou *résidu*). Dans les deux cas (y en fonction de x ; x en fonction de y), ce ne sont pas les mêmes erreurs e que l'on cherche à rendre petites (cf. **Fig. 8.4**).

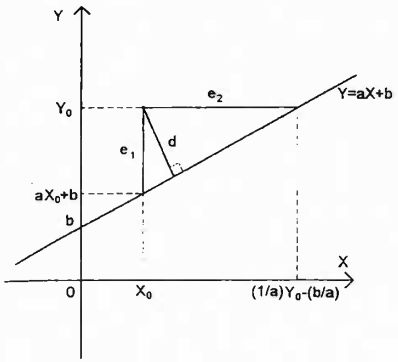


Figure 8.4. *Les différents écarts, dans un plan, entre un point et une droite.*
e_1 : erreur dans la prévision de y à partir d'une fonction linéaire en x ;
e_2 : erreur dans la prévision de x à partir d'une fonction linéaire en y ;
d : distance usuelle entre un point et une droite ;
La méthode des moindres carrés des écarts conduit à des droites différentes selon que l'on utilise l'un ou l'autre de ces écarts.

Cas du rôle symétrique des deux variables

On peut vouloir chercher une droite qui ajuste au mieux le nuage, non pas selon la « distance verticale » de la régression mais selon la distance usuelle entre les points et la droite (cf. **Fig. 8.4**). Dans cette perspective, les deux variables jouent des rôles symétrique : ce n'est pas la régression qu'il faut utiliser. Cet ajustement est obtenu par l'analyse en composantes principales, méthode introduite dans le chapitre suivant et utilisée en pratique uniquement dans l'étude de plus de deux variables ; elle peut se restreindre au cas particulier de deux variables mais il est rare que l'on ait réellement besoin d'une telle droite dans les applications à deux variables.

Régression : modèle explicatif.

On utilise quelquefois la régression pour décrire une liaison à l'aide d'un modèle ; l'intérêt du modèle réside alors essentiellement dans l'interprétation de ses paramètres, ces derniers étant une façon de synthétiser les données, intéressante en soi et/ou facilitant la comparaison entre ensembles de données.

Dans notre exemple comment interpréter $b=5.19$ et $a=0.726$ dans $y=.726x+5.19$?

- $b=5.19$ fait penser à un gain systématique de 5.19 points entre le troisième trimestre et le bac alors que la différence des moyennes n'est que de 2.16. Cette contradiction apparente est encore plus frappante pour la régression inverse ($b=+2.66$ alors que la moyenne de la différence *3ème trim.–bac* est négative !).

- La valeur de a (0.726) est inférieure à 1 ; cela indique que, en moyenne, une augmentation d'un point au troisième trimestre ne se traduit que partiellement (72.6%) pour le bac. Ce coefficient inférieur à 1 suggère une variabilité des notes en mathématiques moindre au bac qu'au 3ème trimestre, ce qui n'est pas en accord avec les écarts-types (3.19 au bac et 2.98 au 3ème trimestre ; cf. **Tab. 7.3**).

En fait ces contradictions ne sont qu'apparentes. Le modèle qui relie x et y est : $y=ax+b+e$; le coefficient $0<a<1$ montre effectivement que la variabilité de x ne se répercute que partiellement sur y ; mais la variabilité de y provient à la fois de celle due à x et de celle de e. D'autre part, il résulte de la relation entre a et $\{s_y, s_x, r(x,y)\}$ que, lorsque les écarts-types des deux variables sont du même ordre, le coefficient a, très proche alors de la corrélation, est toujours inférieur à 1 en valeur absolue (pour raison de simplicité de l'exposé et en conformité avec l'exemple, nous raisonnons avec $r(x,y)>0$) ; dans ce cas :

- étant multipliées par $a<1$, les améliorations de la note x qui sert à prédire ne sont traduites qu'en partie sur la note prédite y ;
- la diminution systématique de la note prédite est « rattrapée » par une forte valeur de b ;
- en conséquence, les faibles notes augmentent proportionnellement plus que les fortes ; autrement dit, les élèves ayant une note exceptionnelle (très basse ou très haute) au 3ème trimestre ont, en moyenne, une note moins exceptionnelle au bac (cet effet automatique est d'ailleurs à l'origine du terme régression).

On peut songer à utiliser a pour mesurer la part de variabilité de x « transférée » par le modèle sur y. Dans cet esprit, on utilise plutôt en pratique le coefficient de détermination $R^2=[r(x,y)]^2$, qui, d'après la relation explicitant a, s'écrit :

$$R^2=[r(x,y)]^2=[a^2 s_x^2]/s_y^2$$

On reconnaît au numérateur la variabilité de x (s_x^2) « transférée » sur y, et au dénominateur la variabilité totale de y (s_y^2). Ainsi, R^2 mesure la part de la variabilité de y « explicable » par l'équation de régression de y en x. Dans l'exemple, on dira que la note en maths au 3ème trimestre « explique » 46% (.678²) de la variabilité de la note en maths au bac.

Qu'apporte donc la connaissance des coefficients a et b par rapport à celle de s_y, s_x et $r(x,y)$? Dans cet exemple, comme dans beaucoup d'autres du même type, la réponse est : pratiquement rien.

En fait, la combinaison des coefficients s_y, s_x et $r(x,y)$ que constitue a n'est véritablement utile que lorsque l'on suspecte une relation causale entre y et x ; l'exemple toujours cité d'une telle relation est celui dans lequel y est le rendement d'une culture et x un facteur de production tel la dose d'engrais. L'intérêt du coefficient a est alors :

- prédictif : a représente l'augmentation de rendement que l'on peut attendre en moyenne d'une augmentation d'une unité de la dose d'engrais ;
- descriptif : a représente la valorisation par la culture d'une unité de la quantité d'engrais.

Un modèle plus simple et plus raisonnable

Si l'on veut exprimer la relation entre la note du bac et celle du troisième trimestre, il est certainement plus clair de dire simplement que, entre les deux épreuves, les élèves améliorent en moyenne leur note de 2.16 points. Se limiter à cela revient implicitement à adopter le modèle $y=x+2.16$, soit $a=1$ et $b=2.16$. Par rapport au modèle obtenu par la

méthode des moindres carrés, ce nouveau modèle présente l'avantage de la simplicité mais aussi deux défauts (pas forcément gravissimes selon l'utilisation escomptée du modèle) :

- il ne minimise pas la somme des carrés des erreurs ; cela étant, comme on peut le voir sur le graphique **8.5** où les deux droites sont tracées, la dégradation, par rapport à la droite de régression, de la qualité de l'ajustement n'est pas frappante ; ainsi, la valeur absolue de l'erreur vaut en moyenne 1.95 pour le modèle $y=x+2.16$ et 1.84 pour le modèle $y=.726x+5.19$;
- il rend mal compte des élèves ayant 18 ou plus au 3ème trimestre et donc censés obtenir plus de 20 le jour du bac.

Cette dernière réflexion suggère que le fait que les notes sont nécessairement comprises entre 0 et 20 pèse fortement sur la forme de la liaison entre x et y et, par voie de conséquence, sur le calcul de a et b ; d'où l'idée de calculer une nouvelle droite de régression en se limitant aux valeurs de x les plus éloignées des bornes 0 et 20 ; en sélectionnant les élèves ayant obtenu entre 8 et 12 au 3ème trimestre on obtient, tous calculs faits, la droite suivante (cf. **Tab. 8.5**) : $y=1.005x+2.47$

$8\leq x\leq 12$	$\bar{x}=10.20$	$s_x=1.20$	$r(x,y)=.44$
$n=464$ élèves	$\bar{y}=12.72$	$s_y=2.72$	$y=1.005x+2.47$

Tableau 8.5. *Quelques statistiques concernant les notes en mathématiques, au bac (y) et au 3ème trimestre (x), pour les 464 élèves tels que : $8\leq x\leq 12$.*

Ce nouvel ajustement renforce l'intérêt descriptif du modèle simple dans lequel $a=1$ et donc $b=\bar{y}-\bar{x}$, même si sur ce sous-ensemble de données le coefficient de corrélation $(r(x,y)=.44)$, et a fortiori le coefficient de détermination $(R^2=.19)$, est faible. Il illustre aussi qu'un calcul statistique peut extraire une structure visiblement non fortuite à partir de données empreintes d'une grande variabilité (cf. le nuage de la figure 8.5 réduit aux 464 points tels que $8\leq x\leq 12$).

Une autre technique de prévision

Pour prévoir la note du bac à partir de celle du troisième trimestre, un autre point de vue consiste à estimer la note au bac d'un individu i par la moyenne des notes du bac des individus ayant au 3ème trimestre une note voisine de celle de i, ce qui revient à :

- subdiviser, pour la note du 3ème trimestre, l'échelle de 0 à 20 en intervalles ;
- pour un individu i, prendre comme estimation de sa note au bac la moyenne des notes au bac des individus appartenant au même intervalle que i ; ces moyennes, calculées pour les individus vérifiant une condition (e.g. $9.5<x\leq 10.5$), sont dites *conditionnelles*.

La figure **8.5** illustre ce principe en choisissant, pour la note du 3ème trimestre, des intervalles centrés sur les notes entières et d'un point d'amplitude. Il est remarquable de constater que ces moyennes conditionnelles sont dans l'ensemble très proches de la droite de régression globale et que celles qui entourent 10 sont situées presque exactement sur la droite du modèle calculé sur les 464 élèves ayant une note voisine de 10 au 3ème trimestre. Ce résultat est général : on montre que si le couple (x,y) est distribué selon une loi normale bidimensionnelle, alors les moyennes conditionnelles sont strictement alignées. Historiquement, l'observation de cet alignement dans un cas particulier a joué un rôle important dans l'apparition de l'idée de droite de régression.

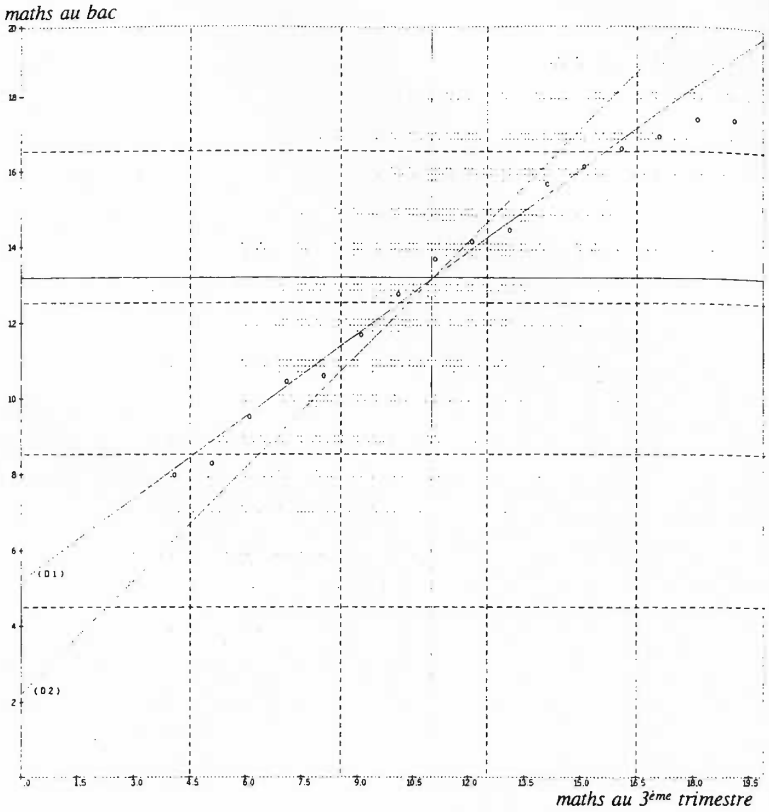


Figure 8.5. *Nuage des 909 élèves défini par les notes en mathématiques au troisième trimestre et au bac.*
Quadrillage (lignes tiretées) associé au tableau 1 ; lignes continues : moyennes générales ; droite de régression (D1) : $y=.726x+5.19$; un modèle raisonnable (D2) : $y=x+2.16$; o : moyenne de y pour les individus appartenant à un intervalle (d'amplitude 1 point) de x (moyenne conditionnelle).

Régression linéaire et régression non linéaire

La régression associée à l'équation $y=ax+b$ est dite *linéaire*. Elle est bien adaptée lorsque la forme du nuage est elliptique. Lorsque, par exemple, la forme du nuage ressemble à un croissant, on cherche plutôt à ajuster le nuage à une parabole c'est-à-dire à une équation dans laquelle intervient le carré de la note trimestrielle, par exemple $y=ax^2+b$. Pour cela, on met en œuvre une régression linéaire $y=au+b$ en prenant comme variable $u=x^2$. La régression linéaire est donc un outil plus général qu'il ne paraît au premier abord.

Régression simple et régression multiple

Jusqu'ici nous avons tenté de prédire la note du bac à l'aide de la seule note du troisième trimestre. Mais nous disposons des autres notes trimestrielles. Dans l'espoir d'améliorer l'estimation, peut-on faire une régression utilisant simultanément les trois trimestres ? Avec plusieurs variables explicatives, la technique se complique un peu mais ses principes restent identiques. Avec une seule variable la régression est dite *simple*, avec plusieurs elle est dite *multiple*. Ainsi, on peut chercher à ajuster les données par une formule du type :

$$bac\ estimé = a_1T_1 + a_2T_2 + a_3T_3 + b$$

dans laquelle T_1, T_2 et T_3 désignent respectivement les notes du premier, deuxième et troisième trimestre et a_1, a_2, a_3 et b des coefficients. Les calculs sont plus complexes que pour la régression simple mais il existe encore une solution qui minimise la somme des carrés des erreurs. On obtient, dans le cas des notes en mathématiques :

$$bac\ estimé = .08\ T_1 + .33\ T_2 + .42\ T_3 + 3.96$$

La formule de prévision qui utilise T_1, T_2 et T_3 est nécessairement meilleure que celle qui n'utilise que T_3, au moins lorsqu'elle est appliquée sur les données qui ont servi à calculer les coefficients ; dans cet exemple, la moyenne des valeurs absolues des erreurs pour les 909 élèves est de 1.78 points en utilisant T_1, T_2 et T_3 (à comparer avec la valeur 1.84 obtenue avec seulement T_3). Une mesure classique de l'adéquation d'un modèle de régression est le coefficient de détermination R^2, introduit à propos de la régression simple et qui s'interprète à la fois comme :

- le carré du coefficient de corrélation (dit *coefficient de corrélation multiple* entre *bac* d'une part et T_1, T_2 et T_3 d'autre part) entre les valeurs réelles et les valeurs estimées par le modèle ;
- la part de la variabilité de y expliquée par le modèle.

Dans cet exemple, le coefficient de détermination du modèle utilisant les 3 trimestres vaut .502 ; comparée au coefficient du modèle incluant le seul 3ème trimestre (.460), cette valeur indique que, une fois le troisième trimestre pris en compte, l'ajout des deux premiers trimestres dans le modèle n'apporte pratiquement rien à « l'explication » de y.

L'interprétation de la formule avec les trois trimestres pose problème avec son coefficient du premier trimestre presque nul qui pourrait faire croire que le premier trimestre n'est pas lié au bac, ce qui est faux (cf. Tab. 9.1). Ceci provient du fait que les 3 notes trimestrielles sont corrélées entre elles : il est donc en quelque sorte difficile de faire la part de chacune. Quelquefois, dans des cas semblables (i.e. lorsque toutes les variables, explicatives et expliquée, sont corrélées fortement et positivement entre elles), on peut même obtenir des coefficients négatifs, ce qui dans notre exemple pourrait faire croire qu'une mauvaise note à l'un des trimestres conduit à une note plutôt bonne au bac !

Conclusion sur la régression simple

La simplicité de la méthode et des calculs ne doit pas entraîner à utiliser la régression simple à tort et à travers. Avant de la mettre en œuvre, il faut se demander si la connaissance des coefficients a et b apporte vraiment quelque chose –en fait, il faut se poser ce type de question pour toutes les méthodes–, en particulier par rapport à s_y, s_x et $r(x,y)$.

La réponse est a priori positive si l'on a de bonnes raisons de suspecter une relation causale entre x et y ou si l'on doit effectuer des prévisions. Sinon, elle est a priori négative.

Synthèse d'un ensemble de variables quantitatives

Dans les chapitres précédents, les notes sont considérées une par une (e.g. histogramme, moyenne, etc.) ou deux par deux (e.g. graphique, coefficient de corrélation, etc.). Nous abordons ici l'Analyse en Composantes Principales (ACP), méthode qui appréhende globalement l'ensemble de toutes les notes dans l'optique d'en faire une synthèse. Nous ne la présentons pas en détail, ce qui impliquerait un exposé trop technique, mais illustrons son intérêt à partir d'un exemple.

9.1 Deux objectifs de l'analyse en composantes principales

Etudier l'ensemble des corrélations entre variables

Nous avons étudié assez minutieusement la liaison entre les notes en mathématiques, au bac et au troisième trimestre, et rapidement les liaisons entre les notes du bac et les notes trimestrielles de chaque matière. Il reste à envisager les liaisons entre matières différentes, avec des questions du type : quelles sont les matières les plus liées entre elles ? Comment évoluent ces liaisons durant l'année et au bac ?

Pour étudier ces liaisons, il est raisonnable, au moins dans un premier temps, de se restreindre à l'étude des coefficients de corrélation : cette démarche est particulièrement justifiée avec ces données pour lesquelles les graphiques croisant les variables deux à deux suggèrent uniquement des liaisons linéaires. Le tableau **9.1** rassemble les coefficients de corrélation calculés pour tous les couples de notes. Comme il y a 20 notes, cela fait 190 couples et donc 190 coefficients de corrélation à étudier ce qui est beaucoup !

Ce tableau est appelé matrice des corrélations. Certaines de ses caractéristiques apparaissent immédiatement : en premier lieu, dans ces données, toutes les corrélations sont positives. Mais, sans techniques appropriées, il est exclu de pouvoir synthétiser ces 190 coefficients dans l'optique des questions posées. L'analyse en composantes principales aide, entre autres, à réaliser cette synthèse.

Recherche de variables synthétiques

Peut-on résumer (approximativement) l'ensemble des 20 notes par un petit nombre de variables, non pas extraites des 20 notes mais les combinant ?

Un exemple de variable synthétique : la moyenne des notes

Un résumé simple et largement employé, notamment au bac et dans la plupart des examens et concours, est la moyenne de toutes les notes.

	Maths				Physique				Sciences nat.				Histoire-géo.				Philosophie			
	M	m1	m2	m3	P	p1	p2	p3	SN	s1	s2	s3	H	h1	h2	h3	P	p1	p2	p3
MAt	1.0																			
mat1	.58	1.0																		
mat2	.66	.77	1.0																	
mat3	.68	.72	.81	1.0																
PHY	.62	.56	.62	.66	1.0															
phy1	.56	.59	.61	.60	.63	1.0														
phy2	.61	.60	.65	.66	.71	.74	1.0													
phy3	.60	.57	.63	.65	.72	.71	.77	1.0												
SN	.34	.25	.31	.31	.37	.35	.37	.38	1.0											
sn1	.36	.34	.37	.37	.43	.41	.44	.45	.42	1.0										
sn2	.33	.39	.43	.43	.42	.39	.42	.44	.43	.58	1.0									
sn3	.39	.39	.41	.44	.45	.41	.48	.48	.43	.55	.53	1.0								
HG	.25	.19	.22	.23	.31	.24	.29	.30	.32	.31	.25	.30	1.0							
hg1	.23	.25	.22	.25	.26	.28	.32	.30	.28	.30	.30	.29	.42	1.0						
hg2	.27	.25	.26	.27	.31	.26	.37	.35	.31	.36	.35	.34	.45	.63	1.0					
hg3	.33	.27	.26	.29	.32	.32	.38	.39	.32	.30	.30	.33	.46	.61	.63	1.0				
PHI	.28	.24	.30	.30	.30	.31	.32	.29	.33	.23	.23	.30	.36	.32	.35	.35	1.0			
phi1	.17	.19	.20	.22	.24	.27	.27	.27	.27	.29	.23	.30	.27	.33	.44	.41	.42	1.0		
phi2	.23	.22	.25	.26	.24	.24	.26	.24	.24	.31	.23	.31	.34	.33	.42	.40	.43	.64	1.0	
phi3	.21	.16	.22	.24	.24	.23	.26	.26	.32	.30	.25	.32	.36	.31	.42	.40	.47	.61	.69	1.0

Tableau 9.1. *Matrice des corrélations.*
Exemple : r(Maths 3ème trim., Maths bac) = .68 ; la matrice est symétrique : seule figure la partie triangulaire basse.

Moyenne arithmétique, moyenne pondérée, poids.

Les notes obtenues au bac servent à déterminer qui obtient le diplôme. Pour cela, on calcule, pour chaque élève, la moyenne de ses notes au bac dans les différentes matières. En réalité d'autres notes que celles de notre fichier interviennent (Français, langues, options, etc.) mais nous raisonnons dans la suite comme si le diplôme était délivré à partir des seules 5 matières dont nous disposons. Pour obtenir la *note du bac* (en abrégé *bac*), on calcule la moyenne des 5 notes. Ce n'est pas la *moyenne arithmétique simple* obtenue en divisant la somme des 5 notes par 5, mais une *moyenne pondérée*. A chaque matière est attribué un coefficient appelé poids de cette matière : mathématiques et physique ont chacune un poids de 5, tandis que sciences naturelles, histoire-géographie et philosophie ont chacune un poids de 2. La somme de ces poids vaut : 5+5+2+2+2=16. La moyenne, notée *bac*, est obtenue par la formule suivante :

$$bac = \frac{5maths + 5physique + 2sciences\ nat. + 2histoire\text{-}géographie + 2philosophie}{16}$$

Pour ce bac scientifique, les poids ont pour objet de donner plus d'importance (dans la moyenne finale) aux matières scientifiques qu'aux matières littéraires.

En quoi la moyenne synthétise-t-elle les notes ?

Si un élève a 18 au bac, il a certainement de très bonnes notes dans toutes les matières. A contrario, un autre qui obtient 2 n'a certainement bien réussi aucune épreuve. La note du bac permet donc de situer approximativement un élève pour l'ensemble des 5 matières (approximation utilisée par l'Education Nationale).

Une seule variable synthétique ne suffit généralement pas

La moyenne d'un élève permet de situer globalement l'ensemble de ses notes parmi celles des autres élèves : c'est un résumé succinct de toutes les notes. Mais utiliser seulement la moyenne peut être insuffisant : par exemple, elle ne permet pas de séparer les élèves moyens partout des élèves très bons quelque part et très mauvais ailleurs ; ni de séparer ces derniers suivant les matières où ils sont les plus forts. Pour affiner l'analyse, d'autres variables sont nécessaires.

Combinaison linéaire de variables quantitatives

La note moyenne au bac est définie pour chaque élève : c'est donc une nouvelle variable. Ses valeurs sont des nombres : il s'agit, comme pour les notes de chaque matière, d'une variable quantitative. Elle est obtenue à partir des 5 notes au bac par la formule simple indiquée plus haut. On dit que la variable *bac* est une combinaison linéaire des 5 variables car on l'obtient par une somme des 5 variables initiales, chacune étant multipliée par un coefficient (par exemple, le coefficient des mathématiques est 5/16).

Le problème de la cohérence des unités des différentes variables

Ici, nous sommes dans un cas simple : les notes sont des variables de même nature, ayant le même intervalle de variation et si l'on peut dire la même unité de mesure. Lorsque les variables sont de natures différentes (poids, taille, prix, etc.), la moyenne n'a pas de sens concret. Cependant rien n'empêche de souhaiter les synthétiser par une variable ayant une répartition aussi "proche" que possible de celle de chacune des variables. Dans ce cas, il peut être intéressant de faire apparaître, dans la formule de la combinaison linéaire, les variables sous leur forme centrée-réduite : en effet, l'opération de centrage-réduction harmonise d'une certaine manière les unités (chaque variable a pour unité son écart-type).

Le problème du sens de variation des variables

Ici, la moyenne est une synthèse assez satisfaisante car les notes varient généralement dans le même sens (les coefficients de corrélation sont tous positifs). Ce n'est pas toujours le cas (cf. **Tab. 9.2**).

	cas 1				cas 2			
élève	*X1*	*Y1*	*M1*	*E1*	*X2*	*Y2*	*M2*	*E2*
1	18	18	18	0	18	2	10	16
2	15	15	15	0	15	5	10	10
3	10	10	10	0	10	10	10	0
4	5	5	5	0	5	15	10	−10
5	2	2	2	0	2	18	10	−16

Tableau 9.2. *Moyenne et variable synthétique : 5 élèves décrits par 2 notes (X, Y). Cas 1 : la moyenne M1 synthétise parfaitement les notes X1 et Y1 ; ce n'est pas le cas de leur différence E1 (=X1−Y1). Cas 2 : la moyenne M2 ne synthétise pas les notes X2 et Y2 ; ce n'est pas le cas de leur différence E2.*

Dans le cas 2 du tableau **9.2**, la moyenne arithmétique vaut 10 partout et donc n'apporte aucune information sur la répartition de chacune des 2 variables *X2* et *Y2*. C'est une bien pauvre synthèse ! Il vaut bien mieux prendre comme synthèse la différence entre les deux notes. Ainsi, selon les cas et plus précisément selon les liaisons qui existent entre les

variables étudiées, on n'utilise pas la même combinaison linéaire pour synthétiser des variables.

Variables synthétiques et composantes principales

L'analyse en composantes principales repose sur la recherche « automatique » d'une suite des « meilleures variables synthétiques » appelées composantes principales. Il reste évidemment à préciser en quoi ces variables sont meilleures que d'autres.

9.2 Première composante principale

Définition

La première composante principale (i.e. la « meilleure » variable synthétique) n'est pas systématiquement, comme on vient de l'évoquer, la moyenne. Dans notre exemple où toutes les notes varient globalement dans le même sens (les corrélations sont positives), elle devrait toutefois en être proche.

On veut construire une variable qui résume aussi bien que possible les 20 notes. Pour des raisons de simplicité de calcul et d'interprétation, on cherche une combinaison linéaire de ces 20 notes (on pourrait imaginer des fonctions plus compliquées). Une variable est un bon résumé si elle est liée à chacune des notes. La mesure de liaison utilisée en ACP est le coefficient de corrélation. Mais le signe du coefficient de corrélation ne doit pas intervenir ici : une variable ayant une corrélation égale à -1 avec une note, la représente très bien (il suffit de savoir que cette corrélation est négative et que, entre la variable et la note, l'ordre des élèves est inversé). Ce sont donc les valeurs absolues des 20 corrélations qu'il s'agit de rendre aussi grandes que possible. En fait, plutôt que les valeurs absolues des corrélations, pour des raisons de simplicité des calculs analogues à celles évoquées à propos de la régression, on cherche à rendre maximum la somme de leurs carrés.

La première composante principale est une variable, combinaison linéaire de toutes les notes, telle que la somme des carrés de ses corrélations avec toutes les notes est maximum.

Ainsi, en notant $r(z, X_k)$ le coefficient de corrélation entre une combinaison linéaire quelconque z et la $k^{ème}$ note, on cherche z telle que :

$$\sum_{k=1,20} [r(z, X_k)]^2 \text{ maximum}$$

La première composante principale représente-t-elle bien les 20 variables ?

Pour apprécier dans quelle mesure une composante principale représente bien les 20 variables, on calcule le coefficient de corrélation entre cette composante principale et chacune des variables. En pratique, on visualise d'un coup l'ensemble des coefficients de corrélation entre chaque variable et les deux premières composantes principales par un graphique (dit plan factoriel) dans lequel :

- chaque axe représente une composante principale (généralement axe horizontal = première composante ; axe vertical = seconde composante) ;
- une variable X_k est représentée par un point dont la coordonnée le long de l'axe horizontal (resp. vertical) est égale au coefficient de corrélation entre X_k et la première (resp. deuxième) composante principale.

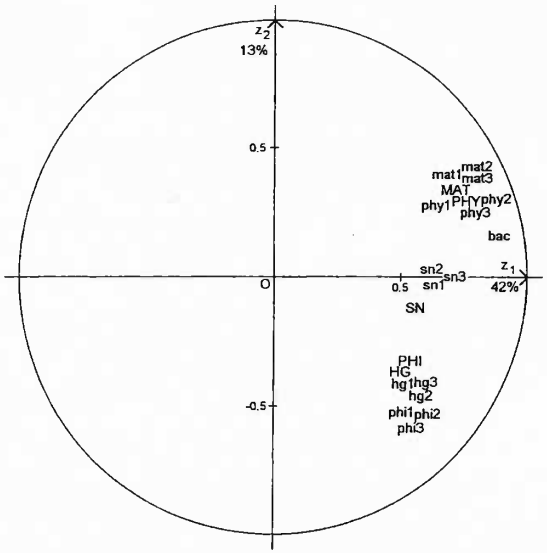


Figure 9.1. *Représentation des notes par leurs coefficients de corrélation avec les deux premières composantes principales.*
En raison de la nature des coordonnées, chaque variable se trouve nécessairement à l'intérieur du cercle de rayon 1 dit cercle des corrélations. En majuscules, les notes du bac ; en minuscules, les notes trimestrielles (ex : hg2 = histoire-géographie au 2ème trimestre) ; bac : moyenne pondérée des notes du bac (avec les coefficients du bac). Par rapport à leur place exacte, les libellés ont été un peu écartés les uns des autres pour éviter les chevauchements.

Appliqué à nos données (cf. **Fig. 9.1**), ce graphique montre que les coefficients de corrélation entre la première composante principale et chaque note sont tous supérieurs à .5 : étant liée à chacune, cette composante principale synthétise l'ensemble des variables.

La quantité (notée λ_1) maximisée par cette 1ère composante est la somme des carrés de ces coefficients de corrélation, soit $\lambda_1 = 8.41$. Si cette composante synthétisait parfaitement chaque variable, ces coefficients de corrélation seraient chacun égal à ±1. La valeur maximum possible de λ_1 est donc égale au nombre de variables soit 20. Le rapport $\lambda_1/20$ donne une idée de la qualité de la synthèse ; ici $\lambda_1/20 = .42$: on dit que la 1ère composante principale rend compte (ou « explique ») 42% de la variabilité des données.

Cette combinaison est la meilleure possible ; à titre de comparaison, la variable *bac* rend compte de 34% des données.

Comment s'analyse cette composante principale ?

Dans le graphique **9.1**, la coordonnée de chaque variable sur l'axe horizontal est égale à sa corrélation avec la première composante principale. Les coordonnées de toutes les notes sont positives car elles sont toutes corrélées positivement avec la première composante

principale : un élève ayant une forte valeur pour cette composante a généralement des bonnes notes dans l'ensemble des matières, au bac et pendant l'année.

On montre que ces coefficients de corrélation sont exactement les coefficients de la combinaison linéaire des notes préalablement centrées et réduites. Cette première composante principale (notée z_1), qui est aussi une variable centrée (i.e. de moyenne nulle), est donc égale à :

$$z_1 = .71\ \frac{maths\ bac - 13.21}{3.19} + .70\ \frac{maths\ 1t - 10.64}{2.85} + \ \ldots\ + .52\ \frac{philo\ 3t - 9.26}{2.33}$$

Cette relation fait apparaître explicitement le centrage et la réduction puisqu'on retire à chaque variable sa moyenne puis on la divise par son écart-type ; par exemple 13.21 (resp. 3.19) est la moyenne (resp. l'écart-type) des notes en mathématiques au bac des 909 élèves (cf. **Tab. 7.1.** et **7.3**).

Un élève ayant dans toutes les matières des notes supérieures à la moyenne de l'ensemble, a des valeurs positives pour toutes les variables centrées ; comme tous les coefficients sont positifs, cet élève a une forte valeur positive pour z_1 (remarquer l'intérêt de faire apparaître les variables centrées, ce qui montre qu'une note influe dans un sens ou dans un autre selon qu'elle est supérieure ou inférieure à la moyenne de l'ensemble). Réciproquement, un élève qui a une forte valeur positive pour z_1 a globalement des notes au-dessus de la moyenne de l'ensemble. En ce sens, cette première composante principale représente le niveau général de l'élève.

Si tous les coefficients étaient égaux entre eux, z_1 serait confondue avec la moyenne m_{20} des 20 notes. Ce n'est pas le cas, même si l'on en est pas très loin dans ces données. A titre indicatif $r(z_1, m_{20}) = .90$

9.3 Deuxième composante principale

Définition

La deuxième composante doit apporter des informations complétant celles apportées par la première. Or, quand deux variables sont corrélées entre elles, les informations qu'elles contiennent sont redondantes : on impose donc à la deuxième composante d'être non corrélée à la première. Pour qu'elle traduise elle aussi l'ensemble des notes, on lui impose la même quantité à maximiser qu'à la première. Finalement, la deuxième composante principale est la combinaison linéaire des notes, non corrélée avec la première composante principale, telle que la somme des carrés de ses corrélations avec toutes les notes est maximum.

Dans le graphique **9.1**, la coordonnée de chaque variable sur l'axe vertical est égale à sa corrélation avec la deuxième composante principale.

La somme des carrés des coordonnées des 20 variables vaut : $\lambda_2 = 2.63$ soit $\lambda_2/20 = 13\%$; cette composante principale exprime 13% de la variabilité des données ; par rapport à la première, elle représente une structure moins forte.

A la différence de la première composante principale, certaines coordonnées sont positives et d'autres sont négatives. Cette composante traduit donc une opposition entre certaines notes. Cette opposition est facilement interprétable puisque l'on trouve en haut les notes de mathématiques et de physique, en bas les notes d'histoire-géographie et de philosophie, et

au milieu les notes de sciences naturelles, dont les coefficients sont proches de zéro : c'est une opposition entre matières scientifiques et matières littéraires, les sciences naturelles se situant à mi-chemin.

Schématiquement cette composante, notée z_2, s'écrit :

$$z_2 = .33\frac{maths\ bac - 13.21}{3.19} + \ldots + .28\frac{phys\ 3t - 11.45}{3.16} - .37\frac{hist\ bac - 10.21}{2.65} - \ldots - .56\frac{philo\ 3t - 9.26}{2.33}$$

Pour un élève, des valeurs supérieures à la moyenne générale dans les matières scientifiques augmentent sa valeur pour z_2 ; il en est de même des notes inférieures à la moyenne dans les matières littéraires. Ainsi, un élève **à la fois** bon scientifique et mauvais littéraire a une forte valeur pour z_2. Plus précisément, une valeur positive pour z_2 correspond à un élève meilleur scientifique que littéraire. Réciproquement, une valeur négative pour z_2 correspond à un élève meilleur littéraire que scientifique. En ce sens, la deuxième composante principale représente le profil, plutôt scientifique ou plutôt littéraire, de l'élève.

Un élève ayant une valeur proche de zéro n'est pas meilleur dans l'un des domaines que dans l'autre : si c'est un bon élève (coordonnée positive sur la première composante), il est bon dans les deux domaines ; s'il est mauvais (coordonnée négative sur la première composante), il est mauvais dans les deux domaines.

Remarque

On peut aussi se demander si la variable *bac* (approximation de la note moyenne officielle) est liée à z_1 : on a : $r(z_1, bac) = .89$; en calculant aussi $r(z_2, bac)$ on peut placer la variable *bac* sur le graphique de la figure **9.1**. Un tel élément, placé sur un plan factoriel qu'il n'a pas contribué à construire, est dit *supplémentaire* ou *illustratif* (par opposition aux autres éléments qui sont dits *actifs* ou *principaux* ; un élément supplémentaire peut être une variable, comme ici, mais aussi un individu). Ici la variable *bac* apparaît comme une synthèse qui privilégie les matières scientifiques, ce qui était attendu compte tenu des coefficients (5 pour mathématiques et physique, 2 pour les autres matières).

9.4 Représentation des élèves et des lycées

Représentation des élèves

Considérer les composantes principales comme des variables synthétiques suggère de représenter les 909 élèves sur le graphique, dit aussi plan factoriel, croisant z_1 en abscisse et z_2 en ordonnée (cf. **Fig. 9.2**).

L'interprétation des axes de ce graphique est par définition celle des composantes principales : l'axe des abscisses représente le niveau général et celui des ordonnées le profil de l'élève, plutôt scientifique ou plutôt littéraire. Cette interprétation peut être illustrée en examinant quelques élèves. Par exemple :

- le premier axe oppose d'une part, 544 et 64 (qui ont globalement de bonnes notes) à, d'autre part, 755 et 748 (qui ont globalement de mauvaises notes) ;
- le second axe oppose d'une part, 544 et 755 (qui ont globalement des notes meilleures dans les matières scientifiques que dans les matières littéraires) à, d'autre part, 64 et 748 (qui ont globalement des notes meilleures dans les matières littéraires que dans les matières scientifiques).

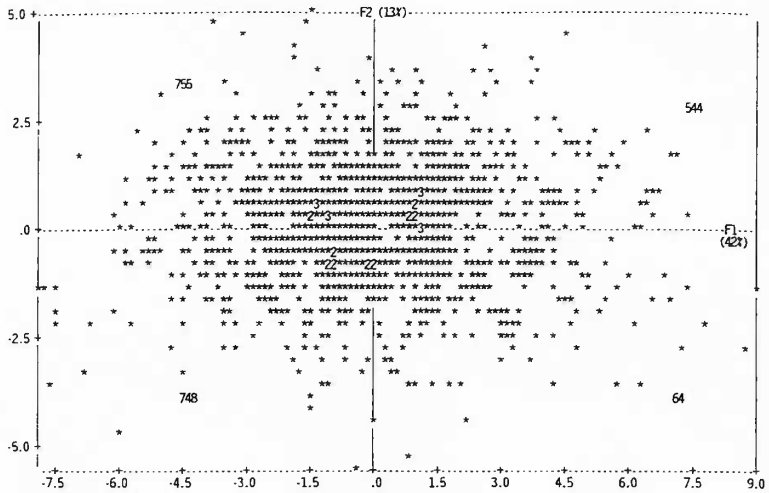


Figure 9.2. *Représentation des individus sur le plan factoriel défini par les deux premières composantes principales : allure générale du nuage de points.*
*Ce graphique est fourni par le logiciel SPADN : un élève =une * ; en cas de superposition, on indique le nombre de points superposés. 4 individus sont identifiés par leur numéro d'ordre dans le fichier : 64, 544, 755, 748 ; leurs notes sont dans le tableau 9.3.*

élève	lycée	maths		physique		sciences nat.		hist.-géo.		philosophie	
		bac	année	bac	année	bac	année	bac	année	bac	année
64	4	17.00	15.33	14.00	13.13	14.00	14.50	15.00	16.20	17.00	15.33
544	17	19.00	18.33	19.00	18.50	16.00	16.33	8.00	14.50	9.00	8.67
755	12	16.00	11.33	10.00	8.50	4.00	5.67	3.00	8.17	5.00	6.00
748	14	7.00	5.40	5.00	6.03	9.00	7.50	13.00	11.33	4.00	10.53
moy.	-	13.21	10.95	11.00	11.17	10.97	11.02	10.21	11.01	7.84	9.09

Tableau 9.3. *Notes des 4 élèves identifiés sur le plan factoriel (cf. Fig. 9.2).*
Année : moyenne des trois notes trimestrielles ; moy : ensemble des 909 élèves.

Ceci peut être vérifié en consultant les notes de ces élèves (cf. **Tab. 9.3**).

Les composantes principales étant centrées (ce sont des combinaisons linéaires de variables centrées), le nuage des 909 élèves est centré : le point qui correspond à l'individu théorique qui aurait obtenu comme notes les moyennes de l'ensemble des 909 élèves, dit point moyen (ou centre de gravité), est confondu avec l'origine. La forme générale du nuage des élèves est globalement elliptique, avec pour plus grande direction d'allongement le premier axe. Ce plus grand allongement ne peut coïncider avec une direction autre que les axes de coordonnées car, par construction, les composantes principales sont non corrélées.

Sur un tel graphique, on peut essayer de distinguer des groupes d'individus. Ici, la répartition des points semble homogène et l'on distingue des tendances plutôt que des

groupes (tendance à avoir globalement des notes plus ou moins bonnes ; tendance à présenter un profil plutôt scientifique ou plutôt littéraire).

Ainsi, la position d'un élève synthétise l'ensemble de ses notes : on ne peut pas en déduire exactement ses 20 notes, mais on peut le situer à peu près dans la population.

L'appartenance au lycée

Il est intéressant d'examiner si les élèves d'un même lycée sont proches entre eux sur le graphique **9.2.** Pour cela, on peut représenter chaque élève par le numéro de son lycée. C'est ce qui a été fait pour les lycées 1, 7 et 9 (cf. **Fig. 9.3**) et qui met en évidence :

- une grande variabilité entre les élèves d'un même lycée : le lycée 1 contient le meilleur élève et l'un des moins bon ; on peut penser que cette variabilité serait moins grande si l'on avait limité l'étude aux notes du bac (les notes trimestrielles sont au moins en partie relatives à chaque classe) ;
- des tendances générales : globalement, les élèves du lycée 9 sont meilleurs que ceux du lycée 1 ; ici encore, on peut penser que ces tendances seraient plus nettes si l'on avait limité l'étude aux notes du bac.

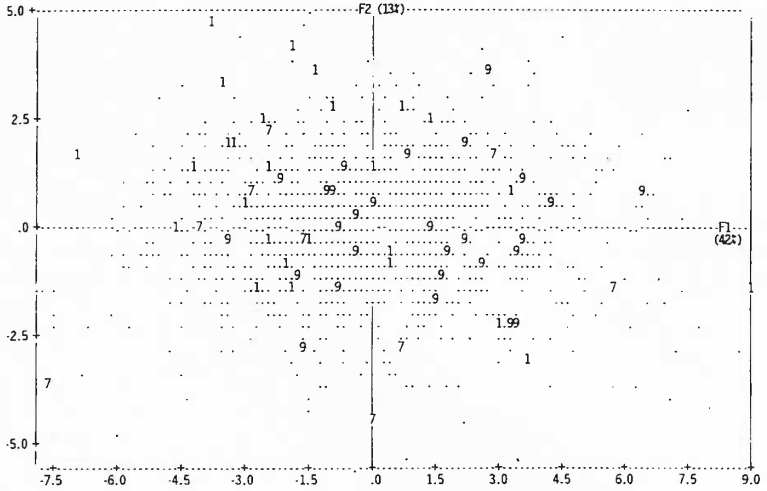


Figure 9.3. *Représentation des individus sur le plan factoriel défini par les deux premières composantes principales.*
Les élèves des lycées 1, 7 et 9 sont identifiés par le numéro de leur lycée.

Compte tenu du nombre de points, il est utile de figurer les points moyens (centres de gravité) des individus d'un même lycée. On obtient ainsi un nuage de 22 points (cf. **Fig. 9.4**). Cette représentation des lycées suggère quelques commentaires.

- La comparaison entre les échelles des figures 9.3 et 9.4 montre que les points moyens des lycées sont assez regroupés au centre du graphique. On retrouve ici le phénomène général selon lequel les moyennes de groupes d'éléments sont moins variables que les

éléments eux-mêmes ; cet effet est ici particulièrement fort du fait de la grande variabilité intra-lycée.

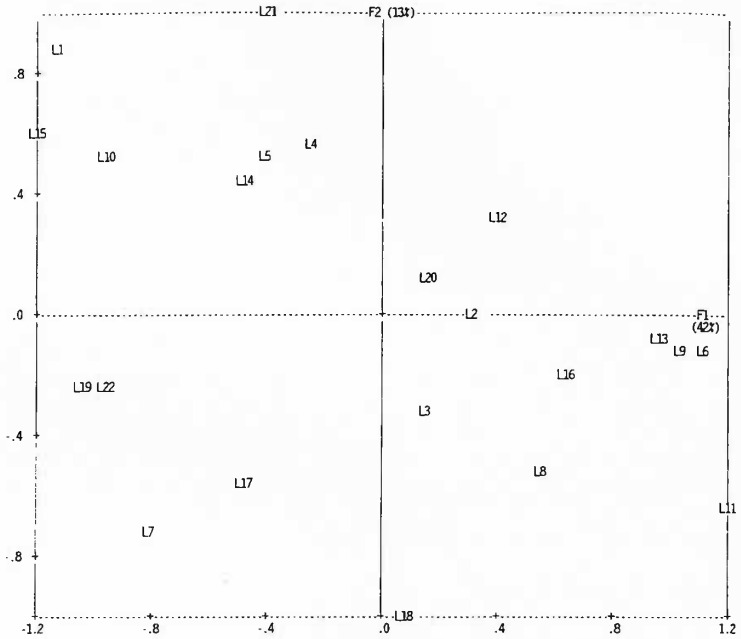


Figure 9.4. *Représentation des lycées sur le plan factoriel défini par les deux premières composantes principales.*
Un lycée apparaît au centre de gravité de ses élèves tels qu'ils apparaissent figures 9.2 ou 9.3 ; les élèves ne sont pas représentés ici.

- Les élèves du lycée 1 ont des notes généralement mauvaises (coordonnée négative sur l'axe 1). Leurs résultats sont cependant meilleurs (ou moins mauvais) dans les matières scientifiques que dans les matières littéraires (coordonnée positive sur l'axe 2).
- Les élèves du lycée 17 sont globalement moins mauvais que ceux du lycée 1, mais se rattrapent surtout dans les matières littéraires.
- Les élèves des lycées 6, 9 et 13 ont en général de bonnes notes ; en moyenne, ils surpassent les autres partout, et légèrement plus dans les matières littéraires.

Remarque : il est clair que la comparaison entre les lycées à partir de ces deux variables synthétiques est beaucoup plus simple qu'à partir de l'étude directe de l'ensemble de toutes les notes. Elle peut être suffisante ou servir de cadre à une analyse plus fouillée.

9.5 Plan des deux premières composantes : bilan des corrélations entre variables

Sur le graphique **9.1** où sont représentées les 20 notes, il apparaît quelque chose de très clair : les 8 notes de mathématiques et de physique sont regroupées entre elles ; les 8 notes littéraires sont regroupées entre elles et nettement séparées des 8 premières ; les 4 notes de sciences naturelles sont aussi groupées entre elles et situées à mi-chemin entre les matières littéraires et les matières scientifiques.

Que peut-on en conclure ?

Trois propriétés techniques de la représentation des variables

1° Intuitivement, les deux premières composantes principales étant par définition non corrélées entre elles, aucune variable ne peut être très fortement corrélée à chacune d'entre elles ; plus précisément, on montre que la somme des carrés des coefficients de corrélation entre une variable X et un ensemble de variables $\{Y_k\ ;\ k=1,K\}$ non corrélées entre elles, est toujours inférieure ou égale à 1 ; ainsi les points représentant les variables, dont la somme des carrés des coordonnées (i.e. la distance à l'origine ; cf. le théorème de Pythagore) est inférieure ou égale à 1, sont toujours situés à l'intérieur ou sur le bord du cercle de rayon 1 dit *cercle des corrélations*.

2° Il résulte de 1° que la somme, pour les 20 variables, des carrés des distances à l'origine est toujours inférieure ou égale à 20 ; toujours du fait du théorème de Pythagore, cette somme est égale à la somme des carrés des coordonnées des 20 variables sur les deux axes, c'est-à-dire à $\lambda_1+\lambda_2$ qui dans ce cas vaut : $8.41+2.63=11.04$; rapportée à 20, cette quantité exprime le pourcentage de la variabilité des données représenté par le plan ; soit ici 55%.

3° Deux variables corrélées de la même manière et de façon importante aux composantes principales sont nécessairement corrélées entre elles. Ainsi deux notes assez éloignées de l'origine et proches entre elles sont corrélées de façon importante.

Bilan des corrélations entre les différentes notes

Utilisant les propriétés précédentes, on peut lire la figure 9.1 dans l'optique d'un bilan des liaisons entre variables.

- Toutes les variables occupent une zone assez restreinte du disque délimité par le cercle des corrélations. Cette proximité relative entre les points suggère que toutes les variables sont corrélées positivement entre elles (ceci est vérifiable sur la matrice des corrélations).
- Dans chaque matière les quatre notes sont assez corrélées entre elles, en tout cas un peu plus corrélées entre elles qu'avec les autres matières. En se reportant à la matrice des corrélations, on constate que cette affirmation se vérifie globalement (e.g. nettement pour l'histoire-géographie et la physique) malgré des exceptions nombreuses (e.g. *maths au bac* est plus liée à *physique au bac* qu'à *maths 1er trimestre*) mais correspondant à des écarts faibles. Le graphique traduit une tendance générale réelle en gommant les irrégularités qui rendent ardue la lecture de la matrice des corrélations.
- Les huit notes de mathématiques et de physique sont plus liées entre elles qu'avec les autres matières (on peut contrôler que les corrélations entre mathématiques et physique sont toujours supérieures à .55 alors que celles entre ces matières et les autres sont toujours inférieures à .48). La différence n'est pas énorme mais ces 8 notes forment

nettement un groupe. Ceci suggère l'existence de qualités communes (ou de goûts communs) pour réussir dans ces deux matières.

- On peut faire des remarques identiques pour la philosophie et l'histoire-géographie. L'écart entre ces deux matières et les précédentes suggère l'existence de qualités différentes (ou de goûts différents) pour réussir dans ces deux groupes de matières.

Cet exemple illustre que l'étude d'un tel graphique, pour peu que l'on en connaisse les principales règles de lecture, est plus simple et plus informative que l'analyse directe de la matrice des corrélations.

9.6 Troisième et quatrième composantes

Si l'on estime que les deux premières composantes ne suffisent pas pour résumer les 20 notes, on peut en chercher d'autres (une troisième, une quatrième, une cinquième, etc.), non corrélées aux composantes déjà trouvées et rendant maximum la somme des carrés des corrélations avec les notes. Chaque nouvelle composante exprime une part de la variabilité des données moins importante que les précédentes et ne fait que les nuancer.

Les 3ème et 4ème composantes principales peuvent être appariées dans un graphique, analogue à la figure **9.1**, représentant leurs corrélations avec les 20 notes (cf. **Fig. 9.5**).

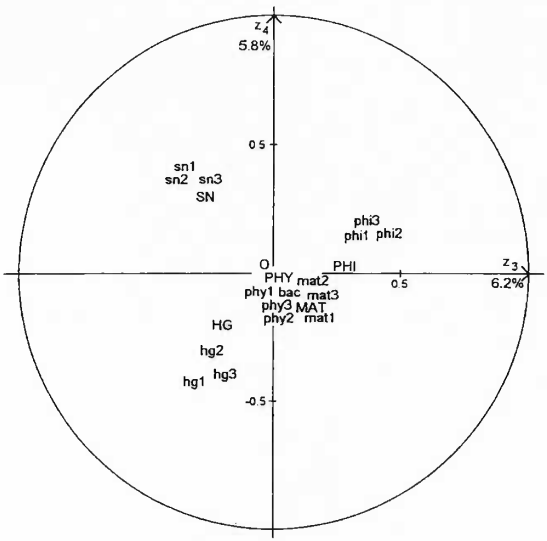


Figure 9.5. *Représentation des notes par leurs coefficients de corrélation avec la 3ème et la 4ème composante principale (cf. légende Fig. 9.1).*

Ce plan sépare entre elles : philosophie, sciences naturelles et histoire-géographie. On peut le commenter par composante.

La troisième composante oppose les élèves meilleurs en philosophie qu'en sciences naturelles et histoire-géographie à ceux qui sont dans une situation opposée. Les deux

premières composantes où histoire-géographie et philosophie étaient très proches (car très liées) ne permettaient pas de distinguer les élèves meilleurs dans l'une de ces matières (à la fois pendant l'année et au bac) que dans l'autre. Même s'ils sont peu nombreux, ces élèves existent.

De même, la quatrième composante oppose sciences naturelles à histoire-géographie.

9.7 Suite du bilan des corrélations (plan 3-4)

L'examen du plan des composantes 3 et 4 permet de compléter un peu le bilan des corrélations entre les notes. Il n'apporte rien de plus sur les mathématiques et la physique, très proches de l'origine (et donc non corrélées avec chacune de ces deux composantes). Par contre, dans les trois autres matières, il montre que les corrélations internes à chaque matière sont plus fortes que les corrélations entre les matières : il existe des élèves nettement meilleurs dans l'une ou l'autre de ces trois matières. Pour les notes au bac, la tendance est un peu moins marquée que pour les notes trimestrielles.

9.8 Cinquième composante

La cinquième composante est encore intéressante (cf. **Fig. 9.6**). Elle oppose pour les 2 matières littéraires et les sciences naturelles, et à un moindre degré pour les mathématiques, les notes du bac à celles de l'année, donc les élèves qui progressent au bac (dans la plupart des matières) à ceux qui au contraire chutent systématiquement. Il est intéressant de voir que certains lycées se différencient beaucoup selon cette composante (cf. **Fig. 9.7**) : ceci suggère que l'interprétation de ce phénomène ne doit pas se référer uniquement à des caractéristiques des élèves (perte de moyens pendant un examen, etc.), mais aussi au fait que certains lycées sous-notent leurs élèves (e.g. lycées 22 et 15) tandis que d'autres (e.g. lycées 7 et 21) les surnotent.

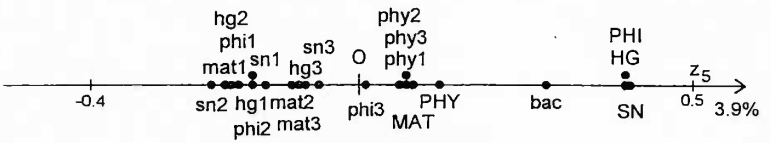


Figure 9.6. *Représentation des notes par leurs coefficients de corrélation avec la 5ème composante principale.*

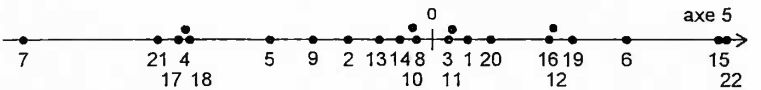


Figure 9.7. *Représentation des lycées par la 5ème composante principale. Comme Fig. 9.4, un lycée est situé au barycentre de ses élèves (ces derniers n'étant pas ici représentés).*

9.9 Conclusion

On peut résumer ce que l'ACP nous apprend sur la structure des données, en quelques phrases respectant l'ordre d'importance des tendances.

- Globalement, lorsqu'un élève a plusieurs bonnes notes, ses autres notes ont aussi tendance à être bonnes. Le meilleur résumé des 20 notes est proche de leur moyenne et classe les élèves des meilleurs aux plus mauvais. Cette seule composante rend compte de 42% de la variabilité des données.
- Il existe une opposition entre élèves à profil plutôt scientifique et élèves à profil plutôt littéraire qui nuance la moyenne générale. Cette seconde composante représente 13% de la variabilité des données. Les deux premières représentent donc 55% de cette variabilité.
- On peut aussi distinguer des réussites assez stables dans l'année pour l'une ou l'autre des trois matières : histoire-géographie, sciences naturelles et philosophie.
- Enfin, certains élèves brillent plus au bac que pendant l'année tandis que d'autres baissent. Mais cette dernière composante ne représente que 3.9% de la variabilité des données.

Caractérisation d'une sous-population : élèves avec données manquantes

Les 3 groupes d'élèves ayant des données manquantes (cf. **Tab 4.3**) présentent-ils des caractéristiques différentes de celles de la population générale ? Autrement dit, sont-ils très typés ? Cette question, évoquée durant l'étude de ces groupes (cf. Ch. 4), sert maintenant pour introduire la comparaison entre une sous-population et la population générale dont elle est extraite.

Ce problème n'est pas sans lien avec la stabilité des notes étudiée en 7.10 : il ne peut y avoir de variation notable des caractéristiques générales lors de l'adjonction (ou de l'élimination) d'un groupe que si les éléments de ce groupe diffèrent des autres. Mais l'inverse n'est pas forcément exact : un petit nombre d'individus, même particuliers, a peu de chances de modifier notablement les moyennes calculées sur un millier.

Nous avons jusqu'ici seulement mentionné (cf. **Tab. 6.1**) la variation des effectifs des élèves dans les lycées en fonction de l'adjonction des groupes présentant des données manquantes ; or, la sur-représentation de certains lycées dans un groupe peut présenter de l'intérêt.

Il n'est pas du tout question ici d'une comparaison systématique entre les élèves avec données manquantes et les autres, mais seulement de déceler des différences importantes entre ces deux types d'élèves, qui justifieraient un commentaire dans la description des données. De telles différences sont ici appelées « anomalies », sous-entendu *par rapport à la situation théorique uniforme dans laquelle les élèves avec données manquantes ne se différencient pas des autres*.

Comparer un groupe et la population dont il est issu est un problème un peu différent pour une variable qualitative (le lycée) et pour une variable quantitative (une note) ; mais certains principes sont communs aux deux cas :

- la comparaison directe des répartitions, dans le groupe et dans la population générale, permet sans aucun outil particulier de repérer des anomalies évidentes ;
- un calcul plus sophistiqué, quoique de conception simple, permet de disposer d'une mesure de l' « importance » d'une anomalie compte tenu des données étudiées.

10.1 Les élèves avec données manquantes proviennent-ils de lycées particuliers ?

Formellement, examiner si un groupe d'élèves provient de lycées particuliers revient à comparer la distribution des modalités de la variable *lycée*, pour le groupe d'une part et pour la population entière d'autre part.

1) *Comparaison directe : détection d'une anomalie*

Pour illustrer comment repérer directement une anomalie, prenons l'exemple du groupe de 15 élèves ayant un bac incomplet, groupe décrit en détail au chapitre 5. Les tableaux du chapitre 5 montrent que ce groupe est composé de 8 candidats libres et de 7 élèves provenant de 5 lycées. Pour analyser des effectifs, il faut faire apparaître dans un même tableau non seulement la sous-population étudiée mais aussi les autres individus (cf. **Tab 10.1**).

	c. lib.	lycéen	total
bac incomplet	8	7	15
bac complet	5	955	960
total	13	962	975

a : effectifs

	c. lib.	lycéen	total
bac incomplet	53.3	46.7	100
bac complet	0.5	99.5	100
total	1.3	98.7	100

b : pourcentages-lignes

Tableau 10.1. *Répartition des 975 élèves selon les variables* "bac complet / bac incomplet" *et* "candidat libre / lycéen" : *tableau croisé.*

Parmi les *bacs incomplets*, les candidats libres sont les plus nombreux. Est-ce "normal" ? Cela le paraîtrait tout à fait s'ils étaient aussi les plus nombreux dans la population générale. Ce n'est bien évidemment pas le cas : ils ne sont que 13 en tout (parmi les 975 élèves qui ne sont pas des fantômes). La différence est flagrante, c'est une anomalie. Comment le voit-on ? Même si le calcul exact n'est pas réalisé, c'est la différence entre les pourcentages des candidats libres dans les deux populations qui étonne.

Transformation des effectifs en fréquence ou pourcentage

Pour comparer deux distributions d'une même variable qualitative, il est donc plus facile de travailler sur les pourcentages que sur les effectifs bruts, qui ne sont pas directement comparables. Ces pourcentages, ou proportions d'individus ayant une modalité particulière, s'appellent fréquences relatives de la modalité (par opposition aux effectifs, appelés fréquences absolues). La proportion de la modalité *candidat libre* parmi les 15 individus est de 8/15, c'est-à-dire .53 ou encore 53 %. Celle de ces mêmes candidats libres dans la population générale est de 13/975, c'est-à-dire .0133 ou encore 1.33 %.

La transformation des effectifs en pourcentages présente un autre avantage : les pourcentages sont plus parlants que les chiffres bruts car on a l'habitude de les manier (1.33% évoque plus directement quelque chose que 13 sur 975).

2) *Comparaison graphique : juxtaposition de diagrammes en bâtons.*

Pour visualiser la différence entre les deux distributions, on peut juxtaposer les diagrammes en bâtons (cf. **Fig. 10.1**). Comme les effectifs totaux sont très différents (960 et 15), il faut représenter les pourcentages. L'ordre des lycées doit être identique dans les deux diagrammes. Nous choisissons l'ordre par effectif décroissant sur la population entière, qui est la population de référence.

Une anomalie intéressante : la sur-représentation des candidats libres parmi les bacs incomplets

La simple comparaison visuelle entre les deux histogrammes (cf. **Fig. 10.1**) met en évidence de façon flagrante la différence pour la modalité *candidat libre* : de .5% pour les élèves ayant un bac complet ce pourcentage passe à 53% parmi ceux qui ont un bac incomplet.

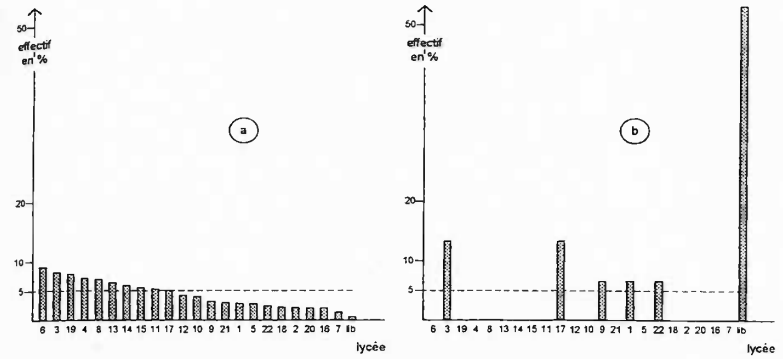


Figure 10.1. *Répartitions des effectifs de la variable* lycée *selon le nombre de notes au bac (a: 5 notes ; b: 2 notes) : diagrammes en bâtons des pourcentages.*

Nous retenons cette anomalie dans la description de l'ensemble des données pour deux raisons complémentaires. L'une se base uniquement sur les chiffres : la différence est très grande et l'effectif concerné (8 élèves) n'est pas trop faible. L'autre raison n'est pas d'ordre statistique mais de celui de l'interprétation : on sait bien que la modalité *candidat libre* de la variable *lycée* est tout à fait particulière. On peut penser que la sur-représentation des candidats libres parmi les bacs incomplets est un phénomène qui a une explication et, si les raisons en sont assez générales, il est vraisemblable qu'il se reproduise ailleurs.

Les autres lycées ; attention aux effectifs faibles

Les 5 lycées qui ont 1 ou 2 représentants parmi les 15 *bacs incomplets* ont tous aussi des pourcentages supérieurs à ceux des *bacs complets*. Mais nous ne le considérons pas comme une "anomalie" méritant un commentaire. En effet, les écarts entre les pourcentages sont beaucoup moins importants que celui des candidats libres, mais surtout les effectifs de 1 ou 2 sont très faibles. Pour les lycées qui n'ont qu'un élève parmi les 15, il suffit que celui-ci disparaisse pour que le pourcentage chute à zéro. Cette instabilité potentielle ne permet pas de donner des conclusions générales.

3) *Chiffrer l'importance des anomalies : probabilité d'une différence aussi importante*

Pour examiner si les 15 élèves étudiés sont particuliers du point de vue de leur lycée, nous avons utilisé une approche qualitative qui, dans le cas présent, est tout à fait suffisante. Il est possible de compléter cette approche en mesurant le « degré » d'une anomalie.

Que chiffre-t-on ?

Précisons d'abord ce que l'on cherche à chiffrer. Dans notre cas, ce qui étonne n'est pas que le nombre de candidats libres soit égal à 8 dans ce groupe de 15, mais qu'il soit aussi élevé alors que dans la population générale les candidats libres sont très peu nombreux.

Premier point de vue : pourcentage de groupes de 15 « au moins aussi exceptionnels »

On considère tous les groupes de 15 élèves qu'il est possible de constituer à partir des 975 et l'on calcule pour chacun d'eux le nombre de candidats libres. Comme le pourcentage global de candidats libres n'est que de 1.33%, on se rend bien compte que bien peu de

groupes auront autant de candidats libres que celui des *bacs incomplets*. Parmi l'ensemble des groupes, ce groupe est donc exceptionnel par son nombre élevé de candidats libres. Pour mesurer ce caractère exceptionnel (ou ce degré d'anomalie), on calcule le pourcentage de groupes de 15 élèves "au moins aussi exceptionnels que lui", c'est-à-dire ayant un nombre de candidats libres au moins aussi élevé.

Le nombre total de groupes différents de 15 élèves parmi les 975 est très élevé ; il faut un nombre de presque 45 chiffres pour l'écrire. Aussi le calcul exact du pourcentage est assez lourd et l'on se contente souvent d'une approximation (cf. fiche 6 § 2.1). Un calcul approché conduit à .000000000004 %. Nous le commentons un peu plus loin.

Second point de vue : probabilité d'extraire un groupe de 15 « au moins aussi exceptionnel »

Le second point de vue utilise un modèle probabiliste simple. Au lieu de parler de la proportion des candidats libres dans la population des 975 élèves, on dit que, si l'on prend « au hasard » un élève parmi les 975, il y a une probabilité de 13/975 que ce soit un candidat libre. Le passage de la proportion (ou fréquence relative) à la probabilité n'est pas seulement un changement de vocabulaire mais aussi de niveau. Dans le premier cas, on décrit des données ; dans le second, on se réfère à un modèle mathématique artificiel (ici, nous étudions la population entière directement et non via un sondage ; à aucun moment nous n'envisageons de réaliser effectivement un tirage au hasard d'élèves) qui obéit à des lois précises et permet des calculs. Malgré ce caractère artificiel, le modèle est une référence utile pour apprécier la situation réelle : le but est d'obtenir par des calculs sur le modèle des indications sur les données.

Probabilités

Nous donnons ici simplement quelques notions très frustes, quelques règles à support intuitif et un peu de vocabulaire ; les ouvrages de probabilités auxquels on peut se référer sont nombreux.

Parler de probabilités implique d'abord un ensemble sur lequel les définir : dans notre cas, c'est l'ensemble des 975 élèves. Puis, il faut spécifier la probabilité de chaque élément de l'ensemble, c'est-à-dire la probabilité "si l'on prend un élève au hasard" d'obtenir Un tel ou Un tel. Ici, nous prenons le cas le plus naturel en donnant à tous les élèves la même probabilité (*probabilité uniforme*). Comme la somme de toutes les probabilités doit faire 1, la probabilité de chaque élève est de 1/975.

On s'intéresse maintenant à la variable *lycée* (simplifiée à deux modalités : *lycéen/candidat libre*) d'un élève pris au hasard. Parmi les 975 élèves, il y a 13 candidats libres. Pour que l'élève tiré au hasard soit un candidat libre, il faut que ce soit l'un ou l'autre d'entre eux. Chacun d'eux ayant une probabilité de 1/975, la probabilité de l'ensemble des 13 est 13/975=.0133 : c'est la probabilité de tirer un candidat libre. On retrouve les valeurs des pourcentages.

Cette formalisation permet de faire des calculs et notamment de répondre aux questions du type suivant : si on prend au hasard 15 élèves parmi les 975, quelle est la probabilité d'obtenir au moins 8 candidats libres ? Le principe du calcul est simple (il est détaillé dans la fiche 9). La probabilité uniforme sur tous les élèves implique que tous les groupes possibles de 15 élèves choisis parmi les 975 ont la même probabilité. La probabilité d'avoir au moins 8 candidats libres dans un groupe de 15 élèves tirés au hasard est égale au pourcentage des groupes de ce type dans l'ensemble de tous les groupes possibles.

Commentaire sur la valeur obtenue.

Cette probabilité (ou ce pourcentage) est de l'ordre de .000000000004, ce que l'on note $4*10^{-12}$, où 10^{-12} signifie $(1/10)^{12}$. Elle est extrêmement faible et il est tout à fait exclu de penser qu'il ne s'agit que d'un hasard. Quelle que soit l'optique choisie, le même chiffre conduit à la même conclusion.

En terme de pourcentage, on dira que ce groupe est tout à fait exceptionnel (parmi tous les groupes possibles). En terme de probabilité, on dira qu'il est tout à fait improbable (d'obtenir autant de candidats libres dans l'hypothèse "tous les élèves ont la même probabilité"). On est donc sans doute très loin de la situation "neutre" que traduit cette hypothèse et, comme nous l'avions déjà noté lors de la simple comparaison des pourcentages, ceci vaut la peine d'être retenu dans la description du fichier.

Les autres lycées

On peut procéder de même pour chaque lycée (comportant n élèves dont n_0 bacs incomplets), en calculant à chaque fois la probabilité d'obtenir par un tirage au hasard de n élèves parmi les 975, au moins n_0 bacs incomplets. On constate qu'aucune de ces probabilités (ou pourcentage) n'est très faible. La plus faible (lycée 17 avec 2 bacs incomplets parmi 51 élèves) est .14 soit plus "d'une chance sur 10" (d'observer au moins 2 *bacs incomplets* en tirant au hasard 51 élèves). Considérer qu'une probabilité est petite dépend des contextes et des enjeux ; en pratique, dans des études telles que la nôtre, l'usage est de considérer comme peu probable un événement d'une probabilité inférieure à .05 (5 chances sur 100 dans le langage courant). Cette limite est tout à fait arbitraire, mais si répandue qu'elle est presque devenue une norme.

4) *Ordonner les anomalies*

On peut procéder de la même façon pour chaque lycée et pour les deux autres groupes d'élèves ayant des données manquantes : comparer les pourcentages, repérer d'un coup d'œil des anomalies, puis mesurer l'intensité de ces anomalies avant de les étudier de plus près. Ce n'est pas très rapide si l'on veut faire une étude systématique sur tous les groupes et tous les lycées ! Le plus simple est de calculer puis d'ordonner toutes les probabilités mesurant le degré d'anomalie. Ainsi, on repère immédiatement les plus remarquables.

On a fait un calcul approché systématique de ce pourcentage (ou probabilité) pour chacun des 3 groupes d'élèves ayant des données manquantes (les 15 élèves avec bac incomplet ; les 13 élèves n'ayant de notes qu'au bac ; les 38 élèves avec quelques données manquantes ; cf. **Tab. 4.3** pour la définition des groupes) et pour chacune des 23 modalités de la variable *lycée* (22 lycées + *candidat libre*). Ceci permet de déceler d'éventuelles sur-représentations importantes de certains lycées dans l'un ou l'autre de ces trois groupes.

Nous examinons donc en tout 69 situations (23×3). Seules 3 d'entre elles dépassent le seuil de 5% fixé arbitrairement (cf. **Tab. 10.2**). Celle dont la probabilité est la plus faible, et de très loin, concerne l'étude que nous venons de faire. L'anomalie suivante est la sur-représentation du lycée 13, avec 12 élèves (sur 61) dans le groupe des 38 élèves ayant quelques notes manquantes dispersées. La troisième et dernière est la sur-représentation des candidats libres (3 sur les 13 candidats libres) dans le groupe des 13 élèves sans notes trimestrielles.

probabilité	groupe d'élèves	lycée	effectif
0.000000000004	bac incomplet (15)	candidats libres (13)	8 élèves
.0000002	quelques dm (38)	lycée 13 (61)	12 élèves
0.001	bac complet seul (13)	candidats libres (13)	3 élèves

Tableau 10.2. *"Anomalies" dans la répartition de la variable* lycée pour *les différents groupes d'élèves ayant des données manquantes (dm).*
Exemple : 8 élèves sont candidats libres et appartiennent au groupe bac incomplet. Seules figurent les "anomalies" associées aux 3 plus faibles probabilités (<.05). Entre parenthèses : effectif de la modalité dans l'ensemble des 975 élèves.

5) *Que fait-on avec les anomalies ?*

On a donc mis en évidence trois anomalies dans la répartition de la variable *lycée* pour les trois groupes d'élèves ayant des données manquantes. Quelles conclusions faut-il en tirer ?

Ce sont des chiffres qui ont attiré l'attention sur ces trois cas : ils incitent à chercher des explications, mais ce ne sont pas eux qui permettent de conclure.

La première réaction devant une anomalie doit systématiquement être de vérifier qu'elle ne provient pas d'une erreur. Penser aux erreurs possibles doit pratiquement être un réflexe conditionné devant tout résultat un tant soit peu curieux.

Ici, après quelques vérifications, les anomalies ne proviennent vraisemblablement pas d'erreurs. Il faut donc chercher d'autres explications. Etudions les trois configurations les plus improbables.

Sur-représentation des candidats libres parmi les bacs incomplets

Nous ne connaissons pas l'explication de ce phénomène. On peut consulter des spécialistes, émettre des hypothèses, faire une enquête pour les vérifier. Si l'anomalie relevée présentait un intérêt fondamental, cela vaudrait la peine d'y consacrer temps et énergie, mais ce n'est pas le cas. Toutefois vu l'importance de la sur-représentation et le statut très particulier des candidats libres, il existe certainement une ou des raisons bien précises.

Sur-représentation du lycée 13 dans le groupe présentant quelques notes manquantes

Pour trouver des explications, on peut s'enquérir auprès du lycée d'un éventuel problème, par exemple une épidémie de grippe très localisée ou bien une exigence particulière de la direction quant au nombre de notes nécessaires pour bénéficier d'une note trimestrielle, etc. L'intérêt ou non de cette enquête est à juger par les demandeurs de l'étude : le statisticien ne fait qu'attirer leur attention.

Sur-représentation des candidats libres dans le groupe avec bac seulement

L'explication est ici simple : parmi les candidats libres, seuls ceux inscrits à certains cours par correspondance peuvent bénéficier de notes trimestrielles. L'étonnant n'est pas ici la sur-représentation des candidats libres, mais plutôt la présence, dans ce groupe, de candidats régulièrement inscrits dans un lycée.

6) *Remarques*

Inférence

L'étude est limitée aux 975 élèves d'un même centre d'examen. Mais on souhaite peut-être à travers cette étude dégager quelques conclusions générales, valables pour l'ensemble de la population des candidats français et peut-être même pour plusieurs années. Il est clair que

pour le faire sans risque d'erreur, il faut disposer de toutes les données. Faire une telle généralisation sans ces données (en termes statistiques on dit que l'on réalise une inférence) comporte un risque.

Dans certains cas il est possible de chiffrer les risques encourus lors d'une telle généralisation : tel est l'objet de la théorie des tests statistiques, qui s'applique lorsque les individus étudiés forment un *échantillon* extrait de façon aléatoire d'une population générale, appelée *population parente*. Ici les 992 élèves n'ont pas été tirés au hasard et on est trop loin d'une telle situation pour que la théorie des tests statistiques soit applicable.

Cela n'empêche pas une inférence empirique, sans quantification des risques encourus, tout à fait raisonnable. **Il faut se résigner à réfléchir et non à appliquer des formules**. Le principe de l'inférence qualitative relève de l'affirmation courante : « les mêmes causes produisent les mêmes effets ».

Ainsi, considérer que la forte proportion d'absences de notes pour les candidats libres n'est pas spécifique de notre ensemble de données mais est un phénomène général, semble tout à fait raisonnable. Pour le groupe n'ayant de notes qu'au bac, nous en connaissons une explication ; pour l'autre groupe, nous ne la connaissons pas mais elle existe certainement. Autrement dit, il est vraisemblable que, dans la France entière, les candidats libres sont plus nombreux que les autres à ne passer que 2 épreuves du bac et à ne pas avoir de notes trimestrielles.

Par contre, faire une généralisation pour le lycée 13 n'est pas raisonnable : une généralisation aux lycées 13 des autres centres d'examen serait proprement ubuesque ; pour une autre année ce serait aussi trop risqué. En effet, les causes de ces notes manquantes peuvent être conjoncturelles : une épidémie de grippe localisée a peu de chance de se reproduire, une attitude de la direction peut changer, etc.

Test

Le calcul d'une probabilité pour mesurer le caractère exceptionnel ou improbable d'un fait statistique est proche de celui utilisé dans les tests statistiques mais n'a pas la même signification.

Nous ne sommes pas ici, comme dans beaucoup d'études, dans une situation de test statistique :

- les données n'ont pas été tirées au hasard ;
- le fait statistique quantifié par une probabilité a été suggéré par l'examen des données ; il serait donc tautologique de se demander si les données sont compatibles avec ce fait. En comparaison, dans les tests statistiques, on formule, indépendamment des données observées, une hypothèse concernant la population parente et l'on examine si les données observées sont compatibles avec cette hypothèse.

10.2 Les élèves avec données manquantes ont-ils des notes particulières ?

Les *bacs incomplets* ont-ils des notes en mathématiques meilleures, analogues ou plus mauvaises que les autres élèves ? On peut se poser la même question pour chaque matière et pour chaque groupe d'élèves. Formellement, nous voulons comparer la distribution d'une variable quantitative (ici la note en maths), d'une part sur la population tout entière et d'autre part sur une sous-population. De même que dans la section précédente, nous relevons principalement les faits les plus remarquables (ou anomalies). On se doute que la

réponse ne peut être simple puisque les notes de mathématiques des 15 élèves avec bac incomplet sont très variables et s'étalent de 5 à 18 (cf. **Tab 5.3**). Nous examinons en premier lieu les graphiques représentant les distributions puis comparons les moyennes de la sous-population à celles de la population générale.

Situer un individu dans la distribution d'une variable quantitative

Nous restons dans l'exemple de la note en mathématiques au bac et du groupe des *bacs incomplets*. Avant de situer globalement les notes de mathématiques de ce groupe de 15 élèves dans celles de la population générale, voyons déjà comment situer celle d'un seul individu.

La note la plus faible en mathématiques pour ce groupe est 5. Ce 5 est sans discussion une mauvaise note. Mais est-elle une mauvaise note par rapport à l'ensemble de la population ?

Pour y répondre, il faut connaître la répartition des notes de l'ensemble des 975 élèves. Un coup d'œil sur le diagramme en bâtons de la figure **10.2.a** (qui ressemble à celui de la figure **7.1** qui ne concerne que les 909 élèves ayant toutes leurs notes) montre que 5 est parmi les plus faibles notes.

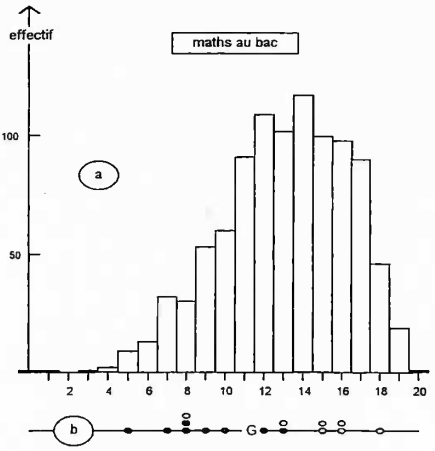


Figure 10.2. *Distribution des notes en mathématiques au bac.*
a) *élèves ayant des notes (975) : diagramme en bâtons ;*
b) *élèves avec bac incomplet (15) : représentation axiale, en distinguant les candidats libres (disques pleins) et les lycéens (cercles).*

Il est toujours utile de chiffrer une assertion. Pour situer ce 5, on peut donner le pourcentage des élèves ayant une note au moins aussi faible. Comme précédemment pour les 8 candidats libres parmi les 15 bacs incomplets, on ne considère pas seulement les 5 élèves qui ont une note égale à 5, mais tous ceux ayant une note au moins aussi mauvaise. Il en existe 14, soit 1.44%. Ce pourcentage permet de conclure que 5 est une valeur extrêmement faible.

A l'autre extrémité, le 18 est une très bonne note ; mais 6.7% des élèves sont au moins aussi bons : elle est donc sensiblement moins remarquable que le 5.

Ces pourcentages résument très bien la situation de ces individus aux valeurs presque extrêmes et sont utiles lorsque l'on s'intéresse à un tel individu en particulier ; mais ils sont moins parlants pour situer un 10 ou un 12 ; pour de telles valeurs, on se limite souvent à les comparer à la moyenne de l'ensemble de la population (13.08 ; cf. **Tab. 7.5**) : ici on constate que ces deux notes lui sont inférieures.

Situer graphiquement une sous-population pour une variable quantitative

Pour situer tout le groupe des *bacs incomplets*, on peut situer chaque individu, mais il reste à faire une synthèse. On peut, à l'œil, repérer des faits remarquables en utilisant quelques artifices graphiques.

- On peut songer (cf. **Fig. 5.1**) utiliser la représentation graphique sur un axe avec un sigle particulier pour la sous-population. Mais cela n'est guère pratique ici compte tenu des 975 points à représenter.
- On peut aussi songer à matérialiser la position des éléments du groupe sur le diagramme en bâtons (ou l'histogramme) de la population entière ; mais cette sous-population est si faible qu'elle ne serait guère visible.
- Finalement, compte tenu du fort effectif de la population et du faible effectif de la sous-population, le mieux semble être de juxtaposer le diagramme en bâtons (ou l'histogramme) de la population entière et le graphique axial de la sous-population (cf. **Fig. 10.2**). Pour pouvoir les comparer, il faut, bien sûr, les représenter à la même échelle.

Sur ces graphiques, rien de très remarquable n'apparaît : ces 15 individus ne forment pas, du point de vue de leur note de mathématiques au bac, une sous-population très particulière : ceci était attendu, vu leur grande variabilité. Ils comportent néanmoins proportionnellement plus de mauvaises notes que l'ensemble de la population.

Chiffrer l'importance des différences : comparer les moyennes

Si l'on veut caractériser de façon chiffrée l'ensemble des 15 élèves, on peut comparer leur moyenne (11.53) à celle de l'ensemble des 975 élèves (13.08 ; cf. **Tab. 7.5**). On conclut alors qu'en moyenne ils sont plus faibles que l'ensemble de la population. Se limiter aux moyennes cache bien évidemment les disparités car si le 5 est faible, le 18 ne l'est pas ! Mais **si l'on veut une réponse simple, il faut se résigner à ne pas tout dire**.

Maintenant peut-on-dire qu'ils sont "nettement plus faibles" ou simplement "un peu plus faibles" ? Autrement dit, une différence de 1.15 entre les deux moyennes est-elle importante ? Est-elle assez importante pour être considérée comme remarquable et être retenue dans le compte rendu de l'étude ?

On peut se contenter de la valeur brute et avoir une opinion a priori sur la valeur 1.15 : on peut juger par exemple que 1.15 points, pour une note comprise entre 0 et 20, est un écart important.

On peut aussi évaluer la valeur 1.15 avec le principe introduit lors de l'étude de la variable qualitative *lycée*, c'est-à-dire en associant une probabilité à cette valeur (ce raisonnement est décrit en détail Fiche 10).

On considère donc tous les groupes différents de 15 élèves que l'on peut constituer à partir des 975. Il est clair qu'à chaque groupe de 15 individus correspond une moyenne et que ces moyennes peuvent être inférieures, égales (très rarement) ou supérieures à la moyenne générale. Quel est le pourcentage de ceux qui par leur moyenne sont au moins aussi « exceptionnels » que le groupe étudié (qui a une moyenne inférieure de 1.15 points à la moyenne générale) ? Ce pourcentage mesure le caractère remarquable (ou le degré d'anomalie) du groupe étudié. Deux calculs sont possibles selon que l'on considère :

- les groupes dont la moyenne est au moins aussi faible ;
- les groupes dont la moyenne diffère au moins autant de la moyenne générale, c'est-à-dire est soit inférieure à 11.53 soit supérieure à 14.59.

Dans un cas comme celui-là, on choisit généralement la première solution (ce que nous faisons ici). Notons que lorsque la distribution est (presque) symétrique, choisir la seconde solution multiplie (à peu près) par 2 les pourcentages obtenus. Ce choix n'est pas crucial pour deux raisons :

- c'est l'ordre de grandeur du pourcentage et non sa valeur exacte qui importe ; on ne fait pas de différence entre 10% et 20%, entre 1% et 2% ;
- on calcule généralement ces pourcentages pour plusieurs sous-populations et plusieurs variables, afin de sélectionner les faits les plus remarquables, c'est-à-dire correspondant aux pourcentages les plus petits ; le classement des pourcentages n'est pas affecté par le choix ci-dessus.

Ce pourcentage s'interprète aussi comme la probabilité d'obtenir un groupe de 15 aussi extrême (que celui effectivement observé) par un tirage au hasard de 15 élèves parmi les 975 ; il vaut à peu près 3%. Le groupe des *bacs incomplets* est quelque peu remarquable par sa moyenne en mathématiques, mais infiniment moins que par sa proportion de candidats libres.

La situation est analogue en physique, où la moyenne du groupe n'est que 9.13 alors que la moyenne générale atteint 10.86. Le pourcentage de groupes possibles de 15 élèves ayant une moyenne au moins aussi faible est aussi de l'ordre de 3%.

Le calcul exact de ces probabilités est formellement possible mais généralement impraticable. On se contente donc d'approximations. Ce type de situation étant extrêmement fréquent, il a été bien étudié et l'on connaît des approximations mathématiques la plupart du temps satisfaisantes (cf. Fiche 7) et faciles à obtenir. On sait aussi, par programme, simuler une suite de tirages « au hasard » d'individus dans une population. On peut alors effectuer quelques milliers de ces tirages et utiliser le pourcentage obtenu à partir de ces tirages comme approximation.

Sélectionner et ordonner les faits remarquables

On peut procéder exactement de la même façon pour les deux autres groupes d'élèves ayant des données manquantes et pour les différentes matières. Ceci conduit à comparer chaque moyenne de chaque groupe à la moyenne générale et à associer un pourcentage (ou une probabilité) à chaque comparaison.

Un nouveau problème se pose, celui de la population de référence. Nous avons situé le groupe précédent (qui n'a de notes au bac qu'en mathématiques et en physique) dans la population entière de 975 élèves. Pour les notes du bac de mathématiques et de physique, ceci est encore possible pour les deux autres groupes, mais cela ne l'est plus pour les 3

autres notes du bac définies uniquement pour 960 élèves. Bien qu'ici les écarts obtenus entre les résultats calculés sur les deux populations soient certainement faibles, il est toujours très dangereux de se référer à des populations différentes (cf. 4.5) pour faire des comparaisons au sein d'une même étude. Aussi, pour étudier les 5 notes et les deux groupes, nous nous limitons à la population de 960 élèves ayant des valeurs complètes.

On peut trier, par ordre croissant des probabilités qui leur sont associées, les moyennes de chaque groupe dans chaque matière (cf. **Tab. 10.3**). Il apparaît que les écarts les plus remarquables vont tous dans le sens de notes inférieures pour les élèves ayant des données manquantes (ceci est en accord avec ce qui a déjà été observé à partir du tableau 7.5).

matière (au bac)	groupe concerné cf. **Tab 4.3.**	moyenne du groupe	moyenne générale	différence moyenne	probabilité associée
histoire-géo.	quelques dm (38)	8.53	10.13	−1.60	$8\ 10^{-5}$
physique	quelques dm (38)	8.82	10.89	−2.07	$2\ 10^{-4}$
maths	quelques dm (38)	11.39	13.11	−1.72	$4\ 10^{-4}$
philosophie	bac complet seul (13)	5.62	7.78	−2.16	$8\ 10^{-3}$
maths	bac complet seul (13)	11.15	13.11	−2.96	10^{-2}

Tableau 10.3. *Caractérisation des groupes d'élèves ayant des données manquantes (dm) par leurs notes aux épreuves du bac.*
Seules figurent les moyennes associées à une probabilité inférieure à .05 ; la population de référence comporte les 960 élèves ayant toutes leurs notes au bac (d'où quelques différences avec le tableau 7.1).

Remarques sur la probabilité associée :

- Bien que la moyenne en mathématiques du groupe avec des notes seulement au bac (11.15) soit encore plus faible que celle du groupe avec quelques notes manquantes (11.39), la probabilité associée est moins petite. Cela vient du fait que ce groupe est moins nombreux que l'autre (13 élèves au lieu de 38). La probabilité associée tient compte du fait que **la moyenne d'un groupe s'éloigne d'autant plus facilement de la moyenne générale que son effectif est faible**.
- Bien que, pour le groupe ayant quelques valeurs manquantes, la différence avec la moyenne générale est plus importante en physique (2.07) qu'en histoire-géographie (1.60), la probabilité associée est plus faible. Cela vient du fait que la dispersion est plus importante pour les notes de physique que d'histoire (les écarts-types de ces deux matières, calculés sur les 960 élèves, valent respectivement 3.65 et 2.69 ; cf. **Tab. 7.6**). La probabilité associée tient compte du fait que **la moyenne d'un groupe s'éloigne d'autant plus facilement de la moyenne générale que la dispersion de l'ensemble des individus est grande**.

En résumé, la probabilité associée à une différence est un indicateur qui tient compte de la différence elle-même mais aussi des effectifs concernés et de la dispersion globale.

10.3 Autres explorations

On pouvait s'attendre à ce que les *bacs incomplets* obtiennent des notes meilleures que les autres ; le raisonnement intuitif sous-jacent est qu'un élève qui n'a que deux matières à

préparer peut y consacrer plus d'énergie et donc, en moyenne, obtenir de meilleurs résultats. Ce n'est pas le cas.

On a déjà insisté sur la grande variabilité de ces 15 élèves. Or, l'examen du tableau 5.3 montre que parmi eux, les huit candidats libres obtiennent presque systématiquement les plus mauvaises notes. D'où l'idée d'étudier séparément, parmi les *bacs incomplets*, les candidats libres et les lycéens. Le tableau **10.4** rassemble les comparaisons entre chacun de ces deux sous-groupes et la population globale (975).

groupe concerné	matière	moyenne du groupe	moyenne générale	différence moyenne	probabilité associée
bac incomplet	maths	14.43	13.08	1.35	.016
lycéens (7)	physique	12.57	10.87	1.70	.062
Bac incomplet	maths	9.00	13.08	−4.08	$1.8\ 10^{-4}$
cand. libre (8)	physique	6.13	10.87	−4.74	$1.2\ 10^{-4}$
candidats	maths	8.46	13.08	−4.62	$6.8\ 10^{-6}$
libres (13)	physique	6.46	10.87	−4.41	$1.2\ 10^{-7}$

Tableau 10.4. *Caractérisation de groupes d'élèves par rapport à la population totale des 975 élèves ayant des notes.*

- Les *bacs incomplets* inscrits dans un lycée obtiennent, en moyenne, des notes supérieures à celles des autres. Mais compte tenu de l'effectif faible de ce groupe et de la variabilité globale des notes, cette différence n'est pas très remarquable (probabilités associées petites mais pas très petites).
- Les *bacs incomplets candidats libres* obtiennent, en moyenne, des notes très inférieures à celles des autres.

D'où l'idée de faire apparaître dans le tableau **10.4** les 13 candidats libres de l'ensemble du fichier ; ils obtiennent aussi, en moyenne, des notes très inférieures à celles des autres.

Remarques

Effet de structure. Les *bacs incomplets* obtiennent, en moyenne, des notes plus basses que les autres. Ce résultat provient d'un effet de structure : les *bacs incomplets* contiennent proportionnellement beaucoup de candidats libres, qui eux obtiennent globalement de mauvais résultats.

Causalité. On peut être tenté de dire que c'est **parce qu**'ils ne bénéficient pas de l'enseignement au lycée que les candidats libres obtiennent de mauvais résultats. En fait, cette interprétation causale n'est qu'une hypothèse, raisonnable il est vrai, parmi d'autres. On peut en effet imaginer que les raisons qui font qu'un élève ne s'inscrive pas dans un lycée, par exemple une mauvaise santé, induisent elles-mêmes des mauvais résultats. Cette hypothèse aussi est raisonnable. Ceci illustre l'intérêt de distinguer le fait statistique, indiscutable sur la vue des chiffres, des interprétations dans lesquelles interviennent des connaissances extérieures aux données.

Comparaison entre plusieurs sous-populations : les élèves d'un même lycée

11.1 Sur quelles variables fonder la comparaison ?

Dans le chapitre 9, la comparaison entre les lycées est abordée en commentant leurs emplacements sur les composantes principales. Cette description des différences entre lycées est qualitative. Dans ce chapitre, on veut aller plus loin et proposer différents points de vue pour quantifier ces différences.

Comparer finement les lycées pour chacune des notes est nécessairement lourd et pas forcément instructif. Il peut être intéressant d'utiliser des variables synthétiques. Là encore le choix ne s'impose pas. Les composantes principales sont les meilleures synthèses du point de vue d'un certain critère : représentant le mieux possible les variables, elles aident à ne rien oublier d'important. Mais ce ne sont pas forcément les meilleures pour répondre à nos questions.

La première composante exprime le niveau général, au bac et pendant l'année. Pour comparer les lycées, il est intéressant de disposer d'une mesure du niveau général limitée aux notes du bac. On pourrait pour cela prendre la première composante d'une analyse en composantes principales des seules notes du bac, mais il est plus simple et plus naturel d'utiliser la note obtenue au bac avec les coefficients imposés aux différentes matières (déjà notée *bac*).

La seconde composante, opposition entre profil plutôt littéraire et plutôt scientifique, ne semble pas particulièrement intéressante pour comparer les lycées, et encore moins la troisième et la quatrième qui séparent histoire, philosophie et sciences naturelles.

En revanche, la cinquième composante, qui montre une différence systématique entre notes du bac et notes trimestrielles, évoque une question posée dès le début de l'étude : certains lycées sous-notent-ils tandis que d'autres surnotent ? Ceci suggère d'utiliser la différence entre la note du bac et celle obtenue au troisième trimestre avec les mêmes coefficients.

Finalement, nous organisons la comparaison entre les lycées principalement à partir des variables suivantes :

- *bac* : moyenne pondérée des notes du bac ;
- *bac*–$3^{ème}t.$: différence entre *bac* et la moyenne pondérée (avec les coefficients du bac) des notes du troisième trimestre ($3^{ème}t.$).

Les candidats libres n'étant que deux et ne formant pas vraiment un lycée, nous ne les faisons pas intervenir dans ce chapitre qui ne concerne donc que les 907 élèves ayant toutes leurs notes et inscrits dans un lycée.

11.2 Que signifie « comparer plusieurs sous-populations » ?

Nous travaillons d'abord avec une seule variable, la plus intéressante à notre avis étant la note au bac. Comparer les lycées avec une seule variable est une problématique générale : cela revient ici à se demander si certains lycées sont meilleurs et d'autres moins bons dans leurs résultats. Nous distinguons 2 sous-objectifs.

- *Faire une partition des lycées* suivant ce critère de niveau : par exemple les « bons », les « moyens », les « mauvais » voire les « très mauvais ». Ceci donnerait à chaque lycée une étiquette quelquefois lourde à porter mais aiderait les parents dans le choix d'un « bon lycée » pour leurs enfants, si tant est que le niveau moyen obtenu dépende plus de l'encadrement du lycée que de la population qu'il touche. Un des aspects de cet objectif est de détecter les lycées « hors norme », c'est-à-dire particulièrement bons ou particulièrement mauvais. Comme nous disposons de plusieurs valeurs pour chaque lycée (une par élève), le problème n'est pas si simple qu'il paraît : 4 paragraphes sont nécessaires pour proposer diverses solutions qui se complètent les unes les autres.
- *Evaluer par un indicateur global* l'ensemble des différences de niveau entre les 22 lycées.

Nous envisageons ci-après la comparaison entre les lycées selon :

- la variable *bac* ;
- deux variables simultanément : *bac* et *bac*–$3^{ème}$*t*.
- toutes les notes, du bac et de l'année.

11.3 Comparaison directe des moyennes sur une variable : la note du bac

L'idée la plus directe (et la plus simple) pour comparer les résultats des lycées au bac est de calculer la moyenne obtenue par les élèves de chaque lycée et de comparer ces moyennes entre elles. On étudie alors pour 22 individus (les lycées) la valeur d'une variable (moyenne de leurs élèves au bac), données que l'on peut regrouper dans un tableau (cf. **Tab. 11.1** qui comprend aussi deux indicateurs explicités plus loin). Pour le décrire, on peut utiliser les méthodes introduites au chapitre 5.

lycée	7	18	10	1	21	17	19	4	15	14	5
effectif	11	20	38	27	28	47	71	65	53	53	25
moyenne	9.13	10.08	10.34	10.35	10.46	10.46	10.54	10.57	10.71	10.77	11.10
probabilité	.002	.018	.012	.031	.050	.015	.008	.014	.061	.087	.419
valeur-test	-2.88	-2.10	-2.27	-1.86	-1.65	-2.18	-2.43	-2.20	-1.54	-1.36	-.20

lycée	G	20	3	2	22	9	13	8	16	12	11	6
effectif	907	19	77	19	23	28	47	65	17	41	51	82
moyenne	11.20	11.27	11.30	11.37	11.41	11.49	11.68	11.80	11.82	12.15	12.33	12.45
probabilité		.451	.352	.380	.335	.260	.080	.018	.143	.005	$3\ 10^{-4}$	$4\ 10^{-7}$
valeur-test		.12	.38	.30	.43	.64	1.41	2.09	1.07	2.59	3.46	4.96

Tableau 11.1. *Moyenne de la note au bac pour chaque lycée : données triées par moyenne croissante (G : ensemble des 907 élèves).*

Les données peuvent être représentées par un graphique (cf. **Fig. 11.1**).

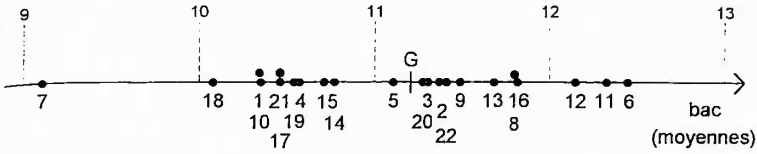


Figure 11.1. *Moyenne de la note au bac pour chaque lycée : représentation axiale. G : ensemble des élèves.*

La moyenne des élèves du lycée 7 est très basse (9.09 à comparer à la moyenne générale 11.20) ; c'est le seul lycée où, en moyenne, les élèves n'obtiennent pas 10. Sur le graphique, ce point se détache très nettement sur la gauche. On peut considérer que, du point de vue de la moyenne au bac, ce lycée est un outlier.

Pour décrire globalement ces données, il peut être pratique de regrouper les lycées selon leurs moyennes. Le choix des limites définissant ces groupes est toujours arbitraire. De toute façon, il est clair que le lycée 7, très mauvais, doit être détaché des autres.

On peut diviser les 21 autres lycées en deux groupes : ceux qui ont une moyenne supérieure à la moyenne générale et ceux qui ont une moyenne inférieure à la moyenne générale.

On peut aussi les diviser en trois groupes en remarquant deux petites « discontinuités » (écarts entre 2 lycées un peu plus importants que les autres) :

- une discontinuité entre les lycées 16 et 12 suggère de définir un groupe de 3 « très bons lycées » comprenant les lycées 12, 11 et 6 ;
- une discontinuité entre 14 et 5 suggère de définir 9 « mauvais lycées » comprenant le lycée 14 et ceux situés à sa gauche ;
- un groupe de 9 « lycées moyens ou bons », compris entre les lycées 5 et 16.

Comme lors de l'étude d'une seule sous-population (cf. 10.2), on peut se contenter de cette description des seules moyennes. Que ceci ait un intérêt nécessite une opinion a priori sur l'importance des valeurs des différences entre les moyennes. En outre, en procédant ainsi, on ne se préoccupe pas des disparités importantes à l'intérieur des lycées (un lycée peut regrouper des élèves de même niveau ou de niveaux très différents) ni du fait que ces moyennes sont calculées à partir d'effectifs très variables d'un lycée à l'autre. Les paragraphes suivants nuancent et enrichissent la simple comparaison des moyennes.

11.4 Probabilité associée à une moyenne calculée pour une sous-population

Les moyennes sont calculées sur des effectifs variant de 11 à 82. Or, il est d'autant plus « facile », pour un lycée, de s'éloigner de la moyenne générale que son effectif est faible (à la limite, on ne s'étonne pas de voir une note très élevée ou très faible pour un seul élève). Pour repérer les lycées dont la moyenne est "anormalement" grande (ou petite) et les séparer de ceux dont la moyenne ne s'écarte de la moyenne générale que d'une valeur en rapport avec les fluctuations banales de la répartition des élèves, on peut utiliser l'indice introduit en 10.2 et la fiche 7. Calculé pour le lycée i comportant n élèves, cet indice est la probabilité d'obtenir, en tirant au hasard les n élèves du lycée i, une différence entre la

moyenne du lycée i et la moyenne générale, au moins aussi importante que celle effectivement observée. L'intérêt de cet indicateur est qu'il tient compte à la fois :

- de la différence entre la moyenne de i et la moyenne générale ;
- de l'effectif n ;
- de la variabilité globale des élèves de l'ensemble des lycées.

Cette probabilité figure dans le tableau **11.1**. Elle nuance le calcul des moyennes. Globalement, les lycées dont la moyenne est très éloignée de la moyenne générale présentent une probabilité associée faible. Dans le détail, on constate quelques inversions. Par exemple, du point de vue de la moyenne, 18 est un lycée plus remarquable que 19 du fait de sa valeur plus faible ; mais, compte tenu des effectifs très différents entre ces deux lycées (20 et 71), le lycée 19 est plus remarquable au sens de non fortuit (i.e. au sens de la probabilité d'observer une moyenne au moins aussi basse dans le cadre d'un tirage au hasard). Ce raisonnement est identique à celui du paragraphe 10.2.

De même que pour les moyennes, il est utile de représenter graphiquement les probabilités associées. Mais une représentation directe conduirait à superposer pratiquement toutes les probabilités les plus faibles : l'écart entre .1 et .15, valeurs ici équivalentes, serait 50 fois plus grand que l'écart entre .001 et .000001 (valeurs correspondant à des degrés d'improbabilité très différents) ; ceci ne peut convenir. On peut songer à utiliser une transformation logarithmique des probabilités, qui revient à considérer équivalents les écarts entre 10^{-1} et 10^{-2} d'une part et 10^{-3} et 10^{-4} d'autre part.

Une idée intéressante consiste à utiliser une transformation fondée sur la distribution normale. A une probabilité p, on associe la valeur $vt(p)$, dite valeur-test, telle que : $P[X \geq vt(p)] = p$ avec X distribuée selon une loi normale centrée-réduite. Par exemple, à la probabilité $p = .025$ on associe $vt(p) = 1.96$. En outre, on affecte à la valeur-test le signe + ou – selon que la moyenne étudiée est supérieure ou inférieure à la moyenne générale. Ces valeurs-tests figurent dans le tableau **11.1**.

Ainsi, la valeur-test de 4.96 (resp –2.88) pour le lycée 6 (resp. 7) signifie que l'écart entre ce lycée et la moyenne générale est du même ordre (en terme de probabilité) que l'observation d'une valeur située à plus de 4.96 (resp. 2.88) écarts-types au-dessus (resp. au-dessous) de la moyenne dans le cadre d'une loi normale.

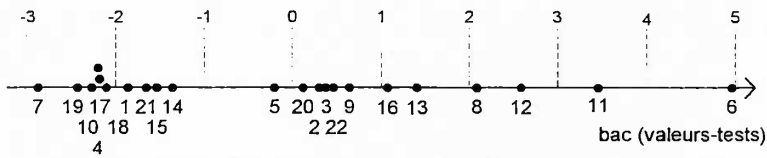


Figure 11.2. *Moyenne de la note au bac pour chaque lycée : représentation axiale des valeurs-tests.*

La représentation graphique de ces valeurs-tests (cf. **Fig. 11.2**) donne une image voisine de celle des moyennes. Mais le point de vue adopté est différent : ainsi le caractère remarquable du lycée 7 est atténué du fait de son faible effectif (11 élèves) ; celui du lycée 6 est amplifié du fait de son fort effectif (82 élèves).

Dans le paragraphe 11.3, nous avons raisonné sur les moyennes, comme s'il s'agissait d'une variable quantitative définie directement sur le lycée et non pas indirectement en calculant les moyennes de sous-populations. Le présent paragraphe a montré comment ce point de vue peut être enrichi.

11.5 Comparaison entre les répartitions des notes des lycées

La moyenne, même en tenant compte des effectifs différents, ne résume que très approximativement la répartition des notes de chaque lycée. Une comparaison complète des lycées implique une comparaison des répartitions. Pour cela, on peut utiliser plusieurs types de graphiques. Citons-en quelques-uns en commentant leur aspect pratique.

Juxtaposition des histogrammes

Il est difficile de comparer 22 histogrammes. Quelle que soit la façon de les juxtaposer, le résultat n'est guère lisible lorsque le nombre de sous-populations dépasse 3 ou 4. En outre, les effectifs faibles de certains lycées rendent leurs histogrammes peu stables. Enfin, un tel ensemble de graphiques est très encombrant et nous ne le donnons pas ici.

Juxtaposition des représentations axiales

Ce point de vue est plus adapté que le précédent lorsque le nombre de points est faible, ce qui est le cas pour plusieurs lycées. Pour les autres lycées, il implique de nombreuses superpositions des points (ceux des élèves d'un même lycée ayant des moyennes identiques ou très proches). De ce fait, il est difficile de représenter la densité des points de façon fidèle. Ici encore, un tel ensemble de graphiques est volumineux et nous ne le donnons pas.

Juxtaposition des boîtes de dispersion

La boîte de dispersion est ici le mode de représentation le plus pratique (cf. **Fig. 11.3**). D'autant plus que l'on peut ordonner les lycées selon un critère qui rend l'exploration visuelle plus efficace. Ici, les lycées ont été ordonnés par moyenne décroissante.

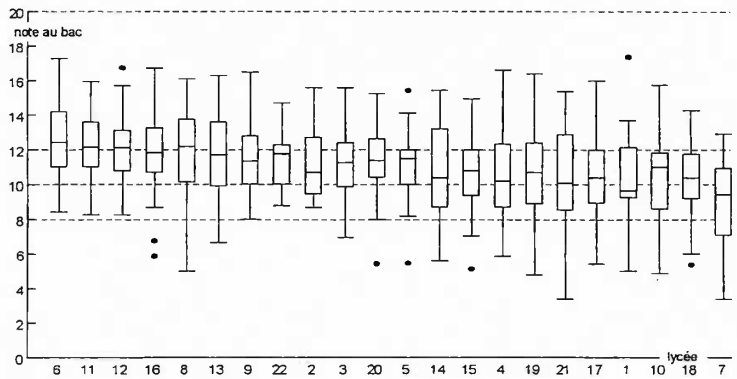


Figure 11.3. *Répartition de la note au bac selon les lycées : boîtes de dispersion juxtaposées.*

Cette juxtaposition met en évidence une grande variabilité intra-lycée. Les minimums et maximums montrent que l'on peut observer de mauvais résultats dans les « bons » lycées et de bons résultats dans les « mauvais » lycées. Mais ces valeurs remarquables ne doivent pas faire oublier la tendance lourde que constitue l'existence de plus ou moins bons lycées : dans les quatre meilleurs lycées, pratiquement les trois quarts des élèves ont obtenu plus de 11 ; dans les trois plus mauvais lycées, moins d'un quart des élèves dépassent 12.

11.6 Comparaison entre les extrema

C'est une chose d'être en moyenne un très bon lycée, c'en est une autre de n'avoir que de bons élèves. Dans le second cas, la note la plus basse au bac est élevée. Inversement, si un lycée n'admet que des mauvais élèves, la note la plus élevée est très basse. Il est donc intéressant d'examiner dans cette optique les valeurs maximum et minimum pour chaque lycée. Ces valeurs sont visibles sur la figure **11.3**.

Les minima sont souvent de l'ordre de 5 (un peu plus faible pour deux lycées : le 7 déjà repéré et le 21 qui n'est pas très bon en moyenne). Pour 6 lycées (2,6,9,11,12 et 22), le minimum dépasse 8 : ils n'ont aucun très mauvais élève. On retrouve d'ailleurs parmi eux les trois "bons" repérés par l'indice du paragraphe 11.3.

Pour les maxima, à l'inverse, ce sont les valeurs faibles qui sont a priori intéressantes, en indiquant par exemple l'absence de bons élèves. Le lycée 7, très faible en moyenne, se fait encore remarquer pour son maximum de 12.94 particulièrement faible.

11.7 Liaison entre une variable quantitative et une variable qualitative

On cherche ici à répondre à une question un peu différente des précédentes. Peut-on mesurer **globalement** la différence entre les lycées du point de vue de la note au bac ?

Cette question renvoie au problème général de la mesure de la liaison entre une variable qualitative (ici *lycée*) et une variable quantitative (ici *bac*). En effet, si un élève provient d'un très bon lycée, il a probablement une bonne note ; réciproquement s'il a une bonne note, il provient probablement d'un bon lycée. Dans ce cas, les deux variables *lycée* et *bac* sont liées puisque la connaissance de la valeur de l'une donne des renseignements sur l'autre. A l'inverse, si la répartition des notes diffère peu d'un lycée à l'autre, les deux variables ne sont pas liées.

Construire un indicateur qui mesure une telle liaison est complexe. La fiche 8 détaille ce sujet. Nous nous contentons ici de dire qu'il existe un indicateur qui mesure le caractère fortuit de la liaison observée, indicateur fondé sur le même principe que celui utilisé pour caractériser une seule sous-population. Il exprime la probabilité d'obtenir des différences (entre les 22 moyennes) globalement au moins aussi grandes que celles effectivement observées si l'appartenance des élèves aux 22 lycées était tirée au hasard (en respectant les effectifs des lycées).

Pour la note au bac, cette probabilité est extrêmement faible (inférieure à 10^{-9}) : la différence globale entre les lycées est importante, autrement dit les deux variables *bac* et *lycée* sont liées. Ce résultat offre ici un intérêt limité après l'étude détaillée des moyennes. Cet indicateur, étant plus global que ceux qui concernent chacun des lycées, est forcément moins riche : en particulier l'« anormalité » d'un seul lycée peut être noyée dans la masse générale. En revanche, cet indicateur global est utile dans des comparaisons entre liaisons (cf. § 11.9).

11.8 Comparaison selon deux variables

Il est intéressant de rapprocher les comparaisons entre lycées fondées sur les deux variables synthétiques : le bac et sa différence avec le troisième trimestre. Pour le bac, l'étude a été détaillée ; l'étude séparée de la différence *bac–3ème trimestre* (cf. **Tab. 11.2**) sera déduite des représentations simultanées des deux variables.

lycée	7	2	4	18	10	21	20	9	1	14	13
effectif	11	19	65	20	38	28	19	28	27	53	47
minimum	-3.39	-2.35	-4.09	-3.03	-2.25	-1.94	-4.84	-3.63	-2.81	-2.69	-2.63
moyenne	-1.89	-.68	-.61	-.38	-.24	-.23	-.15	-.14	-.04	.14	.16
maximum	-.70	1.19	2.91	1.63	2.31	3.19	3.15	1.44	2.91	2.92	3.30

lycée	5	17	G	16	3	15	12	19	6	11	8	22
effectif	25	47	907	17	77	53	41	71	82	51	65	23
minimum	-2.75	-2.06	-4.84	-2.24	-3.00	-1.84	-1.87	-2.13	-3.16	-2.27	-2.59	.13
moyenne	.17	.17	.24	.25	.29	.38	.56	.64	.65	.72	.96	1.26
maximum	2.25	2.28	4.68	2.64	3.31	4.68	2.68	3.07	3.88	3.16	3.60	2.94

Tableau 11.2. *Variable* bac-3t. *: moyenne, minimum et maximum pour chaque lycée. Les lycées sont triés par moyenne croissante. G : ensemble des élèves.*

Le plus simple est de construire deux graphiques, chacun croisant ces deux variables, en représentant chaque lycée par ses deux moyennes d'une part, et par ses deux valeurs-tests d'autre part (cf. **Fig. 11.4.** et **11.5**). En ne considérant qu'un axe de coordonnée, on étudie une seule variable (la coordonnée selon l'axe horizontal fournit les figures 11.1 et 11.2).

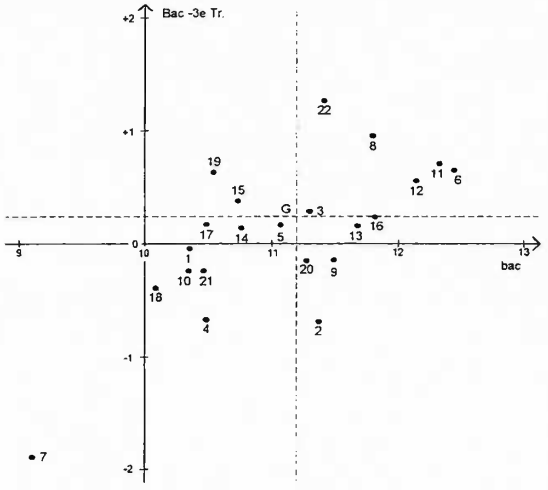


Figure 11.4. *Note au bac et différence* bac-3ème trimestre *: représentation plane des 22 lycées par leurs deux moyennes. G : ensemble des élèves.*

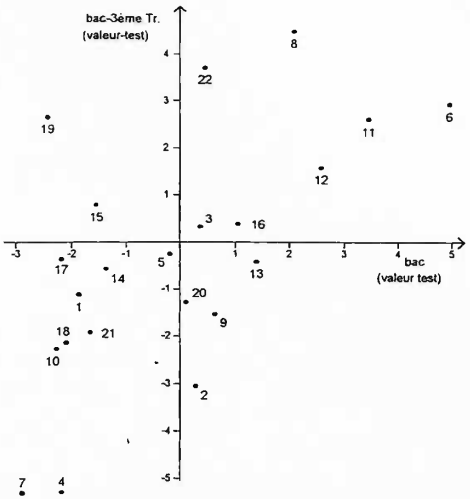


Figure 11.5. *Note au bac et différence* bac–3ème trimestre : *représentation plane des 22 lycées par leurs deux valeurs-tests.*

Comparaison entre les lycées selon la variable bac–3ème trimestre

On se limite ici à la coordonnée sur l'axe vertical.

En moyenne, la différence entre le bac et le 3ème trimestre est de .24 points : il y a une très légère sous-notation au 3ème trimestre.

Les lycées situés en haut ont une note moyenne du bac supérieure à celle du 3ème trimestre : par rapport au bac, leurs élèves ont été sous-notés au 3ème trimestre. Inversement, les élèves des lycées situés en bas ont été surnotés au 3ème trimestre. Finalement, les élèves ont été sous-notés dans 8 lycées et surnotés dans les autres ; le lycée 7 (toujours lui !) s'écarte nettement des autres : ses élèves ont été surnotés au 3ème trimestre de presque 2 points.

Suivant la valeur-test calculée par lycée, on distingue :

- les lycées qui surnotent nettement ; ce sont, dans l'ordre de probabilité croissante, les lycées 7, 4 et 2 ;
- les lycées qui sous-notent nettement : 8, 22, 6, 19 et 11.

On retrouve ici exactement ceux dont les moyennes sont les plus faibles ou les plus fortes ; se baser directement sur la moyenne ou la valeur-test ne change pas ici fondamentalement les conclusions.

Un coup d'œil sur les maxima et les minima (cf. **Tab. 11.2**) permet de repérer :

- un lycée dont la valeur maximum est négative ; le lycée 7 (toujours lui) surnote toujours d'au moins .7 points ;
- un lycée dont la valeur minimum est positive ; le lycée 22 sous-note systématiquement d'au moins .11 points.

Comparaison entre les lycées selon les deux variables simultanément

Le lycée 7 s'écarte beaucoup des autres.

Le nuage est plutôt allongé suivant la première bissectrice (en prenant G comme origine des axes pour le graphique des moyennes) : plus la moyenne d'un lycée est élevée, plus il a tendance à sous-noter ses élèves durant l'année. D'où vient ce phénomène ? Ici encore, ce n'est pas au statisticien de répondre, il n'a fait que le mettre en évidence.

11.9 Comparaison selon l'ensemble des variables : caractérisation d'un lycée

Nous comparons ici les lycées sur les 20 notes et les 2 variables synthétiques.

Indice global

L'indice global de liaison entre une variable quantitative et une variable qualitative introduit en 11.7 peut être calculé pour toutes les variables disponibles ; cela met en évidence les variables pour lesquelles les lycées se différencient globalement le plus (cf. **Tab. 11.3**). Ceci est surtout intéressant pour repérer des anomalies quand on pense a priori que les sous-populations sont analogues (auquel cas aucune note ne serait liée à la variable *lycée*). Les variables "les plus anormales", c'est-à-dire ici les plus liées à la variable *lycée*, sont alors à étudier plus finement avec les techniques des premiers paragraphes (l'intérêt de l'indicateur est principalement de guider le choix des variables à analyser finement).

	proba	v. test	note
1	$<10^{-9}$	13.23	hist. géo. 1er t.
2	$<10^{-9}$	9.48	philo. 2ème t.
3	$<10^{-9}$	9.35	hist. géo. 2ème t.
4	$<10^{-9}$	9.23	*bac−3ème t.*
5	$<10^{-9}$	8.51	hist. géo. bac
6	$<10^{-9}$	8.46	philo. 1er t.
7	$<10^{-9}$	8.01	maths bac
8	$<10^{-9}$	7.77	hist. géo. 3ème t.
9	$<10^{-9}$	6.77	sc. nat. 1er t.
10	$<10^{-9}$	6.76	maths 1er t.
11	$<10^{-9}$	6.50	philo. 3ème t.
12	$<10^{-9}$	6.40	sc. nat. bac

	proba	v. test	note
13	$<10^{-9}$	6.37	*bac* (sans options)
14	$<10^{-9}$	6.26	physique 3ème t.
15	10^{-9}	6.02	physique 1er t.
16	$8\ 10^{-9}$	5.65	maths 3ème t.
17	$3\ 10^{-8}$	5.45	physique 2ème t.
18	$5\ 10^{-8}$	5.25	physique bac
19	$2\ 10^{-7}$	5.11	philo. bac
20	$3\ 10^{-7}$	5.00	sc. nat. 3ème t.
21	10^{-6}	4.74	*3ème t.*
22	$5\ 10^{-4}$	3.31	sc. nat. 2ème t.
23	10^{-3}	3.09	maths 2ème t.

Tableau 11.3. *Notes triées de façon décroissante selon leur degré de liaison avec la variable* lycée.
v. test : P[X>v. test]=proba ; avec X distribuée selon une loi normale centrée-réduite.

Ici, toutes les notes ainsi que les 2 variables synthétiques diffèrent en moyenne sur l'ensemble des lycées. On ne trouve pas de structure forte montrant qu'une matière (via ses 4 notes) différencie plus les lycées que les autres ou bien que le bac (via ses 5 notes) différencie plus ou moins que l'année (via ses 15 notes). Cette absence de structure forte est aussi un résultat. On note quand même quelques tendances. C'est l'histoire-géographie au premier trimestre qui différencie le plus les lycées. On remarque aussi que les 4 notes d'histoire-géographie sont parmi les premières. La différence entre le bac et le troisième trimestre est également dans les premières, bien avant le bac lui-même : d'un lycée à

l'autre, les différences entre sous-notations (ou surnotations) sont donc plus importantes que les différences entre les moyennes du bac.

Exemple de caractérisation d'un lycée

Il est souvent très instructif de chercher pour chaque sous-population, ici chaque lycée, ce qui la différencie le plus de la population totale et qui donc finalement la caractérise. En reprenant l'indice de probabilité défini en 10.2 (et rappelé en 11.4) qui permet de détecter les notes en moyenne anormalement basses ou hautes, et en ordonnant ces probabilité par valeur décroissante pour chaque lycée, on obtient une description synthétique rapide et claire de chaque lycée.

proba.	v.-test	lycée	ensemble	note
$2\ 10^{-6}$	4.64	10.08	9.01	philosophie au 2ème trimestre
$7\ 10^{-5}$	3.82	10.23	9.26	philosophie au 3ème trimestre
$3\ 10^{-4}$	−3.44	10.24	11.04	hist. géo. au 1er trimestre
.003	2.74	8.83	7.85	philosophie au bac
.005	2.60	9.56	8.99	philosophie au 1er trimestre
.006	2.53	11.43	10.64	maths au 1er trimestre

Tableau 11.4. *Notes triées de façon décroissante selon leur degré de caractérisation du lycée 3.*

Ainsi, par exemple, le lycée numéro 3 se caractérise par (cf. **Tab 11.4**) :

- des moyennes en philosophie supérieures à la moyenne générale, au bac et aux 3 trimestres ;
- une moyenne en mathématiques au premier trimestre supérieure à la moyenne générale ;
- une moyenne en histoire-géographie au 1er trimestre inférieure à la moyenne générale.

Chaque lycée peut être ainsi décrit. Nous ne reproduisons pas ces résultats et indiquons seulement que chacun est caractérisé par au moins une note. Les bons lycées repérés par leur moyenne au bac ont évidemment, pour la plupart des notes, des moyennes supérieures à la moyenne générale, mais les résultats sont ici plus fins : le lycée numéro 6 par exemple, déjà signalé pour sa forte valeur pour la variable *bac* (cf. **Fig. 11.2**), est caractérisé par une forte valeur des moyennes de toutes les matières au bac, sauf en sciences naturelles.

11.10 Conclusion

Nous avons abordé la comparaison des lycées suivant beaucoup de points de vue sans prétendre faire un panorama complet des techniques. Nous avons présenté celles qui nous paraissaient les plus simples et les plus maniables.

Une critique encore sur ce travail. Pour le lycée 13 dont 12 élèves sur les 61 n'ont pas été pris en compte car ils avaient des données manquantes (cf. **Tab. 10.2**), les résultats sont à manier avec précaution. Mais que faire de mieux ? En cas de données manquantes, aucune solution n'est parfaite.

Partie 2

Eléments remarquables et éléments aberrants

Introduction à la deuxième partie

La mise en évidence d'éléments remarquables intervient tôt dans l'étude d'un fichier. Nous y faisons d'ailleurs souvent allusion dans les chapitres précédents. Mais sa présentation un tant soit peu détaillée nécessite des notions telles qu'elle ne pouvait s'intégrer dans les premiers chapitres sans rompre le fil de l'exposé.

Qu'est-ce qu'un élément remarquable ?

Lors de l'analyse d'un tableau de données du type de celui de notes, on manipule trois types d'éléments :

- les colonnes, qui représentent les variables, soit ici les épreuves (e.g. *mathématiques au bac*) ;
- les lignes, qui représentent les individus, soit ici les élèves ;
- les valeurs, croisements entre une ligne et une colonne, soit ici les notes.

Toute analyse statistique d'un tableau est une étude de sa variabilité (selon les analyses, on focalise l'attention sur tel ou tel aspect de la variabilité) ; une composante importante de cette variabilité est constituée par les éléments les plus remarquables (c'est-à-dire différents des autres) du tableau. Ainsi, on cherche à mettre en évidence :

- des *valeurs remarquables*, c'est-à-dire très éloignées des autres valeurs de la même variable ;
- des *individus remarquables*, dont les valeurs, dans leur ensemble, diffèrent de celles des autres individus ;
- des *variables remarquables*, dont la distribution diffère de celle des autres variables.

Quelles que soient les données et la problématique, il est prudent d'examiner tour à tour ces trois éléments dans l'étude systématique d'un fichier. Les éléments remarquables sont aussi appelés outliers ; nous utilisons indifféremment les deux termes.

Pourquoi rechercher les éléments remarquables ?

On recherche des éléments remarquables avec trois finalités.

Description des données

La présence d'éléments remarquables est l'une des caractéristiques importantes des données. Leur mise en évidence permet de :

- délimiter le champ de ce qui est *effectivement* possible en comparaison avec le champ de ce qui est *a priori* possible ; par exemple, 20 est une note a priori possible dans toutes les matières mais en pratique attribuée seulement en mathématiques ;
- orienter des études ponctuelles ou des monographies ; c'est par exemple le point de vue du journaliste qui veut réaliser un article sur un élève ayant obtenu un palmarès particulièrement prestigieux.

Détection d'erreurs

Deux exemples.

- En l'absence de détection systématique des valeurs hors plages, la recherche de valeurs remarquables conduit immédiatement à la valeur 22 existant dans le fichier de notes.
- En l'absence d'une interrogation systématique quant aux données manquantes, la présence de nombreuses valeurs 0 chez certains individus aurait conduit d'abord à les mettre en évidence en tant qu'outliers puis à découvrir que cette valeur était utilisée pour coder l'absence de valeur.

Mise en évidence d'éléments aberrants ; redéfinition du champ de l'étude

Certains éléments remarquables peuvent être qualifiés d'aberrants, c'est-à-dire non pas impossibles (ce seraient alors des erreurs manifestes) mais improbables compte tenu de l'ensemble du fichier. Ainsi une note de 20 en mathématiques au bac est tout à fait remarquable mais n'est pas improbable lorsque l'on passe en revue un très grand nombre d'élèves. En revanche, un élève qui obtiendrait 20 en mathématiques au bac après avoir eu 5 de moyenne toute l'année est non seulement remarquable mais aussi très improbable, même en passant en revue un millier d'élèves.

Après examen, il arrive que ces éléments puissent être identifiés comme erronés et être corrigés ; il arrive aussi que des individus soient éliminés du fichier, parce que :

- ils ne font pas partie du champ de l'étude et ont été recueillis indûment ;
- ils incitent à penser que le champ de l'étude est trop hétérogène et doit être redéfini.

Dans le fichier des notes, la mise en évidence d'éléments remarquables n'a conduit à aucune élimination. En revanche, un exemple de redéfinition du champ de l'étude est fourni par l'analyse des données manquantes : à l'issue de cette analyse, le champ initial « élèves ayant passé le bac C » est devenu « élèves ayant passé le bac C et ayant toutes les notes ».

Mise en évidence de valeurs remarquables et de valeurs aberrantes

12.1 Examen systématique des distributions

Pour mettre en évidence les valeurs remarquables, la démarche de base consiste à examiner une par une les distributions des variables. Cet examen se fait principalement à partir des histogrammes. On peut ainsi apprécier l'éloignement de n'importe quelle valeur par rapport aux autres, critère intuitif pour juger du caractère remarquable d'une valeur.

Dans le même esprit, on utilise les boîtes de dispersion, qui individualisent les valeurs remarquables (cf. 7.7).

Cet examen systématique a déjà été recommandé.

12.2 Intérêt du centrage et de la réduction

Principe

La notion de valeur remarquable ne pose pas problème lorsque l'on s'intéresse à une seule distribution. Une difficulté se présente lorsque l'on étudie simultanément plusieurs variables, cas du fichier des notes. On peut alors souhaiter rechercher des valeurs *remarquables toutes variables confondues*. Cette idée implique la comparaison de valeurs appartenant à des variables différentes, notion qui se trouve par exemple dans l'affirmation suivante (souvent entendue) : "Au bac C, 18 en mathématiques est une note moins remarquable que 18 en philosophie".

Cette affirmation implique de savoir comparer la place de la même note 18, dans la distribution des notes de mathématiques d'une part et de philosophie d'autre part ; elle sous-entend les deux faits suivants, souvent tenus pour vrais :

- *l'écart entre 18 et la moyenne des notes obtenues est plus petit en mathématiques qu'en philosophie*. Ce fait est vérifié au moins dans le fichier étudié : la moyenne est de 13.21 en mathématiques et de 7.84 en philosophie. Calculé pour tous les élèves et pour la matière k, cet écart constitue la variable centrée : $x_{ik}-\bar{x}_k$ (x_{ik} : note obtenue par l'élève i à l'épreuve k ; $\bar{x}_k$: moyenne de la variable k).
- *les notes sont plus resserrées autour de la moyenne en philosophie qu'en mathématiques*. La notion de dispersion autour de la moyenne peut être mesurée par l'écart-type qui vaut 3.19 pour les mathématiques et 3.29 pour la philosophie. A l'aune de l'écart-type, ce fait n'est pas vérifié dans ce fichier.

Ainsi, le caractère remarquable d'une note (et plus généralement d'une valeur) recouvre à la fois l'écart entre la valeur et la moyenne, et la dispersion de l'ensemble de la distribution autour de cette moyenne. Pour prendre en compte simultanément ces deux notions dans l'appréciation d'une valeur x_{ik}, on calcule la valeur centrée-réduite $(x_{ik}-\bar{x}_k)/s_k$, qui exprime l'écart à la moyenne en nombre d'écarts-types (s_k : écart-type de la variable k).

Exemple : la valeur centrée-réduite associée à 18 en mathématiques au bac est : (18–13.21)/3.19=1.50 . Pour cette matière, 18 est situé à 1,5 écart-type de la moyenne. La valeur correspondante en philosophie est : (18–7.84)/3.29=3.09 . Pour cette matière, 18 est situé à plus de 3 écarts-types de la moyenne. En ce sens, 18 en philosophie est effectivement plus remarquable (i.e. plus éloigné de la moyenne compte tenu de la dispersion de ces notes) que 18 en mathématiques.

Comparaison de valeurs centrées-réduites provenant de variables différentes

Le centrage et la réduction d'un tableau consiste à centrer et réduire toutes les variables du tableau. Cette opération permet de comparer des valeurs de variables différentes, car ces variables sont normalisées au sens suivant : toutes ont même moyenne (égale à 0 du fait du centrage) et même écart-type (égal à 1 du fait de la réduction). Le tableau **12.1** donne quelques exemples qui illustrent les influences respectives du centrage et de la réduction.

Cas	Matière	note	note centrée et réduite				
1	maths au bac	18	(18–13.21)/3.19	=	4.79/3.19	=	1.50
2	philo. au bac	18	(18–7.84)/3.29	=	10.16/3.29	=	3.09
3	maths au 1er tr.	18	(18–10.64)/2.85	=	7.36/2.85	=	2.58
4	philo. au 1er tr.	18	(18–8.99)/2.00	=	9.01/2.00	=	4.51
5	philo. au bac	13	(13–7.84)/3.29	=	5.16/3.29	=	1.57
6	philo. au 1er tr.	14	(14–8.99)/2.00	=	5.01/2.00	=	2.51

Tableau 12.1. *Exemples de valeurs, brutes et centrées-réduites, extraites du fichier des notes.*

Cas 1 et 3 : 18 en mathématiques est une note plus remarquable au 1er trimestre que le jour du bac ; ceci tient essentiellement au fait que les notes en mathématiques sont, en moyenne, plus élevées au bac qu'au 1er trimestre.

Cas 2 et 4 : 18 en philosophie est une note plus remarquable au 1er trimestre que le jour du bac ; ceci tient essentiellement à la dispersion des notes de philosophie, plus grande le jour du bac qu'au 1er trimestre.

Cas 1 et 5 : le jour du bac, il est à peu près aussi exceptionnel d'avoir 18 en mathématiques que 13 en philosophie (la note exacte de philosophie qui « correspond » à 18 en mathématiques est : 7.84+1.50×3.29=12.77).

12.3 Approche systématique pour mettre en évidence des valeurs remarquables

L'opération de centrage et de réduction d'un tableau permet de comparer entre elles des valeurs de variables différentes. D'où l'idée de rechercher les plus grandes (en valeur absolue et toutes variables confondues) valeurs centrées-réduites d'un fichier.

Si le tableau de données est de petite taille (quelques dizaines de lignes et de colonnes), cette recherche peut se faire visuellement à partir de l'impression (ou de l'affichage) du tableau complet centré-réduit. Si le tableau est de grande taille (cas du tableau de notes), cette recherche se fait par programme.

Le tableau **12.2** rassemble les 20 plus grandes (en valeur absolue) valeurs centrées-réduites du fichier des notes. Sans prétendre à l'exhaustivité, indiquons quelques commentaires.

rang	valeur c.r.	note	épreuve	élève
1	4.25	17.5	philo. 1er tr.	496
2	−3.84	1	maths bac	183
3	3.75	18	philo. 3ème tr.	75
4	3.75	16.5	philo. 1er tr.	75
5	−3.61	3.1	hist. 2ème tr.	134
6	−3.54	1	philo. 3ème tr	752
7	−3.48	1	hist. bac	796
8	−3.46	2.7	sc. nat. 2ème tr.	714
9	−3.45	1.6	maths 2ème tr.	183
10	3.39	19	philo. bac	75
11	3.39	19	philo. bac	142
12	3.39	19	philo. bac	580
13	3.39	19	philo. bac	664
14	−3.37	3	hist. 3ème tr.	785
15	3.32	17	philo. 3ème tr.	64
16	3.32	17	philo. 3ème tr.	124
17	3.30	15.6	philo. 1er tr.	84
18	3.30	15.6	philo. 1er tr.	169
19	3.29	16	philo. 2ème tr.	64
20	3.25	15.5	philo. 1er tr.	142

Tableau 12.2. *Les 20 valeurs les plus remarquables du fichier des notes.*

- Sur les 20 valeurs, 14 concernent la philosophie (note au bac ou note trimestrielle). Cette matière donne lieu à des distributions différentes de celles des autres. Une grande partie des valeurs remarquables a pour origine des distributions "remarquables".
- Quelques élèves apparaissent plusieurs fois : par exemple 75 apparaît trois fois et 183 deux fois. Une partie des valeurs remarquables a pour origine des individus "remarquables".

Si l'on s'intéresse à une valeur particulière, il convient de s'assurer de sa cohérence avec les autres informations disponibles sur l'élève qui l'a obtenue. Ainsi, l'élève 496 a toujours obtenu au-dessus de 13.5 en philosophie, que ce soit au bac ou pendant l'année ; de même, l'élève 183 a toujours obtenu moins de 4 en mathématiques ou en physique, et l'élève 75 a toujours obtenu au-dessus de 14.5 en philosophie ; il n'y a aucune raison de suspecter ces valeurs d'être des erreurs.

Pour la personne qui analyse ce fichier, ces cas peuvent mériter une étude particulière, par exemple lorsque l'on recherche des exemples frappants de bonnes ou de mauvaises copies.

12.4 Limites du centrage et de la réduction dans la recherche de valeurs remarquables

Le classement de l'ensemble des valeurs centrées-réduites d'un tableau ne nécessite aucune hypothèse à caractère probabiliste. Cependant, il faut garder à l'esprit que la réduction ne prend en compte que la dispersion globale d'une distribution, à l'exclusion de toutes autres caractéristiques (en particulier de forme de la distribution). Ainsi, dans le cas de distributions très dissymétriques, le classement obtenu peut ne pas correspondre à l'intuition du caractère remarquable des valeurs.

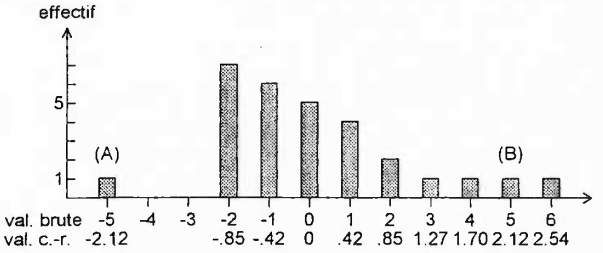


Figure 12.1. *Centrage et réduction dans un cas de distribution dissymétrique. L'échelle des abscisses est exprimée en valeurs brutes et en valeurs centrées-réduites.*

Dans la distribution illustrée **figure 12.1**, le point **A** n'est pas plus remarquable que le point **B** au sens de la valeur centrée-réduite (qui vaut −2.12 pour chacun). Pourtant, sur la vue du diagramme, **A** semble plus particulier que **B** car :

- **A** correspond à la plus petite valeur de la variable (valeur extrême), ce qui n'est pas le cas de **B ;**
- **A** est séparé des autres valeurs par un intervalle vide assez grand qui suggère une valeur remarquable ; ce n'est pas le cas de **B**.

L'origine de cette différence entre le calcul et l'intuition est due en bonne partie au fait que le calcul utilise la même unité de mesure (l'écart-type) pour les valeurs au-dessus et au-dessous de la moyenne, alors que, dans le cas d'une distribution dissymétrique, le même écart à la moyenne semble moins important s'il est du côté de la plus grande queue de distribution.

Cet exemple montre les limites d'un indicateur unique qui ne peut prétendre être parfaitement adapté à tous les cas. On prendra garde toutefois à **ne pas négliger un outil très commode sous prétexte qu'il n'est pas adapté à quelques cas particuliers**.

En pratique, la présence pour une variable d'une queue de distribution d'un seul côté se traduit par l'apparition, en début d'un classement du type de celui du tableau **12.2**, de plusieurs valeurs centrées-réduites (ayant le même signe) de cette variable. Réciproquement l'apparition, en début de classement, d'un ensemble de valeurs de même signe appartenant à une même variable, indique la présence d'une queue de distribution. Ainsi, même dans le cas d'une distribution dissymétrique, l'opération de centrage-réduction est utilisable dans toute première approche d'un tableau.

12.5 Appréciation du caractère aberrant d'une valeur

Probabilité associée à une valeur centrée-réduite et caractère remarquable d'une valeur

A une valeur centrée-réduite, on peut associer une probabilité, celle d'observer une valeur au moins aussi grande.

Une majoration de cette probabilité peut être obtenue, via l'inégalité de Bienaymé (cf. Fiche 3), sans aucune hypothèse sur la forme de la distribution à laquelle appartient la valeur. Mais en pratique, on calcule plutôt cette probabilité à l'aide d'une loi normale de même moyenne et de même variance que la distribution observée. Ainsi, à la valeur 2, on associe .046 (probabilité d'obtenir une valeur x telle que $|x| \geq 2$ dans le cadre d'une loi normale centrée-réduite ; la probabilité de .05 est associée à la valeur 1.96).

La probabilité de .05 est une approximation, la plupart du temps satisfaisante, de la probabilité de valeurs situées, par rapport à la moyenne, au-delà de 2 écarts-types : ainsi, dans le fichier des notes, quelle que soit la variable, cette probabilité (i.e. le pourcentage de notes situées à plus de deux écarts-types de la moyenne) est toujours comprise entre .024 et .059

Par ailleurs, elle est souvent jugée petite ; cela revient alors à considérer comme remarquables les valeurs centrées-réduites supérieures à 2.

Le même raisonnement, appliqué à la valeur 3 (3 écarts-types), conduit aux chiffres suivants (probabilités de valeurs situées au-delà de 3 écarts-types).

- Plage de variation des probabilités observées sur les 20 notes : entre 0 et 0.009
- Probabilité associée à la loi normale : 0.003

Plus généralement, dans un problème donné, on peut se fixer un seuil de probabilité pour définir une valeur comme remarquable.

Cas des boîtes de dispersion

Une boîte de dispersion fait apparaître de façon individualisée les valeurs dont la distance au rectangle central est supérieure à plus d'une fois et demie l'intervalle interquartile (cf. 7.7). De façon implicite cette individualisation incite à les considérer comme remarquable.

Si l'on applique cette règle à une distribution normale, cela revient à individualiser les valeurs situées à plus de 2.68 écarts-types de la moyenne ; on sélectionne ainsi 0,74% des valeurs.

Du remarquable à l'improbable : les valeurs aberrantes

Il est utile de disposer d'un indicateur désignant les valeurs aberrantes, c'est-à-dire présumées être des erreurs ou être hors du champ de l'étude, afin de les examiner de plus près. Un indicateur possible est la probabilité d'apparition de la valeur, ou plus précisément d'une valeur au moins aussi remarquable, dans un ensemble de données ayant la taille de celui qui est étudié. Le calcul réalisé dans les boîtes de dispersion ne suffit pas car il ne tient pas compte de l'effectif des valeurs observées : en effet, si l'on observe beaucoup de valeurs, on peut s'attendre à trouver des valeurs exceptionnelles.

Appliqué au fichier des notes, ce raisonnement conduit à un nouveau calcul de probabilité associée, présenté ci-après à partir de l'exemple de la valeur la plus remarquable du fichier des notes.

- Valeur centrée-réduite la plus remarquable (i.e. de plus grande valeur absolue ; cf. **Tab. 12.2**) : 4.25
- Probabilité d'observer une valeur au moins aussi grande que 4.25 dans le cadre d'un tirage dans une loi normale : 0.00002 ; cette probabilité très faible indique le caractère très remarquable de la valeur 4.25 (en effet, si l'on choisit au hasard un élève, la probabilité que sa moyenne en philosophie au 1er trimestre soit supérieure à 17.5 est très faible).
- Nombre de valeurs observées : 909 (élèves)×20 (notes/élève) = 18180 notes
- Probabilité d'observer une valeur au moins aussi remarquable que 4.25 dans le cadre de 18180 tirages dans une loi normale (on montre que la probabilité P d'observer un événement de probabilité p lorsque l'on réalise n épreuves, peut être approchée par np) : $0.00002 \times 18180 = 0.36$

En conclusion, la moyenne de 17.5 en philosophie au 1er trimestre (valeur centrée-réduite : 4.25) est très remarquable, mais compte tenu du nombre d'élèves et du nombre de notes par élèves, sa présence n'est pas aberrante dans notre fichier ; son observation est presque attendue et ne remet pas en cause l'homogénéité du champ de l'étude.

Valeur aberrante et valeur à éliminer

On peut être tenté de considérer une valeur aberrante, comme devant être automatiquement éliminée de l'étude. Ce point est très délicat car l'analyste est souvent confronté à l'opposition suivante :

- le concept de donnée aberrante est un recours tentant pour justifier l'élimination de données qui ne cadrent pas avec des idées a priori ;
- il est dommage de polluer un ensemble de données par l'ajout d'erreurs ou de données extérieures au champ de l'étude.

La règle d'or est de **ne retirer une valeur que si des raisons autres que statistiques justifient de la retirer**. En ce cas, il faut être capable d'expliciter la modification du champ de l'étude impliquée par le fait de retirer une valeur. Ce n'a pas été le cas dans notre étude.

En conclusion, les valeurs centrées-réduites sont des indicateurs désignant les valeurs à examiner avec attention ; il faut éviter les "règles de décision" spécifiant l'élimination automatique de toute valeur dépassant un seuil fixé.

Place de la loi normale dans la démarche présentée

L'association d'une probabilité à une valeur nécessite le recours à une distribution théorique. En l'absence d'informations a priori (cas général), on choisit la loi normale, car beaucoup de distributions observées s'apparentent suffisamment à une distribution normale pour que celle-ci puisse être utilisée dans le calcul approché d'une probabilité. D'une certaine manière, la probabilité (calculée à partir de la loi normale) associée à une valeur d'une variable mesure l'écart entre cette variable et la loi normale. Mais le fait d'utiliser cette probabilité comme un indicateur (et non comme un seuil de rejet automatique) autorise son calcul dans des situations éloignées de la loi normale. En effet, les principales règles d'interprétation sont les suivantes :

- l'association d'une très faible probabilité à une seule valeur d'une variable incite à examiner de plus près cette valeur ;

- l'association d'une très faible probabilité à plusieurs valeurs d'une même variable indique la présence d'une ou deux queues de distribution. Ces valeurs remarquables gagneront à être considérées ensemble ;
- l'absence, pour une variable, de valeurs associées à une très faible probabilité, indique l'absence de toute queue de distribution. Cette variable n'est pas le siège de valeurs remarquables.

Ainsi, dans la démarche présentée, la loi normale ne constitue absolument pas un modèle que les données doivent vérifier, mais :

- un moyen empirique généralement efficace d'effectuer des calculs approchés de probabilités ;
- une référence commode pour étudier des distributions.

Mise en évidence d'individus remarquables

13.1 Qu'est-ce qu'un individu remarquable ?

Un individu remarquable possède une ou plusieurs valeurs remarquables. Deux références servent pour juger du caractère remarquable de la valeur de l'individu *i* pour la variable *k*.

- *Les valeurs des autres individus pour la variable k* ; on retrouve le point de vue du chapitre précédent, mais ici on considère l'ensemble des valeurs de l'individu. Ainsi, chaque note en philosophie étant considérée séparément, l'individu 496, qui a obtenu (aux trois trimestres puis au bac) les notes (14 ; 17.5 ; 13.6 ; 14), est remarquable du fait de sa note au second trimestre. Mais, du point de vue de l'ensemble des notes en philosophie, l'individu 64, qui a obtenu (aux trois trimestres puis au bac) les notes (16 ; 17 ; 13 ; 16), est plus remarquable que 496 en ce sens que, dans l'ensemble, ses notes sont plus éloignées des moyennes (8.99 ; 9.01 ; 9.26 ; 7.84). Selon ce point de vue, on ne prend pas en compte d'éventuelles liaisons entre variables.
- *Les valeurs de l'individu i pour les variables autres que k* ; en ce sens, est remarquable l'individu qui obtient, dans une même matière, les notes trimestrielles suivantes : 5, 15, 5. Cette approche prend en compte les liaisons entre variables : cet individu est remarquable car il présente des notes disparates dans une même matière, alors que, de façon générale, les notes trimestrielles d'une même matière sont liées.

Ces deux points de vue conduisent à deux approches de la notion d'individu remarquable.

13.2 Cas où l'on ne prend pas en compte les liaisons entre variables

Distance au point moyen dans le cas de deux variables

Lorsque, comme dans le chapitre précédent, on étudie une seule variable à la fois, on appréhende le caractère remarquable d'un individu par sa valeur centrée-réduite, qui mesure l'écart (exprimé en écarts-types) entre l'individu et le point moyen. Dans l'étude simultanée de deux variables, cette idée se généralise immédiatement : on représente les individus sur le graphique croisant les deux variables, chacune étant au préalable centrée et réduite. Sur ce graphique, l'origine représente le point (dit moyen) dont les valeurs sont les moyennes ; la distance entre un point et l'origine, mesurable par exemple à l'aide d'un double décimètre, est un indicateur du caractère remarquable de l'individu.

Exemple de données choisies

Pour illustrer les raisonnements, nous utilisons un exemple numérique de données choisies (i.e. imaginées pour mettre en évidence un phénomène), dans lequel 11 élèves ont obtenu chacun deux notes X et Y (cf. **Tab. 13.1** et **Fig. 13.1**).

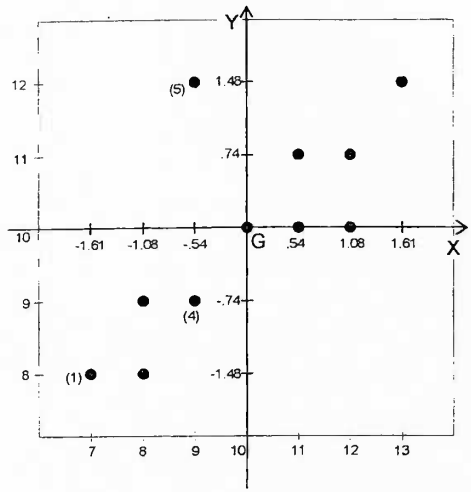


Figure 13.1. *Données choisies ; 11 élèves notés pour deux matières X et Y.*
Le graphique est réalisé à partir des notes centrées-réduites. La conversion en notes brutes figure sur les bords (cf. Tab. 13.1).

élève	note brute		note centrée-réd.			
	X	Y	X	Y	$d(i,G)$	$d_m(i,G)$
1	7	8	−1.614	−1.483	2.192	1.677
2	8	8	−1.076	−1.483	1.832	1.483
3	8	9	−1.076	−.742	1.307	1.077
4	9	9	−.538	−.742	.916	.742
5	9	12	−.538	1.483	1.578	2.775
6	10	10	0	0	0	0
7	11	10	.538	0	.538	.782
8	11	11	.538	.742	.916	.742
9	12	10	1.076	0	1.076	1.563
10	12	11	1.076	.742	1.307	1.077
11	13	12	1.614	1.483	2.192	1.677
Moyenne	10	10	0	0		
Ec.-type	1.859	1.348	1	1		

Tableau 13.1. *Données choisies : 11 élèves notés dans deux matières X et Y. L'indicateur* $d(I,G)$ *(resp.* $d_m(I,G)$*) est défini en 13.2 (resp. 13.3).*

Soient : x_i et y_i les notes de l'élève i ; $\bar{x}$ et $\bar{y}$ les moyennes de X et Y ; s_x et s_y les écarts-types de X et Y.

Nous avons choisi de centrer et de réduire les données. Du fait du centrage, le point moyen est confondu avec l'origine, et le carré de la distance au point moyen est égale à la somme des carrés des coordonnées (théorème de Pythagore) ; soit :

$$d^2(i,G) = \left[\frac{x_i - \bar{x}}{s_x}\right]^2 + \left[\frac{y_i - \bar{y}}{s_y}\right]^2$$

Application numérique pour l'élève 4 : $d^2(4,G) = (-.538)^2 + (-.742)^2 = .839 = (.916)^2$

Exemple extrait du fichier des notes : les notes obtenues au 1er trimestre en mathématiques et en philosophie.

A partir de la représentation des 909 élèves (cf. **Fig. 13.2**) on sélectionne les quatre élèves les plus éloignés de G et on calcule leur distance à G (cf. **Tab. 13.2**).

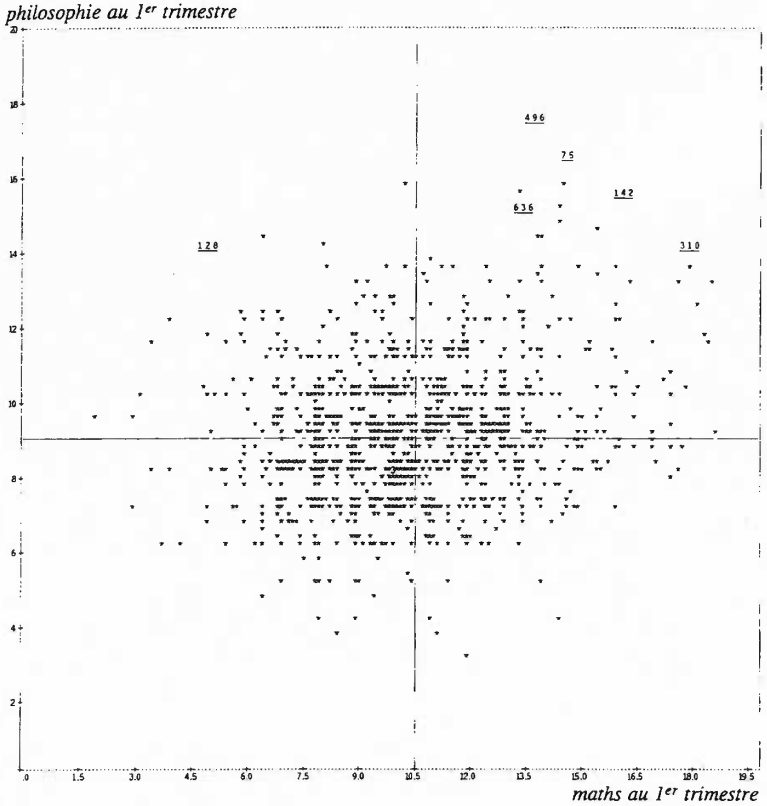


Figure 13.2 *Nuage des 909 élèves défini par les notes, au premier trimestre, en mathématiques et en philosophie.*
6 élèves, commentés dans le texte, sont identifiés par leur numéro d'ordre dans le fichier.

	note brute		note centrée-réduite		
élève	maths	philo.	maths	philo.	d(i,G)
496	13.8	17.5	1.11	4.25	4.39
75	14.7	16.5	1.43	3.75	4.01
142	16.2	15.5	1.95	3.25	3.79
310	18.0	14	2.59	2.50	3.60

Tableau 13.2. *Fichier des notes ; les 4 plus grandes valeurs de* d(*i, G*) *pour le couple de notes : mathématiques et philosophie (au 1er trimestre).*

Exemples numériques : $d^2(496,G) = (1.11)^2 + (4.25)^2 = 1.23 + 18.06 = (4.39)^2$

$d^2(310,G) = (2.59)^2 + (2.50)^2 = 6.71 + 6.25 = (3.60)^2$

Contribution d'une variable à une distance et réduction

Le détail du calcul de la distance à G montre que l'éloignement de 496 est surtout dû à sa note excellente en philosophie, alors que celui de 310 est dû de façon presque égale aux deux matières. Ceci introduit la notion très générale de contribution à un indicateur statistique, qui peut ici être formalisée ainsi : la contribution de la variable X à la distance du point i à G est la part de la variable X dans la distance du point i à G, soit, dans le cas de variables centrées-réduites : $[(x_i - \bar{x})/s_x]^2$.

On peut se demander si, globalement, c'est-à-dire lorsque l'on considère l'ensemble des élèves, les deux matières contribuent également à la distance entre un élève et G. Or, la contribution moyenne d'une variable n'est autre que sa variance. Ainsi, au sens de la contribution moyenne, la réduction, qui rend chaque variance égale à 1, équilibre l'influence des variables dans le calcul de la distance à G.

Individu remarquable et individu ayant une moyenne remarquable

Sur ces données, les individus les plus remarquables ont des notes excellentes au 1er trimestre en mathématiques et en philosophie. Le caractère remarquable ne peut-il pas se mesurer tout simplement par la moyenne des deux notes ? Non, car la distance à G et la moyenne sont deux notions différentes, ainsi que l'illustrent les deux élèves (128 et 636) qui, situés à la même distance de G, diffèrent grandement du point de vue de leur moyenne (cf. **Tab. 13.3**).

	note brute			note centrée-réduite		
élève	maths	philo.	moyenne	maths	philo.	d(*i,G*)
128	5	14	9.5	−1.98	2.5	3.19
636	13.5	15	14.25	1	3	3.16

Tableau 13.3. *Note moyenne et distance au point moyen : deux exemples (cf. Fig. 13.2).*

Cas de plus de deux variables

Dans le cas de plus de deux variables, on ne peut plus recourir au graphique plan. En revanche, la formule de distance au point moyen se généralise très facilement pour un nombre K quelconque de variables. Soit, en notant x_{ik} la valeur de l'élève i pour la matière k et $\bar{x}_k$ (resp. s_k) la moyenne (resp. l'écart-type) de la variable k :

$$d^2(i,G) = \sum_k \left[\frac{x_{ik} - \bar{x}_k}{s_k} \right]^2$$

Le tableau **13.4** rassemble les 4 individus les plus remarquables (pour ce critère) du fichier des notes, en tenant compte des 20 notes.

Elève		75	124	183	142
d(i,*G*)		9.66	9.42	9.29	9.06
Notes					
Maths	Bac	17	17	1	18
	1er trim.	14.7	17.8	3.2	16.2
	2ème trim.	14.8	17	1.6	17
	3ème trim.	16.6	16.6	2	18.2
Physique	Bac	17	19	3	18
	1er trim.	17.3	17.3	4	18
	2ème trim.	18.7	17.3	2.8	18.5
	3ème trim.	18	19	2	17.5
Sci. nat.	Bac	17	16	7	8
	1er trim.	13.5	14.2	13	11.7
	2ème trim.	13.5	15.8	8.8	10.2
	3ème trim.	16.2	17.5	8	12
Hist. Géo.	Bac	13	17	5	12
	1er trim.	14	13	9.5	13.5
	2ème trim.	14.8	15.7	9.5	13.4
	3ème trim.	15.5	14	6	15.2
Philo.	Bac	19	16	5	19
	1er trim.	16.5	13	6	15.5
	2ème trim.	14.5	13.4	9.3	15.6
	3ème trim.	18	17	10.3	15.2

Tableau 13.4. *Fichier de notes : les 4 élèves les plus remarquables du point de vue de leur distance au point moyen (G).*

Ce classement selon la distance à G permet de mettre en évidence les élèves les plus remarquables (pour ce critère) du fichier, c'est-à-dire ici ceux qui présentent un palmarès particulièrement prestigieux ou particulièrement catastrophique (ce calcul de distance à G est réalisé en routine dans les programmes d'ACP).

13.3 Cas de deux variables liées linéairement

Impact de la liaison linéaire entre deux variables sur le caractère remarquable d'un individu

Cas des données choisies

L'allure générale allongée du nuage de points (cf. **Fig.13.1**) suggère une liaison linéaire entre les deux notes X et Y ; le coefficient de corrélation vaut .725 entre ces deux variables.

Du point de vue de la distance à G, le point (1) est plus remarquable que (5). Or, d'un certain point de vue, (5) est plus remarquable que (1). En effet :

- l'élève (1) est proche d'autres élèves ; il a obtenu deux très mauvaises notes, mais s'inscrit dans la logique de l'ensemble des résultats (une mauvaise note dans l'une des matières s'accompagne généralement d'une mauvaise note dans l'autre) ;
- l'élève (5) est remarquable en ce sens qu'il est loin de tous les autres ; sa note légèrement au-dessous de la moyenne en X n'est pas "en accord" avec sa très bonne note en Y.

Exemples extraits du fichier des notes

Considérons cette fois les notes en mathématiques obtenues au troisième trimestre et au bac. Le graphique croisant ces deux variables (cf. **Fig. 8.1**) montre que ces deux notes sont liées (coefficient de corrélation : .68).

Comparons la place sur le graphique des élèves 509 et 6, dont les notes sont tableau **13.5**.

	notes brutes		notes centrées-réduites		
élève	3ème trim.	bac	3ème trim.	3ème bac	d(i,G)
509	5.5	6	−1.86	−2.26	2.93
6	15.8	8	1.59	−1.63	2.28

Tableau 13.5. *Fichier de notes : notes en mathématiques de deux élèves, au dernier trimestre et au bac.*
*d(i,G) est calculé à partir des données centrées-réduites (cf. **Fig. 8.1**).*

Du point de vue de la distance à G, 509 est plus remarquable que 6. Or, le graphique suggère que, d'un certain point de vue, 6 est plus remarquable que 509. En effet,

- l'élève 509 a obtenu deux mauvaises notes, mais s'inscrit dans la logique de l'ensemble des résultats ;
- l'élève 6 est remarquable en ce sens que sa bonne note au 3ème trimestre n'est pas "en accord" avec sa mauvaise note au bac.

Les autres notes en mathématiques de 6 (1er trim. : 15.1 ; 2ème trim. : 13.3 ; 3ème trim. : 15.8 ; bac : 8) laissent à penser que c'est la note au bac qui est atypique. Insistons bien sur le fait que cette valeur n'est pas remarquable du point de vue de la distribution des notes en mathématiques au bac mais relativement aux autres notes en mathématiques de l'élève 6.

A titre d'exemple, indiquons plusieurs interprétations qui peuvent rendre compte d'une telle valeur :

- erreur de saisie ;
- accident individuel ;
- surnotation pendant l'année ; cette interprétation peut être abandonnée compte tenu de la série des moyennes en maths (aux 3 trimestres puis au bac) du lycée 5 auquel l'élève 6 appartient (10.14 ; 11.32 ; 12.02 ; 14).

Les données contenues dans le fichier ne permettent pas de choisir entre les deux premières hypothèses. Cependant, cet exemple illustre l'intérêt de détecter ce type d'anomalie au début de l'analyse d'un fichier.

Nécessité d'un indicateur

De telles anomalies sont facilement mises en évidence à l'aide d'un graphique. Néanmoins l'examen de l'ensemble des graphiques croisant les variables prises deux à deux est inextricable dès lors que le nombre de variables est un tant soit peu grand (20 variables engendrent $(20\times19)/2=190$ graphiques). D'où l'intérêt de formaliser et de quantifier le caractère remarquable de ce type de situation en vue d'une recherche automatique.

Rappelons d'abord que du fait du centrage et de la réduction :

- l'influence des dispersions différentes d'une variable à l'autre est éliminée par le centrage et la réduction ;
- lorsque le nuage de points présente une direction principale d'allongement, cette direction est nécessairement une bissectrice, la première si la liaison est positive, la seconde si la liaison est négative (cf. Fiche 4).

L'exemple précédent suggère, dans le calcul de l'écart entre un point et G, de prendre en compte le fait que le nuage est inégalement allongé dans les directions des bissectrices. Intuitivement, pour un point, un même éloignement par rapport à G confère à ce point un caractère moins remarquable s'il s'exerce dans la direction principale d'allongement que dans l'autre.

*Introduction à la distance d_m (cf. **Fig. 13.3**)*

Pour calculer la distance usuelle entre deux points, il est indifférent d'utiliser les coordonnées initiales ou les coordonnées le long des deux bissectrices (ces dernières sont orthogonales et constituent un nouveau repère).

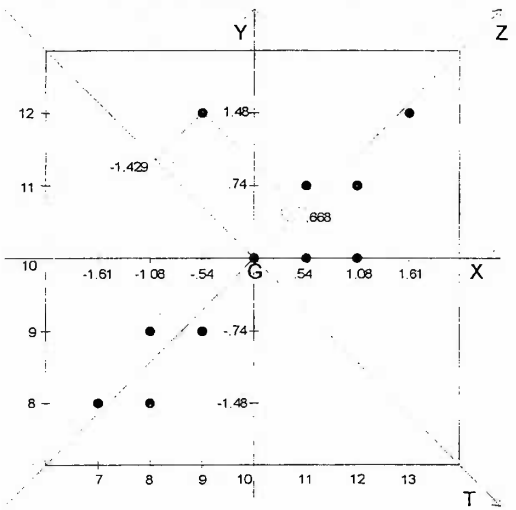


Figure 13.3. *Données choisies ; introduction à la distance* d_m.
Les données, déjà centrée et réduites dans le repère (X, Y), sont exprimées dans le repère (Z, T) des deux bissectrices ***puis*** *à nouveau réduites.*

Soient les coordonnées du point i :

- dans l'ancien repère (notes centrées-réduites) : x_i, y_i
- dans le nouveau repère (bissectrices) : z_i, t_i

La distance d(i,G) s'écrit :

- dans l'ancien repère : $d^2(i,G) = x_i^2 + y_i^2$
- dans le nouveau repère : $d^2(i,G) = z_i^2 + t_i^2$

Exemple numérique : cas du point 5 du jeu de données choisies

$$d^2(5,G) = (-.54)^2 + (1.48)^2 = (.668)^2 + (-1.429)^2 = 2.49 = (1.58)^2$$

Si l'on considère le calcul de la distance à G dans le repère engendré par les bissectrices, on est confronté au problème déjà rencontré de la plus grande dispersion du nuage dans une direction. Pour définir une distance à G dans laquelle les contributions moyennes des deux bissectrices sont équilibrées, on réduit ces deux nouvelles variables, c'est-à-dire que l'on divise chacune des nouvelles coordonnées par son écart-type (soient s_z et s_t les écarts-types des coordonnées le long des bissectrices). La nouvelle distance, appelée distance de Mahalanobis et notée d_m, s'écrit :

$$d_m^2(i,G) = \left[\frac{z_i}{s_z}\right]^2 + \left[\frac{t_i}{s_t}\right]^2$$

Applications numériques

A) *Cas du point 5 du jeu de données choisies*

Les valeurs de s_z et de s_t peuvent être calculées directement : elles peuvent aussi se déduire (cf. Fiche 5) du coefficient de corrélation entre X et Y ($r_{xy} = .725$) ;

$$s_z^2 = (1 + r_{xy}) = 1.725 \ (s_z = 1.313) \ ; \ s_t^2 = (1 - r_{xy}) = .275 \ ; \ (s_t = .524)$$

$$d_m^2(5,G) = (.668/1.313)^2 + (-1.429/.524)^2 = 7.700 = (2.77)^2$$

Comparée aux autres valeurs de d_m pour ces données (cf. **Tab. 13.1**), celle du point 5 est très importante ce qui illustre bien la signification de cet indicateur.

B) *Cas du fichier des notes*

Appliqué aux deux élèves étudiés plus haut, le calcul de cette distance donne :

$$d_m^2(509,G) = 2.3$$

$$d_m^2(6,G) = 4.02$$

Le caractère remarquable de 6 est bien mis en évidence.

Pour mieux saisir l'intérêt de l'indicateur d_m, on peut le calculer pour tous les points et identifier les 5 individus les plus éloignés de G (cf. **Tab. 13.6** et **Fig. 8.1**).

On pourrait aussi reporter, sur le graphique des données centrées-réduites, les ensembles de points équidistants (selon d_m) de G. Ces ensembles ont pour équation :

$$\frac{z_i^2}{s_z^2} + \frac{t_i^2}{s_t^2} = \text{constante}$$

Ils constituent des ellipses, ce qui illustre comment, selon sa direction, un même écart par rapport à G n'a pas la même importance dans le calcul de d_m.

$d_m(i,G)$	élève	notes brutes		notes centrées-réd.	
		3ème trim.	bac	3ème trim.	bac
4.46	533	2.5	15	-2.87	.56
4.02	6	15.8	8	1.59	-1.63
3.87	183	2	1	-3.03	-3.83
3.59	238	8.1	3	-.99	-3.20
3.39	521	3.5	13	-2.53	-.06

Tableau 13.6. *Fichier des notes : les 5 plus grandes valeurs de d_m pour le couple des notes en mathématiques au dernier trimestre et au bac.*

Comparaison de distances d_m calculées à partir de graphiques différents

Les calculs de d_m peuvent être faits pour n'importe quel couple de variables. Or, du fait de la double normalisation (centrage et réduction des variables initiales puis réduction des coordonnées le long des bissectrices), ces valeurs sont comparables d'un graphique à l'autre (de la même façon que l'on peut comparer deux valeurs centrées-réduites provenant de variables différentes).

A titre d'exemple, on peut lister, à partir du graphique représentant les notes obtenues au premier trimestre en mathématiques et en philosophie, les quatre individus les plus éloignés (au sens de d_m) de G (cf. **Tab. 13.7**).

$d_m(i,G)$	élève	notes brutes		notes cent.-réd.	
		maths	philo.	maths	philo.
4.27	496	13.8	17.5	1.11	4.25
3.82	75	14.7	16.5	1.43	3.75
3.54	128	5	14.0	-1.98	2.50
3.52	142	16.2	15.5	1.95	3.25

Tableau 13.7. *Fichier des notes : les quatre élèves les plus éloignés de G (au sens de d_m) pour le couple des notes en maths et en philosophie au 1er trimestre (cf. Fig. 13.2).*

Ces deux variables sont peu liées (coefficient de corrélation : 0.19) ; les dispersions le long des deux bissectrices sont comparables ; le fait de "normaliser" par ces dispersions ne modifie pas sensiblement le nuage de points ; les deux façons de calculer la distance (d et d_m) diffèrent peu entre elles dans ce cas. En effet, à quelques légères différences près, les élèves les plus remarquables sont les mêmes selon chacun de ces deux critères (cf. **Tab. 13.2** et **13.7**).

On peut rapprocher les tableaux **13.6** et **13.7** puisque les valeurs de d_m sont comparables d'un couple de variable à l'autre. Ainsi par exemple, d'après d_m, l'élève 6 est plus remarquable de par son couple de notes en mathématiques au dernier trimestre et au bac (15.8 et 8 ; d'où : d_m=4.02) que l'élève 128 ne l'est de par son couple de notes en mathématiques et en philosophie au 1er trimestre (5 et 14 ; d'où : d_m=3.54). Pourtant, considérées une à une, les deux notes de 128 sont plus remarquables que celles de 6

(comparer les valeurs centrées-réduites : −1.98 et 2.50 pour 128 ; 1.59 et −1.63 pour 6 ; cf. **Tab. 13.6** et **13.7**) ; mais le fait d'observer simultanément une valeur au-dessus et au-dessous de la moyenne est remarquable entre les deux notes de mathématiques ordinairement liées et ne l'est pas entre des notes en mathématiques et en philosophie dans l'ensemble peu liées.

13.4 Procédure de détection systématique d'individus remarquables

Pouvoir comparer des valeurs de d_m calculées à partir de couples de variables différents suggère, dans le cadre d'une exploration automatique de fichier :

- de considérer tous les couples de variables ;
- de calculer pour chaque couple et chaque individu *i* les distances $d_m(i,G)$;
- d'examiner les plus grandes valeurs observées de $d_m(i,G)$.

Appliquée, tous couples de variables confondus et tous individus confondus, au fichier des notes, cette procédure conduit aux deux plus grandes valeurs (cf. **Tab. 13.8**) commentées ci-après.

$d_m(i,G)$	élève	note 1	note 2
5.15	863	15 : h. géo. 2ème tr.	5 : h. géo.3ème tr.
4.85	404	6.5 : sci. nat. 1er tr.	17.5 : sc. nat. 2ème tr.

Tableau 13.8. *Fichier des notes : les deux plus grandes valeurs de* d_m*, quel que soit le couple de notes, pour l'ensemble du fichier.*

Premier cas ; l'élève 863 possède 18 notes sur 20 inférieures ou égales à 10. Ses notes en histoire-géographie sont : 3, 6.5, 15, 5. Cet ensemble suggère une erreur de saisie ou de transcription pour la valeur 15 qui semble bien atypique ; une vérification s'impose.

Deuxième cas ; l'élève 404 a des notes très variables. Il a obtenu en sciences naturelles : 12, 6.5, 17.5, 11. Cet ensemble suggère plus un élève irrégulier qu'une erreur ; on pourra toutefois vérifier ses données par acquis de conscience.

On peut ainsi dérouler cette liste ordonnée par valeurs décroissantes de d_m. Les individus deviennent de moins en moins remarquables au fur et à mesure du déroulement de la liste.

13.5 Prolongements possibles

Probabilité associée à une valeur de d_m

A une valeur d'une variable centrée-réduite, on a associé une probabilité (cf. 12.6), celle d'observer une valeur au moins aussi grande (pour ce calcul, on utilise la loi normale). Ce raisonnement peut être transposé à d_m sous réserve de disposer de la distribution (au moins approchée) de cet indicateur.

A condition de ne pas envisager une distribution trop « tordue » (il s'agit ici de la distribution des points dans le plan et non des distributions marginales), la probabilité d'apparition d'une valeur de d_m^2 supérieure ou égale à une valeur donnée est bien approchée par une distribution théorique connue : la loi du χ^2 à 2 ddl (cette loi est la loi exacte de d_m^2 dans le cas d'une distribution normale bidimensionnelle des points dans le plan). A titre

d'exemple, on peut comparer quelques probabilités issues de cette loi et les pourcentages de valeurs observées correspondant sur le fichier des notes.

valeur	1	2	3	4	5	6	7	8	9
Probabilité (χ^2)	.388	.631	.779	.867	.920	.952	.971	.982	.989
Pourcentage observé	.392	.635	.779	.866	.916	.947	.969	.981	.988

Tableau 13.9. *Loi du χ^2 à 2 ddl.*
Probabilité théorique d'observer une valeur inférieure à une valeur donnée ; pourcentage calculé à partir des 172710 couples de valeurs du fichier des notes.

Ainsi, en appliquant la norme usuelle selon laquelle une valeur est remarquable dès lors qu'elle est associée à une probabilité inférieure à .05, la valeur référence pour d_m^2 est 6 ($\approx 2.45^2$).

De même que l'on a généralisé la distance d^2, d'un couple de variables à un nombre quelconque de variables, peut-on généraliser d_m ?

D'un point de vue théorique, la généralisation de d_m à un nombre quelconque K de variables ne pose pas de problèmes dès l'instant que K est inférieur au nombre d'individus. Mais cette généralisation donne lieu à une formule complexe dont l'interprétation directe est malaisée : autrement dit, il peut être difficile de lire, dans les données d'un individu, ce qui a engendré sa forte valeur de d_m. On sort donc ici du cadre d'une approche simple d'un fichier.

Mise en évidence de variables remarquables

14.1 Introduction

Dans l'optique de cette seconde partie, une variable est remarquable si sa distribution diffère de celle des autres variables.

L'examen de base d'une distribution se fait visuellement à partir d'un histogramme ou d'un diagramme en bâtons. Il est toutefois utile de quantifier différents aspects d'une distribution :

- pour comparer entre elles des distributions, lorsque l'examen visuel n'est pas suffisamment précis ;
- pour détecter de façon automatique des distributions particulières (ou remarquables) dans un fichier comportant un grand nombre de variables.

Nous avons déjà défini deux familles d'indicateurs pour qualifier une distribution :

- indicateurs de tendance centrale (moyenne, médiane, etc.) ;
- indicateurs de dispersion (écart-type, écart interquartile, etc.).

Les aspects de la forme d'une distribution usuellement pris en compte sont (outre la dispersion) :

- son caractère plus ou moins symétrique (ou présence d'une queue de distribution) ;
- son caractère plus ou moins aplati (ou présence de deux queues de distribution).

14.2 Asymétrie

Définition de la symétrie

Il est facile de définir la symétrie d'une distribution : il existe une valeur M telle que à toute valeur observée $X > M$, correspond une valeur observée $Y < M$ telle que $|X-M| = |Y-M|$. Une telle distribution possède beaucoup de propriétés : en particulier, cette valeur M est à la fois la moyenne et la médiane de la distribution qui dans ce cas coïncident.

Un indicateur d'asymétrie

Naturellement, une distribution observée n'est jamais exactement symétrique. L'évaluation de l'asymétrie se fait en choisissant une propriété de la symétrie et en examinant dans quelle mesure la distribution observée vérifie cette propriété. Il existe donc plusieurs façons de mesurer l'asymétrie : par exemple, la comparaison de la moyenne et de la médiane est un

indicateur possible. Lors d'une étude, on peut imaginer un indicateur adapté à tel ou tel cas particulier ; on peut aussi choisir parmi les indicateurs déjà proposés par de nombreux auteurs. Signalons un critère pour choisir entre différents indicateurs : on peut s'intéresser à l'ensemble de la distribution ou, au contraire, focaliser l'attention soit sur ses éléments extrêmes soit sur sa partie centrale.

Un point de vue utile consiste à rechercher la présence d'individus extrêmes d'un seul côté de la distribution, c'est-à-dire d'une queue de distribution. Pour cela, on calcule l'indicateur suivant (en notant $\bar{x}$ et s la moyenne et l'écart-type de la distribution) :

$$\beta_1 = \frac{1}{I}\sum_i \left[\frac{x_i - \bar{x}}{s}\right]^3$$

Il est égal à la moyenne des valeurs centrées, réduites et élevées à la puissance 3. Le centrage revient à ne prendre en compte que les écarts par rapport à la moyenne ; la réduction permet de se ramener à une dispersion standard ; l'élévation au cube conserve le signe des écarts et fait jouer un grand rôle aux valeurs extrêmes. Si la distribution est symétrique, le coefficient est nul ; si les valeurs les plus éloignées de la moyenne sont majoritairement plus grandes que la moyenne, le coefficient est > 0 ; si ces valeurs sont plus petites que la moyenne, le coefficient est < 0.

Exemples de données choisies (cf. Fig. 14.1 et Tab 14.1)

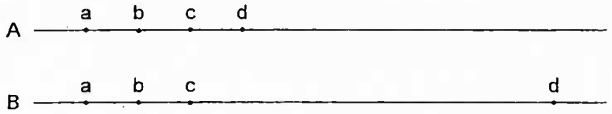


Figure 14.1. *2 variables A, B mesurées sur 4 individus a,b,c,d.*

	val. brutes		valeurs centrées-réduites							
	A	B	X	Y	X^2	Y^2	X^3	Y^3	X^4	Y^4
a	1	1	−1.342	−.849	1.800	.720	−2.417	−.612	3.24	.518
b	2	2	−.447	−.566	.200	.320	−.089	−.181	.04	.102
c	3	3	.447	−.283	.200	.080	.089	−.023	.04	.006
d	4	10	1.342	1.697	1.800	2.880	2.417	4.887	3.24	8.294
somme	10	16	0.	0.	4.000	4.000	0.000	4.071	6.56	8.922
moyenne	2.5	4	0.	0.	1.	1.	0.000	1.018	1.64	2.230

Tableau 14.1. *Calcul du coefficient d'asymétrie pour deux distributions particulières. β_1=moyenne des valeurs centrées-réduites élevées à la puissance 3. $\beta_1(A)=0$, $\beta_1(B)=1.018$ L'intérêt de l'élévation à la puissance 4 apparaît au § 14.3.*

La distribution A est parfaitement symétrique ($\beta_1=0$) ; la distribution B est asymétrique ($\beta_1=1.018$) : elle présente une queue de distribution à droite.

X et Y sont les valeurs de A et de B après centrage et réduction : cette normalisation ne modifie pas la forme des distributions.

X et Y, étant centrées, ont chacune une moyenne nulle. La variance est alors égale à la moyenne des carrés des valeurs. X et Y étant réduites, ces variances valent 1.

β_1 est la moyenne des valeurs centrées-réduites élevées à la puissance 3. Lors de l'élévation à la puissance 3, les valeurs les plus grandes (en valeur absolue) s'accroissent plus que les autres. Si ces valeurs se trouvent d'un seul côté de la moyenne, elles donnent leur signe à β_1. Ainsi, pour X^3, les valeurs −2.417 et 2.417 se compensent ($\beta_1=0$) ; pour Y^3, rien ne contrebalance la valeur 4.887 qui engendre à elle seule le coefficient d'asymétrie de 1.018 : ce coefficient positif indique une queue de distribution à droite.

Exemples extraits du fichier des notes

Les coefficients d'asymétrie des notes au bac en mathématiques et en philosophie (cf. **Tab. 14.2**) indiquent la présence d'une queue de distribution à gauche pour les mathématiques et à droite pour la philosophie. Les écarts-types de ces distributions étant très proches, ces asymétries sont bien visibles sur les diagrammes en bâtons (cf. **Fig. 7.1**). En revanche, seule celle de philosophie est suffisamment importante pour apparaître sur les boîtes de dispersion (cf. **Fig. 7.5**), mais rappelons que ces boîtes sont mal adaptées aux données présentant de nombreux ex æquo..

	β_1	moyenne	ec. type	minimum	maximum
Maths	−.41	13.21	3.19	1	20
Philosophie	.87	7.84	3.29	2	19

Tableau 14.2. *Comparaison des coefficients d'asymétrie pour deux notes du bac.*

Remarquons que, pour ces deux matières, l'intervalle des notes possible est borné, que pratiquement toutes les notes sont utilisées et que la moyenne est sensiblement différente de 10. Ainsi il existe de très faibles notes en mathématiques, situées à plus de 7 points de la moyenne, qu'aucune bonne note ne peut contrebalancer. De même il existe de très bonnes notes en philosophie, situées à plus de 8 points de la moyenne, qu'aucune mauvaise note ne peut contrebalancer.

Selon ces coefficients, l'asymétrie est plus importante en philosophie qu'en mathématiques : les très mauvaises notes en mathématiques sont moins nombreuses que les très bonnes notes en philosophie.

On peut trier les coefficients β_1 pour l'ensemble des variables du fichier (cf. **Fig. 14.2.a**).

- Parmi l'ensemble des notes, les deux distributions précédentes sont les plus dissymétriques, à droite (*philosophie au bac*) et à gauche (*mathématiques au bac*). Ces deux notes sont aussi celles dont la moyenne s'écarte le plus de 10 ce qui renforce l'idée selon laquelle l'origine de ces dissymétries réside dans ces moyennes "décentrées" alors que l'intervalle de variation est borné. Mais ce "décentrage" n'explique pas tout puisque, pour ces deux matières, la variable la plus dissymétrique (*philosophie au bac*) est la moins "décentrée".
- Les notes de philosophie présentent les asymétries les plus marquées (par une queue de distribution à droite). Hormis cela, les notes ne se regroupent ni par matière ni par date.

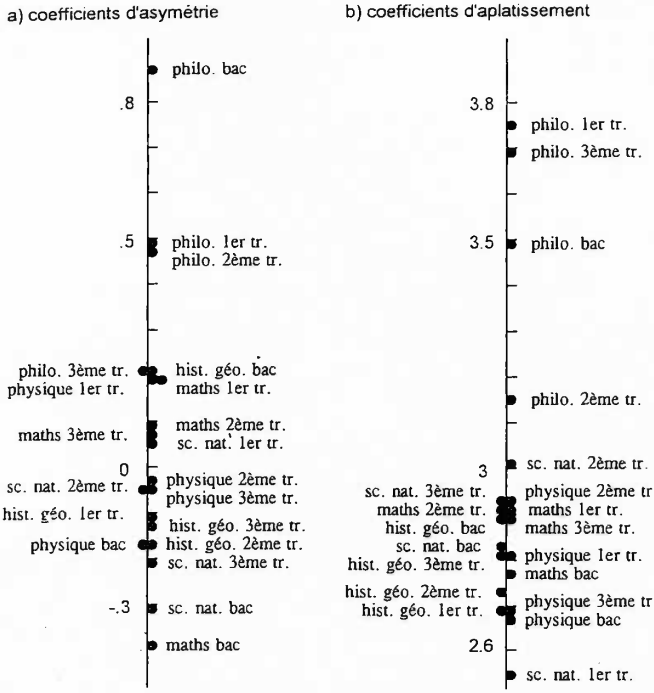


Figure 14.2. *Représentation axiale des valeurs des coefficients d'asymétrie (β_1) et d'aplatissement (β_2) pour les variables du fichier de notes.*

14.3 Aplatissement

Qu'est-ce que l'aplatissement ?

Cette notion est souvent mal perçue car, bien que consacré par l'usage, le terme « aplatissement » n'évoque pas très bien ce que représentent divers indicateurs de forme qui ont été proposés. On oppose souvent des distributions pointues à des distributions aplaties au sens courant du terme, c'est-à-dire étalées. Or, cet aspect de la forme d'une distribution est pris en compte par l'écart-type. Comme pour l'asymétrie, il convient de comparer les formes de distributions de même variance.

Le problème est de mesurer l'importance des queues de distribution ; à variance constante, la présence de queues de distribution s'accompagne nécessairement d'une partie centrale de la distribution très resserrée.

Un indicateur d'aplatissement

Pour détecter la présence d'individus extrêmes, à la fois au-dessus et au-dessous de la moyenne, il existe plusieurs coefficients dont le plus connu est le suivant :

$$\beta_2 = \frac{1}{I}\sum_i \left[\frac{x_i - \bar{x}}{s}\right]^4$$

Il est égal à la moyenne des valeurs centrées, réduites et élevées à la puissance 4. Le centrage et la réduction jouent le même rôle que pour le coefficient β_1 ; l'élévation à la puissance 4 annule le signe des écarts et fait jouer un grand rôle aux valeurs extrêmes. β_2 est toujours supérieur à 1, valeur atteinte pour une distribution limitée à 2 valeurs de même effectif : cette situation reflète parfaitement l'absence de queues de distribution puisque tous les individus sont équidistants de la moyenne. Pour une loi normale, β_2 vaut 3, ce qui explique que l'on mesure quelquefois l'aplatissement par le coefficient : $\gamma_2 = \beta_2 - 3$. Autre référence : β_2 vaut 1,8 pour une loi uniforme.

Exemples de données choisies (cf. Fig. 14.3 et Tab. 14.3)

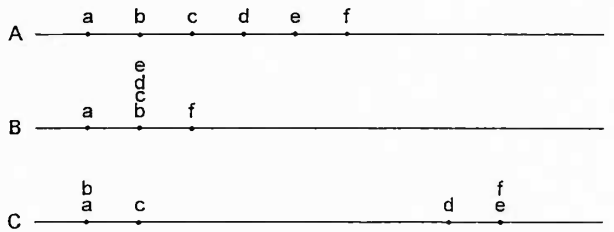


Figure 14.3. *3 variables A, B, C mesurées sur 6 individus a, b, c, d, e, f.*

	valeurs brutes			valeurs centrées-réduites		
	A	B	C	*X*	*Y*	*Z*
a	1	1	1	−1.464	−1.732	−1.082
b	2	2	1	−.878	.000	−1.082
c	3	2	2	−.293	.000	−.812
d	4	2	8	.293	.000	.812
e	5	2	9	.878	.000	1.082
f	6	3	9	1.464	1.732	1.082
moyenne	3.5	2	5	0.	0.	0.
ec. type	1.71	.58	3.70	1.	1.	1.
β_1	0.	0.	0.	0.	0.	0.
β_2	1.73	3.00	1.06	1.73	3.00	1.06

Tableau 14.3. *3 variables A, B, C mesurées sur 6 individus a, b, c, d, e, f*

X, *Y* et *Z* sont les distributions centrées-réduites correspondant à A, B et C. Le centrage et la réduction normalisent les distributions au sens de l'égalisation des moyennes à 0 et des variances à 1 mais conserve la forme des distributions au sens de l'asymétrie et de l'aplatissement.

Ne pas confondre dispersion et aplatissement. La distribution C a un fort écart-type (elle est très étalée) mais un faible aplatissement (pas de queues de distribution : tous les points sont à peu près équidistants de la moyenne). La distribution B a un faible écart-type (elle est peu

étalée) mais un fort aplatissement (présence de queues de distribution : deux points sont beaucoup plus éloignés de la moyenne que ne le sont les autres).

Exemples extraits du fichier des notes

	β_1	β_2	moyenne	ec. type	minimum	maximum
Physique bac	−.16	2.66	11.01	3.61	1	19
Philosophie 3ème tr.	.23	3.70	9.26	2.33	1	18

Tableau 14.4. *Comparaison entre deux coefficients d'aplatissement. Les histogrammes de ces distributions se trouvent* ***Fig. 7.1 et 7.6***

Les deux notes du tableau **14.4** ont une distribution pratiquement symétrique. La répartition des notes en physique au bac ne présente pratiquement pas de queue de distribution ; celle des notes en philosophie au 3ème trimestre en présente deux. On remarquera que, comparée à celle de physique au bac, la distribution de la note en philosophie au 3ème trimestre est plus "aplatie" alors qu'elle présente une écart-type plus petit.

On peut visualiser les coefficients β_2 des différentes notes (cf. **Fig. 14.2.b**).

- La distribution la plus aplatie est *philosophie au 1er trim.* (β_2=3.76 ; γ_2=.76), qui possède les queues de distribution les plus importantes.
- La distribution la moins aplatie est *sciences naturelles au 1er trim.* (β_2=2.55 ; γ_2=−.45), qui possède des queues de distribution moins importantes que celles de la loi normale.
- Plus généralement, seules les distributions des notes en philosophie présentent des queues de distribution importantes. Ceci est bien visible sur les histogrammes (cf. **Fig. 7.6**) et les boîtes de dispersion (cf. **Fig. 7.5**). Hormis cette matière, les distributions sont moins aplaties que ne l'est la distribution normale, ce qui peut être relié, au moins en partie, au fait que les notes sont nécessairement comprises dans un intervalle borné.

Attention : on ne peut comparer directement les valeurs absolues des coefficients γ_2 lorsqu'ils sont de signes différents (alors qu'avec β_1 on peut comparer directement les intensités d'une dissymétrie droite et d'une dissymétrie gauche) ; ainsi, sur la vue de β_2 (ou γ_2) on ne peut pas dire que *philosophie au 1er trim.* diffère plus d'une loi normale que *sciences naturelles au 1er trim.*

14.4 Probabilité associée à un coefficient de forme

Cas de β_1

Le calcul de β_1 permet de trier des distributions de même effectif, par exemple par dissymétrie croissante. Mais ce seul calcul est insuffisant lorsque l'on souhaite comparer des distributions d'effectifs différents. Présentons deux points de vue.

- Une même valeur de β_1 est beaucoup **plus stable** – donc interprétable – si elle est obtenue à partir de 1000 individus que si elle est obtenue à partir de 10 individus. Concrètement, si l'on retire au hasard quelques individus de l'échantillon, la valeur de β_1 changera beaucoup moins dans le cas de 1000 individus que dans le cas de 10.

- On peut s'interroger sur le caractère **plus ou moins fortuit** d'une valeur observée de β_1. Intuitivement, une valeur observée de β_1 sera considérée comme d'autant plus fortuite qu'elle a une plus grande probabilité d'apparaître dans le cadre d'un tirage au hasard dans une distribution symétrique.

Il est donc utile d'imaginer un indicateur lié à β_1 et intégrant l'effectif observé. Pour cela, on associe, à toute valeur observée de β_1 (issue d'une distribution de taille n), la probabilité d'apparition, dans le cadre d'un échantillon de taille n issu d'une loi normale, d'une valeur au moins aussi grande que celle effectivement observée (un tel calcul de probabilité requiert de spécifier une loi ; la loi normale est choisie en raison de son caractère de référence). Ainsi, cette probabilité mesure l'écart, du point de vue de l'asymétrie et en tenant compte des effectifs observés, entre la distribution observée et une situation de référence symétrique. Ce nouvel indicateur, s'interprétant comme une probabilité, permet de comparer des situations différentes, en particulier du point de vue des effectifs. A la limite, on pourra porter dans certains cas une appréciation absolue à partir de cet indicateur, en considérant par exemple qu'une asymétrie associée à une probabilité de 0.5 ne mérite pas attention car elle correspond à une situation proche de celles que l'on obtient couramment par un tirage au hasard dans une loi symétrique.

Exemples de données choisies (Tab. 14.5)

	n	β_1	$P(\beta_1)$
Cas 1	20	.4	.190
Cas 2	909	.2	.007

Tableau 14.5. *Données choisies : coefficient d'asymétrie et probabilité associée.*

La distribution 1 présente une asymétrie droite plus importante que la distribution 2 (.4 > .2). Mais, son effectif étant beaucoup plus faible, son asymétrie peut être considérée comme fortuite ce qui n'est pas le cas de la distribution 2.

Cas de β_2

On peut associer une probabilité à chaque valeur observée de β_2, à l'aide d'un raisonnement analogue à celui tenu pour l'asymétrie. En conservant la loi normale comme distribution de référence, la valeur de référence pour β_2 est 3. Ainsi, on associe à toute valeur observée de β_2 (à partir d'une distribution de taille n) la probabilité d'apparition, dans le cadre d'un échantillon de taille n issu d'une loi normale, d'une valeur de β_2 au moins aussi éloignée de 3 que ne l'est celle effectivement observée.

Outre la possibilité de comparer des distributions d'effectifs différents, cette probabilité peut aussi être considérée comme une évaluation du caractère remarquable de l'aplatissement d'une distribution.

Exemple : Aplatissement de *philosophie 1er trim.* et *sciences nat 1er trim.* (cf. **Tab. 14.6**).

Ces deux écarts à la loi normale sont comparables en terme de probabilité. En ce sens, la présence de valeurs extrêmes en *philosophie au 1er trim.* est aussi remarquable que leur absence en *sciences nat au 1er trim.* Cette opposition se lit sur les histogrammes (cf. **Fig. 14.4**), mais de façon malaisée du fait de la différence entre les dispersions qui s'y superpose.

	ec. type	asymétrie		aplatissement	
		β_1	probabilité	β_2	probabilité
maths bac	3.19	-.41	$5\ 10^{-7}$	2.77	.07
physique bac	3.61	-.16	.02	2.66	$7\ 10^{-3}$
sc. nat. 1er tr.	2.41	.05	.3	2.55	$2\ 10^{-4}$
sc. nat. bac	2.88	-.31	$9\ 10^{-5}$	2.83	.1
philo. 1er tr.	2.00	.50	$2\ 10^{-9}$	3.76	$2\ 10^{-4}$
philo. bac	3.29	.87	$< 10^{-9}$	3.49	$5\ 10^{-3}$

Tableau 14.6. *Quelques coefficients d'asymétrie et d'aplatissement accompagnés de leur probabilité associée.*

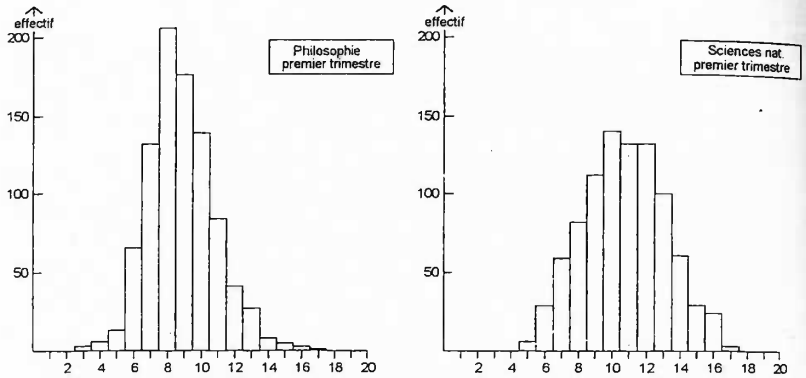


Figure 14.4. *Distributions de* philosophie *et* sciences naturelles *au premier trimestre : histogrammes.*

Comparaison entre les probabilités associées aux deux coefficients de forme (cf. Tab. 14.6)

La distribution de *philosophie au bac* est plus remarquable par son asymétrie que celle de *philosophie au 1er trim.* ne l'est par son aplatissement.

Plus généralement, les distributions remarquables le sont plus par leur asymétrie que leur aplatissement.

Conclusion sur le fichier des notes. La plupart des distributions s'apparentent d'assez près à la distribution normale. L'écart le plus notable est constitué par les notes en philosophie qui présentent des queues de distribution importantes à droite.

Partie 3

Fiches techniques

Construction du tableau de données, type de variables, codage

Dans l'exemple du fichier des notes, le tableau de données était fourni et sa structure s'imposait : les élèves sont les individus au sens statistique du terme, les notes sont des variables quantitatives et le lycée une variable qualitative. Ce n'est pas toujours le cas : la construction d'un tableau rectangulaire à partir de données disponibles n'est pas forcément évidente (ni unique).

On ne parle pas ici du problème encore plus fondamental du recueil des données : choix des variables à mesurer dans une expérience, des questions à poser dans une enquête, définition précise de la population étudiée, assurance de la qualité des données recueillies, etc. Disons simplement que pour étudier un problème donné, on doit tenter de déterminer des variables qui permettent d'en faire une approche aussi complète que possible. Naturellement, la qualité des données est primordiale car tous les résultats en dépendent. Il faut donc que les erreurs, presque toujours inévitables, ne soient pas trop nombreuses, que les mesures soient faites assez précisément, que les réponses à un questionnaire soient assez fiables (clarté des questions, sérieux de l'enquêteur et de l'enquêté), etc. Lorsqu'on maîtrise le recueil des données, ces préoccupations doivent intervenir dans la toute première phase de l'étude ; lorsqu'on ne le maîtrise pas, il faut tout de même évaluer la fiabilité des données, pour nuancer éventuellement les conclusions et même, s'il le faut, renoncer à l'étude.

On ne parle ici que du choix du type des variables et de leur codage ou recodage. L'ensemble des types possibles est presque illimité. Nous nous restreignons aux types les plus courants, auxquels on peut la plupart du temps se ramener, notamment les variables qualitatives et quantitatives largement commentées dans le corps de l'ouvrage.

1. Variable qualitative (ou nominale)

Définition

Les variables qualitatives prennent des valeurs, appelées modalités, dans un ensemble quelconque. Exemples : le sexe (2 modalités), la situation matrimoniale (e.g. 4 modalités : *célibataire*, *marié*, *divorcé*, *veuf*), la réponse à la question « *Aimez-vous Brahms ?* » (*oui*, *non*, *ne connaît pas Brahms*, *non-réponse*), etc. La plupart du temps, le nombre de modalités est assez faible : le cas des lycées où il atteint 22 est exceptionnel (une autre exception, classique dans les grandes enquêtes, est le numéro du département).

Codage

Généralement, on code dans le tableau chacune des n modalités par un nombre, ce qui simplifie le traitement informatique. Le codage numérique le plus simple consiste à numéroter les n modalités de 1 à n ou de 0 à $n-1$.

Données manquantes

En présence de données manquantes, on crée une modalité particulière : *donnée manquante*. Par exemple, la variable *situation matrimoniale* aura 5 modalités : les 4 modalités déjà citées et *non-réponse*.

Etude préliminaire

Au début d'une étude, on calcule toujours les effectifs de chacune des modalités de chaque variable qualitative, effectifs que l'on représente souvent par des diagrammes en bâtons. C'est la première description statistique du tableau de données. Ces chiffres permettent aussi des vérifications : il faut être très attentif aux résultats inattendus.

Recodage par regroupement de modalités

L'examen des effectifs peut suggérer des recodages, en particulier si certaines modalités ont des effectifs très faibles. Par nature, les modalités rares concernent une faible partie de la population et, en ce sens, sont un peu marginales. Les conserver telles quelles leur confère la même importance qu'aux autres modalités, ce qui peut occulter des tendances plus globales, que ce soit pour présenter des tableaux d'effectifs ou pour étudier des liaisons entre variables. Aussi est-il souvent judicieux de *regrouper des modalités*. Par exemple, les résultats synthétiques d'élections regroupent la plupart du temps les partis peu représentés sous les vocables *divers droites* et *divers gauches*.

Il n'y a pas de règle systématique pour ce regroupement : dans certains cas il peut être logique de regrouper les "divers droites". Mais si pour les autres variables ces électeurs diffèrent, cela pose un problème de fond dans l'étude des liaisons avec ces variables. Au cas par cas on décidera entre :

- créer une modalité *autre* regroupant des modalités rares ;
- regrouper chaque modalité rare avec la modalité non rare qui a priori lui ressemble le plus.

Création de variables croisées

Au cours de l'étude, il s'avère souvent intéressant de raffiner les classes données par des variables qualitatives, en considérant le croisement de 2 variables : étudier séparément le sexe (2 modalités) et les classes d'âge (4 modalités par exemple) est beaucoup moins riche que d'étudier la variable croisée *sexe×classes d'âge* qui présente 8 modalités.

2. Variable qualitative ordonnée

Définition

Une variable qualitative est dite ordonnée lorsque l'ensemble de ses modalités est ordonné. Exemples : un questionnaire où l'on demande de choisir parmi les quatre positions *pas du tout d'accord*, *pas d'accord*, *d'accord*, *tout à fait d'accord* à propos de l'affirmation : *il*

faut augmenter le salaire des infirmières ; un autre questionnaire où l'on demande à des dégustateurs de noter de 1 à 5 l'acidité ou la qualité d'un vin ; ou encore, un questionnaire dans lequel à la question *combien de fois par mois, en moyenne, allez-vous au cinéma ?*, on propose 4 réponses possibles : *jamais*, *1 fois*, *2 fois*, *3 fois et plus*. Dans ces trois exemples les modalités de réponses sont ordonnées. Il n'y a sans doute pas le même écart entre *pas du tout d'accord* et *pas d'accord* d'une part et entre *pas d'accord* et *d'accord* d'autre part ; même dans le cas des vins où la valeur donnée est une note, il n'y a vraisemblablement pas le même écart entre 1 et 2 d'une part et entre 3 et 4 d'autre part.

Codage et recodage

Comme pour les variables qualitatives non structurées, le plus simple est le codage numérique allant de 1 au nombre n de modalités (ou de 0 à $n-1$) en respectant, bien sûr, l'ordre des modalités. Pour regrouper des modalités rares avec d'autres modalités, l'ordre est primordial : il est logique de regrouper *jamais* avec *une fois* et absurde de le regrouper avec *trois fois et plus*.

Traitement en tant que variable qualitative

Ce traitement ne diffère pas essentiellement de celui des variables non ordonnées. Les différences apparaissent au niveau de la présentation et surtout de l'interprétation des résultats. Les diagrammes en bâtons doivent respecter l'ordre des modalités (cf. **Fig. 4.1**). Il en est de même pour les tableaux croisant cette variable avec d'autres (cf. **Tab. 8.1**).

Traitement en tant que variable quantitative

Considérer les variables qualitatives à modalités ordonnées comme des variables quantitatives est critiquable sur le plan du principe puisque les données ne sont pas alors traitées dans le total respect de leur recueil : en particulier, le codage de 1 à n implique alors l'« équidistance » des modalités adjacentes, problème soulevé plus haut. Néanmoins, lorsque la structure des données est assez forte, elle est peu perturbée par ce changement de point de vue, bien pratique lorsque l'on est confronté à des données volumineuses.

- Pratique pour comparer des sous-populations entre elles : par exemple, les réponses des hommes et des femmes (ou des différentes catégories socioprofessionnelles) sur l'accord avec la proposition *il faut augmenter le salaire des infirmières*. Garder la nature qualitative de la variable implique de comparer les répartitions, des hommes et des femmes dans les 4 modalités : ce n'est pas tout à fait immédiat. En la considérant comme une variable quantitative, il suffit de comparer les moyennes des sous-populations, ce qui permet de repérer immédiatement une différence notable quand elle existe.
- Pratique aussi pour comparer rapidement les liaisons entre variables, en calculant des coefficients de corrélation plutôt que d'analyser les tableaux croisés.
- Pratique encore pour comparer globalement les réponses à différentes variables qui ont les mêmes modalités. Ceci se produit notamment dans les questions avec une même échelle pour la réponse, une intensité d'accord comme dans l'exemple du salaire des infirmières ou une note comme dans l'exemple de l'évaluation de l'acidité d'un vin. La comparaison des moyennes de ces variables a alors un sens et est beaucoup plus simple que la comparaison des ensembles de fréquences.
- Pratique enfin pour regrouper des sous-populations. Soit, par exemple, 15 vins notés par 50 dégustateurs sur une échelle de 1 à 5 suivant 10 critères ; pour comparer les vins, on

peut regrouper les jugements des 50 dégustateurs. En considérant la note comme une variable numérique, on calcule, pour chaque vin, la moyenne des notes qui lui ont été attribuées pour chaque critère. On obtient alors un nouveau tableau, croisant 15 individus (les vins) et 10 variables (les moyennes des notes données par les dégustateurs pour chaque critère).

Variable quantitative ou qualitative ?

Les deux traitements sont donc possibles. Ils ne mettent pas en évidence exactement les mêmes faits et ne font pas double emploi. Le principal intérêt du traitement en variable numérique est de permettre des comparaisons rapides, précieuses lorsque les variables sont nombreuses. Le traitement en variable qualitative est plus lourd mais plus complet.

Données manquantes

Pour les variables qualitatives, il est possible d'ajouter une modalité *donnée manquante*. Il est encore possible de le faire avec des modalités ordonnées, mais la structure d'ordre n'est plus respectée totalement. Ceci n'est pas très gênant au niveau de la présentation et de l'interprétation des résultats lorsque l'on considère les variables comme qualitatives (diagramme en bâtons, tableaux croisés, etc.), mais cela devient un réel problème si l'on veut la considérer comme une variable quantitative.

3. Variable indicatrice

Définition

Une variable indicatrice code la présence (ou l'absence) d'un caractère. Exemples : pour une plante, le caractère *a des fleurs* ; dans un questionnaire où l'on demande de citer des hommes politiques, le fait d'*avoir cité Un tel*. Ces variables sont généralement codées par 1 quand le caractère est présent et 0 quand il ne l'est pas. En fait, il s'agit de variables qualitatives à 2 modalités seulement (présence ou absence d'un caractère).

Eclatement de variables qualitatives, indicatrices floues

Notons qu'une variable qualitative à n modalités peut être « éclatée » en n variables indicatrices, chacune correspondant à une modalité. Chaque individu présente alors la valeur 1 une et une fois seulement sur l'ensemble des variables indicatrices défini par une même variable. Ce codage est dit *disjonctif complet*. Il est précieux pour introduire le traitement des tableaux *individus×variables qualitatives* par l'analyse des correspondances multiples (méthode analogue, jusqu'à un certain point, à l'analyse en composantes principales, mais adaptée aux variables qualitatives).

Cette présentation des données permet une généralisation utile lorsqu'il est nécessaire de nuancer la présence ou l'absence des modalités. Ainsi, pour un individu donné qui se partage entre plusieurs modalités, on peut affecter à chacune de ces modalités un nombre compris entre 0 et 1, la somme de ces nombres devant être égale à 1 (par exemple, pour 2 modalités, on mettra 1/2 à chacune pour exprimer un partage équilibré). Cette situation se présente dans le cas de réponses multiples à une question, de recodage de variable quantitative, etc.

Ce dernier codage est dit *en indicatrices floues*. Il possède plusieurs propriétés des indicatrices usuelles. Ainsi, la somme des valeurs d'une modalité (l'effectif pour une

indicatrice) est un nombre qui peut ne pas être entier, puisque chaque individu qui la possède ne compte pas forcément pour 1, mais qui a un sens. Ce codage est surtout utilisé dans le cadre d'une variante de l'analyse des correspondances multiples.

4. Variable quantitative

Définition

C'est une variable dont les valeurs possibles sont des nombres.

Variable faussement quantitative

Exemple : un commerçant distribue des tuyaux de 3 diamètres différents. Le diamètre est formellement une variable quantitative : c'est une mesure. Mais elle n'est pas mesurée sur chaque tuyau et, fondamentalement, il s'agit d'une variable qualitative à 3 modalités ; pour décrire sa répartition, on voit bien qu'il suffit de donner les trois effectifs ou les trois pourcentages des 3 modalités ; pour étudier ses liaisons avec une variable quantitative quelconque, il est bien plus significatif de regarder les moyennes de cette variable pour chaque groupe de tuyaux que de calculer des coefficients de corrélation, etc.

Autre exemple. Dans l'étude des notes, nous analysons la variable *nombre d'épreuves passées au bac* qui est formellement quantitative mais que nous considérons comme qualitative : elle ne prend que deux valeurs - *2 notes* ; *5 notes* - qui représentent surtout le fait de posséder ou non déjà un bac (en comparaison, le nombre exact d'épreuves est secondaire).

Recodage en variable qualitative par division en classes

Plus généralement, pour les variables quantitatives qui prennent un très petit nombre de valeurs, il est souvent judicieux, pour beaucoup de traitements statistiques, de les considérer comme des variables qualitatives. Mais, même lorsque le nombre de valeurs est élevé, on peut, technique très fréquemment utilisée (cf. 7.8), transformer une variable quantitative en variable qualitative ordonnée en divisant son intervalle de variation en classes.

L'intérêt de cette transformation est multiple.

- Elle permet de présenter la distribution d'une variable de manière plus riche qu'avec les indices définis pour les variables quantitatives (moyenne, écart-type) et plus synthétique au niveau des graphiques (comparer un histogramme et une représentation axiale, en particulier lorsqu'il y a beaucoup d'individus). C'est un très bon compromis entre richesse et synthèse. De même, cela permet de présenter la liaison entre deux variables à l'aide d'un tableau de contingence, de façon plus riche qu'avec la corrélation (inadaptée si la liaison n'est pas linéaire) et de manière plus quantifiée et plus synthétique qu'avec un graphique (cf. Tab. 8.1).
- Elle est utile, sinon indispensable, pour étudier des variables dont la distribution est très irrégulière. Par exemple, des comptages de bactéries dans des échantillons où les nombres varient de quelques unités à quelques dizaines de milliers. Pour ces variables très irrégulières, les moyennes et les corrélations n'ont guère de sens ; il est bien plus efficace de recoder en quelques modalités ordonnées comme : *rare*, *présence faible*, *présence moyenne*, *présence très importante*.

- Elle est nécessaire si l'on veut rendre homogène un tableau de données mixte (comportant à la fois des données qualitatives et quantitatives) afin d'en faire une synthèse globale par une analyse des correspondances multiples.
- Elle permet quelquefois de donner une solution au problème de données manquantes très nombreuses et relativement homogènes : on ajoute la modalité *donnée manquante*.
- Elle permet, par le biais de croisements entre variables qualitatives, de mettre en évidence de façon assez simple des liaisons complexes (exemple : X lié à Y uniquement pour certaines valeurs de Z).

Comment choisir les limites et le nombre des classes ?

- en choisissant, quand elles s'imposent, des limites "naturelles" ou de référence (le SMIC pour le revenu mensuel, 18 ans pour l'âge, etc.) ;
- en regardant sur des histogrammes si des discontinuités dans la distribution délimitent des sous-populations ;
- sinon, **en divisant en classes d'effectifs à peu près égaux plutôt qu'en intervalles égaux**.

Le nombre de classes dépend à la fois de la forme de la distribution et de l'effectif de la population. Si la population est nombreuse, on peut se permettre de prendre plus de classes car on risque moins d'obtenir beaucoup de classes d'effectif nul dans les croisements entre variables. En pratique, il suffit souvent de prendre 5 ou 6 classes.

Par cette technique, les individus se trouvant proches d'une limite de classe se trouvent affectés brutalement à une classe donnée exactement de la même façon que ceux qui sont situés au milieu de l'intervalle. Pour nuancer cette répartition, on peut utiliser des indicatrices floues en affectant les valeurs 1/2 aux deux classes ou même des valeurs tenant compte de la distance à la limite des classes. Ce codage, beaucoup plus lourd à effectuer et à traiter, ne présente d'intérêt que si le nombre d'individus est faible : sur 1000 individus, les tendances générales ne seront guère modifiées.

Recodage par centrage et réduction

Le centrage et surtout le centrage-réduction (cf. 7.6) permettent de rendre homogènes les variations des différentes variables quantitatives. Il permet aussi d'explorer rapidement un tableau de données en repérant notamment les valeurs situées très loin de la moyenne (cf. Ch. 12).

Création de nouvelles variables par combinaison de variables

Il est souvent très utile de créer de nouvelles variables quantitatives par combinaison des variables de départ : ainsi, dans l'exemple des notes, nous avons créé la variable *bac* (cf. 9.1) et la variable *différence entre note du bac et troisième trimestre en maths* (cf. 11.1). Quelquefois ce sont des rapports entre variables ou des transformations plus compliquées qui permettent d'obtenir des indices utiles et/ou classiques. Ces nouvelles variables ne se substituent pas aux variables initiales mais apportent des perspectives différentes dont certaines peuvent être démonstratives. Il ne faut pas hésiter à user et abuser de cette possibilité facile à mettre en œuvre à l'aide d'un tableur.

5. Variable ordinale

Si on demande à un jury de 20 personnes de classer 10 vins par ordre de préférence, les données brutes sont 20 classements de 10 vins. Les classements sont généralement difficiles à obtenir du fait du grand nombre de comparaisons qu'ils impliquent ; en outre, le classement par ordre de préférence n'est pas très simple car l'individu interrogé doit arbitrer entre plusieurs critères. Il existe des méthodes spécifiques pour traiter ces données, mais elles sortent du cadre de cet ouvrage.

On peut cependant obtenir un certain nombre de résultats en associant à chaque objet classé une variable quantitative, le numéro d'ordre donné par chaque individu (ce point de vue revient à considérer les classements comme des notes). La moyenne et l'écart-type des classements d'un même objet ont un sens facilement interprétable : ils peuvent être comparés directement d'un objet à l'autre. Les corrélations entre ces variables ou avec d'autres variables numériques définies sur les mêmes objets apportent aussi des informations.

On cherche souvent un "classement moyen" agrégeant tous les classements donnés. Une méthode un peu fruste consiste à prendre l'ordre des objets induit par la moyenne des classements. Elle est satisfaisante lorsqu'il y a un (relatif) consensus entre les individus (i.e. quand les classements ne diffèrent pas trop entre eux) ; autrement, les rangs moyens représentent évidemment très mal les données.

6. Fréquence et tableau de contingence

Du comptage au tableau de contingence

Imaginons que l'on ajoute au fichier de notes la nouvelle variable suivante : nombre de notes supérieures ou égales à 15 quelle que soit l'épreuve. Chaque élève ayant 20 notes, cette nouvelle variable varie entre 0 et 20.

On peut la traiter comme une variable quantitative ordinaire en la considérant comme un indicateur d'excellence de l'élève. Le fait que chaque individu possède le même nombre de notes est ici bien pratique. Mais ce n'est en général pas le cas ; par exemple, si l'on avait inclu les options, ce nombre aurait varié d'un individu à l'autre.

Il est alors naturel de considérer le pourcentage (de notes supérieures à 15), et non le nombre, pour se ramener à un indicateur ayant le même intervalle de variation pour chaque individu. En fait, en procédant ainsi, on ne prend pas en compte tous les aspects spécifiques des comptages. Pour cela, il est nécessaire de considérer une variable comptage comme une colonne d'un tableau de contingence, mis en évidence en adoptant la démarche suivante :

- identifier la nature des éléments comptés (ici des notes) et l'ensemble de tous les éléments comptés (ici l'ensemble des notes, toutes épreuves et tous élèves confondus) ;
- considérer un ensemble de variables de comptage tel que chaque note apparaisse une et une fois dans les données ; ici, cela est possible en considérant une seconde colonne : nombre de notes inférieures à 15 ; la nécessité de faire apparaître toutes les notes peut suggérer d'utiliser un découpage plus fin, en distinguant par exemple les notes très basses (<5), intermédiaires et très bonnes (≥ 15) ;
- expliciter dans cet ensemble de colonnes la structure d'un tableau de contingence, à savoir un ensemble d'éléments décrits par deux variables qualitatives ; ici, l'ensemble

des notes est ventilé selon le niveau de la note (variable qualitative à 3 modalités : *très bas*, *intermédiaire*, *très bon*) et l'élève auquel elle a été attribuée (variable qualitative à 909 modalités).

Spécificité du tableau de contingence

A la différence des données considérées jusqu'ici, un tableau de contingence ne croise pas des individus et des variables mais les modalités de deux variables qualitatives (ou quantitatives codées en classes). Contrairement aux tableaux précédents, les lignes et les colonnes jouent des rôles symétriques.

Le cœur de la problématique associée au tableau de contingence respecte cette symétrie : étude de la liaison entre deux variables qualitatives. Dans l'exemple, cela revient à étudier la liaison entre l'élève et le niveau des notes : par rapport à l'approche qui ne considère que la moyenne, on peut distinguer ici, parmi les élèves moyens, ceux qui sont « uniformément moyens » de ceux qui ont obtenu à la fois de bonnes et de mauvaises notes.

Etude

Les notions de base utilisées dans l'étude de tels tableaux sont introduits en 8.3. Rappelons-les rapidement. On calcule les sommes des lignes et des colonnes (marges) qui contiennent la distribution de chacune des variables considérées séparément ; on étudie le tableau à travers ses lignes (ou à travers ses colonnes) en calculant les pourcentages en lignes (ou en colonnes), qui représentent les distributions conditionnelles d'une variable, l'autre étant fixée ; on compare ces distributions conditionnelles entre elles ou à la marge en mettant en évidence les écarts les plus importants. Signalons ici une méthode plus sophistiquée, particulièrement bien adaptée à l'étude de ces tableaux, qui permet d'en faire rapidement une synthèse globale en extrayant les éléments les plus marquants de la liaison entre les deux variables qualitatives. Cette méthode, qui s'apparente à l'analyse en composantes principales, est l'*analyse des correspondances*.

Tableau de contingence ou variables numériques

Un exemple extérieur au fichier des notes permet d'illustrer l'intérêt d'aborder des variables de comptage au travers d'un tableau de contingence.

Dans les analyses de données textuelles, on construit souvent de très grands tableaux de contingence croisant I textes et J mots en indiquant, à l'intersection de la ligne i et de la colonne j, le nombre x_{ij} de fois que le mot j est utilisé dans le texte i. Les textes peuvent être des réponses à des questions ouvertes, des tracts politiques, des œuvres littéraires de différents auteurs, etc.

A première vue, on pourrait penser que ce type de comptage peut être étudié comme des variables numériques en prenant par exemple comme individus les textes et comme variables les mots ou l'inverse. Déjà, on peut avoir un soupçon sur cette façon de considérer les données en voyant que l'on peut permuter le rôle des lignes et des colonnes. Mais, le problème est beaucoup plus fondamental. Que veut-on étudier à travers un tel tableau ? On veut étudier et comparer le vocabulaire utilisé dans les textes. Or, ces textes peuvent être plus ou moins longs : les questionnés ou les hommes politiques peuvent être plus ou moins bavards, les auteurs plus ou moins prolixes. Dans la comparaison entre les textes, ce ne sont pas les différences dues à la longueur qui sont essentielles. Cette longueur, nombre de mots d'un même texte, apparaît dans l'une des marges du tableau (marge-colonne si les textes constituent les lignes du tableau).

Pour éliminer l'effet de cette longueur, on considère, pour chaque texte, non pas les effectifs de chaque mot mais les *pourcentages d'utilisation de chaque mot* obtenus en divisant ces effectifs par la longueur totale du texte. On peut penser qu'il suffit d'effectuer la transformation en pourcentages et de considérer ces pourcentages comme des variables numériques. Mais ce n'est pas encore la meilleure solution car ce qui vient d'être dit pour les textes est aussi valable pour les mots. On veut comparer leur fréquence d'utilisation dans les différents textes. Or, certains (comme *le*, *la*, *les*, *un* ou *une* par exemple) peuvent être systématiquement très fréquents alors que d'autres sont beaucoup plus rares. L'influence des effectifs totaux de chaque mot (contenus dans la marge-ligne si les mots constituent les colonnes du tableau) est éliminée en considérant pour chaque mot, au lieu des effectifs d'utilisation par chaque texte, ses *pourcentages d'utilisation par chaque texte* obtenus en divisant ces effectifs par l'effectif total du mot. Cette formulation symétrique en termes de lignes et de colonnes correspond exactement à l'approche des données en tant que *tableau de fréquence* et non en tant que *tableau individus×variables*.

Juxtaposition de tableaux de contingence

La prise en compte de plusieurs variables de comptage peut conduire à considérer simultanément plusieurs tableaux de contingence juxtaposés. Cela revient à étudier la liaison entre plusieurs variables qualitatives. Selon les données et les objectifs, plusieurs juxtapositions sont possibles. L'étude de ces tableaux est complexe et nécessite le recours à l'analyse des correspondances.

Données manquantes

Les données manquantes sont toujours très délicates à manier : aucune solution n'est idéale. Leur prise en compte dépend de beaucoup de facteurs qui s'imbriquent :

- la nature de la donnée manquante (oubli, sans objet, hors mesure, etc.) ;
- le nombre et la répartition des données manquantes ;
- la nature de la variable concernée (quantitative ou qualitative) ;
- le traitement statistique envisagé.

Nous ne prétendons pas faire une étude exhaustive de tous les cas possibles ni indiquer de solution à appliquer systématiquement ; le chapitre 4 décrit un exemple de démarche vis-à-vis d'un fichier comportant un type de données manquantes : nous en présentons ici d'autres types et quelques idées illustrées par des exemples.

1. Quelques types de données manquantes

Pour montrer la variété du problème, commençons par décrire quelques types de données manquantes avec des propositions de traitement. C'est dans les enquêtes que l'on en trouve peut-être le plus grand éventail.

1.1 Donnée n'existant pas

Enquête avec questions hiérarchisées

Les questions sont dites hiérarchisées lorsqu'une ou plusieurs questions ne concernent que les individus ayant fourni une certaine réponse à une question précédente.

Exemple 1. Dans un questionnaire distribué dans un cinéma, les deux questions suivantes s'enchaînent :

Connaissez-vous la « classification Art et essai » du cinéma ? Si oui, êtes-vous intéressé ?

Seuls les spectateurs ayant répondu *oui* à la première question peuvent répondre à la seconde. Une première solution consiste à mettre une modalité particulière *non-réponse* à la seconde question. Mais cette solution n'est pas très bonne. D'une part la modalité *non-réponse* serait fortement redondante avec la réponse *non* de la première question, ce qui est un défaut gênant lorsque l'on étudie les liaisons entre variables (par exemple en réalisant une analyse des correspondances multiples) ; d'autre part, cette *non-réponse* regrouperait

deux réalités complètement différentes, la réponse *non* à la première question et la réponse *oui* cumulée avec une absence de réponse à la seconde. Une meilleure solution consiste à recoder l'ensemble des deux variables par une seule comportant 3 réponses : *ne connaît pas*, *connaît et est intéressé*, *connaît et n'est pas intéressé*. Pour éviter un recodage toujours lourd, il vaut mieux prévoir cela dès l'élaboration du questionnaire.

Exemple 2. Dans un questionnaire général sur le transport, les questions suivantes s'enchaînent :

Quel moyen de transport utilisez-vous ?

Si vous utilisez une voiture, quelle est sa puissance ? Quel est le nombre de personnes dans la voiture ? etc.

Le problème est différent du précédent puisque toute une série de questions n'ont pas de réponses si le moyen de transport n'est pas la voiture : un recodage simple est difficile à définir. Comme ci-dessus, une première solution, valable uniquement pour les variables qualitatives ou codées en qualitatives, est de créer une modalité *non-réponse* ; le problème de redondance y est encore plus marqué puisqu'il concerne plusieurs questions. On peut aussi faire des études partielles en se restreignant pour l'étude de ces sous-questions à la population concernée et en se restreignant aux autres questions pour la population entière. Il est clair que ces différentes approches se complètent.

Individus non concernés par une ou plusieurs variables

Ce cas s'apparente au précédent.

Exemples 3 et 4. Dans le fichier des notes, les candidats passant au bac uniquement les épreuves de mathématiques et de physique ne sont pas concernés par les autres notes ; dans une enquête auprès de la clientèle d'une boîte de nuit, les personnes venant pour la première fois ne peuvent répondre aux questions concernant l'évolution de cette boîte de nuit.

L'attitude à suivre dépend à la fois du nombre d'individus non concernés, de leur typicité par rapport aux autres individus et du nombre de variables touchées. Si ces individus sont peu nombreux, on peut restreindre le champ de l'étude aux autres individus. Sinon, comme dans le cas précédent, il ne reste guère que deux solutions :

- créer des modalités *réponse manquante*, avec les problèmes de redondance déjà évoqués ;
- limiter certaines études à un sous-ensemble des variables.

Dans ces deux cas. les individus non concernés sont un peu particuliers : il peut être intéressant de les comparer au reste de la population pour les autres variables (notes connues, autres questions).

Exemple 5 : les morts ou disparus. Il est fréquent de suivre l'évolution dans le temps d'un ensemble d'individus (par exemple des animaux) et de vouloir analyser le tableau dans lequel chaque variable est une mesure effectuée à une date donnée. Dans ce tableau, les animaux qui meurent au cours du suivi ont des valeurs manquantes à partir du jour de leur mort. Comme dans les exemples précédents, ces animaux sont particuliers et généralement un volet de l'étude consiste à les détecter à partir des premières valeurs relevées. Le

problème est que les animaux ne meurent pas tous le même jour et que le relevé des données s'arrête à des dates différentes selon l'animal. On ne peut comparer entre eux les animaux qu'à partir des variables communes, c'est-à-dire avant le premier décès. On peut aussi éventuellement, pour faire cette comparaison, séparer l'ensemble de ces animaux morts en plusieurs sous-populations suivant la date de décès.

Non-réponses motivées

Exemple : dans un examen sous forme de QCM (questions à choix multiples), un étudiant peut ne pas connaître la réponse à une question.

Dans ce cas, la non-réponse correspond à une attitude particulière devant la question : il est tout à fait justifié de créer une modalité *non-réponse* qui traduit cette attitude. Pour que cette modalité soit utilisée, les examinateurs peuvent adopter un barème qui sanctionne plus une mauvaise réponse qu'une non-réponse.

1.2 Donnée existante mais non connue

C'est typiquement les cas des données perdues, ou non recueillies par erreur ou impossibilité. On peut ranger ici les questions jugées trop personnelles qui engendrent des refus de répondre.

La plupart des commentaires précédents s'appliquent ici. En particulier on doit toujours se demander d'abord si la donnée manquante n'a pas une signification particulière. Exemple : une valeur trop grande ou trop petite pour être mesurée avec l'appareil disponible.

La spécificité ici est que l'estimation des données non disponibles a un sens.

2. Prise en compte dans les traitements statistiques

2.1 Etude d'une seule variable

Dans la description d'une seule variable, le traitement des données manquantes s'impose assez facilement. On calcule d'abord l'effectif des individus avec données manquantes et l'étude se fait sur la population réellement décrite. Exemple : les résultats d'élections exprimés en pourcentages par rapport au nombre de votants.

2.2 Etude simultanée de plusieurs variables

Il est dangereux de comparer entre elles des variables ayant des valeurs manquantes (cf. 4.6). L'étude simultanée de plusieurs variables implique de travailler sur les mêmes individus pour chaque variable. D'ailleurs, cette contrainte est techniquement indispensable pour mettre en œuvre une analyse en composantes principales.

Pour la satisfaire, plusieurs actions sont possibles :

- supprimer les individus (ou éventuellement les variables) ayant des valeurs manquantes ; dans ce cas, attention à expliciter le changement du champ de l'étude ;
- pour une variable qualitative, créer une modalité particulière *donnée manquante* ;

- pour une variable quantitative, la recoder en qualitative et créer une modalité *donnée manquante* ;
- recoder simultanément plusieurs variables qualitatives (cf. exemple 1 de cette fiche) ;
- affecter une valeur, avec le danger de traiter des données erronées : par exemple, une modalité au hasard ou « plausible » pour une variable qualitative, la moyenne générale (ou une "estimation" plus fine) pour une variable numérique.

Ces solutions sont les plus fréquentes.

2.3 Estimation et type de la variable concernée

A) *Pour les variables qualitatives* on peut créer une modalité *donnée manquante*. Cette solution simple et parfaite sur un plan informatique n'est pas la panacée quant à l'analyse statistique des résultats. Elle est bien adaptée seulement lorsque la donnée manquante est fondamentalement une modalité en elle-même (donnée normalement absente ou réponse traduisant une attitude particulière). Sinon, cette modalité particulière et complètement artificielle risque de perturber les résultats.

Si la donnée manquante existe mais n'est pas connue, on peut essayer d'affecter à l'individu l'une des modalités. L'affectation peut se faire au hasard ; on peut aussi affecter la modalité la plus fréquente ou tenir compte des autres valeurs de l'individu. Lorsque le nombre d'individus est très grand et le nombre de données manquantes faible, ce sont des solutions possibles si l'on veut des résultats globaux sans s'intéresser à chaque individu.

B) *Pour une variable numérique*, une solution particulièrement simple pour combler le vide dû à une valeur manquante est d'y mettre la moyenne. Si cette valeur manquante n'a pas de signification particulière et si le pourcentage de valeurs manquantes dans la variable est faible, il n'y a guère de danger, même si cette solution biaise toujours un peu les données : l'écart-type notamment est systématiquement sous-évalué dès lors que des valeurs sont remplacées par la moyenne. Par contre, si la donnée manquante a une signification, il est clair que l'on déforme notablement les données réelles : la particularité des individus concernés est effacée.

3. Conclusion

La prise en compte de données manquantes est toujours délicate ; elle s'appuie sur une analyse de leur signification et un dénombrement systématique par individu et par variable. Quels que soient les choix faits, on n'obtient jamais un traitement parfait des données ; il faut se résigner à un certain empirisme et en tenir compte pour nuancer ses conclusions.

Il est prudent de se méfier des logiciels qui « gèrent les données manquantes », évitant à l'utilisateur cette réflexion en imposant une solution unique, souvent non précisée, quelquefois inadaptée voire totalement absurde (on a même vu un logiciel remplaçant les valeurs manquantes par le numéro d'ordre de l'individu dans le fichier sans que les utilisateurs s'en émeuvent le moins du monde puisque « la machine s'en occupait »).

Mesure de la dispersion d'une variable quantitative

1 Pourquoi mesurer la dispersion ?

Lorsque, pour une variable, tous les individus présentent la même valeur, l'étude de la variable se résume à la donnée de la valeur. Si, au contraire, les individus ne présentent pas la même valeur, il existe une dispersion (ou variabilité) ; l'objet de l'analyse des données en général est précisément d'étudier des dispersions.

Lorsque l'on s'intéresse à un individu en particulier, on le situe en calculant l'écart entre sa valeur et la moyenne de l'ensemble des valeurs. Lorsque l'on s'intéresse à l'ensemble des individus, on examine l'ensemble des écarts à la moyenne. Un histogramme met bien en évidence la dispersion des individus, en particulier autour de la moyenne.

On peut chercher à quantifier cette dispersion principalement dans deux optiques :

- répondre directement à une question du type « quel est l'écart moyen à la moyenne ? »
- comparer deux distributions du point de vue de leur dispersion ; la comparaison visuelle d'histogrammes selon un seul point de vue (ici la dispersion) n'est pas toujours facile car les histogrammes peuvent différer selon plusieurs points de vue souvent difficiles à démêler (par exemple, outre la dispersion, deux distributions peuvent différer par la présence d'asymétrie ou de queues de distribution ; cf. Ch. 14).

2 Ecart absolu moyen (E_m)

Notations

I : nombre d'individus sur lesquels on a mesuré la variable quantitative X ; x_i : valeur de l'individu i pour la variable X ; $\bar{x}$: moyenne de la variable X.

Définition

E_m : moyenne des valeurs absolues des écarts à la moyenne.

$$E_m = \frac{1}{I}\sum_i |x_i - \bar{x}|$$

Remarques

La valeur absolue évite aux écarts (à la moyenne) de signes contraires de s'annuler (sans la valeur absolue, la somme des écarts à la moyenne est toujours nulle).

La division par I élimine l'influence du nombre des valeurs prises en compte.

E_m est une mesure directe et intuitive, utilisée pour répondre à la question : quel est l'écart moyen à la moyenne ? En dehors de cela, on lui préfère l'écart-type.

La formule ci-dessus est donnée pour des individus de même poids. Si on affecte le poids p_i à l'individu i, la définition devient :

$$E_m = \frac{1}{\sum p_i} \sum_i p_i |x_i - \bar{x}|$$

*Exemples extraits du fichier des notes (cf. **Tab. 1**)*

	Moyenne	Ec. abs. moyen	Ecart-type	Ec. interquart.
Mathématiques	13.21	2.61	3.20	5.00
Philosophie	7.84	2.56	3.29	3.00

Tableau 1. *Moyenne et indicateurs de dispersion de deux notes au bac*

Les moyennes de ces deux matières diffèrent très sensiblement. Mesurée par l'écart absolu moyen, la dispersion autour de la moyenne est pratiquement identique dans ces deux cas.

3 Ecart-type (s) et variance (s^2)

Définitions

s^2 (variance) : moyenne des carrés des écarts à la moyenne (s : écart-type).

$$s^2 = \frac{1}{I} \sum_i (x_i - \bar{x})^2$$

Remarques

L'élévation au carré des écarts ($x_i - \bar{x}$) évite aux écarts de signes contraires de s'annuler dans la sommation ; elle joue un rôle analogue à celui de la valeur absolue dans l'écart absolu moyen, mais accentue l'influence des valeurs extrêmes (cf. **Fig. 1**).

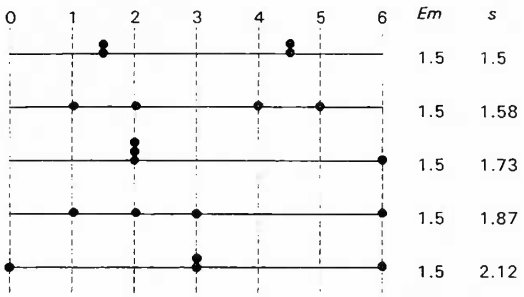


Figure 1. *En lignes, 5 distributions de 4 points ayant même moyenne (3) et même écart absolu moyen (1.5). De haut en bas, l'écart-type croît au fur et à mesure de l'apparition de points extrêmes.*

La division par I élimine l'influence du nombre de valeurs prises en compte ; la variance est une moyenne de carrés.

L'intérêt de l'écart-type s, par rapport à la variance s^2, est de s'exprimer dans les mêmes unités de mesure que la variable.

Lorsque l'on multiplie chaque valeur x_i par une constante k, les indicateurs s et E_m sont multipliés par k ; ces indicateurs dépendent donc des unités de mesure : on ne peut comparer des dispersions que si les valeurs sont exprimées dans les mêmes unités. D'où l'idée du coefficient de variation ($CV = s/\bar{x}$ ou $100s/\bar{x}$), qui exprime la dispersion en « pourcentage de la moyenne $\bar{x}$ ».

On peut montrer que l'écart-type est supérieur ou égal à l'écart absolu moyen.

La formule ci-dessus est donnée pour des individus de même poids. Si on affecte le poids p_i à l'individu i, la définition devient :

$$s^2 = \frac{1}{\sum p_i} \sum_i p_i (x_i - \bar{x})^2$$

Exemples extraits du fichier des notes

Les dispersions de *mathématiques* et de *philosophie* (cf. **Tab. 1**) sont très proches, quelle que soit la mesure de dispersion utilisée. Remarquons que l'écart absolu moyen et l'écart-type sont à peu près dans le même rapport pour ces deux matières. En fait, pour l'ensemble des 20 notes du fichier, le rapport E_m/s est toujours compris entre .77 et .82 ; plus généralement, en pratique, ce rapport s'écarte peu de .80, valeur qu'il prend pour une loi normale.

Sur cet exemple, écart-type et écart absolu moyen ne classent pas les deux matières dans le même ordre car les distributions diffèrent selon leur forme (cf. Ch. 14).

Pourquoi utiliser l'écart-type ?

L'interprétation directe de l'écart-type est moins intuitive que celle de l'écart absolu moyen. Quelle est donc l'origine de l'utilisation systématique de l'écart-type ?

Réponse : la variance, et par voie de conséquence l'écart-type, se prête mieux aux calculs théoriques que l'écart absolu moyen ; de ce fait, elle possède bon nombre de propriétés précieuses pour une mesure de dispersion. Nous en donnons les principales.

• On peut songer à mesurer la dispersion d'un ensemble de valeurs à partir de l'écart entre deux valeurs quelconques (ce point de vue est utilisé pour juger de la reproductibilité d'un appareil de mesure). On peut montrer que :

$$\frac{1}{I^2} \sum_i \sum_j (x_i - x_j)^2 = 2s^2$$

Autrement dit, en considérant les carrés des écarts, il revient au même de mesurer la dispersion en utilisant les écarts par rapport à la moyenne ou les écarts entre les valeurs prises deux à deux. Cette propriété n'est pas vérifiée par l'écart absolu moyen.

• La moyenne $\bar{x}$ est la valeur de v qui rend minimum :

$$\sum (x_i - v)^2$$

Ainsi la moyenne est la valeur autour de laquelle la dispersion de l'ensemble des valeurs est minimum, à condition de mesurer cette dispersion par la somme des carrés des écarts. En ce sens la moyenne représente l'ensemble des valeurs.

• Lorsque deux variables X et Y sont non corrélées, la variance de leur somme $S=X+Y$ est égale à la somme de leurs variances (cf. Fiche 11 § 3.2). Cette additivité des variances correspond bien à l'intuition de l'additivité des dispersions, par exemple dans le cas d'erreurs de mesure.

• L'écart-type s est plus facile à relier à la répartition des valeurs que l'écart absolu moyen E_m.

. Quelle que soit la distribution, on peut montrer que (inégalité de Bienaymé) :

$$P[|x - \bar{x}| \geq a] \leq \frac{s^2}{a^2}$$

Ainsi, par exemple, au moins 3/4 des valeurs d'une distribution sont situées à moins de deux écarts-types de la moyenne. Cette proportion de 3/4 est un minimum atteint dans un cas très particulier (distribution dont 80% des valeurs sont confondues avec la moyenne $\bar{x}$, le reste se répartissant également sur les valeurs $\bar{x}-2s$ et $\bar{x}+2s$) ; en pratique cette proportion est sensiblement plus élevée ; dans le cas de la loi normale elle est de 95% (plus précisément, cette probabilité correspond à l'intervalle $[\bar{x}-1.96s\ ,\ \bar{x}+1.96s]$)

. La distribution normale, dont le rôle théorique et pratique est important, s'exprime de façon simple en fonction de son écart-type.

• La variance n'est pas un concept purement artificiel puisqu'elle s'interprète physiquement comme une inertie : celle du nuage des points-individus par rapport à leur centre de gravité.

• La variance intervient dans le coefficient de corrélation, précisément pour harmoniser les dispersions de deux variables avant d'étudier leur liaison.

Pour ces raisons, la dispersion des variables est presque toujours prise en compte au travers de la variance ou de l'écart-type.

Restriction à l'utilisation de l'écart-type

Plaçons-nous dans le cas où l'on veut comparer deux distributions du point de vue de leur dispersion.

Lorsque les distributions ont la même forme, E_m et s sont dans le même rapport et il est indifférent d'utiliser l'un ou l'autre de ces deux critères. Lorsque les distributions ont des formes différentes la non-équivalence entre les deux critères traduit que, dans ce cas, l'objectif de comparaison des dispersions est insuffisamment spécifié et qu'il convient de préciser le point de vue (i.e. l'indicateur) selon lequel la comparaison doit s'effectuer.

En pratique, il est rarement nécessaire de comparer finement les dispersions de distributions dont les formes sont très différentes (finement signifie ici : au point d'être gêné par l'influence de la nature de l'indicateur choisi sur le résultat).

4 Etendue et écart interquantile

Définitions

Une appréciation directe de la dispersion consiste à donner une plage de valeurs contenant un pourcentage donné d'observations. Plusieurs indicateurs sont possibles.

Le plus intuitif est l'étendue (parfois appelée amplitude), notée a, longueur du plus petit intervalle qui contient 100% des valeurs observées :

$$a = \text{maximum} - \text{minimum}$$

L'intervalle interquartile (IQR) est celui qui contient 50% des valeurs, les autres se répartissant également au-dessus et au-dessous. Cette mesure de dispersion ne prend en compte que la partie centrale de la distribution. Cette définition vaut lorsqu'il n'y a pas d'ex æquo. Dans le cas contraire, on doit utiliser la définition plus générale suivante :

IQR = Q3−Q1 avec :

- Q1 (premier quartile), valeur telle que 25% des observations lui sont inférieures ou égales ;
- Q3 (troisième quartile), valeur telle que 75% des observations lui sont inférieures ou égales.

L'IQR est la longueur des boîtes de dispersion (cf. § 7.7).

De même que les quartiles fragmentent une distribution en 4 parties de même effectif (le second quartile est la médiane), on définit les déciles (resp. centiles) qui fragmentent une distribution en 10 (resp. 100) parties de même effectif. Plus généralement, on appelle *quantile d'ordre p* $(0<p<1)$, la valeur telle que les observations qui lui sont inférieures sont en proportion p.

Exemples extraits du fichier des notes

Les dispersions globales des notes dans les deux matières du tableau 1, mesurées aussi bien par l'écart-type que par l'écart absolu moyen, sont comparables ; la comparaison entre les écarts interquartiles suggère que la majorité des notes est toutefois moins dispersée en philosophie qu'en mathématiques. En fait, la différence entre les écarts interquartiles est difficilement interprétable ici car, du fait des ex æquo, cet écart contient 63% des valeurs en mathématiques et 51% des valeurs en philosophie.

Remarques

On peut imaginer d'autres intervalles du même type, contenant une proportion donnée des valeurs observées, par exemple 95%. Cette idée rejoint celle qui conduit à considérer l'intervalle $(\bar{x}-1.96s,\ \bar{x}+1.96s)$, qui contient 95% des valeurs dans le cas d'une distribution normale. Ce type d'intervalle est précieux pour décrire une distribution car il s'adapte bien à une distribution donnée : ainsi, on peut dire que, au bac, 95.3% des notes en mathématiques sont comprises entre 7 et 18 et que 94,2% des notes en philosophie sont comprises entre 3 et 14. En revanche la comparaison, par ce type d'indicateur, de distributions comportant de nombreux ex æquo n'est pas toujours aisée du fait de l'impossibilité de calculer un intervalle comportant exactement un pourcentage donné d'observations.

De tels intervalles ne prennent en compte que la partie centrale de la distribution.

5 Niveau d'un échantillon ou d'une population

Les calculettes usuelles et les tableurs fournissent souvent deux valeurs pour la variance d'une série de I nombres :

s^2 : valeur définie dans cette fiche ;

$$s'^2 = \frac{1}{I-1}\sum_i (x_i - \bar{x})^2$$

Cette seconde relation ne donne pas l'écart-type de l'ensemble des valeurs entrées dans la calculette. Elle prend un sens dans le cas où l'ensemble des valeurs entrées constitue un échantillon résultant d'un tirage au hasard dans une population dite alors parente.

Dans ce cas, chaque valeur x_i est la réalisation d'une variable aléatoire notée classiquement X_i. La distribution de X_i est l'ensemble des valeurs possibles pour la $i^{ème}$ valeur (celles auxquelles on peut s'attendre avant le tirage) ; x_i est celle d'entre elles qui résulte du $i^{ème}$ tirage aléatoire.

Pour obtenir une estimation de la variance de la population parente (variance inconnue notée classiquement σ^2), il est tentant d'utiliser la variance de l'échantillon (s^2). Toujours du fait du caractère aléatoire de l'échantillon, la valeur s^2 (dite estimation) est la réalisation de la variable aléatoire S^2 correspondante (dite estimateur) :

$$S^2 = \frac{1}{I}\sum_i (X_i - \bar{X})^2 \quad \text{avec } \bar{X} = \frac{1}{I}\sum_i X_i$$

Or, on montre facilement que la valeur s^2 a tendance à sous-estimer la vraie valeur σ^2 ; en d'autres termes, l'estimateur S^2 fournit en moyenne une estimation plus petite que σ^2. Plus précisément, l'espérance de S^2 diffère de σ^2, valeur que l'on cherche à estimer (un tel estimateur est dit biaisé), et vaut :

$$E[S^2] = \frac{I-1}{I}\sigma^2$$

D'où l'idée, pour obtenir une estimation de σ^2, d'utiliser s'^2, réalisation de la variable aléatoire S'^2 qui s'écrit :

$$S'^2 = \frac{1}{I-1}\sum_i (X_i - \bar{X})^2$$

et qui constitue un estimateur non biaisé (i.e. fournissant **en moyenne** la bonne valeur).

En fait, dès que le nombre de valeurs est grand, la différence entre les deux estimations est très faible et sans importance pratique.

Représentation simultanée de deux variables quantitatives

1. Exemple dans lequel le choix des unités de mesure s'impose

Dans le chapitre 8, on étudie la liaison entre les notes en mathématiques, au bac et au troisième trimestre. Un des éléments importants de cette étude est la représentation graphique dans laquelle chaque élève est représenté par un point sur un plan défini par deux axes : l'axe horizontal correspond à la note au troisième trimestre, l'axe vertical correspond à la note au bac. Ces deux axes sont gradués de 0 à 20 (cf. **Fig. 8.1**).

L'étude de la liaison est faite à partir de la forme du nuage de points. En effet la forme du nuage met en évidence (cf. 8.1 et 8.2) :

- d'éventuels outliers ;
- d'éventuelles sous-populations ;
- les distributions conditionnelles.

2. Exemple dans lequel le choix des unités de mesure pose problème

2.1 Données

Une hypothèse est quelquefois formulée : les grands lycées, c'est-à-dire ceux dont l'effectif est important, obtiennent de meilleurs résultats que les petits. Nous abordons ici un aspect de cette hypothèse en étudiant la liaison entre :

- la moyenne en mathématiques obtenue au bac par les élèves d'un même lycée ;
- l'effectif des élèves de ce lycée se présentant au bac C.

Pour cela, on représente les 22 lycées sur un graphique dont l'axe horizontal correspond à la moyenne en mathématiques au bac et l'axe vertical à l'effectif.

2.2 Le problème des unités de mesure

Pour tracer ce graphique, se pose d'emblée un problème d'échelle. Dans le cas de l'exemple 1 ce problème n'apparaît pas : chaque axe est gradué de 0 à 20 et il est naturel de prendre la même unité pour chacun. Or, dans ce nouveau cas :

- les moyennes par lycée en mathématiques varient de 9.45 à 15.27 ;
- les effectifs des lycées varient de 11 à 82.

Si l'on prend comme unité (par exemple 1 cm) :

- un point de moyenne pour l'axe horizontal,

- un élève pour l'axe vertical,

le nuage de points va s'étirer sur une bande verticale étroite de 71 cm de long (71=82–11) et de 5.82 cm de large (5.82=15.27–9.45). Un tel graphique est peu commode car le nuage de points présente une forme très allongée due uniquement à la plus grande dispersion le long de l'axe vertical ; cette plus grande dispersion est elle-même liée au fait que la même longueur représente un effectif de 1 élève et une note de 1 point.

Pour équilibrer les dispersions horizontale et verticale, une idée intuitive est d'exprimer les effectifs en dizaines d'élèves ; les effectifs varient alors non plus de 11 à 82 mais de 1.1 à 8.2 ; cette dispersion est comparable à celle de la variable *moyenne en mathématiques* et le graphique est techniquement plus facile à réaliser.

2.3 Equilibrer les intervalles de mesure

Une façon d'exploiter cette idée plus systématiquement est d'équilibrer les intervalles de mesure. En pratique, on fixe la taille du graphique, carré de préférence, puis on choisit les unités de façon telle que les intervalles de mesure occupent l'ensemble de la place disponible. Ainsi, dans la Figure 1.A, les abscisses varient de 0 à 20 et les ordonnées de 0 à 100. Ce graphique présente l'intérêt, au moins pour les moyennes en mathématiques, de visualiser l'étendue des valeurs observées par rapport aux valeurs possibles. Mais la faible étendue des valeurs observées gêne l'étude de la liaison entre les deux variables.

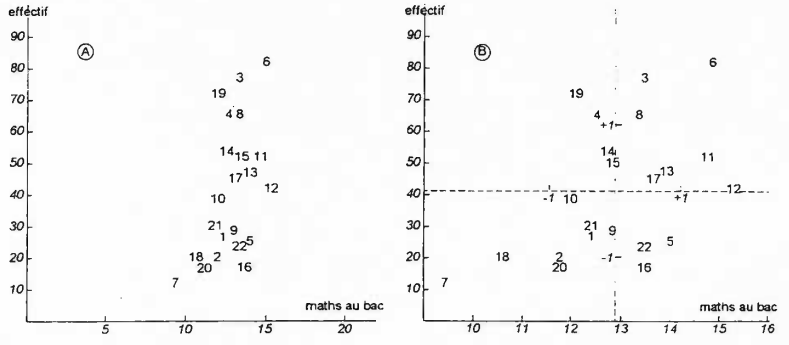


Figure 1. *Les 22 lycées représentés par leur moyenne en mathématiques et leur effectif. **A** : équilibre des intervalles de mesure ; **B** : équilibre des dispersions ; les données sont centrées-réduites puis représentées dans le repère des axes en pointillés ; les unités de mesure initiales sont reportées sur le cadre.*

2.4 Equilibrer les variabilités observées

A partir d'un graphique carré de taille donnée, on peut choisir les abscisses et les ordonnées en faisant coïncider les limites du graphique avec les minimums et les maximums observés. Ce faisant, les variabilités des deux variables sont équilibrées au sens de l'étendue.

Une idée apparentée est à l'origine du centrage et de la réduction des valeurs avant de les représenter ; cela revient à prendre :

- du fait du centrage, comme origine des axes le point moyen du nuage, correspondant au lycée (imaginaire) dont l'effectif est l'effectif moyen et la note en mathématiques la moyenne des notes des lycées ; ainsi ce point moyen est mis en évidence, ce qui permet de visualiser immédiatement les lycées supérieurs ou inférieurs à la moyenne du point de vue de l'un et/ou de l'autre des axes ;
- du fait de la réduction, comme unité l'écart-type ; l'unité de mesure d'une variable est ainsi d'autant plus grande que cette variable présente une grande dispersion ; les nouvelles variables ont alors chacune un écart-type de 1, et de ce fait exactement la même dispersion.

Naturellement, il convient de reporter aussi les unités de mesure originelle, par exemple sur le bord du cadre (cf. **Fig. 1.B**).

2.5 Direction d'allongement et bissectrices

Sur ce graphique, la relation entre les deux variables paraît assez lâche : on peut remarquer toutefois que c'est le lycée d'effectif le plus faible (lycée 7) qui a obtenu la moyenne en mathématiques la plus faible.

Dans l'étude d'un nuage de points, on cherche souvent à reconnaître une direction d'allongement. Ainsi le nuage des lycées semble présenter une forme plutôt allongée dans la direction de la première bissectrice. Cet allongement exprime que, en général, les lycées d'effectif faible obtiennent des résultats en mathématiques inférieurs à ceux obtenus par les lycées d'effectif important.

La quantification de cet aspect de la forme d'un nuage peut se faire à l'aide du coefficient de corrélation qui vaut ici .47, valeur à laquelle on peut associer, compte tenu du nombre de points (22), la probabilité .03 (cf. 8.4 et Fiche 5).

Remarquons que si le nuage centré-réduit présente une forme allongée, alors cet allongement se fait nécessairement dans la direction de l'une des deux bissectrices. Ce résultat est dû à la réduction simultanée des deux variables, qui égalise les variabilités le long de l'axe horizontal et vertical (une forme allongée dans une autre direction que les bissectrices est incompatible avec l'égalité des dispersions le long des axes de coordonnées).

Liaison entre deux variables quantitatives

1 Du graphique au coefficient de corrélation

1.1 Importance du graphique pour étudier la liaison entre deux variables quantitatives

Les notes en mathématiques obtenues au troisième trimestre et le jour du bac sont-elles liées ? Autrement dit, les élèves qui obtiennent une bonne note dans un cas obtiennent-ils généralement aussi une bonne note dans l'autre et réciproquement ? Pour répondre à cette question, le meilleur outil possible est le graphique qui croise ces deux variables. Ce type de graphique est décrit ailleurs (cf. 8.4 et Fiche 4) ; réalisé pour ces deux notes (cf. **Fig. 8.1**), il met en évidence leur liaison.

La valeur pratique du graphique tient à ce que, en général, la relation entre deux variables ne peut pas être résumée par une formule simple. L'examen du graphique permet ainsi, pour chaque valeur de l'une des notes, de mettre en évidence la répartition des valeurs pour l'autre note. Par exemple, parmi les élèves ayant eu 6 en mathématiques au troisième trimestre, la plupart ont obtenu entre 3 et 5 le jour du bac et deux seulement ont eu la moyenne.

1.2 Nécessité d'un indicateur

On peut se poser la même question à propos d'autres notes, celles en mathématiques et en philosophie au 1er trimestre (cf. **Fig. 13.2**). L'examen du graphique croisant ces deux matières met en évidence une forme assez confuse. On a bien l'impression que le nuage de points présente une forme allongée, mais cette impression est peu nette et l'on peut craindre de ne pas bien prendre en compte visuellement les superpositions de points.

On atteint ici les limites d'une appréciation visuelle et il apparaît nécessaire de recourir à un indicateur de forme de nuage de points.

1.3 Définition du coefficient de corrélation linéaire

L'indicateur de forme le plus répandu est, sans conteste, le coefficient de corrélation linéaire (introduit en 8.4). Il mesure la parenté de forme entre un nuage de points et une droite. La référence à une droite est naturelle :

- c'est la forme de nuage la plus simple ;
- la droite exprime une relation très simple entre deux variables, à savoir une relation de la forme $Y=aX+b$, appelée de ce fait relation linéaire (la forme la plus simple de cette relation est obtenue pour $a=1$ et $b=0$, ce qui conduit à l'équation $Y=X$ correspondant à la première bissectrice).

Le coefficient de corrélation se calcule à partir des variables centrées-réduites. Cette transformation préalable le rend insensible aux changements d'unité de mesure. Elle conduit à considérer un nuage dont les variabilités le long des axes de coordonnées sont identiques, ce qui facilite l'étude de la forme du nuage de points. En outre, l'égalité des variances des variables ainsi transformées implique que, si le nuage de points présente une forme allongée, la direction d'allongement ne peut être que l'une des bissectrices (cf. Fiche 4).

En notant I le nombre de points, X et Y les deux variables et x_i et y_i les valeurs centrées-réduites de X et Y pour le point i, le coefficient de corrélation, noté r, s'écrit :

$$r = \frac{1}{I}\sum_i x_i y_i$$

Il est égal à la moyenne des produits des coordonnées centrées-réduites.

On peut faire apparaître explicitement le centrage et la réduction dans la formule ci-dessus en exprimant r en fonction des données brutes. Soit, en notant x_i et y_i les valeurs brutes, $\bar{x}$ et s_x (resp. $\bar{y}$ et s_y) la moyenne et l'écart-type de X (resp. Y) :

$$r = \frac{1}{I}\sum_i \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\frac{1}{I}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{\mathrm{COV}(X,Y)}{\sqrt{\mathrm{V}(X)\mathrm{V}(Y)}}$$

Dans les deux dernières expressions de r, apparaît la covariance des deux variables X et Y, moyenne des produits des valeurs centrées. Cet indicateur dépend de la liaison entre les deux variables mais aussi de leur dispersion ; en pratique, il n'est utilisé que comme intermédiaire de calcul.

1.4 Deux exemples de données choisies

Nous appliquons ces calculs à deux nuages de points choisis (cf. **Tab. 1** et **Fig. 1**).

	Exemple 1				
	A	*B*	*X*	*Y*	*XY*
a	1	20	-1.18	-.78	.923
b	2	10	-.78	-1.18	.923
c	6	70	.78	1.18	.923
d	7	60	1.18	.78	.923
moy.	4	40	0.00	0.00	**.923**
e.t.	2.55	25.5	1.00	1.00	

Exemple 2				
A	*B*	*X*	*Y*	*XY*
1	3	-1.34	.45	.600
3	1	-.45	-1.34	.600
5	7	.45	1.34	.600
7	5	1.34	.45	.600
4	4	0.00	0.00	**.600**
2.24	2.24	1.00	1.00	

Tableau 1. *Exemples de calcul de coefficient de corrélation.*
A, B : variables brutes ; X, Y : variables A et B centrées-réduites ; XY : variable produit des variables X et Y (moyenne de XY = coefficient de corrélation).

Dans les deux graphiques (cf. **Fig. 1**), le nuage de points est allongé dans la direction de la première bissectrice (notée Z) ; ici, tous les points i sont tels que le produit $x_i y_i$ est positif (lorsque le nuage est allongé dans la direction de la seconde bissectrice (notée T), la plupart des produits $x_i y_i$ sont négatifs). Le signe de r indique donc la direction d'allongement.

En outre, r est d'autant plus grand que la forme du nuage est linéaire, c'est-à-dire, que le nuage centré-réduit est resserré autour d'une bissectrice. Ainsi r est plus grand dans l'exemple 1 que dans l'exemple 2.

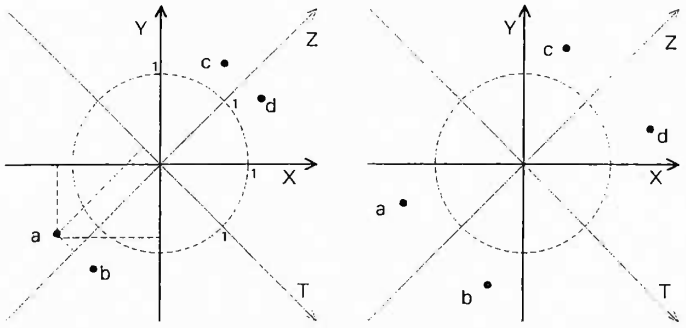


Figure 1. *Représentation graphique des données choisies centrées-réduites (cf.* Tab. 1*) Le cercle de rayon 1 indique l'échelle dans toutes les directions ; les lignes pointillées illustrent la coordonnée du point a sur les axes X, Y, T et Z.*

1.5 Coefficient de corrélation et dispersion le long des bissectrices

Pour appréhender le lien entre r et la forme du nuage centré-réduit, on peut exprimer chaque point i en fonction de ses coordonnées t_i et z_i le long des bissectrices. Alors, on montre que (Z = première bissectrice ; T = seconde bissectrice) :

$$\mathrm{Var}(Z) = 1 + r$$

$$\mathrm{Var}(T) = 1 - r$$

Ces propriétés sont illustrées numériquement sur les données de l'exemple 1 (cf. **Tab. 2**).

	Exemple 1			
	X	*Y*	*Z*	*T*
a	-1.18	-.78	-1.39	-.28
b	-.78	-1.18	-1.39	.28
c	.78	1.18	1.39	-.28
d	1.18	.78	1.39	.28
moy.	0.00	0.00	0.00	0.00
var.	1.00	1.00	1.92	.08

Tableau 2. *Exemple de calcul de dispersion le long des bissectrices (données du tableau 1) La coordonnée sur Z (resp. T) s'obtient par :* $Z=(X+Y)/\sqrt{2}$ *(resp.* $(X-Y)/\sqrt{2}$ *)*

Il découle de ces propriétés que :

- $-1 \leq r \leq 1$ car une variance est toujours positive ;
- si $r=1$, la variance le long de T est nulle et tous les points sont situés sur la première bissectrice ;
- si $r=-1$, la variance le long de Z est nulle et tous les points sont situés sur la seconde bissectrice.

2 Peut-on apprécier le caractère plus ou moins fortuit d'un coefficient de corrélation ?

Cet aspect est détaillé dans la fiche 10 où il sert à illustrer la problématique générale d'association d'une probabilité à un indicateur statistique.

2.1 Comment associer une probabilité à un coefficient de corrélation

Le coefficient de corrélation entre les notes de mathématiques et de philosophie au 1er trimestre est $r = .19$; cela indique une relation linéaire positive entre les deux matières.

On peut toutefois se demander si une telle valeur n'est pas fortuite, sachant bien qu'avec des données réelles on n'obtient jamais très exactement la valeur 0. Pour cela, on suit le raisonnement ci-après :

- on considère les deux ensembles de 909 notes de mathématiques et philosophie sans tenir compte des élèves qui les ont obtenues ;
- on considère tous les appariements possibles entre ces deux ensembles de notes ; un appariement est un ensemble de 909 couples de notes, chaque couple comportant une note en mathématiques et une en philosophie, utilisant toutes les notes réelles ;
- pour chaque appariement, on calcule le coefficient de corrélation ;
- on compte le pourcentage de coefficients de corrélation ainsi obtenus qui dépassent, en valeur absolue, la valeur réelle (ici .19) ; ce pourcentage peut aussi être vu comme la probabilité d'observer un coefficient de corrélation au moins aussi grand en tirant au hasard un appariement parmi les notes observées ; il s'interprète donc comme une mesure du caractère plus ou moins fortuit du degré de linéarité observé.

2.2 Exemple des 4 points a,b,c,d

Si l'on applique ce raisonnement aux données de l'exemple 1 (cf. **Tab. 1**), on obtient les valeurs suivantes :

- nombre d'appariements possibles : 24 (=4×3×2)
- nombre d'appariements conduisant à un coefficient de corrélation supérieur à .923 (valeur observée) : 8

Autrement dit, en appariant au hasard les deux notes, on obtient dans 33% des cas un coefficient de corrélation au moins aussi grand (en valeur absolue) que celui effectivement observé ; le degré de linéarité de ce petit nuage de points est, en ce sens, de l'ordre de ce que l'on observe avec des tirages au hasard.

2.3 Application à la liaison entre philosophie et mathématiques au 1er trimestre

Calculé à partir des 909 notes de mathématiques et philosophie, ce même pourcentage est de l'ordre de 10^{-7}. Le nuage des 909 points présente effectivement une forme allongée que l'on ne doit pas considérer comme fortuite. On peut donc conclure que, dans l'ensemble, les élèves qui obtiennent une forte note pour l'une de ces deux matières ont plutôt tendance à obtenir une forte note pour l'autre.

Le caractère non fortuit de la relation entre ces deux notes ne doit pas être confondu avec l'intensité de cette relation. La valeur .19 du coefficient de corrélation rappelle que le nuage est très loin de se réduire à une droite, ce qui est conforme à son examen visuel. En résumé, dans ce cas, la liaison est lâche mais bien réelle.

2.4 Approximation de la probabilité

Réaliser tous les appariements implique généralement des calculs irréalisables. Une première solution consiste à calculer le pourcentage (de coefficients de corrélation supérieurs à celui observé) à partir d'une série de tirages au hasard parmi tous les appariements possibles ; le pourcentage obtenu est une approximation du pourcentage réel ; cette approximation est d'autant meilleure que l'on effectue un grand nombre de tirages au hasard (en pratique, quelques milliers).

On peut aussi approcher la distribution des valeurs possibles du coefficient de corrélation par une loi théorique et calculer la probabilité cherchée à l'aide de cette loi théorique. Pour cela, on dispose de trois propriétés concernant la distribution du coefficient de corrélation r lorsque l'on réalise le tirage de n couples (x_i, y_i) **dans une population où les deux variables X et Y sont linéairement indépendantes.**

- Lorsque, dans la population, les deux variables X et Y sont distribuées selon une loi normale, la variable :

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

  est distribuée selon une loi de Student à n–2 ddl. Ainsi, par un changement de variable, on peut calculer exactement la probabilité cherchée.

- Lorsque les variables ne sont pas distribuées normalement, le changement de variables précédent conduit à une approximation de la probabilité cherchée, d'autant meilleure que la taille de l'échantillon grandit (tout en restant petite vis-à-vis de la taille de la population). En pratique, cette convergence est rapide et conduit à une approximation satisfaisante dès que l'effectif de l'échantillon atteint quelques dizaines.

- Lorsque n croît, quelles que soient les distributions des x_i et des y_i, la distribution de r s'approche d'une distribution normale centrée de variance $1/(n-1)$. Cette approximation est utilisable dès que n dépasse quelques dizaines.

Ces trois propriétés suggèrent l'utilisation de la loi de Student ou de la loi normale pour approcher la distribution réelle d'un coefficient de corrélation dans l'hypothèse selon laquelle le tirage des couples (x_i, y_i) est réalisé dans une population où les deux variables X et Y sont linéairement indépendantes.

Application numérique aux notes en mathématiques et philosophie au 1er trimestre

$n=909$; $r=.19$; $r\sqrt{n-2}/\sqrt{1-r^2} \approx 5.8$; $r\sqrt{n-1} \approx 5.7$

Pour une aussi grande valeur de n, la loi de Student à n–2 ddl peut être confondue avec une loi normale centrée-réduite. Les deux approximations conduisent à une probabilité associée de l'ordre de 10^{-7}.

Remarque : approcher une probabilité aussi faible par quelques milliers de tirages au hasard conduit en pratique à un résultat égal à 0.

3 Coefficient de corrélation et forme du nuage de points associé

A partir d'un nuage de points, on sait calculer un coefficient de corrélation. Autrement dit, à une forme de nuage est associée une valeur du coefficient de corrélation. Mais, hormis le cas $r=\pm 1$, à une valeur du coefficient de corrélation peuvent correspondre plusieurs formes

de nuage. La figure suivante présente deux cas dans lesquels $r=0$. Dans le premier, on a une liaison parfaite (mais non linéaire) entre Y et X ($Y=X^2$) ; dans le second, l'ensemble des individus peut être fragmenté en deux groupes au sein desquels la liaison entre Y et X est linéaire.

Ceci montre que le calcul du coefficient de corrélation complète l'examen du graphique à deux dimensions mais ne le remplace pas.

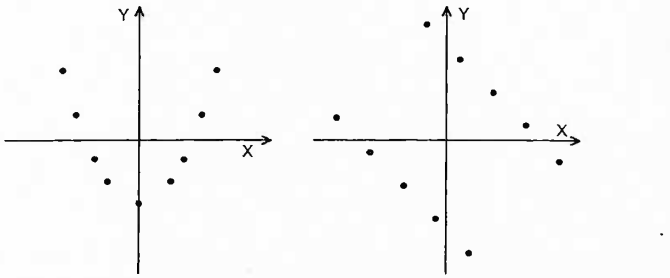


Figure 2. *Deux nuages de points associés à une valeur 0 du coefficient de corrélation.*

4 Matrice des corrélations

Lorsque l'on dispose de plusieurs variables définies sur le même ensemble d'individus, on souhaite souvent calculer l'ensemble des coefficients de corrélation entre les variables prises deux à deux. Il est commode de rassembler ces coefficients dans un tableau (=matrice) unique dont :

- les lignes et les colonnes représentent les variables ;
- le terme général, à l'intersection de la ligne l et de la colonne k, vaut $r(l,k)$, coefficient de corrélation entre les variables l et k.

Ce tableau est symétrique : $r(k,l)=r(l,k)$; on en figure généralement que la moitié. Sa diagonale ne comporte que la valeur 1 : $r(k,k)=1$. Un exemple est donné Tableau 9.1.

Liaison entre deux variables qualitatives

1. Tableau des données et tableau de référence

1.1 Exemple de données dans lequel on étudie la liaison entre deux variables qualitatives présentant chacune deux modalités

1.1.1 Données

Les données étudiées ici sont extraites du chapitre 10 dans lequel on examine un groupe d'élèves n'ayant que deux notes au bac (appelés souvent ici *candidats partiels*). Parmi eux, se trouvent beaucoup de candidats libres. Les données peuvent être regroupées dans un tableau dit tableau croisé (ou tableau de contingence) dans lequel les individus (ici les 975 élèves) sont répartis selon leurs valeurs pour deux variables qualitatives (ici *nombre d'épreuves* et *inscription*).

	cand. libres	cand. inscrits	total
2 notes au bac	8	7	15
5 notes au bac	5	955	960
total	13	962	975

Tableau 1. *Répartition des élèves selon le nombre d'épreuves passées et l'inscription ou non dans un lycée. Exemple : 8 élèves sont à la fois candidats libres et ne possèdent que deux notes au bac. La ligne et la colonne total sont appelées marges du tableau (marge-ligne en bas et marge-colonne à droite).*

1.1.2 Problème

Il est clair que, dans les données, la proportion de candidats libres est beaucoup plus importante parmi les élèves n'ayant que deux notes au bac que parmi ceux ayant toutes leurs notes. En d'autres termes, du point de vue de la variable *inscription*, les deux sous-populations *2 notes au bac* et *5 notes au bac* ne paraissent pas homogènes. Il semble que les étudiants qui n'ont que deux matières à préparer ont tendance, plus que les autres, à ne pas s'inscrire dans un lycée.

Il est clair aussi que, à partir de données réelles, on ne peut pas s'attendre à observer exactement la même proportion de candidats libres dans les deux groupes d'étudiants. D'abord, parce qu'un nombre observé d'élèves est toujours entier alors qu'un calcul de proportion ne conduit pas nécessairement à un nombre entier (c'est le cas ici : aucun sous-ensemble de 15 élèves ne peut respecter exactement la proportion de la population totale

13/975). Ensuite, parce que même si les élèves qui n'ont que deux matières à préparer ne diffèrent pas des autres du point de vue de leur propension à décider ou non de s'inscrire dans un lycée, ceci n'implique pas que la somme de ces décisions soit strictement égale pour l'ensemble des *élèves à 2 notes* et l'ensemble des *élèves à 5 notes*.

La difficulté consiste à apprécier dans quelle mesure les données observées diffèrent de la situation théorique dans laquelle la proportion de candidats libres est la même pour les *élèves à 2 notes* et les *élèves à 5 notes*.

1.1.3 Homogénéité et indépendance

Dans l'étude de ce type de données, on peut aussi adopter le point de vue de la liaison entre les deux variables qualitatives. La situation théorique à laquelle on se réfère dans l'examen des données observées est alors celle de l'indépendance entre les deux variables. En pratique, ce point de vue est équivalent au précédent : l'indépendance entre deux variables qualitatives implique que la répartition des individus pour une variable (e.g. *l'inscription*) est identique si l'on fixe la modalité des individus pour l'autre (*2 notes* ou *5 notes*). Cette répartition (ou distribution) est dite conditionnelle (*cf.* Fiche 9). Cette égalité des distributions conditionnelles coïncide avec l'homogénéité entre les sous-populations.

1.2 Construction du tableau théorique servant de référence

1.2.1 Point de vue de l'homogénéité entre sous-populations

L'égalité des proportions de candidats libres entre les élèves qui n'ont que deux notes et les autres, constitue une situation théorique (i.e. un modèle), référence à laquelle on compare les données observées. Pour réaliser concrètement cette comparaison, on associe à ce modèle de référence le tableau, dit théorique, qui respecte strictement l'égalité des proportions précédentes. Ce tableau théorique doit aussi respecter les marges du tableau observé : le nombre total de candidats libres, par exemple, ne dépend pas de leur répartition entre les deux sous-populations (*2 notes* et *5 notes*) et doit être considéré comme fixe.

Pour construire ce tableau, on calcule la proportion de candidats libres dans la population totale : 13/975 . On calcule alors, dans chaque sous-population (dont l'effectif total est fixé), l'effectif de candidats libres qui respecte exactement cette proportion ; soit, pour les *élèves à 2 notes* : $15\times(13/975)=.2$; on construit ainsi un tableau dont les marges sont identiques à celles du tableau 1 et dont chaque ligne est proportionnelle à la marge-ligne (cf. **Tab. 2**).

	cand. libres	cand. inscrits	total
2 notes au bac	.2	14.8	15
5 notes au bac	12.8	947.2	960
total	13	962	975

Tableau 2. *Répartition des élèves selon la situation de référence dans laquelle le pourcentage d'inscrits dans un lycée est strictement constant quel que soit le nombre d'épreuves passées ; cette situation respecte les marges du tableau réel (cf. Tab. 1).*

On constate facilement que, dans un tel tableau, lorsque les lignes sont proportionnelles à la marge-ligne les colonnes le sont à la marge-colonne. Cette situation de référence fait jouer un rôle symétrique aux lignes et aux colonnes, ce qui n'apparaît pas d'emblée.

1.2.2 Point de vue de l'indépendance entre deux variables qualitatives

La relation qui définit l'indépendance entre deux événements A et B est :

$$P[A \text{ et } B] = P[A] \times P[B]$$

Appliquée par exemple à la première case du tableau, elle donne :

$$P[2 \textit{ notes au bac} \text{ et } \textit{candidat libre}] = (15/975) \times (13/975)$$

ce qui correspond bien à l'effectif théorique du paragraphe précédent :

$$975 \times (15/975) \times (13/975) = .2$$

On dit qu'un tableau croisé satisfait au modèle d'indépendance lorsque chaque case vérifie la relation d'indépendance (i.e. le tableau est proportionnel au produit de ses marges).

2. Cas de deux variables ayant chacune deux modalités

Le cas particulier de deux variables ayant chacune deux modalités est en pratique très fréquent (exemple : relation entre la présence et/ou l'absence de deux caractéristiques dans une population d'individus). Il est commode pour introduire le critère du χ^2, qui s'applique à des situations plus générales.

2.1 Approche fondée sur l'un des effectifs

2.1.1 Deux points de vue pour comparer un tableau d'effectifs observés à un tableau d'effectifs théoriques

Partant du principe qu'il y a toujours un écart entre un tableau observé et un tableau théorique, le problème est alors d'apprécier cet écart : peut-on le considérer comme fortuit ? Pour cela, deux présentations peuvent être adoptées :

- situer le tableau observé parmi tous les tableaux possibles, c'est-à-dire respectant les marges observées ;
- situer le tableau observé dans le cadre d'un tirage au hasard.

Ces deux présentations décrivent la même démarche en des termes différents (cette double présentation d'une même démarche est générale : cf. Fiche 10).

2.1.2 Situer le tableau observé parmi tous les tableaux possibles

La mesure de l'écart entre les effectifs observés et le modèle de référence s'effectue ainsi. On considère :

- l'ensemble des 975 élèves dont 13 sont candidats libres (cela revient à fixer la marge-ligne : 13 ; 962) ;
- le nombre M de tous les sous-ensembles de 15 élèves qu'il est possible de constituer à partir des 975 élèves (ce faisant on fixe la marge-colonne : 15 ; 960) ; à chaque sous-ensemble correspond un tableau présentant les mêmes marges que le tableau observé ;
- parmi les M sous-ensembles de 15 élèves, le nombre M' de ceux comportant au moins 8 candidats libres.

Le rapport M'/M est la proportion des tableaux (respectant les marges observées) au moins aussi éloignés du tableau de référence que ne l'est le tableau effectivement observé.

Cette proportion est un indicateur d'éloignement entre les données observées et le modèle de référence ; elle mesure le caractère plus ou moins fortuit de la liaison observée entre les deux variables qualitatives (sachant que l'on n'observe jamais exactement l'indépendance). Son intérêt est de prendre en compte tous les éléments du problème : le nombre total d'élèves et le nombre de candidats libres, dans la population totale et parmi les candidats *à 2 notes*.

2.1.3 Situer le tableau observé dans le cadre d'un tirage au hasard

On considère le modèle suivant : les candidats *partiels* ont été tirés au hasard (au moins du point de vue de la variable *inscription dans un lycée*). Le modèle du tirage au hasard assure à chaque individu la même probabilité d'être tiré (on précise quelquefois : tirage au hasard uniforme), ce qui formalise le fait que les individus choisis ne possèdent pas de caractéristique particulière en commun. Ce modèle implique qu'il y a **potentiellement** la même proportion de candidats libres parmi les candidats *partiels* et parmi les autres. Concrètement, il correspond au tirage au hasard de 15 individus parmi 975 dont 13 sont candidats libres.

Dans le cadre de ce modèle de référence, on calcule la probabilité d'observer au moins 8 candidats libres parmi les 15. Cette probabilité indique dans quelle mesure des valeurs au moins aussi différentes de l'équiproportion que le sont les données observées sont susceptibles d'apparaître du seul fait du hasard (et non du fait du comportement particulier des candidats *partiels*). Plus cette probabilité est petite et plus on considère que les données observées sont éloignées de la situation de référence. En ce sens, cette probabilité constitue un indicateur du caractère plus ou moins fortuit de la spécificité observée (ou de l'hétérogénéité vis-à-vis des autres élèves) des candidats partiels du point de vue de leur inscription dans un lycée, et donc de la liaison observée entre les deux variables qualitatives.

2.1.4 Lien entre les deux points de vue et application

Les deux approches conduisent au même calcul, lié au problème classique du tirage de 15 boules dans une urne contenant 975 boules, dont 13 blanches et 962 noires. La proportion cherchée apparaît alors comme étant la probabilité d'obtenir une valeur au moins aussi grande que la valeur observée (dans l'exemple : 8) dans le cadre de la loi hypergéométrique (cf. Fiche 9 § 3.2) : H(N=975 ; N_1=13 ; n=15), avec les notations :

- N : effectif total de la population dans laquelle on tire un échantillon de taille n ;
- N_l : nombre d'individus, parmi les N, présentant la caractéristique étudiée ; $p=N_1/N$.

Appliqué à l'exemple, le calcul de cette proportion donne : $.4\ 10^{-12}$

En conclusion, les proportions observées sont très éloignées du modèle de référence et suggèrent, de la part des élèves qui n'ont que deux épreuves à préparer, un comportement différent de celui des autres.

2.1.5 Calcul approché de la probabilité issue de la comparaison entre un tableau d'effectifs et un tableau théorique

Le recours à la loi hypergéométrique pour calculer la probabilité (ou proportion) précédente est en général impraticable. Aussi est-il nécessaire de recourir à une approximation. Plusieurs sont possibles. L'une d'entre elles consiste à approcher la loi hypergéométrique par une loi normale ayant même moyenne et même variance que la loi exacte.

Notations :

- X : le nombre d'individus, parmi les n, présentant la caractéristique étudiée ;
- E_H et S_H la moyenne et l'écart-type de la loi hypergéométrique H(N=975 ; N_1=13 ; n=15) ; cf. Fiche 9 Tab. 1.

On calcule la valeur centrée-réduite : $Q=(X-E_H)/S_H$. L'approximation cherchée est fournie par le calcul de Prob($Y \geq Q$), Y étant distribué selon une loi normale centrée-réduite.

Application numérique

N=975 ; N_1=13 ; n=15

E_H=.2 (on retrouve l'effectif du tableau théorique)

$(S_H)^2$=.2×.9866×.9856=.1944=.441²

Q=(8−.2)/.441=17.69

La probabilité cherchée est celle d'observer une valeur plus petite que 17.69 dans le cadre d'une loi normale centrée-réduite. Cette probabilité est si petite qu'elle n'est pas accessible dans les tables et logiciels courants qui indiquent seulement qu'elle est inférieure à 10^{-9}. Cet ordre de grandeur est bien sûr suffisant pour considérer cette probabilité comme très petite.

2.1.6 Conclusion sur la probabilité

Cette probabilité très petite indique que la liaison observée entre les deux variables *nombres d'épreuves* et *inscription* n'est certainement pas fortuite, ou, ce qui revient au même, que l'hétérogénéité entre les deux sous-populations définies par les lignes (ou par les colonnes) du tableau croisé n'est certainement pas fortuite. On est en présence d'un fait statistique que l'on peut chercher à interpréter.

2.2 Approche fondée sur le critère du χ^2

2.2.1 Principe

Pour comparer globalement un tableau d'effectifs observés et un tableau de référence, on utilise couramment le critère du χ^2. Cette approche peut être décrite de la façon suivante.

1) On cherche à apprécier l'écart entre le tableau observé et le tableau théorique correspondant au modèle de référence.

2) Le critère (ou indicateur) du χ^2 formalise l'écart entre les deux tableaux ; il est défini par (avec E_i et T_i les effectifs observés et théoriques de la case i du tableau) :

$$\chi^2 = \sum_i \frac{(E_i - T_i)^2}{T_i}$$

Ce critère prend en compte les différences E_i-T_i entre effectifs observés et effectifs théoriques. Chaque différence est élevée au carré et pondérée par $1/T_i$: une même différence a ainsi d'autant plus d'influence qu'elle concerne une case de faible effectif. On obtient la valeur 0 si la situation de référence est parfaitement respectée (i.e. si $E_i=T_i$).

La pondération par $1/T_i$ peut sembler arbitraire (pourquoi pas $1/T_i^2$?). Une autre écriture de ce même critère peut sembler plus naturelle :

$$\chi^2 = \sum_i E_i \left[\frac{E_i}{T_i} - 1 \right]$$

L'écart entre E_i et T_i est cette fois mesuré par leur rapport, chaque écart entre ce rapport et 1 étant pondéré par l'effectif observé E_i.

3) On associe au critère du χ^2 une proportion ou une probabilité, en suivant un raisonnement analogue à celui tenu pour l'effectif observé des candidats partiels non inscrits dans un lycée.

- On considère l'ensemble des tableaux possibles, c'est-à-dire qui respectent les marges du tableau observé.
- Pour chaque tableau possible, on calcule la valeur de l'indicateur χ^2 ; on obtient ainsi une distribution de cet indicateur.
- A l'aide de cette distribution, on calcule la proportion des tableaux possibles présentant une valeur du critère du χ^2 supérieure ou égale à la valeur calculée sur les données observées (en d'autres termes : proportion des tableaux possibles au moins aussi éloignés du tableau théorique que ne l'est le tableau effectivement observé).
- En pratique, cette proportion est calculée par approximation en approchant la distribution exacte des valeurs du critère du χ^2 par une loi théorique dite loi de χ^2. Cette loi dépend d'un paramètre dit nombre de degrés de liberté (en abrégé ddl), noté ν, qui est ici le nombre de cases que l'on peut remplir librement lorsque l'on construit un tableau dont les marges sont fixées ; il dépend donc du nombre de lignes I et du nombre de colonnes J du tableau et vaut : $\nu=(I-1)(J-1)$; pour un tableau à 2 lignes et deux colonnes (cas de l'exemple) ce degré de liberté est égal à 1. Ainsi, la proportion cherchée est approchée en calculant $\text{Prob}(Y \geq \chi^2)$ avec Y distribué selon une loi de χ^2 à 1 degré de liberté. Cette approximation est d'autant meilleure que l'effectif total est grand.

2.2.2 Application numérique

En appliquant la définition du critère du χ^2 on obtient :

$$\chi^2 = \frac{(8-.2)^2}{.2} + \frac{(7-14.8)^2}{14.8} + \frac{(5-12.8)^2}{12.8} + \frac{(955-947.2)^2}{947.2}$$

$$\chi^2 = 304.2 + 4.11 + 4.75 + .06 = 313.13$$

La probabilité cherchée est celle de dépasser la valeur 313.13 dans le cadre d'une loi de χ^2 à 1 ddl. Elle est si petite qu'elle n'est pas accessible dans les tables et logiciels courants qui indiquent seulement qu'elle est inférieure à 10^{-9}. Cet ordre de grandeur est bien sûr suffisant pour considérer cette probabilité comme très petite.

2.3 Lien entre le critère du χ^2 et l'approche fondée sur un seul effectif

Dans le cas particulier d'un tableau croisant deux variables présentant chacune deux modalités, les relations entre les différents effectifs permettent d'exprimer le critère du χ^2 en fonction de l'effectif d'une seule case (et des effectifs marginaux). Le tableau 3 récapitule les notations et relations.

	cand. libres	cand. lycée	total
2 notes au bac	X		n
5 notes au bac			N-n
total	N_1	N-N_1	N

Tableau 3. *Notations et relations dans un tableau 2×2*

Tous calculs faits, le χ^2 peut s'écrire :

$$\chi^2 = \left[X - \frac{nN_1}{N}\right]^2 \Big/ \left[\frac{n(N-n)N_1(N-N_1)}{N^3}\right]$$

Lorsque l'on considère l'ensemble des tableaux ayant des marges données, les valeurs de N, N_1 et n sont fixées ; il en résulte qu'il revient au même de classer les tableaux selon la valeur de X ou selon la valeur du χ^2. Ainsi, dans le cas d'un tableau croisant deux variables ayant chacune deux modalités, les deux approches sont strictement équivalentes.

L'intérêt du critère du χ^2 est qu'il se généralise à un tableau de taille quelconque.

Remarque technique

L'approximation de la probabilité à l'aide d'une loi du χ^2 est très proche de l'approximation présentée dans l'approche fondée sur un effectif. Ceci tient aux deux résultats ci-après.

1. Au coefficient $(N-1)/N$ près, en pratique très proche de 1, le critère du χ^2 est le carré de l'indicateur Q utilisé lors de l'approximation de la loi hypergéométrique par une loi normale (cf. 2.1.5).

2. Lorsqu'une variable X est distribuée selon une loi normale centrée-réduite, la variable X^2 est distribuée selon une loi du χ^2 à 1 ddl (*cf.* Fiche 9 § 4.3). Ainsi, les deux probabilités suivantes sont égales :

$\text{Prob}(X \geq Q)$ avec $\text{loi}(X) = N(0,1)$

$\text{Prob}(Y \geq Q^2)$ avec $\text{loi}(Y) = \chi^2(1)$

3. Généralisation au cas de deux variables ayant un nombre quelconque de modalités

3.1 Données et tableau de référence

3.1.1 Présentation d'un exemple : données, problème

Les données analysées ici sont proches de celles commentées au chapitre 10. On se limite aux 945 élèves inscrits dans un lycée et présentant soit peu de données manquantes (38 élèves) soit aucune (907 élèves). On répartit ces élèves dans le tableau (dit de contingence) croisant les deux variables : *présence ou non de données manquantes* et *lycée*.

La répartition des 38 élèves présentant des données manquantes n'est pas homogène d'un lycée à l'autre (cf. **Tab. 4**) : par exemple, le lycée 13 en contient beaucoup plus que les autres.

Lycée	1	2	3	4	5	6	7	8	9	10	11
sans dm	27	19	77	65	25	82	11	65	28	38	51
avec dm	0	1	1	2	1	3	2	1	1	1	1
total	27	20	78	67	26	85	13	66	29	39	52

Lycée	12	13	14	15	16	17	18	19	20	21	22	total
sans dm	41	47	53	53	17	47	20	71	19	28	23	907
avec dm	0	12	3	0	3	1	0	4	1	0	0	38
total	41	59	56	53	20	48	20	75	20	28	23	945

Tableau 4. *Répartition des élèves selon la présence de données manquantes (dm) et le lycée dans lequel ils sont inscrits (tableau de contingence).*

C'est cette hétérogénéité que l'on veut quantifier. Pour cela, on se réfère à la situation théorique (ou modèle) parfaitement homogène dans laquelle la proportion d'individus ayant des données manquantes est la même dans chaque lycée (et se retrouve donc dans la population totale). On souhaite alors construire un indicateur mesurant l'écart entre les effectifs observés et le modèle théorique.

Ce problème peut aussi être formulé en terme de liaison entre les deux variables : *lycée* et *présence de données manquantes*. Selon ce point de vue, on cherche à quantifier la liaison entre ces deux variables. Pour cela, on mesure l'écart entre les données observées et la situation théorique (ou modèle) d'indépendance. Le modèle d'indépendance entre les deux variables est équivalent au modèle d'homogénéité des lycées.

3.1.2 Tableau théorique de référence

De même que dans le cas de variables à deux modalités, on construit le tableau d'effectifs théoriques (cf. **Tab. 5**) qui respecte à la fois les marges observées et l'homogénéité des sous-populations (ce qui revient à respecter le modèle d'indépendance ; cf. 1.2.2).

Le modèle de référence spécifie que la proportion d'élèves présentant des données manquantes est identique dans tous les lycées ; dans ce cas, cette proportion est aussi égale à celle observée pour l'ensemble des élèves, soit 38/945 (=0.04). Par exemple, selon ce modèle, le nombre théorique d'élèves présentant au moins une donnée manquante pour le lycée 6 est : 85×38/945=3.42 élèves.

Lycée	1	2	3	4	5	6	7	8	9	10	11
sans dm	25.91	19.20	74.86	64.31	24.95	81.58	12.48	63.35	27.83	37.43	49.91
avec dm	1.09	.80	3.14	2.69	1.05	3.42	.52	2.65	1.17	1.57	2.09
total	27.00	20.00	78.00	67.00	26.00	85.00	13.00	66.00	29.00	39.00	52.00

Lycée	12	13	14	15	16	17	18	19	20	21	22	total
sans dm	39.35	56.63	53.75	50.87	19.20	46.07	19.20	71.98	19.20	26.87	22.08	907
avec dm	1.65	2.37	2.25	2.13	.80	1.93	.80	3.02	.80	1.13	.92	38
total	41.00	59.00	56.00	53.00	20.00	48.00	20.00	75.00	20.00	28.00	23.00	945

Tableau 5. *Répartition théorique des élèves (selon la présence ou non de données manquantes – dm – et le lycée dans lequel ils sont inscrits), vérifiant le modèle d'indépendance.*

3.2 Comparaison entre un tableau d'effectifs observés et un tableau d'effectifs théoriques

3.2.1 Démarche

Comme précédemment, on mesure le caractère non fortuit de l'hétérogénéité observée entre lycées, en s'appuyant sur un indicateur d'écart au modèle d'indépendance. Pour cela :

- on considère l'ensemble des 945 élèves dont 38 présentent une donnée manquante ;
- on considère l'ensemble de toutes les répartitions possibles (c'est-à-dire respectant les marges observées) de ces 945 élèves dans les 22 lycées, les effectifs globaux des lycées étant égaux aux effectifs observés ;
- parmi ce dernier ensemble, on compte le nombre de répartitions au moins aussi éloignées de la répartition théorique que ne l'est la répartition observée.

3.2.2 Indicateur χ^2

Plusieurs mesures de l'écart entre le tableau observé et le tableau théorique sont possibles. Le plus classique est le critère du χ^2, introduit à partir de l'exemple à 2 lignes et 2 colonnes et qui se généralise sans peine à la situation présente :

- le nombre de cases du tableau vaut ici $N=22\times2=44$, car il y a 22 lycées et 2 catégories d'élèves (*avec données manquantes* ou *sans données manquantes*) ;
- l'écart entre le tableau observé et le tableau théorique peut être mesuré par le critère du χ^2 précédemment défini, en effectuant la sommation sur les N cases du tableau.

Application numérique. A partir des tableaux 4 et 5, on obtient :

$$\chi^2 = \frac{(27-25.91)}{25.91} + \frac{(0-1.09)^2}{1.09} + \frac{(19-19.20)^2}{19.20} + \ldots = 64.17$$

3.2.3 Probabilité associée au χ^2

La valeur du χ^2 n'est pas directement interprétable : elle tend à croître avec la taille du tableau. Il est nécessaire de lui associer une proportion ou une probabilité.

Le calcul exact, par énumération, du pourcentage de tableaux ayant une valeur du χ^2 supérieure ou égale à une valeur donnée est impraticable. Ces calculs peuvent être réduits de deux façons :

- en procédant par simulation ; on génère, par programme, un grand nombre de tableaux ; le pourcentage que l'on observe sur cet échantillon de tableaux est une valeur approchée du pourcentage exact ;
- en utilisant une loi de probabilité théorique conduisant à une probabilité proche de celle cherchée ; la distribution des valeurs de l'indicateur χ^2 est approchée par une loi du χ^2. Cette approximation est bonne si le tableau théorique ne comporte pas trop de cases d'effectif très faible (ce qui, soit dit en passant, n'est pas le cas du tableau étudié dans cette fiche, et pour lequel on devra se contenter d'une approximation grossière).

Dans l'exemple, la distribution qui sert à l'approximation est une loi de χ^2 à $\nu=21$ ddl (cf. 2.2.1 ; 2 lignes et 22 colonnes : $\nu=(2-1)\times(22-1)=21$). La proportion cherchée est donc approchée en calculant $P(Y\geq64.17)$ sachant que la loi de Y est une loi de χ^2 à 21 ddl. Soit, numériquement : $P(Y\geq64.17)=0.000003=3\ 10^{-6}$

Bien que la qualité de l'approximation soit a priori mauvaise (plusieurs effectifs théoriques sont très faibles), cette probabilité est si petite qu'elle suggère que l'irrégularité d'un lycée à l'autre de la proportion d'élèves avec données manquantes n'est sûrement pas fortuite.

4. Quelques prolongements

4.1 Mesurer une liaison et étudier une liaison

Dès lors que le tableau croisant deux variables comporte plus de 4 cases, considérer qu'une liaison est non fortuite incite à préciser la nature de cette liaison. Pour cela, on cherche à mettre en évidence :

- des associations entre modalités des deux variables (i.e. dont l'effectif conjoint observé est sensiblement plus grand que l'effectif correspondant du tableau théorique) ;
- des modalités présentant une distribution exceptionnelle ; ici, la distribution d'une modalité est celle des individus qui la possèdent, c'est-à-dire une distribution conditionnelle ; cette distribution est exceptionnelle si elle est très différente de celle de l'ensemble de la population, c'est-à-dire de la distribution marginale ;
- des ressemblances entre modalités d'une même variable (i.e. des ensembles de modalités d'une même variable ayant des distributions voisines).

Ces aspects peuvent être cernés en comparant directement les tableaux observés et théoriques, en détaillant pour chaque case sa contribution au χ^2 et en sommant ces contributions par ligne et par colonne. C'est ce que fait, à propos de l'exemple, le tableau suivant. Il montre immédiatement que les lycées les plus particuliers sont le 13 et, à un moindre degré, le 16.

Lycée	1	2	3	4	5	6	7	8	9	10	11
sans dm	.045	.002	.061	.007	.000	.002	.175	.043	.001	.009	.024
avec dm	1.086	.048	1.455	.179	.002	.051	4.175	1.031	.024	.206	.569
total	1.131	.050	1.516	.186	.002	.053	4.349	1.074	.025	.215	.593

Lycée	12	13	14	15	16	17	18	19	20	21	22	total
sans dm	.069	1.637	.010	.089	.251	.019	.034	.013	.002	.047	.039	2.58
avec dm	1.649	39.068	.249	2.131	5.995	.448	.804	.321	.048	1.126	.925	61.59
total	1.718	40.705	.259	2.221	6.246	.467	.838	.335	.050	1.173	.964	64.17

Tableau 6. *Contribution au χ^2 de chaque case, de chaque ligne et de chaque colonne.*

L'Analyse Factorielle des Correspondances, non traitée dans cet ouvrage, permet d'aborder tous ces aspects de manière synthétique et systématique.

4.2 Cas d'un grand nombre de variables qualitatives

Lorsque l'on souhaite étudier un grand nombre de tableaux croisés, une démarche méthodique consiste à calculer la probabilité associée au χ^2 pour chaque couple de variables et à étudier en priorité les tableaux associés aux probabilités les plus petites. Si ces tableaux constituent l'ensemble des tableaux obtenus en croisant 2 à 2 toutes les variables qualitatives disponibles, on aura alors recours à l'Analyse des Correspondances Multiples, non abordée dans cet ouvrage, qui fournit un bilan des liaisons existant à l'intérieur d'un ensemble de

variables qualitatives (de façon analogue à l'Analyse en Composantes Principales qui fournit un bilan des liaisons entre plusieurs variables quantitatives).

4.3 Test classique du χ^2

Le calcul du χ^2 a été présenté dans le cadre de la description des données. Ce même calcul est usuellement présenté dans le cadre de l'inférence statistique classique. La démarche de cette inférence, exposée plus généralement fiche 10, s'articule ici autour des points suivants.

- Les individus constituent un échantillon tiré au hasard dans une population.
- Question : les deux variables étudiées peuvent-elles être considérées comme indépendantes au niveau de la population ?
- A partir du tableau issu de l'échantillon, on calcule le critère du χ^2 et sa probabilité associée P ; P est ici la probabilité d'obtenir, dans le cadre d'un tirage au hasard dans une population dans laquelle les deux variables qualitatives sont indépendantes, un échantillon au moins aussi éloigné (au sens du critère du χ^2) de la situation d'indépendance que ne l'est l'échantillon effectivement observé (en effet, dans le cadre de l'hypothèse d'indépendance, le critère du χ^2 suit – approximativement : cf. § 3.2.3 – la loi du χ^2 servant à calculer P).
- Si $P \leq \alpha$ (α seuil fixé a priori généralement à .05), on considère que les deux variables ne sont pas indépendantes au niveau de la population ; autrement dit, P est suffisamment faible pour considérer comme improbable que l'échantillon a été extrait d'une population dans laquelle les deux variables qualitatives sont indépendantes ; on dit que l'on rejette l'hypothèse d'indépendance. L'indicateur χ^2 est alors dit *significatif*. Sinon (si $P > \alpha$), les données sont considérées compatibles avec l'indépendance dans la population.
- α est la probabilité, fixée a priori, de rejeter à tort l'hypothèse d'indépendance

Dans nos données, qui ne sont pas issues d'un tirage au hasard, χ^2 et P ne sont que de simples indicateurs, sans référence explicite à une population plus vaste que celle étudiée.

4.4 Indépendance entre trois variables

4.4.1 Notations pour deux variables

Soient deux variables qualitatives V1 et V2 définies sur n individus : V1 ayant pour modalités $\{1, \ldots, i, \ldots, I\}$ et V2 ayant pour modalités $\{1, \ldots, j, \ldots, J\}$. Soit x_{ij} le nombre d'individus possédant à la fois la modalité i de V1 et j de V2 (les x_{ij} sont usuellement rangés dans un tableau –cf. Tableaux 1 et 4–) ; soit p_{ij} la probabilité observée correspondante : $p_{ij}=x_{ij}/n$. Les effectifs marginaux et les probabilités marginales sont :

$$\text{nombre d'individus ayant la modalité } i : x_{i.} = \sum_j x_{ij}$$

$$\text{nombre d'individus ayant la modalité } j : x_{.j} = \sum_i x_{ij}$$

$$P[\text{avoir la modalité } i] = p_{i.} = x_{i.}/n = \sum_j p_{ij} \quad \text{et} \quad P[\text{avoir la modalité } j] = p_{.j} = x_{.j}/n = \sum_i p_{ij}$$

Les deux variables sont indépendantes (cf. 1.2.2) si pour tout couple (i, j) : $p_{ij} = p_{i.} \times p_{.j}$

4.4.2 Cas de trois variables

Par rapport à la situation précédente, on dispose en plus, pour les n individus, d'une 3ème variable qualitative V3 ayant pour modalités $\{1, \ldots, k, \ldots, K\}$. Soit x_{ijk} le nombre d'individus possédant à la fois la modalité i de V1, j de V2 et k de V3 (cf. **Tab. 7** : exemples où les x_{ijk} sont rangés dans une juxtaposition de deux tableaux) ; soit p_{ijk} la probabilité observée correspondante : $p_{ijk}=x_{ijk}/n$.

Il y a maintenant deux sortes d'effectifs marginaux. Ceux qui correspondent aux marges dites « unaires » :

nombre d'individus ayant la modalité i : $x_{i..} = \sum_j \sum_k x_{ijk}$

nombre d'individus ayant la modalité j : $x_{.j.} = \sum_i \sum_k x_{ijk}$

nombre d'individus ayant la modalité k : $x_{..k} = \sum_i \sum_j x_{ijk}$

avec les probabilités marginales observées : $p_{i..}=x_{i..}/n$; $p_{.j.}=x_{.j.}/n$; $p_{..k}=x_{..k}/n$

et ceux qui correspondent aux marges dites « binaires » :

nombre d'individus ayant la modalité i **et** la modalité j : $x_{ij.} = \sum_k x_{ijk}$

nombre d'individus ayant la modalité j **et** la modalité k : $x_{.jk} = \sum_i x_{ijk}$

nombre d'individus ayant la modalité i **et** la modalité k : $x_{i.k} = \sum_j x_{ijk}$

avec les probabilités marginales observées : $p_{ij.}=x_{ij.}/n$; $p_{.jk}=x_{.jk}/n$; $p_{i.k}=x_{i.k}/n$

Définition. V1, V2 et V3 sont indépendantes si pour tout triplet $i\,j\,k$: $p_{ijk} = p_{i..} \times p_{.j.} \times p_{..k}$

Cette relation implique l'indépendance entre les variables prises deux à deux mais la réciproque n'est pas vraie ainsi que le montrent les deux exemples du tableau 7.

Cas 1

Données		V3 : k=1		V3 : k=2	
		V2		V2	
		j=1	j=2	j=1	j=2
V1	i=1	336	224	84	56
	i=2	144	96	36	24

Marges unaires

V1		V2		V3	
i=1	700	j=1	600	k=1	800
i=2	300	j=2	400	k=2	200

cas 2

Données		V3 : k=1		V3 : k=2	
		V2		V2	
		j=1	j=2	j=1	j=2
V1	i=1	320	240	100	40
	i=2	160	80	20	40

Marges unaires

V1		V2		V3	
i=1	700	j=1	600	k=1	800
i=2	300	j=2	400	k=2	200

Tableau 7. *1000 individus décrits par 3 variables qualitatives à 2 modalités : V1, V2 et V3. Exemple : dans le cas 1, 224 individus sont tels que V1=1, V2=2 et V3=1 ; en tout, 400 individus sont tels que V2=2. Les deux jeux de données ont les mêmes marges unaires et binaires. Les données du cas 1 satisfont l'indépendance entre les 3 variables ; les données du cas 2 satisfont seulement l'indépendance entre les variables prises deux à deux.*

L'indicateur χ^2 se généralise à ce type de données dites souvent *cubiques*.

Comparaison entre deux moyennes

1. Cas de moyennes d'une même variable définie sur deux ensembles d'individus

1.1 Présentation d'un exemple

1.1.1 Données (issues du Ch. 5 et du § 10.2)

Parmi les 975 élèves ayant des notes, 15 n'ont, pour le bac, que des notes en mathématiques et en physique. On veut apprécier ici dans quelle mesure ils diffèrent des autres élèves, du point de vue des notes qu'ils possèdent en mathématiques (cf. **Tab. 1**).

		mathématiques		physique	
catégories d'élèves	effectif	moyenne	éc.-type	moyenne	éc.-type
élèves ayant toutes leurs notes au bac	960	13.11	3.23	10.89	3.65
élèves ayant des notes seulement en mathématiques et en physique	15	11.53	3.83	9.13	4.65
ensemble	975	13.08	3.24	10.87	3.67

Tableau 1. *Moyenne et écart-type de la note au bac en mathématiques et en physique de deux sous-ensembles d'élèves.*

1.1.2 Problème

Les élèves qui n'ont que deux matières à présenter à l'examen sont-ils particuliers du point de vue de la note en mathématiques ? Cette question est posée dans une double perspective.

- Du fait de leurs données manquantes, ces 15 élèves ont été éliminés du fichier. Cette suppression engendre-t-elle une modification importante du fichier ? Dans ce cas particulier, compte tenu des effectifs en présence, on se doute que la suppression de 15 individus parmi 975 est de peu d'influence. De fait, pour la note en mathématiques, la moyenne n'a pratiquement pas été affectée par l'élimination des 15 individus.
- L'étude des élèves qui n'ont que deux matières à présenter est intéressante en soi. Ces élèves obtiennent-ils des notes plutôt moins bonnes, équivalentes ou plutôt meilleures que les autres ?

1.2 Calcul et utilisation d'une probabilité associée

1.2.1 Pourquoi une probabilité associée ?

Les données montrent une différence de plus d'un point et demi entre les moyennes des deux groupes d'élèves (les 960 candidats "totaux" et les 15 candidats "partiels"), différence dont l'importance reste à apprécier, sachant que l'on observe rarement deux moyennes exactement égales. Un aspect de cette appréciation consiste à se demander si l'écart observé n'est pas fortuit. Plus précisément, on se pose la question suivante : l'écart observé est-il du même ordre de grandeur que les écarts que l'on obtiendrait couramment en réalisant le tirage au hasard de 15 élèves parmi les 975 ? Pour cela, on associe une probabilité à la différence entre les deux moyennes.

1.2.2 Principe d'obtention de la probabilité associée ; modèle intérieur aux données

(cf. présentation plus formalisée dans la fiche 10)

- On considère toutes les partitions possibles des 975 élèves, en deux sous-ensembles d'effectifs respectifs 15 et 960.
- Pour chaque partition, on calcule les moyennes $\bar{x}_{15}$ et $\bar{x}_{960}$ des deux sous-ensembles.
- On compte les partitions pour lesquelles l'écart $\bar{x}_{15}-\bar{x}_{960}$ est supérieur, en valeur absolue, à l'écart correspondant pour la partition effectivement observée (il revient au même de considérer l'écart entre la moyenne de l'un des groupes et la moyenne générale : $\bar{x}_{15}-\bar{x}_{975}$; ci-après le terme "différence de moyennes" correspond à l'un quelconque de ces deux écarts).

Exprimé en proportion, ce comptage s'interprète comme la probabilité d'observer une différence au moins aussi grande que celle effectivement observée en tirant au hasard le groupe de 15 parmi les 975 (ce modèle de tirage au hasard est appelé dans ce livre modèle intérieur aux données). On mesure ainsi le caractère fortuit de la différence observée.

1.2.3 Application

Compte tenu de l'impraticabilité du calcul exact de la probabilité associée, on se contente presque toujours d'une approximation que l'on peut obtenir soit par une série de tirages au hasard soit par des lois théoriques. Cette deuxième possibilité a été choisie ici (la façon de réaliser cette approximation est décrite ci-après dans le paragraphe *valeur-test*) et a conduit à la valeur approchée : .062 (ou 6.2%).

La valeur .062 n'est ni franchement grande (auquel cas on serait conduit à considérer l'écart entre les deux moyennes comme fortuit) ni franchement petite ; il est prudent ici, avant de tirer une conclusion définitive, d'aller plus avant dans l'analyse, par exemple en examinant les données cas par cas et/ou en considérant la note de physique, autre note disponible pour ces élèves.

En effet, on observe en physique un phénomène analogue : la moyenne des élèves ayant seulement deux notes au bac est plus basse que celle des autres élèves et la probabilité associée à cet écart de moyennes vaut .066 ; cette coïncidence n'est certainement pas fortuite : au moins dans ce fichier, les élèves ayant seulement deux notes au bac ont des moyennes plus basses que les autres. Cela étant, ce résultat cache des disparités, visibles Fig. 5.2 et analysées au § 10.3.

1.2.4 L'indicateur prend en compte les effectifs observés

Pour mieux cerner ce qu'apporte le pourcentage 6.2% à l'écart 11.53–13.08=–1.55, il suffit d'envisager une situation dans laquelle un même écart de 1.55 aurait été observé, non pas à partir d'un groupe de 15 élèves, mais d'un groupe de 3. Or, parmi tous les groupes possibles de 3 élèves, le pourcentage de groupes dont la moyenne est extérieure à l'intervalle 13.08±1.55 est 40%. Cette situation aurait conduit à considérer ce même écart de 1.55, mais obtenu sur 3 individus, comme fortuit. Ceci illustre le fait que le pourcentage défini ci-dessus prend en compte les effectifs mis en jeu.

1.3 Valeur-test

1.3.1 Objectifs ; données

Pour un calcul approché de la probabilité associée à une différence entre deux moyennes, on adopte le point de vue de la comparaison de l'une d'entre elles à la moyenne générale.

On dispose :

- d'un ensemble de N valeurs dont la moyenne est notée $\bar{x}$ et l'écart-type s (dans l'exemple N=975 ; $\bar{x}$ =13.08 ; s=3.24).
- d'un sous-ensemble (parmi les N) de n valeurs dont la moyenne est $\bar{x}_n$ (dans l'exemple n=15 ; $\bar{x}_n$ =11.53).

On considère le tirage au hasard, parmi les N valeurs, d'un échantillon de taille n. Si l'on recense tous les échantillons possibles de taille n, en calculant à chaque fois la moyenne $\bar{x}_n$, on obtient la distribution des valeurs possibles de $\bar{x}_n$. La moyenne de cette distribution est $\bar{x}$; son écart-type, noté $s_{\bar{x}}$, vaut (cf. Fiche 11 § 1.2) :

$$s_{\bar{x}}^2 = \left(\frac{N-n}{N-1}\right)\frac{s^2}{n}$$

Dans l'exemple, 11.53 est l'une des valeurs possibles de $\bar{x}_n$ et $s_{\bar{x}}$ vaut .831 puisque : $s_{\bar{x}}^2$ =(960/974)(3.24²/15)=.690

Dans le cadre de ce tirage au hasard, on cherche quelle est la probabilité d'obtenir une moyenne au moins autant éloignée de $\bar{x}$ que ne l'est $\bar{x}_n$.

1.3.2 Approximation

Le calcul exact de cette probabilité est impraticable : il implique de compter toutes les valeurs possibles de $\bar{x}_n$ au moins autant éloignées de $\bar{x}$ que ne l'est la valeur effectivement observée de $\bar{x}_n$. Une approximation de cette probabilité est obtenue à l'aide d'une loi théorique dont la distribution n'est pas trop éloignée de la distribution réelle de $\bar{x}_n$; s'agissant de la distribution d'une moyenne, la loi normale convient très bien (cf. Fiche 11).

Le problème revient à calculer la probabilité d'observer :

- une valeur au moins autant éloignée de $\bar{x}$ que ne l'est $\bar{x}_n$ dans le cadre d'une loi normale de moyenne $\bar{x}$ et d'écart-type $s_{\bar{x}}$;
- ou, ce qui revient au même, une valeur supérieure ou égale, en valeur absolue, à la quantité $(\bar{x}_n - \bar{x})/s_{\bar{x}}$, dans le cadre d'une loi normale centrée-réduite.

1.3.3 Application numérique

Dans l'exemple, $(\bar{x}_n - \bar{x})/s_{\bar{x}} = (11.53 - 13.08)/.831 = -1.866$; la probabilité d'observer une valeur supérieure ou égale (en valeur absolue) dans le cadre d'une loi normale centrée-réduite se lit dans les tables usuelles de la loi normale centrée-réduite et vaut .062

1.3.4 Définition de la valeur-test

La quantité $(\bar{x}_n - \bar{x})/s_{\bar{x}}$ est appelée valeur-test. Elle exprime, en nombre d'écarts types, l'écart entre la différence effectivement observée et la différence que l'on observe en moyenne lorsque l'on réalise des tirages au hasard.

On peut considérer la valeur-test comme un simple intermédiaire de calcul de la probabilité associée. Mais, avec un peu d'habitude, la valeur-test est aussi parlante que la probabilité associée, en particulier pour des valeurs élevées (supérieures à 3 en valeur absolue), associées à des probabilités très faibles.

1.4 Test *t* classique

A bien des égards, la démarche précédente ressemble à la démarche classique connue sous le nom de test *t* de Student. Ce test a été conçu pour aider la comparaison entre deux moyennes dans le cas où les données sont obtenues à partir de deux échantillons d'individus tirés au hasard dans deux populations. Il se situe dans une perspective différente de la précédente : les moyennes des échantillons ne sont pas intéressantes en elles-mêmes mais uniquement parce qu'elles apportent des informations sur les populations, ces dernières étant le véritable objet de l'étude.

Le test *t* conduit aussi au calcul d'une probabilité associée à la différence entre les moyennes ; mais cette probabilité s'interprète différemment de la précédente car elle fait référence aux populations dont sont extraits les échantillons.

1.4.1 Données ; modèle

On observe :

- un échantillon de taille n_1, de moyenne $\bar{x}_1$ et d'écart-type s_1, issu d'un tirage au hasard dans une population infinie de moyenne μ_1 et d'écart-type σ_1 inconnus ;
- un échantillon de taille n_2, de moyenne $\bar{x}_2$ et d'écart-type s_2, issu d'un tirage au hasard dans une population infinie de moyenne μ_2 et d'écart-type σ_2 inconnus.

(En comparaison, dans le modèle intérieur aux données, le tirage au hasard est réalisé dans une population connue : l'ensemble des données observées.)

1.4.2 Question

Peut-on considérer les deux populations comme identiques du point de vue de leur moyenne (c'est-à-dire $\mu_1 = \mu_2$) ? Cette égalité constitue l'hypothèse de référence (classiquement notée H_0) par rapport à laquelle on situe les données observées : on se demande si les données observées sont compatibles avec une telle hypothèse.

(En comparaison, le modèle intérieur aux données ne considère que les individus observés, sans référence à un ensemble plus vaste.)

1.4.3 Distribution approchée des valeurs possibles de la différence $d=\bar{x}_1-\bar{x}_2$

On se place dans le cadre de l'hypothèse de référence H_0 : $\mu_1=\mu_2$. On souhaite connaître la probabilité P d'obtenir une différence entre les deux moyennes au moins aussi grande que la différence $d=\bar{x}_1-\bar{x}_2$ effectivement observée. Un tel calcul nécessite la connaissance de la distribution des valeurs possibles de d (en raccourci *distribution de d*), connaissance qui elle-même nécessite celle des distributions des populations échantillonnées, en général inconnues. On se contente donc d'une distribution approchée selon les principes suivants.

- forme de la distribution de d : dès que n_1 et n_2 sont grands, on choisit une loi normale ;
- moyenne (ou espérance) : elle vaut $\mu_1-\mu_2$; dans le cadre de l'hypothèse de référence : $\mu_1-\mu_2=0$;
- variance : elle vaut $\sigma_d^2=\sigma_1^2/n_1+\sigma_2^2/n_2$; σ_1 et σ_2 sont en général inconnus ; une idée simple et naturelle consiste à estimer σ_d^2 par son équivalent, noté s_d, dans les données soit : $s_d^2=s_1^2/n_1+s_2^2/n_2$.

Dans ces conditions, la probabilité cherchée est celle d'obtenir une valeur plus grande (en valeur absolue) que $t=d/s_d$ dans le cadre d'une loi normale centrée-réduite.

1.4.4 Autres approximations usuelles de la distribution de d

La qualité de l'approximation précédente ne pose pas de problèmes dès lors que les effectifs observés sont grands, disons au-delà de 20 ou 30 pour fixer les idées. En deçà, l'approximation peut être améliorée en utilisant une autre valeur de s_d et une loi de Student comme distribution approchée des valeurs possibles de $t=d/s_d$.

A) Des considérations théoriques conduisent à estimer σ_d^2 par $s_d^2 = s_1^2/(n_1-1)+s_2^2/(n_2-1)$ (cf. Fiche 3 § 5) et à utiliser, comme distribution approchée des d/s_d, une loi de Student ayant un nombre de ddl compris entre $\inf\{(n_1-1),(n_2-1)\}$ et n_1+n_2-2 ; (le calcul exact de ce nombre de ddl sort du cadre de cette présentation ; disons simplement que la valeur inférieure est obtenue quand n_1 est très différent de n_2, cas dans lequel le résultat du calcul dépend surtout du plus petit des deux ensembles de valeurs) ; la probabilité calculée de cette façon est très proche de celle obtenue à partir de la loi normale dès que $\inf\{(n_1-1),(n_2-1)\}$ est grand (la loi de Student converge vers une loi normale lorsque le nombre de degrés de liberté croît ; cf. Fiche 9 § 4.3).

B) Quelquefois, les populations étudiées n'ont aucune raison de différer par leur variance, ce qui peut être confirmé par des valeurs observées de s_1 et s_2 voisines. Dans ce cas, lorsque n_1 est assez différent de n_2, on peut se protéger de l'influence prépondérante du plus petit des deux ensembles de valeurs ; on ajoute alors au modèle l'hypothèse $\sigma_1^2=\sigma_2^2$ et l'on estime globalement les variances σ_1^2 et σ_2^2 ; d'où :

$$s_d^2=\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}\left[\frac{1}{n_1}+\frac{1}{n_2}\right]$$

La distribution des d/s_d est alors approchée par une loi de Student à n_1+n_2-2 ddl.

Les calculs dans les cas **A)** et **B)** se confondent lorsque $n_1=n_2$.

1.4.5 Démarche du test

La démarche du test proprement dit est la suivante :

- on calcule la probabilité P d'obtenir, avec l'hypothèse $\mu_1=\mu_2$, une valeur au moins aussi grande que t en valeur absolue (en utilisant selon les cas une distribution normale ou de Student) ;
- si P est petite, on considère que les données observées sont incompatibles avec l'hypothèse $\mu_1=\mu_2$. Le seuil (classiquement fixé à 5%) à partir duquel on considère P petit (et donc à partir duquel on rejette l'hypothèse $\mu_1=\mu_2$) correspond à un risque (ce risque, classiquement noté α, correspond à l'erreur, dite de première espèce, qui consiste à rejeter à tort l'hypothèse $\mu_1=\mu_2$) maximum que l'on se fixe à l'avance.

(En comparaison, dans le modèle intérieur aux données, la probabilité associée, bien que calculée de façon analogue, ne fait pas référence explicitement à une population plus vaste que les données observées.)

1.4.6 Application numérique

L'exemple des notes en mathématiques des candidats "partiels" et "totaux" ne se rapporte pas au test t puisque ces élèves n'ont pas été tirés au hasard. Il est néanmoins possible de réaliser numériquement (cf. **Tab. 2**) les 3 calculs approchés présentés en 1.4.3 (cas 1) et 1.4.4 (cas 2).

Cas	s_d^2	s_d	$t=d/s_d$	Loi utilisée	Probabilité
1	$\left[\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right]$	.994	1.56	normale	.1189
2	$\left[\frac{s_1^2}{n_1-1}+\frac{s_2^2}{n_2-1}\right]$	1.029	1.51	Student 13 ddl	.1550
3	$\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}\left[\frac{1}{n_1}+\frac{1}{n_2}\right]$	.844	1.84	Student 973 ddl	.0661

Tableau 2. *Trois approximations de la probabilité associée à une différence entre deux moyennes ; la probabilité est celle d'obtenir une valeur au moins aussi grande que d/s_d (appelé classiquement t uniquement dans les cas 2 et 3).*

Du fait des effectifs très inégaux ($n_1=960$; $n_2=15$) :

- dans les cas 1 et 2, s_d ne dépend pratiquement que de s_2 et n_2 (la variabilité du petit échantillon marque le pas ; on saisit ici l'intérêt de disposer, lorsque cela est possible, d'effectifs voisins sinon égaux) ;
- dans le cas 3, qui suppose l'égalité des variances des populations étudiées, la variabilité exprimée par s_2 est diluée dans celle exprimée par s_1.

1.4.7 Calcul d'une valeur-test et test t

Le fait que souvent la valeur-test et l'indicateur t conduisent en pratique à des probabilités voisines ne doit pas faire oublier que ces deux calculs correspondent à deux points de vue différents sur les données : les données sont analysées en référence à des populations plus vastes dans le second cas et non dans le premier.

2. Cas de moyennes de deux variables définies sur le même ensemble d'individus

2.1 Spécificité des données appariées

Pour les mathématiques, entre le troisième trimestre et le bac, les 909 élèves améliorent leur note en moyenne de 2.16 points (cf. **Tab. 3**).

	mathématiques		
	3ème trim.	bac	bac−3ème trim.
moyenne	11.05	13.21	2.16
écart-type	2.98	3.19	2.49

Tableau 3. *Note en mathématiques, au bac et au 3ème trimestre, pour les 909 élèves : moyenne et écart-type.*

Dans le même esprit que précédemment, on se demande si l'on peut considérer ce résultat (i.e. l'augmentation des notes entre le 3ème trimestre et le bac) comme fortuit ou non. Pour cela, on calcule ici encore la probabilité d'observer une différence au moins aussi grande lorsque l'on effectue des tirages au hasard.

Mais cette comparaison entre deux moyennes est différente de celle étudiée précédemment. En effet, on compare ici les moyennes de deux variables différentes pour un même ensemble d'individus (les données sont dites appariées, une paire correspondant aux deux notes d'un même individu) alors que précédemment on comparait les moyennes de deux ensembles d'individus pour une même variable. Ceci a des répercussions dans la définition du tirage au hasard servant au calcul de la probabilité associée.

2.2 Calcul et utilisation d'une probabilité associée

2.2.1 Exemple de données choisies

élève	3ème trim.	bac	différence
A	1	4	3
B	9	11	2
C	17	18	1
moyenne	9	11	2

Tableau 4. *Exemple fictif : données « observées ».*

Dans l'exemple fictif (cf. **Tab. 4**), tous les élèves ont progressé au bac : ce qui est étudié est la différence "bac−3ème trimestre" c'est-à-dire l'évolution des notes pour un même élève (variabilité intra-élève). On doit s'affranchir ici de l'existence d'élèves globalement bons ou globalement mauvais (variabilité inter-élèves) ; par construction, les données sont appariées et cet appariement ne doit pas être remis en cause dans les calculs.

2.2.2 Définition du tirage au hasard servant de référence

On effectue toutes les permutations possibles des deux notes de chaque élève. On respecte ainsi l'appariement : ne pas respecter l'appariement reviendrait à considérer comme possible le couple de notes (1, 18) qui réunit la note au bac d'un bon élève et la note au 3ème

trimestre d'un mauvais. Une permutation est un ensemble de progressions possibles (une par élève) compte tenu des notes observées.

Le même résultat peut être obtenu plus simplement en ne considérant que l'ensemble des différences effectivement observées {3, 2, 1} et en permutant tour à tour les signes de chaque différence. Il y a deux possibilités par élève ce qui conduit à $2^3=8$ (plus généralement 2^n pour n individus) permutations globales possibles (cf. **Tab. 5.**)

Considérer le résultat observé comme fortuit est équivalent à considérer qu'il a été tiré au hasard parmi l'ensemble des permutations possibles.

élève	cas 1	cas 2	cas 3	cas 4	cas 5	cas 6	cas 7	cas 8
A	3	3	3	3	–3	–3	–3	–3
B	2	2	–2	–2	2	2	–2	–2
C	1	–1	1	–1	1	–1	1	–1
moyenne	2	1.33	.67	0	0	–.67	–1.33	–2

Tableau 5. *Exemple fictif : les 8 permutations de signe des différences entre les deux notes.*

2.2.3 Distribution de la moyenne des différences

Notons $\bar{x}$ et s la moyenne et l'écart-type des n différences observées. Dans l'exemple fictif, l'ensemble des différences observées est {3 , 2 , 1} ; d'où : $\bar{x}=2$ et $s^2=2/3$;

La distribution de la moyenne des différences est, par définition, l'ensemble des moyennes des différences possibles ; soit, dans l'exemple fictif :

$$\{2\ ,\ 1.33\ ,\ .67\ ,\ 0\ ,\ -.67\ ,\ -1.33\ ,\ -2\}$$

Par raison de symétrie, cette distribution a une moyenne nulle. On montre en outre que son écart-type $s_{\bar{x}}$ peut être calculé à partir des différences observées ; soit :

$$s_{\bar{x}}^2=(1/n^2)\sum x_i^2=(s^2+\bar{x}^2)/n$$

Dans l'exemple fictif : $s_{\bar{x}}^2=14/9$

2.2.4 Calcul de la probabilité associée à la différence entre deux moyennes

La probabilité associée est celle d'obtenir, en permutant aléatoirement les données, une moyenne des différences au moins aussi grande en valeur absolue que celle effectivement observée.

Deux permutations (1 et 8) sont dans ce cas. La probabilité associée à la différence observée ($\bar{x}=2$) est $2/8=.25$: la différence peut être considérée comme fortuite. Remarquons au passage que, compte tenu des données, cette probabilité est aussi la plus petite possible, ce qui est une façon de constater qu'avec seulement trois observations cette procédure ne peut mettre clairement en évidence aucune différence entre deux moyennes.

2.2.5 Approximation

Dès que le nombre d'individus est important, on ne peut effectuer l'ensemble des permutations. On peut réaliser une série de tirages au hasard dans les permutations possibles. On peut aussi approcher la distribution exacte de la moyenne des différences par une loi normale ayant la même moyenne (i.e. 0) et la même variance $s_{\bar{x}}^2=(s^2+\bar{x}^2)/n$.

Application numérique

Dans l'exemple des notes du bac : $s_{\bar{x}}^2 = (2.49^2 + 2.16^2)/909 = .109^2$; $\bar{x} / s_{\bar{x}} = 19.73$

La probabilité de dépasser la valeur 19.73 dans le cadre d'une loi normale centrée-réduite est quasiment nulle. Une telle différence de moyennes n'est certainement pas fortuite.

2.3 Test *t* classique dans le cas de données appariées

2.3.1 Démarche

Lorsque les données appariées sont obtenues à partir d'individus tirés au hasard dans une population, il existe une procédure, analogue au test *t* précédent, qui permet de calculer de façon adaptée à ce cas une probabilité associée à une différence de moyennes. La démarche de cette procédure utilise les mêmes principes que celle du *t* non apparié ; nous en indiquons les principales étapes :

- à chaque individu, on associe la différence entre ses valeurs pour les deux variables ; la variable étudiée est donc la différence entre les variables initiales ; pour l'échantillon observé, la moyenne des différences est noté $\bar{x}_0$ et l'écart-type s_0.
- le modèle est celui du tirage au hasard, dans une population de taille infinie, de n valeurs de la variable *différence* ; dans la population échantillonnée, cette variable *différence* est distribuée selon une loi de moyenne μ et d'écart-type σ ;
- l'hypothèse de référence est que les deux variables initiales ont même moyenne, ou ce qui revient au même, que la variable *différence* a une moyenne nulle ($\mu=0$) ;
- à chaque échantillon de n différences, on associe la moyenne $\bar{x}$ de ces n différences ; l'ensemble des valeurs possibles de $\bar{x}$ constitue la distribution de la moyenne des différences ;
- l'hypothèse de référence ne suffit pas pour définir la distribution des valeurs de $\bar{x}$; on choisit une loi approchée selon une démarche analogue à celle du *t* non apparié ; **A)** on estime (cf. Fiche 3 § 5) la variance σ^2 par $s_{\bar{x}}^2 = s_0^2/(n-1)$; **B)** on approche la distribution (sous l'hypothèse $\mu=0$) des valeurs possibles de $t_a = \bar{x} / s_{\bar{x}}$ par une loi de Student, ici à $n-1$ ddl ;
- finalement on calcule, pour cette distribution, la probabilité d'observer une valeur supérieure (en valeur absolue) à $\bar{x}_0 / s_{\bar{x}}$.

2.3.2 Application numérique

Bien que les données du bac ne proviennent pas d'un tirage au hasard, on peut calculer l'indicateur t_a ; soit (cf. **Tab. 3**) : $s_{\bar{x}} = .0826$; $\bar{x}_0 / s_{\bar{x}} = 26.14$

La probabilité de dépasser 26.14 dans le cadre d'une loi de Student à 908 ddl (en pratique confondue avec une loi normale centrée-réduite) est quasiment nulle : la différence entre les moyennes est dite significative. Si les données provenaient d'un tirage au hasard dans une population, on conclurait à l'existence d'une différence entre les moyennes de ces deux variables dans la population.

2.4 Données appariées et graphique

Dans le cas de données appariées, on dispose, par définition, de deux valeurs numériques pour chaque individu. Cela suggère de représenter les individus sur le graphique usuel à deux dimensions. Pour les notes en mathématiques au 3ème trimestre et au bac, c'est bien ce qui a été fait **Fig. 8.1** et **8.5**.

Comme toujours, le graphique contient beaucoup plus d'informations : le progrès « moyen » est en fait hétérogène selon la note au 3ème trimestre, on peut distinguer des outliers, etc. L'indicateur statistique a ici pour principal intérêt est de préciser un aspect du graphique rapporter le décalage moyen à la variabilité.

Liaison entre une variable quantitative et une variable qualitative

1 Données, problématique

On reprend ici l'un des exemples du chapitre 11. Pour chacun des 907 élèves (les deux candidats libres sont exclus ici), on dispose des informations suivantes :

- le numéro du lycée auquel appartenait l'élève (22 lycées) ;
- la note au bac, moyenne pondérée avec les coefficients officiels des notes obtenues le jour du bac.

Question : la note au bac varie-t-elle beaucoup d'un lycée à l'autre ? Ceci revient à étudier la liaison entre une variable quantitative (*note au bac*) et une variable qualitative (*lycée*).

Pour y répondre, on examine si les élèves de tel lycée obtiennent, en moyenne, de meilleures notes que ceux de tel autre lycée. Selon ce point de vue, la question revient à comparer plusieurs moyennes entre elles. Ainsi formulé, le problème est une généralisation de la comparaison entre deux moyennes.

Les notations sont rassemblées dans le tableau 1.

Lycée	Notes					Moyennes
1	y_{11}	...	y_{1j}	...	y_{1n_1}	$\bar{y}_1$
i	y_{i1}	...	y_{ij}	...	y_{in_i}	$\bar{y}_i$
I	y_{I1}	...	y_{Ij}	...	y_{n_I}	$\bar{y}_I$

Tableau 1. *Notes au bac des élèves répartis selon les 22 lycées : notations.*
I : nombre de lycées (ici I=22) ; y_{ij} : note obtenue par le $j^{ème}$ élève du lycée i ;
n_i : nombre d'élèves du lycée i ; n : nombre total d'élèves ;
$\bar{y}_i$: note moyenne des n_i élèves du lycée i ; $\bar{y}$: note moyenne de l'ensemble des n élèves.

2 Approche graphique

Comme toujours, on s'efforce d'illustrer le problème étudié par quelques graphiques : ici, on peut juxtaposer les boîtes de dispersion de chaque lycée (cf. **Fig. 11.3**).

Un simple coup d'œil sur ces boîtes de dispersion montre qu'aucune règle absolue du type "n'importe quel élève du lycée i est meilleur que n'importe quel élève du lycée l" ne peut

être énoncée. Et pourtant, globalement, certains lycées obtiennent de meilleures notes que d'autres. Tel est le type de fait que l'on cherche à mettre en évidence : des tendances qui concernent l'ensemble des individus d'un lycée - et non chaque individu du lycée - du type "les élèves du lycée *i* obtiennent en général de meilleures notes que ceux du lycée *l*".

3 Trois variabilités en présence : totale, inter-classes et intra-classe

Considérons, pour simplifier, 3 lycées ayant 4 élèves chacun (cf. **Fig. 1**).

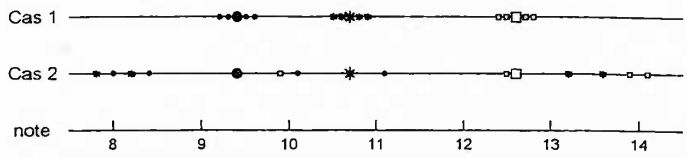


Figure 1. *Exemples fictifs : variabilités de 12 élèves répartis dans 3 lycées. Chaque élève est représenté par le symbole de son lycée (●✱□) ; les symboles agrandis représentent les moyennes par lycée.*

Dans ces deux cas de la figure 1, les moyennes par lycée sont identiques. Pourtant la relation entre la variable *lycée* et la variable *note au bac* semble beaucoup plus forte dans le premier cas que dans le second. Cette différence tient à la variabilité des notes à l'intérieur de chaque lycée, beaucoup plus faible dans le premier cas que dans le second. Pour évaluer la liaison entre ces deux variables, il importe donc de prendre en compte :

- la variabilité entre les lycées, i.e. la variabilité des moyennes des lycées ;
- la variabilité à l'intérieur des lycées.

Pour le $j^{ème}$ individu appartenant au lycée i, on peut écrire :

$$(y_{ij} - \bar{y}) = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})$$

Ainsi, pour cet individu, l'écart entre sa note y_{ij} et la moyenne générale $\bar{y}$ (variabilité totale) peut être décomposé en deux parties :

- l'écart $y_{ij} - \bar{y}_i$ entre sa note et la moyenne de son lycée (variabilité intra-lycée) ;
- l'écart $\bar{y}_i - \bar{y}$ entre la moyenne de son lycée et la moyenne générale (variabilité inter-lycées).

En élevant au carré les deux termes de l'égalité précédente et en les additionnant pour tous les élèves, on obtient l'équation dite d'analyse de variance :

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - \bar{y})^2$$

soit, en écrivant SCE pour Somme des Carrés des Ecarts :

$$\text{SCE totale} = \sum_i [\text{SCE intra-groupe } i] + \text{SCE inter-groupes}$$

$$= \text{SCE intra-groupes} + \text{SCE inter-groupes}$$

La SCE est une mesure de variabilité liée à la variance. Ainsi :

- la SCE totale vaut n fois la variance des notes de tous les élèves (tous lycées confondus) ;
- la SCE inter-groupes vaut n fois la variance des moyennes $\bar{y}_i$ affectées des poids n_i ;
- la SCE intra-groupe i vaut n_i fois la variance des données du groupe i.

Cette relation, toujours vraie, formalise la décomposition de la variabilité totale des notes en deux parties. L'étude de la liaison entre les variables *lycée* et *note au bac* s'appuie sur cette décomposition : cette liaison est d'autant plus étroite que, à SCE totale constante, la SCE inter-groupes est grande (ou, ce qui revient au même, la SCE intra-groupes est petite).

4 Comparaison entre les variabilités : rapport de corrélation

4.1 Définition

Pour comparer les variabilités mises en évidence par l'équation d'analyse de variance, on peut calculer le rapport suivant :

$$\eta^2 = \frac{\text{SCE inter - groupes}}{\text{SCE totale}}$$

η^2 est appelé *rapport de corrélation*. Il varie entre 0 et 1. Il vaut 0 s'il n'existe aucune différence entre les moyennes des différents lycées (i.e. aucune liaison entre les variables *lycée* et *note au bac*) et 1 s'il n'existe aucune différence entre les élèves d'un même lycée (liaison « parfaite » entre les deux variables). En ce sens, η^2 peut être considéré comme un indicateur de liaison entre une variable quantitative et une variable qualitative.

Ce rapport s'interprète comme le pourcentage de la variabilité totale de y conservé lorsque l'on remplace la note y_{ij} de chaque élève par la moyenne $\bar{y}_i$ de son lycée. En ce sens, η^2 mesure la part de variabilité « explicable » par l'appartenance au lycée ; η^2 est l'adaptation, à la liaison entre une variable quantitative et une variable qualitative, du coefficient de détermination R^2 introduit en régression (cf. § 8.7) ; c'est aussi le carré du coefficient de corrélation calculé, pour les 907 élèves, entre leur note et la moyenne de leur lycée.

Attention : le rapport de corrélation ne peut être interprété qu'en regard de l'effectif total n et du nombre de groupes I. Ceci apparaît en considèrant les deux cas limites :

- si $I=n$, il y a un seul élève par lycée : η^2 vaut 1 ;
- si $I=1$, tous les élèves appartiennent au même lycée : η^2 vaut 0.

4.2 Probabilité associée au rapport de corrélation

Comme pour tout indicateur statistique, il est précieux d'associer à η^2 une probabilité afin :

- d'évaluer le caractère plus ou moins fortuit d'une valeur observée ;
- de comparer des rapports de corrélation calculés dans des situations comportant des nombres de classes et des effectifs différents.

Le calcul d'une probabilité associée au rapport de corrélation η^2 résulte d'une démarche analogue à celle suivie lors de la comparaison entre deux moyennes (cf. Fiche 10).

- On considère toutes les partitions possibles des 907 élèves en 22 groupes respectant les effectifs des lycées. Ces partitions sont celles que l'on peut obtenir en tirant au hasard

l'affectation des individus aux groupes ; ce tirage au hasard est un modèle intérieur aux données, en ce sens qu'il ne fait intervenir aucun individu autre que ceux observés.

- Pour chaque partition, on calcule les moyennes $\bar{y}_i$ ainsi que le rapport de corrélation η^2 ; on obtient ainsi la distribution des valeurs possibles de η^2 ; on montre que cette distribution a pour espérance : $(I-1)/(n-1)$, ce qui fait apparaître la nécessité de prendre en compte le nombre d'individus n et le nombre de classes I dans l'évaluation d'un rapport de corrélation.
- On compte les partitions dont le rapport de corrélation est supérieur ou égal à la valeur de η^2 effectivement observée.

Exprimé en proportion (ou en pourcentage), ce comptage s'interprète comme la probabilité P d'obtenir un rapport de corrélation au moins aussi grand que celui effectivement observé, en tirant au hasard l'affectation des élèves dans les lycées. On mesure ainsi le caractère fortuit de la liaison observée. Cette mesure prend en compte les valeurs n et I ; elle peut servir à comparer des situations qui ne présentent pas le même effectif global et/ou le même nombre de classes.

4.3 Approximation de la probabilité associée au rapport de corrélation

Du fait du nombre de permutations qu'il implique, le calcul exact de la probabilité P est impraticable. Une approximation de P peut être obtenue en réalisant des tirages au hasard. On peut aussi utiliser une loi théorique en exploitant la propriété suivante.

Si l'on tire aléatoirement et indépendamment n valeurs dans I populations distribuées selon des lois normales identiques (i.e. ayant la même moyenne et la même variance), alors l'indicateur (cf. aussi § 5.3) :

$$F = \frac{\text{SCE inter - groupes}}{\text{SCE intra - groupes}} \times \frac{n-I}{I-1} = \frac{\eta^2}{1-\eta^2} \times \frac{n-I}{I-1}$$

est une variable aléatoire distribuée selon une loi F de Fisher à $I-1$ et $n-I$ ddl.

Les indicateurs η^2 et F étant liés de façon monotone (quand l'un croît l'autre croît), on peut calculer la probabilité associée à l'un à partir de celle associée à l'autre. La propriété ci-dessus suggère une possibilité de calcul approché (la loi F de Fisher est bien connue), sachant bien que la situation réelle diffère de cette situation théorique car :

- les valeurs observées ne sont pas distribuées selon une loi normale ;
- la situation théorique spécifie des tirages indépendants.

4.4 Exemple numérique

Le tableau 2 récapitule les données et les calculs correspondants aux données de la figure 1.

La variance inter-lycées est la même dans les deux cas. Mais dans le cas 2, une forte variabilité intra-lycées engendre une forte variabilité totale et donc un rapport de corrélation plus faible.

Dans ces deux exemples, il y a en tout 34650 façons d'affecter les 12 données aux 3 classes (la formule qui conduit à 34650 est explicitée à propos de la loi binomiale Fiche 9 § 3.2) ; tous calculs faits, le nombre d'affectations qui conduisent à une valeur de η^2 au moins aussi grande que celle effectivement observée est 1 (cas 1) et 6408 (cas 2), d'où les probabilités associées (ou pourcentages) de $3\ 10^{-5}$ et .1849 ; cette seconde valeur indique que, dans le

cas 2, la valeur observée de η^2 est du même ordre que ce que l'on observe couramment (i.e. grossièrement une fois sur cinq) lors de tirages au hasard.

	Cas 1			Cas 2		
	• ●	* ✳	▫ □	• ●	* ✳	▫ □
Données	9.2	10.5	12.4	8.0	7.8	9.9
	9.3	10.6	12.5	8.4	8.2	12.5
	9.5	10.8	12.7	10.1	13.2	13.9
	9.6	10.9	12.8	11.1	13.6	14.1
moyenne	9.4	10.7	12.6	9.4	10.7	12.6
écart-type	.158	.158	.158	1.259	2.707	1.676
SCE tot.	21.02			67.62		
SCE inter	20.72			20.72		
SCE intra	.30			46.9		
η^2	.986			.306		
F	310.797			1.988		
P exacte	$1/34650=3\ 10^{-5}$			$6408/34650=.1849$		
P approchée	$<10^{-9}$			.1889		

Tableau 2. *Exemples fictifs (cf. Fig. 1) : calcul de l'indicateur de liaison η^2 et de la probabilité associée P.*

L'approximation fournie par la distribution F est suffisante dans ces cas, malgré les effectifs très faibles.

5 Test *F* de Fisher

Le test F a pour objet d'aider la comparaison entre I moyennes dans le cas de données issues d'échantillons d'individus tirés au hasard dans I populations. Bien que conçu dans un cadre différent, il conduit au même calcul que celui de l'approximation de la probabilité associée à η^2. Mais la portée de ce calcul est différente puisqu'il se réfère aux populations d'où proviennent les échantillons.

5.1 Données ; le modèle

On observe I échantillons. L'échantillon i, de taille n_i ($n=\Sigma n_i$), est issu d'un tirage au hasard dans une population infinie de moyenne μ_i et d'écart-type σ_i. A ce modèle, on ajoute les hypothèses techniques supplémentaires suivantes : les I populations sont distribuées suivant des lois normales de même variance ($\sigma_i=\sigma$ pour tout i) ; l'hypothèse d'égalité des variances rend plus aisée l'interprétation de l'indicateur F et permet de calculer sa loi exacte.

5.2 Question

Elle concerne les populations. Peut-on considérer les I populations identiques du point de vue de leur moyenne ? Plus précisément, on se demande si l'on peut considérer les données observées comme compatibles avec l'hypothèse notée classiquement H_0 : $\mu_i=\mu$ pour tout i.

5.3 Choix d'un indicateur statistique : le F

Remarque technique préliminaire : à une somme de carrés de termes, on associe son nombre de degrés de liberté, à savoir son nombre de termes indépendants, égal au nombre total de termes diminué du nombre de relations existant entre eux. Ainsi, la SCE intra-groupes :

$$\sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$

« a » $n-I$ ddl car elle contient n termes $(y_{ij} - \bar{y}_i)$ reliés par les I relations : $\sum_j (y_{ij} - \bar{y}_i) = 0$

De même, la SCE inter-groupes :

$$\sum_i n_i (\bar{y}_i - \bar{y})^2$$

« a » $I-1$ ddl car elle contient I termes $(\bar{y}_i - \bar{y})$ reliés par la relation : $\sum_i (\bar{y}_i - \bar{y}) = 0$

On synthétise les variabilités en présence à l'aide de l'indicateur F introduit au paragraphe précédent et qui peut s'écrire aussi (CM = abréviation pour *carré moyen*) :

$$F = \frac{\text{SCE inter - groupes}}{I - 1} \Big/ \frac{\text{SCE intra - groupes}}{n - I} = \frac{\text{CM inter - groupes}}{\text{CM intra - groupes}}$$

Ce rapport confronte les variabilités inter et intra-groupes. Les individus étant tirés au hasard, les y_{ij} et par suite les Carrés Moyens sont des variables aléatoires. On montre que, en notant E[X] l'espérance de X :

- E[CM intra-groupes] = σ^2, variance supposée commune de chaque population ; selon un autre point de vue, ce Carré Moyen estime cette variance (remarquer ici l'importance de l'hypothèse d'égalité des variances intra-populations pour interpréter la démarche) ;
- E[CM inter-groupes] = $\sigma^2 + Q$ avec $Q = \sum n_i(\mu_i - \mu)^2$ et $\mu = (1/n)\sum n_i \mu_i$; la quantité Q est la SCE **inter-populations** ; elle est d'autant plus grande que les μ_i diffèrent entre eux.

Ainsi, la variabilité inter-groupes est engendrée à la fois par la variabilité intra-populations et par l'hétérogénéité des moyennes des populations. D'où l'idée de rapporter la variabilité inter-groupes observée à ce quelle serait si elle était due à la seule variabilité intra-populations. C'est ce que fait l'indicateur F. En l'absence de différences entre les moyennes des populations, on peut s'attendre à des valeurs de F voisines de 1.

5.4 Exemples numériques ; tableau d'analyse de variance

Classiquement, les résultats des calculs précédents sont présentés dans une forme standard, dite tableau d'analyse de variance, qui met en évidence la décomposition de la SCE totale et des degrés de liberté correspondants ; les carrés moyens quant à eux ne s'additionnent pas mais se comparent l'un à l'autre à l'aide du rapport F. Le tableau 3 reprend les exemples du tableau 2 sous cette forme : pour se situer dans le cadre classique, on suppose que les 4 élèves de chaque lycée ont été tirés au hasard dans leurs établissements respectifs.

Le Carré Moyen intra-lycées estime la variabilité intra-lycées dans les populations parentes, supposée commune à tous les lycées ; cette estimation de la variabilité des populations parentes est légèrement supérieure aux variances intra-lycées des échantillons observés (cf. **Tab. 2**), les variances des échantillons fournissant en moyenne une sous-estimation de la variance de la population parente, conformément au raisonnement tenu Fiche 3 § 5.

Source de variation	Cas 1				Cas 2			
	SCE	ddl	CM	*F* (*P*)	SCE	ddl	CM	*F* (*P*)
totale	21.02	11			67.62	11		
inter-lycées	20.72	2	10.36 $(3.22)^2$	310.8 ($<10^{-9}$)	20.72	2	10.36 $(3.22)^2$	1.988 (.189)
intra-lycées	.30	9	.033 $(.183)^2$	.	46.9	9	5.21 $(2.28)^2$	

Tableau 3. *Tableaux d'analyse de la variance correspondant aux données du tableau 2. SCE : somme des carrés des écarts ; ddl : degrés de liberté ; CM : carré moyen ; F (P) : rapport F et probabilité associée.*

Dans le cas 1, le Carré Moyen inter-lycées est largement supérieur au Carré Moyen intra-lycées, ce que le *F* (310.8) et la probabilité associée ($<10^{-9}$) montrent bien : un tel écart entre carrés moyens est pratiquement impossible à observer en permutant les données au hasard. Dans le cas 2, le Carré Moyen inter-lycée n'est que peu supérieur au Carré Moyen intra-lycées ; la probabilité associée au *F* indique qu'un écart au moins aussi grand entre carrés moyens s'observe environ une fois sur 5 en permutant les données au hasard.

5.5 Test F proprement dit

On se place dans le cadre de l'hypothèse de référence H_0 : $\mu_i=\mu$ pour tout i.
On calcule la probabilité *P* d'obtenir, dans le cadre de cette hypothèse, une valeur de *F* au moins aussi grande que celle effectivement observée. Ce calcul de *P* nécessite la connaissance de la loi de distribution de l'indicateur *F* dans le cadre de l'hypothèse de référence. Ceci est possible grâce aux hypothèses techniques de normalité et d'égalité des variances : dans ce cadre, la distribution de l'indicateur *F* est celle d'une loi F de Fisher à $(I-1)$ et $(n-I)$ degrés de liberté ; si ces hypothèses techniques ne sont pas vérifiées, le calcul ne fournit qu'une approximation (généralement suffisante en pratique) de la vraie valeur *P*.

On se fixe a priori une probabilité α (en général .05). La règle de décision est alors :

- si $P>\alpha$ la variabilité entre les moyennes observées est dite non significative ; on considère que toutes les moyennes μ_i sont égales (= on considère comme vraie l'hypothèse de référence) ; cette situation correspond au cas 2 des tableaux 2 et 3 pour lequel on considère que les lycées dont les échantillons d'élèves ont été extraits ne diffèrent pas de par leur moyenne ;
- sinon ($P\leq\alpha$), la variabilité entre les moyennes observées est dite significative (on précise quelquefois « au seuil α ») ; on considère comme fausse l'hypothèse de référence (on dit qu'on la rejette) ; cette situation correspond au cas 1 des tableaux 2 et 3 pour lequel on considère que les lycées dont les échantillons d'élèves ont été extraits diffèrent de par leur moyenne.

Dans une telle règle de décision, α représente la probabilité, fixée a priori, de considérer à tort les populations comme différentes : c'est le risque dit de 1ère espèce. Des éléments généraux concernant ce type de procédure se trouvent dans la fiche 10.

5.6 Suite à donner à un F significatif

Lorsque l'on rejette l'hypothèse d'égalité des moyennes, il reste encore à préciser quelles moyennes diffèrent entre elles et lesquelles peuvent être considérées comme identiques. La

procédure la plus naturelle consiste alors à réaliser un ensemble (voire l'ensemble) de(s) comparaisons entre les moyennes prises deux à deux. Cette procédure a été maintes fois critiquée en particulier parce qu'elle génère un nombre important de tests statistiques qui, même appliqués à des tableaux de nombres tirés au hasard, finissent toujours par conduire à des résultats significatifs (à la limite, en proportion α). En d'autres termes, si l'on envisage **l'ensemble des comparaisons**, le risque α –dit alors *global*– de déceler au moins une différence entre deux moyennes alors qu'elles sont toutes égales est très élevé. Pour diminuer ce risque global, différentes solutions techniques ont été proposées, dont la plus simple consiste à diminuer le risque α *local* de chaque comparaison (dans la procédure dite de Bonferroni on divise α par le nombre maximum $I(I-1)/2$ de comparaisons réalisables) ; leur caractère spécialisé (en particulier la justification de l'intérêt de maîtriser un risque global n'est pas si aisée) exclut ces techniques du champ de cet ouvrage ; dans les ouvrages de statistique classique, elles sont généralement regroupées sous le terme de « comparaisons multiples de moyennes ».

Indépendamment de la distinction *risque global / risque local*, une amélioration simple de la procédure naturelle (réaliser les comparaisons deux à deux) consiste à remplacer, au dénominateur du test t, la variance estimée à partir des deux seuls échantillons comparés par la variance estimée à partir de l'ensemble des échantillons (i.e. le Carré Moyen résiduel de l'analyse de variance), ce qui a un sens compte tenu de l'hypothèse technique $\sigma_i=\sigma$. L'amélioration provient de ce que l'on utilise plus de données pour estimer la variance intra-populations. Concrètement cela revient à remplacer (cf. Fiche 7 § 1.4.4.B)

$$s_d^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left[\frac{1}{n_1} + \frac{1}{n_2}\right] \text{ par } s_d^2 = \frac{\sum_i n_i s_i^2}{n - I}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

Exemple numérique

En reprenant la répartition des 907 élèves en 22 lycées, l'analyse de variance conduit au tableau 4.

Source de variation	SCE	ddl	CM	F (P)
totale	5220.853	906 n-1		
inter-lycées	517.359	21 I-1	24.636 $(4.96)^2$	4.635 (P=5 10^{-7})
intra-lycées	4703.494	885 n-I	5.315 $(2.31)^2$	

Tableau 4. *Tableau d'analyse de la variance de la note au bac des 907 élèves répartis dans les 22 lycées.*

La racine du carré moyen intra-lycées, ici 2.31, est en accord avec la variabilité intra-lycées que l'on peut observer figure **11.3**. Par ailleurs, sur ces même données, la figure **11.1** suggère une partition des lycées en trois groupes. Cette partition est-elle en harmonie avec l'analyse quantitative proposée par l'analyse de variance ? La comparaison entre les moyennes des lycées 15 et 5 conduit, selon la procédure qui utilise la variance intra-lycées calculée à partir de tous les lycées, au calcul suivant (cf. données **Tab. 11.1**) :

$$t = \frac{|10.78 - 10.98|}{2.31\sqrt{\frac{1}{53} + \frac{1}{25}}} = .357 \text{ ; Proba associée } .721$$

Les lycées 15 et 5 ne diffèrent pas significativement : compte tenu de la variabilité intra-lycées, la différence entre les deux moyennes observées peut être considérée comme fortuite.

Le même calcul, appliqué aux lycées 17 et 13 donne :

$$t = \frac{|10.48 - 11.68|}{2.31\sqrt{\frac{1}{47} + \frac{1}{47}}} = 2.52 \text{ ; Proba associée : } .012$$

Les lycées 17 et 13 diffèrent significativement : compte tenu de la variabilité intra-lycées, la différence entre les deux moyennes observées ne peut pas être considérée comme fortuite.

Cette démarche peut conduire à des contradictions apparentes du type : A=B ; B=C ; A≠C. En fait, si l'on envisage la constitution de groupes comme un moyen de décrire des données de façon synthétique, même si les groupes n'ont pas vraiment une réalité physique, on privilégiera l'approche descriptive, ne calculant les indicateurs statistiques qu'à titre indicatif. On maintiendra donc ici la subdivision en 3 groupes, tout en sachant très bien que le meilleur lycée (15) du groupe le plus mauvais n'est pas très différent du plus mauvais lycée (5) du groupe moyen.

Remarque. Lorsque les n_i sont égaux, on calcule une fois pour toute $t_{1-\alpha/2}\, s_d$ (avec $t_{1-\alpha/2}$ définie par Proba$[|X| > t_{1-\alpha/2}] = \alpha$, X étant distribuée selon une loi de Student à $n-I$ ddl). On compare alors cette quantité, dite Plus Petite Différence Significative (PPDS), directement aux différences de moyennes.

6 Vers l'analyse de variance à plusieurs facteurs

6.1 Modèle dans le cas de l'analyse de variance à un facteur

L'analyse de variance décrite ci-dessus, dite à un facteur (de variabilité –soit, dans l'exemple, le lycée–) fournit un cadre d'analyse de la liaison entre une variable quantitative et une variable qualitative. A ce titre, elle fait partie des méthodes de base, utilisables dans un très grand nombre de circonstances. En revanche, l'analyse de variance à plusieurs facteurs se situe dans un autre perspective : l'« explication » de la variabilité d'une variable quantitative généralement notée y (dite souvent variable « réponse ») par un ensemble de variables qualitatives (dites « facteurs ») au travers d'un modèle. Cette méthodologie est fondamentale en expérimentation : on fait varier les facteurs (selon un protocole judicieusement choisi) et l'on déduit du comportement de la réponse ses relations avec les facteurs, relations synthétisées sous la forme d'un modèle.

Dans le cas d'un facteur, ce modèle, décrit en 5.1, s'écrit de façon plus formelle :

$$y_{ij} = \mu_i + e_{ij} \text{ avec } \sum_j e_{ij} = 0$$

La contrainte sur les résidus e_{ij} assure l'interprétabilité de μ_i comme étant le niveau moyen du lycée i.

Dans l'exemple, ce modèle exprime que la note du $j^{ème}$ élève du lycée i « s'explique » par son appartenance au lycée i (μ_i) et par un écart dit « résiduel » qui rassemble la variabilité non prise en compte par des variables explicitées dans le modèle (niveau des élèves, humeur du correcteur lorsqu'il corrige telle copie, etc.). On retrouve une formalisation proche de la régression simple, les deux techniques étant effectivement étroitement liées : dans les deux cas on explique une variable quantitative y, à partir de variables quantitatives en régression et à partir de variables qualitatives en analyse de la variance (ces deux techniques sont souvent regroupées sous le terme générique de « modèle linéaire »).

Une autre écriture de ce modèle est :

$$y_{ij}=\mu+\alpha_i+e_{ij} \text{ avec } \sum_j e_{ij}=0 \text{ et } \sum_i \alpha_i=0$$

La contrainte sur les α_i assure la non-indétermination de μ et de α_i (par exemple, μ_1 et μ_2 étant fixés, il existe a priori une infinité de triplet (μ, α_1, α_2) tels que $\mu_1=\mu+\alpha_1$ et $\mu_2=\mu+\alpha_2$) ; μ représente un effet général et α_i l'effet du lycée i ; avec cette écriture, l'hypothèse de référence ne s'écrit plus H_0 : $\mu_i=\mu$ pour tout i mais H_0 : $\alpha_i=0$ pour tout i.

L'estimation des paramètres ne pose pas de problème dans ce cas simple : on estime μ_i par la moyenne de la classe i ($\bar{y}_i$).

Dans le cadre de l'analyse de variance à un facteur, cette formalisation n'est pas vraiment utile ; son principal intérêt réside dans sa généralisation à plusieurs facteurs.

6.2 Deux analyses de variance à un facteur appliquées aux mêmes données

Exemple. Le professeur de maths du lycée 10 étudie la variation des notes de ses 38 élèves au cours des trois trimestres (cf. **Tab. 9**). Il dispose au total de 38×3=114 notes ; les moyennes calculées pour chacun des trois trimestres sont : 9.2368, 11.7631, 11.0526. L'enseignant se demande si ces variations de moyennes peuvent être considérées comme fortuites ou, au contraire, dues par exemple à des niveaux de difficulté (ou de préparation !) différents entre les épreuves (on assimile un trimestre à une épreuve). Pour cela, il réalise dans un premier temps une analyse de variance à un facteur (les 114 notes sont réparties selon les trois trimestres) selon le modèle :

$$y_{ij}=\mu+\alpha_i+e_{ij} \text{ avec } \sum_j e_{ij}=0 \text{ et } \sum_i \alpha_i=0$$

note au trimestre i de l'élève j = effet général + niveau de l'épreuve i + résidu

Les résultats sont récapitulés tableau **5**.

Source de variation	SCE	ddl	CM	F (P)
totale	1586.631	113 n-1		
inter-trimestres	129	2 I-1	64.5 $(8.03)^2$	4.912 (P=.0088)
intra-trimestres	1457.631	111 n-I	13.13 $(3.62)^2$	

Tableau 5. *Notes en maths aux trois trimestres des 38 élèves du lycée 10 : tableau de l'analyse de la variance selon le facteur « trimestre ».*

Ils aboutissent à une probabilité associée (au *F*) de .0088 qui indique une variabilité significative (i.e. ne pouvant être considérée comme fortuite) entre entre les moyennes observées, ce qui conduit à considérer les niveaux d'épreuves comme différents. Cela étant, le fait de considérer les écarts entre les moyennes (d'environ 2 points entre le premier et les deux autres trimestres) comme grands ou petits (et non plus comme significatifs ou non) n'est plus un problème statistique.

Le carré moyen intra-trimestres (dit aussi résiduel) est ici 13.13 ; c'est lui qui sert de base - via le *F*- à l'évaluation de la variabilité inter-trimestres. Or la variabilité intra-trimestres, qui exprime que tous les élèves n'ont pas eu la même note au même examen, est sans doute due en bonne part à l'hétérogénéité du niveau des élèves ; à la limite, si le niveau des élèves est très hétérogène, le carré moyen résiduel est très grand et empêche de mettre en évidence une variabilité inter-trimestres même importante ; réciproquement, un niveau des élèves très homogène permet de mettre en évidence des différences inter-trimestres même petites. Or ici, étant donné que ce sont les mêmes élèves qui ont subi les épreuves, il est possible de mesurer le niveau de chaque élève, par exemple à l'aide de sa moyenne annuelle. Plus formellement, cela revient à considérer les 114 notes non plus comme trois ensembles de 38 notes mais comme 38 ensembles (un par élève) de 3 notes, et à faire l'analyse de variance correspondante dont le modèle est :

$$y_{ij}=\mu+\beta_j+e_{ij} \quad \text{avec} \quad \sum_i e_{ij}=0 \quad \text{et} \quad \sum_j \beta_j=0 \ ;$$

note au trimestre i de l'élève j = effet général + niveau de l'élève j + résidu

Les résultats sont récapitulés tableau **6**.

Source de variation	SCE	ddl	CM	*F* (*P*)
totale	1586.631	113 *n*-1		
inter-élèves	1285.298	37 *I*-1	34.74 $(5.89)^2$	8.773 ($P<10^{-9}$)
intra-élèves	301.333	76 *n-I*	3.96 $(1.99)^2$	

Tableau 6. *Notes en maths aux trois trimestres des 38 élèves du lycée 10 : tableau de l'analyse de la variance selon le facteur « élève ».*

Cette analyse met en évidence une variabilité inter-élèves hautement significative, terme employé lorsque la probabilité associée est très petite. Les écarts entre élèves peuvent en outre être qualifiés d'importants puisque les moyennes des élèves varient entre 2.67 pour le plus mauvais et 18.67 pour le meilleur (cf. **Tab. 9**). La variabilité résiduelle –ici intra-élèves– inclut, dans ce modèle, l'hétérogénéité des difficultés des trois épreuves, la plus ou moins bonne forme de l'élève le jour de tel épreuve, etc.

6.3 Exemple d'analyse de variance à deux facteurs

En fait, la structure des données analysées ici peut se mettre sous la forme d'un tableau à deux entrées ayant 38 lignes et 3 colonnes (cf. **Tab. 9**) dont les notations générales sont récapitulées tableau 7.

Elève	Trimestre 1	2	I=3	Moyenne
1	y_{11}	y_{21}	y_{31}	$\bar{y}_{1.}$
		...		
j	y_{1j}	y_{2j}	y_{3j}	$\bar{y}_{j}$
		...		
J=38	y_{1J}	y_{2J}	y_{3J}	$\bar{y}_{J}$
Moyenne	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\bar{y}_{3.}$	$\bar{y}_{..}$

Tableau 7. *Notes aux trois trimestres des 38 élèves du lycée 10.*
y_{ij} : note de l'élève j au trimestre i.
$\bar{y}_{j}$: moyenne de l'élève j (sur les I=3 trimestres).
$\bar{y}_{i.}$: moyenne du trimestre i (sur les J=38 élèves).
$\bar{y}_{..}$: moyenne tous élèves et tous trimestres confondus.

Cette structure de données suggère de prendre en compte simultanément l'effet élève et l'effet trimestre au sein d'une même analyse statistique. C'est ce que fait l'analyse de variance à 2 facteurs qui considère le modèle :

$$y_{ij}=\mu+\alpha_i+\beta_j+e_{ij} \text{ avec } \sum_j e_{ij}=0 \ ; \ \sum_i e_{ij}=0 \ ; \ \sum_i \alpha_i=0 \ ; \ \sum_j \beta_j=0$$

note à l'épreuve i de l'élève j = effet général + niveau de l'épreuve i + niveau de l'élève j + résidu

Le modèle est dit additif.

Pour le $j^{ème}$ individu appartenant au lycée i, on peut écrire :

$$(y_{ij}-\bar{y}_{..})=(\bar{y}_{i.}-\bar{y}_{..})+(\bar{y}_{.j}-\bar{y}_{..})+(y_{ij}-\bar{y}_{i.}-y_{.j}+y_{..})$$

En élevant au carré les deux termes de l'égalité précédente et en les additionnant pour tous les élèves, on obtient l'équation d'analyse de variance (toujours vraie) :

$$\sum_i\sum_j(y_{ij}-\bar{y}_{..})^2=\sum_i J(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_j I(\bar{y}_{.j}-\bar{y}_{..})^2+\sum_i\sum_j(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2$$

soit :

SCE totale = SCE inter-trimestres + SCE inter-élèves + SCE résiduelle (=interaction)

La variabilité résiduelle représente la variabilité qui demeure lorsque l'on a pris en compte le niveau de l'élève et celui de l'épreuve. Ainsi, un bon élève (β_j grand) peut obtenir une note mauvaise (y_{ij} petit) lors d'une épreuve pourtant facile (α_j grand) ; en fait, ce phénomène n'est pas véritablement de la variabilité résiduelle (la vraie variabilité résiduelle comporte la « forme » de l'élève le jour de telle épreuve, l'humeur du correcteur lorsqu'il corrige telle copie, etc. ; sa mesure implique, en toute rigueur, que l'on fasse passer plusieurs fois le même examen aux mêmes élèves ce qui n'est pas concevable pratiquement), mais ce que l'on appelle, en référence à l'expérimentation, une interaction. Formellement, on dit que deux facteurs sont en interaction (sous-entendu vis-à-vis d'une réponse) si l'effet de l'un (sur la réponse) dépend de la modalité de l'autre. On retrouve bien cette idée lorsque l'on dit que tel élève réussit bien tel type d'épreuve et non tel autre : l'effet élève dépend alors de l'épreuve. Ici, pour simplifier, nous supposons qu'il n'y a pas d'interaction (ce qui implique que les épreuves mesurent le même type de compétence et que le travail d'un élève dans une matière est homogène pendant l'année, ce qui est raisonnable) et donc que l'écart au modèle est constitué par une vraie variabilité résiduelle.

Appliquée à nos données, cette méthodologie conduit aux résultats rassemblés dans le tableau **8**. Pour les mêmes raisons que celles explicitées en 5.3 et de la même manière, chaque Somme des Carrés des Ecarts est affectée d'un nombre de degrés de liberté et le

Carré Moyen d'un effet est évalué en référence au Carré Moyen résiduel au travers du rapport F.

Source de variation	SCE	ddl	CM	F (P)
totale	1586.631	113 $n-1$		
inter-élèves	1285.298	37 $I-1$	34.74 $(5.89)^2$	14.92 ($P<10^{-9}$)
inter-trimestres	129	2 $J-1$	64.5 $(8.03)^2$	27.70 ($P=10^{-9}$)
résiduelle=interaction	172.333	74 $n-I-J+1$	2.328 $(1.53)^2$	

Tableau 8. *Notes en maths aux trois trimestres des 38 élèves du lycée 10 : tableau de l'analyse de la variance selon les deux facteurs « élève » et « trimestres ».*

La comparaison des tableaux 5, 6 et 8 montre comment une même variabilité -dite totale- peut être décomposée de plusieurs façons différentes mais bien sûr liées entre elles. On remarque que :

- les sommes des carrés des écarts (SCE) et les nombre de degrés de liberté (ddl) des effets « trimestres » et « élèves » sont les mêmes que l'on considère ces facteurs séparément (**Tab. 5** et **6**) ou simultanément (**Tab. 8**) ; Cette propriété est toujours vraie dès lors qu'il y a le même nombre d'observations dans le tableau qui croise les modalités des deux facteurs (concrètement, ici, le tableau **7** (ou **9**) contient une observation par case).
- la variabilité inter-trimestres est beaucoup plus significative (associée à une probabilité beaucoup plus faible) dans l'analyse à deux facteurs (**Tab. 8** : $P=10^{-9}$) que dans l'analyse à un facteur (**Tab. 5** ; $P=.0088$). Par rapport à la première analyse, le même carré moyen inter-trimestres est rapporté à un carré moyen résiduel beaucoup plus petit car on en a retiré la variabilité inter-élèves. L'analyse à deux facteurs est donc plus fine.

Dans cette analyse, les deux facteurs (élève et trimestre) sont significatifs. On conclut que, compte tenu des aléas déjà cités, les niveaux des élèves sont différents (pour ces épreuves) et que les niveaux des épreuves sont différents (pour ces élèves). On ne peut généraliser au-delà de ce lycée et de ces épreuves, ce qui aurait nécessité de tirer au hasard les élèves dans une population plus vaste (à laquelle nous aurions pu alors généraliser les résultats) ainsi que les sujets (dans un ensemble de sujets). Un effet dont les modalités sont tirées au hasard dans un ensemble plus vaste (le cas typique serait celui des élèves) est dit *effet aléatoire* (par opposition à *effet fixe*) : ce type d'effet nécessite un modèle et un traitement adaptés.

6.4 Conclusion

Nous avons esquissé l'une des techniques qui prolongent celles présentées dans cet ouvrage. On remarque que, dans l'exemple, la technique plus sophistiquée (à deux facteurs) ne bouleverse pas les conclusions de la technique simple (à un facteur) mais les affine. Cette situation est en fait fréquente, ce qui explique que l'on obtient, dans les cas usuels, l'essentiel des résultats à partir de techniques simples. Selon un autre point de vue, lorsqu'une structure est forte dans les données, elle apparaît même avec une technique rudimentaire. Toutefois, un raffinement du type de celui présenté ci-dessus peut être décisif, notamment dans des contextes d'expérimentation ou chaque essai est coûteux.

élève	Trim. 1	Trim. 2	Trim. 3	Moyenne	Ecart-type	Variance
1	2	3	3	2.67	0.471	0.222
2	3	3	4	3.33	0.471	0.222
3	4	6	4	4.67	0.943	0.889
4	5	7	6	6	0.816	0.667
5	6	9	6	7	1.414	2
6	7	7	9	7.67	0.943	0.889
7	6	9	8	7.67	1.247	1.556
8	8	7	10	8.33	1.247	1.556
9	6	10	10	8.67	1.886	3.556
10	8	10	8	8.67	0.943	0.889
11	7	12	8	9	2.16	4.667
12	9	11	9	9.67	0.943	0.889
13	9	12	8	9.67	1.7	2.889
14	9	11	10	10	0.816	0.667
15	8	10	12	10	1.633	2.667
16	9	12	10	10.33	1.247	1.556
17	6	12	13	10.33	3.091	9.556
18	9	12	11	10.67	1.247	1.556
19	9	11	12	10.67	1.247	1.556
20	11	12	9	10.67	1.247	1.556
21	10	13	9	10.67	1.7	2.889
22	7	14	13	11.33	3.091	9.556
23	12	12	11	11.67	0.471	0.222
24	8	15	13	12	2.944	8.667
25	10	12	14	12	1.633	2.667
26	9	12	15	12	2.449	6
27	10	13	14	12.33	1.7	2.889
28	11	15	12	12.67	1.7	2.889
29	14	13	12	13	0.816	0.667
30	11	15	13	13	1.633	2.667
31	12	15	13	13.33	1.247	1.556
32	12	15	13	13.33	1.247	1.556
33	10	16	14	13.33	2.494	6.222
34	10	17	14	13.67	2.867	8.222
35	14	12	16	14	1.633	2.667
36	16	16	17	16.33	0.471	0.222
37	16	17	18	17	0.816	0.667
38	18	19	19	18.67	0.471	0.222
Moyenne	9.237	11.76	11.05	10.68	1.45	2.643
Ecart-type	3.437	3.579	3.706	3.357		
Variance	11.81	12.81	13.73	11.27		

Tableau 9. *Notes en maths des 38 élèves du lycée 10. Les moyennes, écarts-types et variances sont donnés par ligne et par colonne. Les élèves sont rangés par moyenne croissante*

Distribution de variables quantitatives, observées ou aléatoires

1. Distribution observée, distribution théorique, variable aléatoire

Dans le chapitre 7, on étudie de façon détaillée la distribution d'une variable quantitative observée sur une population, c'est-à-dire l'ensemble des valeurs (chacune pondérée par sa fréquence) prises par la variable.

Il est souvent utile de disposer de distributions théoriques. Ceci dans des optiques diverses :

- décrire une distribution ; exemple : dire d'une distribution observée qu'elle est très proche d'une distribution de référence est un moyen de la caractériser ;
- peser la vraisemblance d'hypothèses qui impliquent une distribution théorique ; pour cela on confronte la distribution observée à la distribution théorique attendue ;
- calculer une probabilité approchée, à l'aide d'une distribution théorique qui ne diffère pas trop de la distribution exacte du phénomène étudié ; c'est le cas de la plupart des probabilités associées à un indicateur statistique ;
- réaliser divers calculs théoriques ; par exemple, on peut vouloir prévoir, à partir de la distribution de plusieurs variables, la distribution d'une combinaison (e.g. la moyenne) de ces variables.

Ces modèles de référence sont proposés par la *théorie des probabilités*, dans laquelle la grandeur étudiée est appelée *variable aléatoire* et la distribution associée *loi de la variable aléatoire*. Le mot aléatoire vient du fait que la plupart d'entre elles découlent de la modélisation du résultat d'un tirage "au hasard" dans une population.

2. Distribution de variables observées

Nous reprenons quelques notions élémentaires concernant les variables observées avant d'introduire les notions équivalentes pour les variables aléatoires.

2.1 Représentations graphiques

Diagramme en bâtons

Il visualise la fréquence de chaque valeur prise par la variable. Il n'est possible que si la variable est discrète (i.e. en pratique prend un petit nombre de valeurs différentes). Au-dessus de chaque valeur de la variable, figure un bâton dont la longueur est proportionnelle

au nombre (et donc au pourcentage) d'individus prenant cette valeur. L'ordre et même les écarts entre les valeurs de la variable doivent être respectés (cf. exemple **Fig. 4.1**).

Histogramme

L'histogramme dérive du diagramme en bâtons ; il est utilisé lorsque l'on désire regrouper des valeurs successives pour avoir une vue plus "lissée" de la répartition. Ceci est nécessaire quand le nombre de valeurs prises par la variable est grand. Au lieu de faire figurer chacune des valeurs prises par la variable, on s'intéresse aux intervalles obtenus en divisant le domaine de variation de la variable (cf. exemples **Fig. 7.3**). L'histogramme se compose de rectangles ayant chacun pour base un de ces intervalles et pour **surface** le nombre (ou la proportion) d'individus prenant une valeur située dans cet intervalle. La hauteur de ces rectangles, égale au rapport entre le nombre (ou la proportion) d'individus prenant ses valeurs sur l'intervalle et la longueur de l'intervalle, est la *densité moyenne d'individus sur l'intervalle*.

Les intervalles ont le plus souvent la même longueur : la hauteur d'un rectangle est alors proportionnelle au nombre d'individus qui prennent leur valeur dans l'intervalle correspondant (cf. 7.2).

2.2 Moyenne et variance

Pour décrire une distribution, moyenne et écart-type sont les deux indices les plus utilisés.

Moyenne

La moyenne d'un ensemble de I valeurs x_i est notée $\bar{x}$ et définie par :

$$\bar{x} = \frac{1}{I}\sum_{i=1,I} x_i = \sum_{i=1,I}\frac{1}{I}x_i$$

Dans cette somme, chaque valeur x_i a le même poids ($1/I$) et la somme des poids vaut 1.

Lorsqu'il y a des ex æquo, on peut les regrouper et considérer les n valeurs x_k distinctes, chacune affectée du coefficient f_k, proportion d'individus ayant la valeur x_k. L'ensemble des couples (x_k, f_k) constitue la distribution observée. La moyenne précédente peut s'écrire en fonction de cette distribution :

$$\bar{x} = \sum_{k=1,n} f_k x_k$$

Ceci revient à affecter le poids f_k à la valeur x_k, la somme de ces poids étant égale à 1.

Variance et écart-type

Variance : moyenne des carrés des écarts à la moyenne ; écart-type : racine carrée de la variance.

La variance d'un ensemble de I valeurs x_i est notée s^2 et définie par :

$$s^2 = \frac{1}{I}\sum_{i=1,I}(x_i - \bar{x})^2$$

En fonction de la distribution (x_k, f_k), cette variance s'écrit :

$$s^2 = \sum_{k=1,n} f_k (x_k - \bar{x})^2$$

Remarque : toutes ces définitions se généralisent aisément en accordant des poids distincts aux individus i (cf. Fiche 3 § 3).

2.3 Distribution conjointe de deux variables

Considérer simultanément plusieurs variables revient à s'intéresser à leur distribution conjointe, c'est-à-dire, dans le cas de deux variables x et y, à la distribution du couple (x,y). La représentation de la distribution du couple (x,y) se fait :

- pour les variables présentant un petit nombre de valeurs, à l'aide d'un tableau croisé (cf. Tab. 8.1) dans lequel on trouve, à l'intersection de la ligne x_i et de la colonne y_j le nombre (ou le pourcentage) d'individus ayant à la fois la valeur x_i pour x et la valeur y_j pour y ;
- pour les variables présentant un grand nombre de valeurs, à l'aide d'un graphique dont chaque axe de coordonnée correspond à l'une des variables (cf. **Fig. 8.1**).

A partir de la distribution conjointe, on définit les notions de :

distribution marginale, distribution de l'une des variables en regroupant les individus qui présentent la même valeur pour cette variable (quelle que soit la valeur de l'autre variable) ; dans un tableau croisé, ces distributions figurent dans les marges (cf. **Tab. 8.1**) ;

distribution conditionnelle, distribution de l'une des variables lorsque l'on se restreint aux individus présentant une valeur fixée pour l'autre variable ; dans un tableau croisé, chaque ligne (et chaque colonne) constitue une distribution conditionnelle.

2.4 Indépendance entre deux variables

On étudie la loi conjointe de deux variables X et Y pour examiner la liaison entre ces deux variables. Cet examen se fait en référence à une situation d'indépendance parfaite, laquelle stipule que les distributions conditionnelles de X (resp. de Y) sont égales entre elles et à la distribution marginale de X (resp. de Y). Concrètement, dans un tableau croisé satisfaisant ces conditions, les lignes (resp. les colonnes) sont proportionnelles entre elles et à la marge-ligne (resp. marge-colonne). Un tel tableau est utilisé Fiche 6 (§ 1.2).

3. Distribution de variables aléatoires discrètes

Les variables discrètes prennent un nombre fini, généralement assez petit, de valeurs différentes. La distribution observée d'une variable discrète est caractérisée par le pourcentage d'individus prenant chacune des valeurs. Pour une variable aléatoire discrète, ces pourcentages sont remplacés par des probabilités : nombres compris entre 0 et 1 dont la somme vaut 1. Dans les deux cas, la distribution de la variable (ou encore loi de probabilité pour la variable aléatoire) est représentée par un diagramme en bâtons.

3.1 Loi uniforme

C'est l'exemple le plus simple de distribution. Supposons que l'on jette un dé dont les 6 faces sont numérotées de 1 à 6. Les dés ont été inventés pour obtenir "au hasard" l'un des 6 premiers nombres, avec "autant de chance" pour chacun d'eux. La variable aléatoire discrète qui modélise cette situation prend pour valeurs les nombres de 1 à 6 avec, pour chacun, une probabilité égale à 1/6 (cf. **Fig. 1.a**).

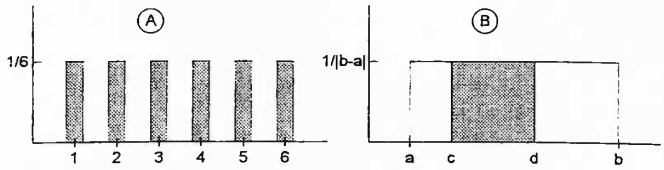


Figure 1. *Exemples de lois uniformes.*
a) Cas discret : variable prenant les valeurs entières de 1 à 6
b) Cas continu : variable prenant les valeurs de l'intervalle [a,b] ; la probabilité d'appartenir à [c,d] est égale au rapport des longueurs [c,d]/[a,b].

Bien évidemment, on ne rencontre pas souvent de dés dans nos problèmes pratiques ; mais le jet de dés est une situation type dans laquelle tous les événements possibles ont la même « chance » de se produire. En fait, la loi uniforme modélise une situation dans laquelle :

- n événements peuvent se produire ;
- chacun d'eux a la même chance de se produire et est donc affecté de la probabilité $1/n$.

3.2 Loi binomiale et loi hypergéométrique

Situation type : n lancers de pièces

Une situation type dans laquelle intervient la loi binomiale est celle où l'on jette n fois une pièce de monnaie et où l'on compte le nombre de fois où *face* est obtenue. Ce nombre est forcément un entier compris entre 0 et n. L'usage premier des pièces de monnaie n'est pas de les lancer en l'air pour voir si elles retombent sur *pile* ou sur *face* mais, lorsqu'on le fait, on suppose qu'il y a autant de "chances" d'obtenir un résultat que l'autre. Ceci se traduit, dans le modèle, par une probabilité égale à 1/2 pour *pile* (et 1/2 pour *face*) pour chaque lancer. On peut supposer aussi, pour plus de généralité, que la probabilité de *face* n'est pas forcément 1/2 mais un nombre p quelconque compris entre 0 et 1.

Loi binomiale

On considère les n lancers. La grandeur étudiée est le nombre de *face* obtenu. Cette grandeur est associée à un tirage au hasard : c'est une variable aléatoire.

Soit k une valeur possible du nombre de *face*. On a : $0 \leq k \leq n$. A chaque valeur de k on associe la probabilité, notée $P(k)$, d'observer k fois *face*. Les règles mathématiques de calcul des probabilités ont été formalisées notamment pour obtenir ce genre de résultat. Ici on trouve (p : probabilité de *face*) :

$$P(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Notations : $n! = n \times (n\text{-}1) \times (n\text{-}2) \times \ldots \times 3 \times 2 \times 1$; se lit « factorielle n ». On note souvent $C_n^k = n!/[k!(n-k)!]$ le nombre d'ensembles comportant k éléments que l'on peut extraire d'un ensemble de n éléments.

L'ensemble des valeurs de $P(k)$ s'appelle *loi binomiale*. En réalité, il ne s'agit pas d'une loi mais plutôt d'une famille dont chaque loi correspond à une valeur de n et de p. La figure **2** donne la distribution de la loi binomiale pour deux couples de valeurs de n *et* p.

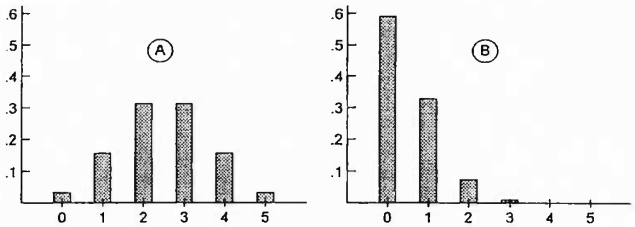


Figure 2. *Distribution de la loi binomiale. A) $n=5$ et $p=1/2$; B) $n=5$ et $p=1/10$.*

Plus n est grand et plus le calcul de P(k) est compliqué. Des tables donnent ces probabilités, généralement pour des valeurs de n inférieures à 30 ; au-delà, on utilise des approximations.

- Lorsque np et $n(1-p)>5$, on utilise la loi normale de même moyenne et de même variance (cf. plus loin).
- Sinon, en particulier quand p est petit ($p<0.1$) , on utilise une autre loi, dite loi de Poisson, qui dépend d'un seul paramètre noté λ et est définie par :

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{en utilisant } np \text{ comme valeur de } \lambda$$

Loi hypergéométrique : n tirages sans remise

La loi binomiale modélise ce que l'on appelle le tirage « avec remise » par opposition au tirage « sans remise » dont le modèle est illustré en 10.1. Nous reprenons ici cet exemple, dans lequel on cherche à calculer la probabilité d'obtenir au moins 8 candidats libres en prenant au hasard 15 élèves parmi les 975, sachant qu'il y a en tout 13 candidats libres. Pour obtenir 15 élèves, on peut commencer par en tirer un au hasard. La probabilité que ce soit un candidat libre vaut 13/975. On peut ensuite en tirer un second au hasard mais la probabilité est moins simple à exprimer : si le premier n'est pas un candidat libre elle vaut 13/974 puisqu'il ne reste que 974 élèves et si le premier est un candidat libre elle vaut 12/974. Le premier nombre est très proche de 13/975, le second en est un peu plus éloigné (0.012 au lieu de 0.013). Pour le tirage du troisième élève et des suivants, la situation se complique encore.

La loi correspondant à un tirage sans remise s'appelle loi hypergéométrique. Elle est plus complexe que la loi binomiale et dépend de 3 paramètres : la taille de la population (N), le nombre d'individus (parmi les N) ayant le caractère étudié (N_1) et la taille de l'échantillon (n). Par analogie avec la loi binomiale, on note $p=N_1/N$.

Pour simplifier et obtenir tout de même un résultat satisfaisant bien qu'approché, on peut considérer qu'à chaque tirage successif des 15 élèves, la probabilité d'obtenir un candidat libre reste de 13/975. Ce serait le cas si, avant de tirer au hasard le deuxième élève, le premier était « remis » dans la population et de même pour les suivants (d'où les termes *tirage avec remise* et *tirage sans remise*). Ce raisonnement conduit à utiliser la loi binomiale en tant qu'approximation de la loi hypergéométrique. Cette approximation est d'autant meilleure que le tirage sans remise « déforme » peu la population, condition réalisée lorsque le nombre n de tirages est très petit vis-à-vis de la taille N de la population.

3.3 Moyenne et variance

Les notions de moyenne et de variance (et donc d'écart-type) définies pour les variables observées se transposent directement pour les variables aléatoires. La moyenne d'une variable aléatoire est aussi appelée *espérance mathématique*. Pour une variable aléatoire discrète X prenant n valeurs notées x_k avec la probabilité p_k, la moyenne ou espérance mathématique de X, notée $E(X)$, s'écrit :

$$E(X) = \sum_{k=1,n} p_k x_k$$

Le parallélisme avec la moyenne $\bar{x}$ d'une distribution observée est évident.

On vérifie facilement que la moyenne de la variable aléatoire associée au jet d'un dé vaut 3.5 ; remarquons que, lors d'une série de lancers, on obtient « en moyenne » 3.5 mais que l'on obtient jamais 3.5 . On montre que la moyenne d'une loi binomiale, ainsi que celle d'une loi hypergéométrique, de paramètres n et p vaut np. Exemple : en tirant au hasard (avec ou sans remise) un échantillon de 1000 personnes dans une population comportant autant d'hommes que de femmes, on observe en moyenne $1000\times(1/2)=500$ femmes.

De même la variance $V(X)$ de la variable aléatoire X s'écrit :

$$V(X) = \sum_{k=1,n} p_k (x_k - E(X))^2$$

Le parallélisme avec la variance s^2 d'une distribution observée est évident.

La distribution de la loi uniforme sur les entiers de 1 à n, qui a pour moyenne $(n+1)/2$, a pour variance $(n-1)(n+1)/12$.

Les variances des lois binomiale et hypergéométrique figurent tableau **1**. Sur ce point, les deux lois diffèrent par le coefficient $[(N-n)/(N-1)]$, toujours inférieur à 1 et d'autant plus proche de 1 que la taille de l'échantillon est petite relativement à celle de la population.

La loi de Poisson de paramètre λ a pour espérance et pour variance λ. En confrontant les espérances et variances des lois binomiale et de Poisson, on retrouve les conditions de l'approximation de la première par la seconde : $\lambda=np$ et p petit.

Loi	espérance	variance
binomiale	np	$np(1-p)$
hypergéométrique	np	$[np(1-p)]\times[(N-n)/(N-1)]$
de Poisson	λ	λ

Tableau 1. *Espérance et variance des lois binomiale, hypergéométrique et de Poisson.*

3.4 Distribution conjointe de deux variables

Dans l'étude simultanée de deux variables aléatoires discrètes, il faut définir pour toute valeur x_i de X **et** toute valeur y_j de Y la probabilité d'obtenir à la fois $X=x_i$ **et** $Y=y_j$ (probabilité notée $P[X=x_i$ et $Y=y_j]$). L'ensemble de ces probabilités constitue la distribution conjointe de X et de Y. En pratique, on présente souvent ces nombres dans un tableau à deux entrées, analogue du tableau croisé introduit dans le cas de données observées (cf. 8.3) à condition de diviser l'effectif de chaque case par l'effectif total. La représentation graphique se fait à l'aide d'un histogramme à deux dimensions, dont un exemple dans le cas de données observées se trouve **Fig. 8.2**.

Les notions introduites à propos du tableau croisant deux distributions observées se transposent directement à un couple de variables aléatoires. Ainsi, on définit les :

distributions marginales, qui ne considèrent que l'une des variables ; dans le tableau à deux entrées, on les obtient en additionnant tous les termes d'une même ligne ou d'une même colonne ; ainsi, la distribution marginale de X est reliée à la distribution conjointe par :

$$P[X=x_i] = \sum_{y_j} P[X=x_i \text{ et } Y=y_j]$$

distributions conditionnelles, décrivant la loi de l'une des variables, l'autre étant fixée. Dans le tableau à double entrée, cela revient à se limiter à une ligne (ou à une colonne), en divisant chaque terme par la probabilité marginale de la ligne (ou de la colonne). La probabilité de $X=x_i$ conditionnellement à $Y=y_j$ est notée $P[X=x_i/Y=y_j]$; elle s'obtient à partir des distributions conjointes et marginales par la relation :

$$P[X=x_i/Y=y_j] = P[X=x_i \text{ et } Y=y_j] \,/\, P[Y=y_j]$$

Ces notions se généralisent à un nombre quelconque de variables aléatoires.

3.5 Indépendance entre deux variables

Intuitivement, on peut dire que deux variables aléatoires X et Y sont indépendantes si la connaissance de la valeur prise par l'une n'apporte aucun renseignement sur la valeur prise par l'autre. Autrement dit, pour des variables discrètes, en fixant la valeur y_j prise par Y, la distribution de probabilité de X (distribution conditionnelle) est la même pour tout y_j et est identique à la distribution marginale. Soit, pour tout couple (i, j) : $P[X=x_i/Y=y_j] = P[X=x_i]$

Si cette relation est vraie quels que soient i et j, il en résulte que, pour tout couple (x_i,y_j), la probabilité d'obtenir $X=x_i$ **et** $Y=y_j$ est égale au produit de la probabilité associée à $X=x_i$ et de la probabilité associée à $Y=y_j$. Soit, pour tout couple (i, j) :

$$P[X=x_i \text{ et } Y=y_j] = P[X=x_i] \times P[Y=y_j]$$

Cette propriété est la définition de l'indépendance entre les deux événements $X=x_i$ et $Y=y_j$.

L'indépendance se définit aussi pour un nombre quelconque de variables aléatoires (cf. Fiche 6 § 4.4). Attention : l'indépendance entre plusieurs variables implique l'indépendance entre les variables prises deux à deux mais cette condition n'est pas suffisante.

4. Distribution de variables aléatoires continues

Les variables dites "continues" prennent généralement toutes les valeurs d'un intervalle donné, borné ou non, de l'ensemble des nombres réels. L'ensemble des valeurs possibles est dit « domaine de variation ».

Pour décrire la distribution d'une variable aléatoire continue, on ne peut donner la probabilité d'obtenir chaque valeur précise : d'une part il y en a une infinité et, d'autre part, cette probabilité est la plupart du temps nulle. Par contre, on peut donner la probabilité d'appartenir à un intervalle donné. Et ceci pour n'importe quel intervalle appartenant au domaine de variation.

On est proche des histogrammes de variables quantitatives qui donnent la densité moyenne de la population sur des intervalles du domaine de définition. Mais pour les variables aléatoires continues qui prennent toutes les valeurs d'un intervalle, on peut prendre des

intervalles aussi petits que l'on veut et définir la densité sur chacun de ces tout-petits intervalles. On peut ainsi définir la densité (dite de probabilité) en chaque point comme étant la limite de la densité d'un intervalle contenant x, lorsque cet intervalle est pris aussi petit que possible. On obtient ainsi une *fonction de densité* $f(x)$ définie pour chaque valeur x du domaine de variation. Comme pour toute fonction, on peut tracer son graphe.

Cette densité (ou fonction de densité) définie en chaque point généralise la densité sur un intervalle représentée par le haut des rectangles d'un histogramme : la probabilité d'appartenir à un intervalle donné est égale à la surface délimitée par la courbe au-dessus de cet intervalle (cf. **Fig. 1** et **3**). La surface délimitée par la courbe toute entière vaut 1 : elle représente la probabilité d'appartenir au domaine de définition.

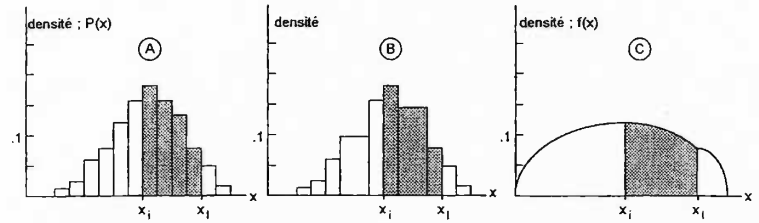


Figure 3. *Représentation de la probabilité P d'appartenir à un intervalle* (x_i, x_l).
P : surface délimitée par la fonction de densité au dessus de l'intervalle.
***A** et **B**) Cas discret, histogramme avec intervalles égaux (**A**) et inégaux (**B**) ; le haut des rectangles représente la densité moyenne sur chaque intervalle ; la fonction de densité est discontinue ; dans **A**, la densité moyenne sur un intervalle est proportionnelle à la probabilité de l'intervalle. **C**) Cas continu.*

Les calculs présentés dans le cas discret se généralisent aux variables continues en utilisant des notions mathématiques plus complexes : les sommes sont remplacées par des intégrales, ainsi que le montre le tableau **2**. L'écriture :

$$\int_{x_i}^{x_l} f(x)dx$$

se lit « intégrale de $f(x)$ entre x_i et x_l » et désigne la surface délimitée par la courbe $f(x)$ au-dessus de l'intervalle (x_i, x_l)

	cas discret	cas continu
$P[X \in (x_i, x_l)]$	$\sum_{k \in (i,l)} p_k$	$\int_{x_i}^{x_l} f(x)dx$
$E(X)$ (Espérance de X)	$\sum_{k=1,n} p_k x_k$	$\int_{-\infty}^{+\infty} xf(x)dx$
$V(X)$ (Variance de X)	$\sum_{k=1,n} p_k (x_k - E(X))^2$	$\int_{-\infty}^{+\infty} [x - E(X)]^2 f(x)dx$

Tableau 2. *Parallélisme entre les cas discret et continu dans quelques calculs fondamentaux.*

4.1 Loi uniforme sur un intervalle

Prenons d'abord l'exemple très simple de la loi uniforme sur un segment $[a,b]$. Comme pour toute loi continue, la distribution uniforme est définie par une fonction de densité $f(x)$. Cette fonction vaut 0 en dehors du segment $[a,b]$. En un point quelconque de $[a,b]$ elle est égale à $1/|b-a|$, inverse de la longueur de $[a,b]$. Le terme uniforme exprime que la densité est constante sur le segment (on qualifie aussi cette distribution de « rectangulaire »). Sa courbe représentative est donc confondue avec l'axe des x en dehors de l'intervalle $[a,b]$; pour $[a,b]$, c'est un segment de droite situé à la hauteur $1/|b-a|$ (cf. **Fig. 1.b**).

La probabilité d'appartenir à un segment qui ne recoupe pas $[a,b]$ est nulle : sur un tel segment, la surface définie par la courbe au-dessus de l'axe des x est nulle.

La probabilité d'appartenir à un segment $[c,d]$ contenu dans $[a,b]$ vaut : $|d-c|/|b-a|$. Elle est proportionnelle à la longueur du segment $[c,d]$ et ne dépend pas de sa position. Si ce segment est $[a,b]$ tout entier, cette probabilité vaut 1 (cf. **Fig. 1.b**).

En accord avec l'intuition, l'espérance d'une loi uniforme vaut $(b-a)/2$, valeur située au milieu de $[a,b]$. La variance vaut $(b-a)^2/12$ et donc l'écart-type $|b-a|/3.46$; ce dernier est un peu plus grand que l'écart absolu moyen : $|b-a|/4$.

4.2 Loi normale

Intérêt

La loi normale apparaît comme un modèle adapté dans beaucoup de situations concrètes :

- représenter la distribution d'une même mesure faite plusieurs fois avec des erreurs ;
- représenter la distribution de la moyenne de n variables aléatoires de même loi ; c'est un résultat fondamental : la distribution de cette moyenne, quelle que soit la loi initiale, converge en pratique rapidement vers une loi normale ;
- approcher certaines lois (comme la loi binomiale pour n grand) qui sont elles-mêmes des modèles naturels de certaines situations.

Bien souvent elle est utilisée, alors que la réalité en est éloignée, car elle seule permet d'effectuer certains calculs. Bien entendu, plus la réalité s'écarte du modèle, plus les conclusions obtenues à partir du modèle doivent être considérées avec prudence.

Fonction et courbe de densité

La fonction de densité d'une loi normale de moyenne m et d'écart-type s s'écrit :

$$f(x) = \frac{1}{s\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-m}{s}\right)^2\right]$$

Sa courbe représentative a une forme de cloche (cf. **Fig. 4**). La densité est surtout importante autour de la moyenne, puis décroît de façon symétrique d'autant plus rapidement que l'écart-type est petit. Cette allure générale est fréquemment observée (cf. par exemple les histogrammes de la **Fig. 7.1**), mais, pour la loi théorique, la densité n'est jamais exactement nulle, ce qui n'est jamais le cas des valeurs observées (toujours bornées).

De sa fonction de densité on déduit qu'une distribution normale peut avoir une moyenne et un écart-type quelconque, mais ces deux paramètres étant fixés, la loi est entièrement

définie. Deux distributions normales de même écart-type et de moyennes différentes sont décalées. Un écart-type plus grand correspond à une courbe plus étalée, c'est-à-dire diminue la probabilité d'appartenir à un intervalle donné $[a,b]$ centré sur le sommet de la cloche.

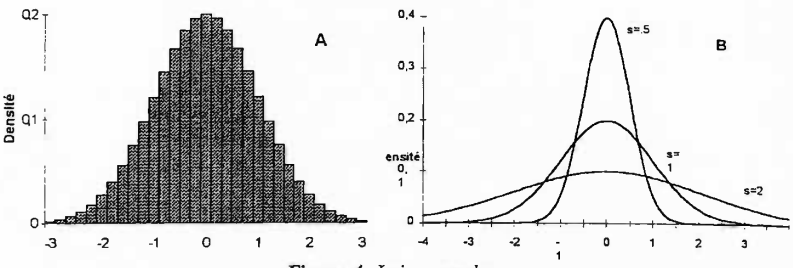


Figure 4. *Lois normales.*
A *: histogramme donnant la densité moyenne sur des intervalles (loi centrée-réduite).*
B *: fonction donnant la densité en chaque point (lois normales centrées ; s=écart-type).*

Loi normale centrée-réduite

Si la variable X est distribuée selon une loi normale d'espérance m et d'écart-type s, alors la variable $(X-m)/s$ est distribuée selon une loi normale centrée-réduite (de moyenne nulle et d'écart-type égal à 1). Il en résulte que la probabilité pour la variable X d'appartenir à l'intervalle $[a,b]$ est égale à la probabilité pour la variable centrée-réduite $(X-m)/s$ d'appartenir à l'intervalle $[(a-m)/s,(b-m)/s]$. Ainsi, on peut déduire les probabilités relatives à une loi normale quelconque à partir de la loi normale centrée-réduite, ce qui explique que cette dernière soit la seule tabulée.

Exemple : X distribué selon une loi normale de moyenne 2 et d'écart-type 3 ;

$$P[-4\leq X\leq 8] = P[(-4-2)/3\leq (X-2)/3\leq (8-2)/3] = P[-2\leq Y\leq 2] = .954$$

.954 est lue dans une table de la loi normale centrée-réduite (loi de Y).

Loi normale et seuils de 5% et de 1%

On prend souvent comme seuil (tout à fait arbitraire mais si utilisé que ces valeurs sont devenues des normes) du peu probable une probabilité inférieure à 0.05 voire à 0.01.

Pour une loi normale, le seuil de 0.05 correspond à 1.96 écarts-types : la probabilité d'être à plus de 1.96 écarts-types de la moyenne est de 0.05. Ainsi, pour une variable supposée être distribuée au moins approximativement selon une loi normale, une valeur située à plus de 1.96 (ou 2 pour simplifier) écarts-types de la moyenne peut être considérée comme exceptionnelle (cf. Ch. 12).

De même, le seuil de 0.01 correspond à 2.58 écarts-types : la probabilité d'être à plus de 2.58 écarts-types est de 0.01. La probabilité d'obtenir une valeur située à plus de a écarts-types de la moyenne diminue très vite lorsque a augmente (cf. **Tab. 3**).

a	1	2	3	4	5	6
$P[\lvert X\rvert \geq a]$	.327	.046	$.27\ 10^{-2}$	$.63\ 10^{-4}$	$.57\ 10^{-6}$	$.2\ 10^{-8}$

Tableau 3. *Quelques probabilités issues de la loi normale centrée-réduite*

4.3 Quelques autres lois

Un certain nombre de lois théoriques, qui dérivent de la loi normale, sont très utilisées pour calculer des probabilités approchées de phénomènes réels. Ces lois sont définies en combinant plusieurs variables aléatoires indépendantes (l'indépendance entre variables aléatoires introduite dans le cas discret en 3.5 est transposée au cas continu en 4.4).

Loi du χ^2

C'est la loi d'une variable aléatoire continue Y, somme des carrés de n variables aléatoires X_i normales, centrées, réduites et indépendantes.

$$Y = X_1^2 + X_2^2 + \ldots\ldots + X_n^2$$

Elle dépend d'un paramètre, noté usuellement υ, appelé *nombre de degrés de liberté* (en abrégé : ddl). Sa moyenne est égale à υ et sa variance à 2υ.

Le terme *degré de liberté* exprime le nombre de composantes indépendantes de Y ; dans la définition ci-dessus, l'indépendance entre les n variables X_i entraîne : $\upsilon = n$; plus généralement, la somme des carrés de n variables aléatoires, chacune distribuée selon une loi normale centrée-réduite mais telles qu'il existe entre elles m relations linéaires indépendantes, est distribuée selon une loi de χ^2 avec un nombre de ddl $\upsilon = n-m$.

On rencontre cette loi lorsque l'on additionne des carrés de valeurs, opération fréquente en statistique (e.g. calcul de somme de carrés d'écarts en analyse de variance cf. Fiche 8 § 3).

Cas particulier : le carré d'une variable aléatoire distribuée selon une loi normale centrée-réduite est distribuée selon une loi du χ^2 à 1 ddl.

Loi t de Student

C'est la loi d'une variable aléatoire continue Z définie par le rapport : $Z = \dfrac{X}{\sqrt{Y/n}}$

dans lequel X est une variable normale centrée-réduite et Y une variable distribuée, indépendamment de X, suivant une loi du χ^2 à n degrés de liberté. Elle dépend du paramètre n, appelé ici encore nombre de degrés de liberté ; sa moyenne est 0 et , lorsque $n_2 > 2$, sa variance vaut $n/(n-2)$.

Remarque : lorsque n croît, la distribution de Student converge vers une loi normale. En pratique, les deux distributions sont très proches dès que $n > 30$.

Loi F de Fisher (dite aussi de Fisher-Snedecor)

C'est la loi d'une variable aléatoire continue Z définie par le rapport : $Z = \dfrac{Y_1/n_1}{Y_2/n_2}$

dans lequel les deux variables aléatoires Y_1 et Y_2 sont indépendantes et distribuées selon une loi de χ^2 à respectivement n_1 et n_2 degrés de liberté. Elle dépend des paramètres n_1 et n_2, appelés ici encore nombres de degrés de liberté ; sa moyenne vaut $n_2/(n_2-2)$ lorsque $n_2 > 2$: elle ne dépend pas de n_1 et est en pratique voisine de 1.

Cas particulier : le carré d'une variable aléatoire distribuée selon une loi de Student à n ddl est distribué selon une loi de Fisher à $n_1 = 1$ et $n_2 = n$ ddl.

4.4 Distribution conjointe et indépendance entre deux variables

Distribution conjointe de deux variables aléatoires

Pour décrire la distribution d'un couple de variables aléatoires continues, on ne peut donner la probabilité d'obtenir chaque couple de valeurs : d'une part, il y a une (double) infinité de couples ; d'autre part, cette probabilité est la plupart du temps nulle. Par contre, il faut définir, pour tout couple d'intervalles $[a,b]$ et $[c,d]$, la probabilité d'obtenir à la fois X appartenant à $[a,b]$ et Y appartenant à $[c,d]$, c'est-à-dire la probabilité associée au rectangle induit par les intervalles $[a,b]$ et $[c,d]$ dans le plan engendré par X et Y (cf. **Fig.** 5 et, pour une distribution observée, Fig. 8.2).

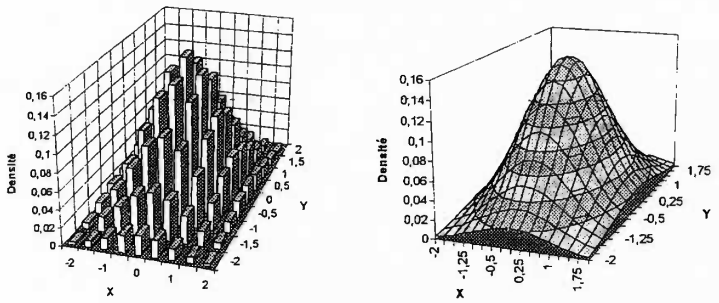


Figure 5. *Loi normale centrée-réduite bidimensionnelle.*
A gauche : histogramme à 2 dimensions (=stéréogramme) donnant la densité moyenne pour chaque élément d'un carroyage du plan (X,Y) défini par les deux variables X et Y. A droite : surface donnant la densité en chaque point du plan (X,Y). Remarques : une distribution conditionnelle, obtenue en coupant cette surface par un plan parallèle à l'axe des X (ou des Y) est normale ; l'ensemble des points de même densité, obtenu en coupant cette surface par un plan parallèle au plan (X,Y), est une ellipse.

On est proche des histogrammes à deux dimensions donnant la densité d'une population sur des rectangles du plan défini par deux variables. En considérant des rectangles de plus en plus petits, on obtient à la limite une surface, représentation graphique d'une fonction de densité à deux variables (dite aussi à deux dimensions). Cette fonction, définie pour chaque couple de points, est notée $f(x,y)$. La probabilité associée au couple d'intervalles $[a,b]$ et $[c,d]$ est égale au volume délimité par la surface au-dessus du rectangle associé au couple $\{[a,b]\ ,\ [c,d]\}$. Le volume délimité par l'ensemble de la surface vaut 1.

De même que dans le cas discret, on définit les :

- *distributions marginales*, qui ne considèrent que l'une des deux variables ; on les obtient en sommant (techniquement, cette « somme continue » est une intégrale) tous les termes pour une même valeur de X (ou de Y) ; ainsi, la densité marginale de X, notée $f_1(x)$, et celle de Y, notée $f_2(y)$ sont reliées à la densité conjointe par :

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy \quad \text{et} \quad f_2(y) = \int_{-\infty}^{+\infty} f(x,y)dx$$

- *distributions conditionnelles*, décrivant la loi de l'une des variables, l'autre étant fixée. Cela revient à se limiter à une seule valeur x de X (ou y de Y), en divisant chaque densité par la densité marginale de x (ou de y). La densité de X conditionnellement à $Y=y$ est notée $f(x/y)$; elle s'obtient à partir des distributions conjointes et marginales par la relation :

$$f(x/y)=f(x,y)/f_2(y)$$

Ces notions se généralisent à un nombre quelconque de variables aléatoires.

Cas particulier : loi normale bidimensionnelle

La figure 8.1 montre un exemple de distribution conjointe observée. La densité des points est maximum aux alentours du point moyen ; elle diminue quand on s'éloigne de ce point, mais pas de la même façon dans toutes les directions. La densité semble à peu près constante sur des ellipses concentriques autour du point moyen (cf. aussi **Fig. 5**).

Ce type de distribution observée peut être modélisé par la loi normale bidimensionnelle, dans laquelle les deux distributions marginales (de X et de Y) sont normales, condition nécessaire mais non suffisante. Dans le cas où les distributions marginales sont centrées-réduites, la fonction de densité s'écrit (en notant r le coefficient de corrélation entre les distributions marginales) :

$$f(x,y)=\frac{1}{2\pi\sqrt{1-r^2}}\exp\left[-\frac{x^2-2rxy+y^2}{2(1-r^2)}\right]$$

Outre les deux distributions marginales normales, la distribution normale bidimensionnelle possède plusieurs propriétés dont en particulier :

- toute combinaison linéaire de X et de Y est distribuée normalement ;
- les distributions conditionnelles sont normales (cf. **Fig. 5**) ;
- la moyenne de la distribution conditionnelle de Y est une fonction linéaire de x, propriété illustrée dans une distribution observée (cf. **Fig. 8.5**) ;
- si les distributions marginales sont non corrélées ($r=0$), elles sont indépendantes.

La loi normale se généralise à plus de deux dimensions : la loi normale multidimensionnelle.

Indépendance entre deux variables aléatoires

L'indépendance définie dans le cas discret (§ 3.5) se transpose au cas continu, en considérant les densités au lieu des probabilités. Deux variables aléatoires continues sont indépendantes si les densités conditionnelles sont égales entre elles et égales aux densités marginales. Soit :

$$f(x/y)=f_1(x) \quad \text{et} \quad f(y/x)=f_2(y)$$

Il résulte de cette définition et de celle d'une distribution conditionnelle, que deux variables aléatoires continues sont indépendantes si et seulement si leur densité conjointe est égale au produit des densités marginales. Soit :

$$f(x,y)=f_1(x)\times f_2(y)$$

L'indépendance se définit aussi pour un nombre quelconque de variables aléatoires par l'égalité entre la distribution conjointe et le produit des distributions marginales. Attention : l'indépendance entre plusieurs variables implique l'indépendance entre les variables prises deux à deux mais cette condition n'est pas suffisante (cf. dans le cas discret Fiche 6 § 4.4).

5. Modèle et variable observée

Il est rare que la distribution d'une variable observée sur une population se confonde avec un modèle. Elle ne fait que s'en approcher. Le cas des variables discrètes et celui des variables continues sont différents comme on peut le voir sur les deux exemples ci-dessous.

Loi uniforme discrète

Si l'on jette un dé bien équilibré un certain nombre de fois et si l'on calcule les pourcentages des résultats correspondant à chacun des 6 nombres, on obtient une distribution pas trop éloignée de celle de la variable aléatoire discrète, surtout si le nombre de jets du dé est grand. Il est rare que l'on obtienne la distribution théorique exacte : mais le modèle est tout à fait admissible pour décrire exactement la réalité.

Loi normale

Beaucoup de distributions observées ressemblent à celle de la loi normale : dans les histogrammes, la "courbe" formée par le haut des rectangles (et leur jonction verticale) s'approche de la "vraie" courbe en cloche. Mais il demeure toujours des différences fondamentales.

A) La densité définie par l'histogramme est constante sur chaque intervalle (la portion de courbe correspondante est horizontale) alors que la densité du modèle varie en chaque point.

B) Dans le modèle, un intervalle aussi petit soit-il n'a jamais une probabilité nulle. Pour la variable observée, si la population est très nombreuse, on peut tracer des histogrammes en choisissant des intervalles plus petits, ce qui permet de se rapprocher de la courbe. Mais par nature la population étudiée étant finie, si on diminue trop la taille des intervalles on finit par engendrer beaucoup d'intervalles vides et la courbe formée par le haut des rectangles devient chaotique. Autrement dit, la distribution normale ne peut jamais être atteinte dans la réalité où l'on ne peut observer qu'un nombre fini de valeurs ; c'est une limite théorique qui suppose la population infinie, ce qui n'est pas vraiment réaliste. Cette différence fondamentale entre réalité et modèle est valable pour toutes les variables aléatoires continues.

C) Dans le modèle normal, la courbe de densité va jusqu'à l'infini à gauche et à droite (la probabilité d'appartenir à un segment très loin du haut de la cloche est beaucoup plus faible que celle d'appartenir à un segment de même longueur situé plus près, mais elle n'est jamais exactement nulle) ; en comparaison, une distribution observée est limitée par sa valeur maximum et sa valeur minimum.

En conclusion, le modèle normal ne peut jamais prétendre décrire exactement une situation concrète ; mais il en est souvent assez proche pour que des calculs sur ce modèle donnent des ordres de grandeur suffisants en pratique.

Indicateur statistique et probabilité associée

1. Pourquoi utiliser des probabilités dans l'examen d'un ensemble de données ?

Les indicateurs statistiques ont pour objet de synthétiser des données dans une perspective déterminée. Ainsi :

- dans la comparaison entre deux populations (par exemple, les élèves de deux lycées) du point de vue d'une variable (par exemple, les résultats en mathématiques au bac), on calcule pour chaque lycée la moyenne des notes obtenues par ses élèves, puis l'indicateur « différence entre ces deux moyennes » ;
- dans l'étude de la liaison entre deux variables (par exemple les notes en mathématiques et en physique au bac), on calcule le coefficient de corrélation entre ces deux variables.

En pratique, l'indicateur statistique n'est pas toujours suffisant. En effet, on n'accorde pas la même importance à la valeur prise par un indicateur statistique selon le contexte dans lequel elle est observée. Ainsi, pour interpréter un indicateur, il faut tenir compte :

- de la variabilité des données (e.g. une différence de deux points n'a pas le même sens si elle concerne une note sur 20 ou une note sur 5) ;
- des effectifs observés (e.g. une différence de deux points semble banale si elle est observée entre deux élèves, moins banale si elle est observée entre deux lycées).

Selon un autre point de vue, pour apprécier la valeur prise par un indicateur, on se pose souvent des questions du type suivant : le résultat observé est-il fortuit ? Est-il comparable à ce que l'on observe en réunissant des données « au hasard » ?

Ces interrogations conduisent à associer une probabilité à la valeur d'un indicateur.

2. Modèle de tirage au hasard intérieur aux données

2.1 Deux exemples pour illustrer la démarche

Tout au long de cette fiche, nous illustrons les raisonnements à l'aide de deux cas.

Cas 1 : comparaison entre deux moyennes

On compare la note moyenne $\bar{x}_1$ des n_1 élèves du lycée 1 avec la note moyenne $\bar{x}_2$ des n_2 élèves du lycée 2. Le tableau **1.a** présente un exemple de petite taille ($n_1 = 3$; $n_2 = 2$).

	Lycée 1	Lycée 2
notes	11	7
	13	12
	14	
moyennes	12.67	9.5
écart	3.17	

a) Exemple fictif 1

	X	Y
individu 1	1	1
individu 2	3	3
individu 3	5	7
individu 4	7	5
coef. correl.	.8	

b) Exemple fictif 2

Tableau 1. *Deux exemples fictifs de données observées*
a) *notes dans une matière pour 5 élèves de deux lycées ;* **b)** *notes dans 2 matières (X, Y) pour 4 élèves.*

Cas 2 : étude de la liaison linéaire entre deux variables quantitatives

On étudie la liaison entre les variables X (e.g. note en mathématiques) et Y (e.g. note en physique) pour un ensemble de N élèves. Le tableau **1.b** contient un exemple de petite taille (N=4).

2.2 Calcul d'une probabilité associée via un modèle intérieur aux données

La question « le résultat observé peut-il être considéré comme fortuit ? » se réfère implicitement à un modèle de tirage au hasard qu'il faut spécifier. Nous qualifions ce modèle d'*intérieur aux données* car il ne dépend que des données observées.

Nous raisonnons pas à pas. A chaque **pas** nous mettons en parallèle le cas de la comparaison entre deux moyennes (*cas 1*) et celui du coefficient de corrélation (*cas 2*).

Pas 0. Spécification du problème

Cas 1 : on compare deux groupes d'individus selon leur moyenne pour une variable.

Cas 2 : on étudie la linéarité de la liaison entre deux variables.

Pas 1. Spécification de la structure (= mode d'organisation) des données

Cas 1 : on dispose de N valeurs d'une même variable, réparties en 2 classes ayant respectivement n_1 et n_2 valeurs ; la structure est l'appartenance des valeurs aux 2 classes.

Cas 2 : on dispose de deux séries de N valeurs, observations de deux variables X et Y sur les mêmes individus ; la structure est l'appariement des valeurs des deux séries.

Pas 2. Choix d'un indicateur statistique selon le problème

Cas 1 : pour comparer deux moyennes, l'indicateur naturel est la différence entre ces deux moyennes ou, ce qui revient au même, la différence entre la moyenne d'une classe (par exemple la première) et la moyenne toutes classes confondues.

Cas 2 : pour apprécier le caractère plus ou moins linéaire de la liaison entre X et Y, l'indicateur usuel est le coefficient de corrélation.

Pas 3. Définition du modèle de tirage au hasard

Principe : on considère le tirage au hasard, dans l'ensemble des valeurs effectivement observées, d'une structure de même type que celle des données.

Cas 1 : l'ensemble des données est constitué des N valeurs. Dans cet ensemble, on tire au hasard l'appartenance aux classes. Cela revient à tirer au hasard n_1 valeurs parmi N, ces valeurs étant affectées à la classe 1 ; les valeurs restantes sont affectées à la classe 2.

Cas 2 : l'ensemble des données est constitué de deux séries de N valeurs. Dans cet ensemble, on tire au hasard l'appariement ; cela revient à tirer au hasard, parmi les N valeurs de Y, celle que l'on apparie à la première valeur de X, puis, parmi les N-1 valeurs restantes de Y, celle que l'on apparie à la seconde valeur de X, etc.

Pas 4. Distribution de l'indicateur statistique

Principe : à une répartition des données selon la structure étudiée, correspond une valeur de l'indicateur statistique. A l'ensemble des répartitions susceptibles d'être tirées au hasard, correspond l'ensemble des valeurs possibles de l'indicateur, c'est-à-dire sa distribution. Cette distribution contient la valeur effectivement observée.

Cas 1 : on énumère, par permutation, toutes les répartitions possibles des N valeurs en deux sous-ensembles de n_1 et n_2 valeurs. Pour chaque répartition, on calcule la différence entre les deux moyennes. On obtient ainsi la distribution de cette différence. La différence effectivement observée est l'une de ces valeurs.

Cas 2 : on énumère, par permutation, tous les appariements possibles entre les valeurs de X et celles de Y. Pour chaque appariement, on calcule le coefficient de corrélation. On obtient ainsi la distribution de ce coefficient de corrélation. Le coefficient de corrélation effectivement observé est l'une de ces valeurs.

Pas 5. Calcul d'une probabilité associée

Principe : le caractère improbable (ou non fortuit) de la valeur effectivement observée est mesuré par la proportion des valeurs possibles au moins aussi éloignées de la moyenne de la distribution que ne l'est la valeur effectivement observée. Cette proportion est la probabilité d'obtenir, dans le cadre d'un tirage au hasard, une répartition au moins aussi remarquable que celle effectivement observée (remarquable signifie ici : éloignée, du point de vue de l'indicateur statistique, de ce que l'on obtient en moyenne par un tirage au hasard).

Cas 1 : par raison de symétrie, la moyenne de la distribution de la différence entre les deux moyennes vaut 0. On calcule donc la probabilité d'obtenir par tirage au hasard une appartenance aux classes telle que la différence entre les moyennes est au moins aussi éloignée de 0 que celle effectivement observée (en raccourci : probabilité d'obtenir une différence entre les deux moyennes supérieure ou égale en valeur absolue à celle observée).

Cas 2 : par raison de symétrie, la moyenne de la distribution du coefficient de corrélation vaut 0. On calcule donc la probabilité de tirer un appariement dont le coefficient de corrélation est, en valeur absolue, supérieur ou égal à celui effectivement observé (en raccourci : probabilité d'obtenir un coefficient de corrélation supérieur ou égal en valeur absolue à celui observé).

2.3 Illustrations numériques

Exemple numérique fictif 1 : comparaison entre deux moyennes

A partir de l'ensemble des données observées, i.e. ici {7,11,12,13,14}, on construit tous les cas possibles qui respectent les effectifs réels des lycées : il y a ici 10 cas possibles, c'est-à-dire 10 façons de fragmenter 5 nombres en un groupe de 3 et un groupe de 2 (cf. **Tab. 2**).

cas 1	
L1	L2
7	13
11	14
12	
−3.5	

cas 2	
L1	L2
7	12
11	14
13	
−2.67	

cas 3	
L1	L2
7	11
12	14
13	
−1.83	

cas 4	
L1	L2
11	7
12	14
13	
1.5	

cas 5	
L1	L2
7	12
11	13
14	
−1.83	

cas 6	
L1	L2
7	11
12	13
14	
−1	

cas 7	
L1	L2
11	7
12	13
14	
2.33	

cas 8	
L1	L2
7	11
13	12
14	
−.17	

cas 9	
L1	L2
11	7
13	12
14	
3.17	

cas 10	
L1	L2
12	7
13	11
14	
4	

Tableau 2. *Exemple fictif 1 : ensemble des permutations possibles des données. Le cas 9 correspond à la situation effectivement observée.*

L'ensemble des valeurs possibles de la différence entre les deux moyennes peut être représenté sur un axe (cf. **Fig. 1**).

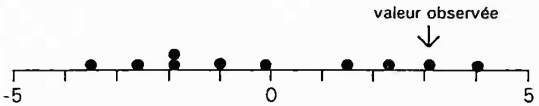


Figure 1. *Exemple fictif 1 : distribution des valeurs possibles de la différence des moyennes lorsque l'on réalise toutes les permutations possibles des données.*

Une fois située parmi l'ensemble des valeurs possibles, la différence observée (3.17) n'est en rien exceptionnelle puisque 30% des situations possibles conduisent à une différence au moins aussi grande (en valeur absolue). Autrement dit : .3 est la probabilité d'obtenir, par un tirage au hasard, une différence au moins aussi grande (en valeur absolue) que celle effectivement observée.

Remarque. Il revient au même de trier toutes les situations possibles selon l'écart entre les deux moyennes (ce qui a été fait) ou selon l'écart entre la moyenne de l'un des lycées et la moyenne générale des deux lycées. Autrement dit, en notant $\bar{x}$ la moyenne générale tous lycées confondus (dans l'exemple $\bar{x} = 11.4$), on montre facilement que les deux critères $\bar{x}_1 - \bar{x}_2$ et $\bar{x}_1 - \bar{x}$ sont équivalents. Cette équivalence fournit une autre interprétation au pourcentage précédent : 30% est le pourcentage de sous-ensembles de 3 valeurs au moins aussi éloignés de la moyenne générale que ne l'est celui effectivement observé.

Exemple numérique fictif 2 : étude de la liaison linéaire entre deux variables quantitatives

Il y a 24 ($=4\times3\times2\times1=4!$ qui se lit *factorielle* 4) façons d'apparier les valeurs de X et de Y ; concrètement, on fixe l'affectation de X et on permute les valeurs de Y. Pour chaque appariement, on calcule le coefficient de corrélation r (cf. **Tab. 3**).

X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
1	7	1	7	1	5	1	5	1	7	1	5	1	1	1	3	1	3	1	3	1	1
3	5	3	5	3	7	3	3	3	1	3	1	3	7	3	5	3	1	3	1	3	3
5	3	5	1	5	1	5	7	5	3	5	7	5	5	5	1	5	7	5	5	5	5
7	1	7	3	7	3	7	1	7	5	7	3	7	3	7	7	7	5	7	7	7	7
r=-1		r=-.8		r=-.6		r=-.4		r=-.2		r=0		r=.2		r=.4		r=.6		r=.8		r=1	

Tableau 3. *Exemple fictif 2 : 11 exemples de permutations des données du tableau 1 b.*

Un diagramme en bâtons décrit la distribution des 24 valeurs possibles de r (cf. **Fig. 2**).

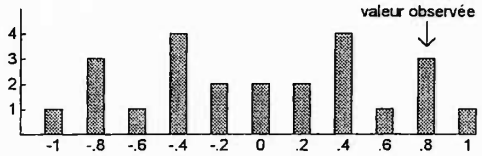


Figure 2. *Exemple fictif 2 : distribution des valeurs possibles du coefficient de corrélation lorsque l'on réalise toutes les permutations des données du tableau 1 b.*

Une fois située parmi les valeurs effectivement possibles compte tenu des distributions de X et de Y, la valeur observée .8 n'est pas exceptionnelle en ce sens que 8 valeurs sur 24 (33%) sont au moins aussi grandes en valeur absolue. 33% est le pourcentage associé à la valeur observée ; .33 est la probabilité d'obtenir, par un tirage au hasard, un coefficient de corrélation au moins aussi grand (en valeur absolue) que celui effectivement observé.

2.4 Mise en œuvre pratique du calcul d'une probabilité associée

Deux voies pour réaliser les calculs

En pratique, l'ensemble des situations possibles est trop grand pour être recensé. La probabilité cherchée ne peut être qu'approchée. Deux démarches sont possibles.

1- Réaliser une série de tirages au hasard dans le cadre du modèle intérieur aux données. La proportion de tirages qui conduisent à un indicateur statistique supérieur ou égal à la valeur effectivement observée est une valeur approchée de la probabilité cherchée. La précision de l'approximation augmente avec le nombre de tirages réalisés. En pratique, quelques milliers de tirages sont un compromis entre une précision suffisante (si l'on ne cherche pas à distinguer entre elles des probabilités très petites) et un temps de calcul raisonnable.

2- Utiliser un modèle théorique qui permet de déterminer une distribution théorique que l'on peut considérer comme une approximation satisfaisante de celle de l'indicateur, moyennant quelques hypothèses techniques que l'on espère pas trop irréalistes. Cette solution très économique du point de vue du calcul est utilisée lorsque l'on calcule systématiquement la probabilité associée d'un grand nombre d'indicateurs (cf. 11.9).

Utilisation d'une distribution approchée

Nous illustrons dans les deux cas précédents (différence entre deux moyennes et coefficient de corrélation) le choix d'une loi théorique pour approcher la distribution de l'indicateur.

Cas 1. On adopte le point de vue de la différence entre la moyenne d'une classe et la moyenne toutes classes confondues. Cette dernière étant fixe (et connue) le problème est ramené à l'étude de la distribution de la moyenne d'un échantillon de taille fixée.

Or, quelle que soit la distribution de la population dans laquelle on effectue le tirage au hasard, la distribution de la moyenne d'un échantillon converge vers une loi normale lorsque la taille de l'échantillon grandit (tout en restant petite vis-à-vis de la taille de la population). En pratique, cette convergence est rapide et conduit à des approximations satisfaisantes dès que l'effectif de l'échantillon atteint quelques dizaines (cf. Fiche 11).

Cette propriété suggère l'utilisation de la loi normale pour approcher la distribution réelle d'une moyenne. Dans l'exemple fictif, le nombre d'individus est trop faible pour appliquer ce calcul approché. Des exemples de mise en œuvre pratique de cette approximation sont décrits dans la fiche 7.

Cas 2. Le problème est celui de la distribution des valeurs possibles du coefficient de corrélation r lorsque les deux variables sont non-corrélées. Or, si n est grand, on peut utiliser comme approximation de la distribution de r une distribution normale, de moyenne nulle et de variance $1/(n-1)$.

Dans l'exemple fictif, le nombre d'individus est trop faible pour appliquer ce calcul approché de r. Un exemple de mise en œuvre pratique de cette approximation (ainsi que d'autres approximations) est décrit dans la fiche 5.

En conclusion

La démarche présentée et illustrée par deux exemples s'applique à la plupart des situations concrètes. Toutefois, lorsque le problème étudié est complexe, on rencontre des difficultés de divers types :

- la définition de l'ensemble des situations possibles peut être difficile (quoique toujours possible) ;
- le cas peut ne pas avoir été prévu dans les logiciels disponibles, auquel cas l'écriture d'un programme spécifique, quoique toujours possible, peut être difficile ;
- l'approximation par une loi théorique peut être de médiocre qualité ou ne pas être connue.

Heureusement on rencontre rarement toutes ces difficultés simultanément. On utilisera :

- de préférence, la probabilité associée exacte (il existe des logiciels spécialisés qui réalisent ces calculs lorsque les données sont peu nombreuses) ;
- sinon, l'approximation fournie par une série de tirages au hasard (il existe des logiciels spécialisés) ;
- en dernier recours, l'approximation fournie par une loi théorique (ce type d'approximation est employé dans la plupart des logiciels courants).

Si, pour des raisons pratiques, on doit se contenter d'une approximation par une loi théorique, on ne sera pas catastrophé pour autant ; en effet la qualité de ce type d'approximation est le plus souvent suffisante car seul l'ordre de grandeur de la probabilité importe ; par exemple, on fait généralement une différence entre .2 et .02 ou .002 mais non entre .2 et .5 , entre .02 et .05 , entre .002 et .005 .

3. Utilisation de la probabilité associée

3.1 Portées respectives de l'indicateur statistique et de la probabilité associée

Utilisation conjointe de l'indicateur et de la probabilité associée

L'indicateur statistique mesure l'intensité d'un phénomène. Ce que l'on exprime par des phrases du type : du point de vue de la moyenne au bac la différence est plus importante entre les lycées 1 et 2 qu'entre les lycées 4 et 5 ; les notes trimestrielles sont plus liées entre elles en mathématiques ($r_{t1t2}=.77$; $r_{t2t3}=.81$) qu'en philosophie ($r_{t1t2}=.64$; $r_{t2t3}=.69$).

La probabilité associée mesure le caractère « non fortuit » du phénomène observé. Ce que l'on exprime par des phrases du type : la différence observée entre les lycées 1 et 2 peut être considérée comme fortuite car elle est du même ordre de grandeur que ce que l'on obtient lors de tirages au hasard ; les liaisons observées entre les notes trimestrielles de mathématiques ne sont vraisemblablement pas fortuites.

Une difficulté provient du fait que ces deux notions se recouvrent partiellement : à nombre de données constant, plus un phénomène observé est intense et moins on a tendance à le considérer comme fortuit. C'est le cas des exemples de corrélation entre mathématiques et philosophie cités plus haut qui s'appuient tous sur le même nombre d'élèves : dans ces cas la relation d'ordre induite par les indicateurs statistiques est la même que celle induite par la probabilité associée. Cependant, ceci n'est plus vrai dès que l'on ne se place plus à effectif constant (cf. **Tab. 4**).

Un cas limite

On mesure deux variables sur deux individus ; le coefficient de corrélation calculé à partir de ces données vaut toujours 1. De fait le nuage, réduit à deux points, est toujours parfaitement situé sur une droite ; en ce sens la liaison entre les deux variables est parfaitement linéaire. Cependant la probabilité d'observer une valeur au moins aussi élevée du coefficient de corrélation dans un tel cas (2 individus) vaut 1 : le moins que l'on puisse dire est que cette valeur (pourtant très grande !) du coefficient de corrélation n'est dans ce cas ni exceptionnelle ni improbable.

Exemples fictifs

	effectif	r	proba
cas 1	10	.4	.2
cas 2	10	.8	.004
cas 3	100	.3	.004

Tableau 4. *Trois exemples fictifs de coefficients de corrélation et de probabilité associée*

Cas 1 et 2 : les effectifs sont identiques ; dans le cas 2, la liaison est plus intense et donc plus exceptionnelle.

Cas 1 et 3 : le nuage de points ressemble plus à une droite dans le cas 1 que dans le cas 3 ; autrement dit, les variables sont plus liées linéairement dans le cas 1 que dans le cas 3 ; mais cette liaison possède un caractère fortuit dans le cas 1 qu'elle ne possède pas dans le cas 3.

Conclusion

La probabilité associée est utile dans la comparaison de situations impliquant des effectifs différents et même des variables différentes. Ainsi, par exemple, on considérera que, dans tel lycée, l'écart entre la note du 3ème trimestre et celle du bac en mathématiques est moins fortuit que ce même écart mesuré en physique dans tel autre lycée. Cette idée est à la base d'un tri automatique d'un ensemble d'indicateurs selon leur probabilité associée (cf. 11.4).

3.2 Peut-on porter un jugement absolu sur une probabilité associée ?

Appréciation d'une probabilité associée

On peut toujours classer des probabilités associées et donc des indicateurs selon leur caractère plus ou moins fortuit. Peut-on aller plus loin ?

Les valeurs 0 (impossibilité) et 1 (certitude) ont le mérite d'être claires mais n'apparaissent jamais. En pratique, il est naturel d'assimiler à 0 les valeurs très petites et à 1 les valeurs très grandes. Se pose alors le problème de la définition de limite(s) sur l'échelle des probabilités. Or, il n'existe aucune limite naturelle qui implique, c'est-à-dire décide à notre place, de considérer tel phénomène comme étant ou n'étant pas fortuit.

Il n'en reste pas moins la nécessité de faire le lien entre le caractère continu de la probabilité et le caractère discontinu de la décision de considérer comme fortuit ou non un résultat, même si l'on ajoute une troisième possibilité, par exemple « différer une conclusion en attendant des informations supplémentaires ».

En pratique, cette décision dépend de plusieurs éléments pas tous quantifiables :

- l'intensité du phénomène observé (i.e. la valeur de l'indicateur) ;
- son caractère plus ou moins fortuit (i.e. la probabilité associée) ;
- son interprétabilité (i.e. la façon dont il s'intègre à d'autres résultats ou plus généralement aux connaissances disponibles sur le sujet) ;
- les enjeux.

Ces éléments ne vont pas obligatoirement dans le même sens et la démarche de l'utilisateur sera toujours teintée d'empirisme. Donnons deux exemples :

Cas 1. La valeur de l'indicateur ne peut être considérée comme fortuite (par exemple, probabilité associée =.0001) mais on ne sait pas l'interpréter. Il est nécessaire de pousser plus loin l'investigation en vérifiant qu'il n'y a pas d'erreur dans les données, en recoupant ce résultat avec d'autres et/ou en imaginant, quitte à les valider par la suite, d'autres raisons rendant compte de l'observation.

Cas 2. La valeur de l'indicateur est du type de celles produites par le hasard (par exemple, probabilité associée =.15) alors que l'on sait l'interpréter et qu'elle est cohérente avec nos connaissances sur la question. Il est nécessaire ici aussi de pousser plus loin l'investigation en vérifiant les données, en recoupant ce résultat avec d'autres et/ou en imaginant, quitte à les valider par la suite, d'autres raisons rendant compte de l'observation. Une des raisons est peut-être que le nombre d'observations est insuffisant pour mettre en évidence un phénomène certain mais ténu.

4. Cas de données obtenues à partir d'un échantillon tiré au hasard dans une population

La démarche développée jusqu'ici met en œuvre des calculs réalisés aussi dans le cadre de l'inférence statistique classique. Aussi n'est-il pas inutile de présenter quelques éléments de cette inférence pour situer ces deux démarches.

4.1 Domaine de l'inférence statistique classique

Comparé aux problèmes abordés jusqu'ici, le domaine de l'inférence statistique classique possède deux caractéristiques.

A- Les données sont obtenues à partir d'un échantillon (ou plusieurs) tiré au hasard dans une population (la façon précise de tirer l'échantillon, définie en rapport avec les objectifs, peut être plus ou moins complexe). Cette situation est typiquement celle d'un sondage (l'étude des dossiers scolaires entrerait dans ce cadre si, par exemple, les 993 élèves avaient été tirés au hasard dans toute la France). Dans ce cas, l'échantillon ne présente pas d'intérêt en lui-même mais uniquement dans la mesure où il donne une idée de la population dont il est extrait, cette dernière étant l'ensemble véritablement étudié. Se pose alors la question d'extrapoler ou non à l'ensemble de la population ce que l'on observe sur l'échantillon. Tel est le problème général de l'inférence. L'inférence statistique classique fournit un cadre formel à une telle extrapolation en s'appuyant sur le fait que l'échantillon étudié provient d'un tirage au hasard. En revanche, lorsque l'échantillon ne provient pas d'un tirage au hasard, l'inférence statistique classique ne s'applique pas et l'extrapolation ne peut s'appuyer que sur des critères qualitatifs : l'inférence est alors dite empirique, non formalisée ou non statistique.

B- Les résultats sur l'extrapolation desquels on s'interroge concernent un questionnement qui doit avoir été spécifié **avant** l'étude des données (à tel point que ce questionnement sert à définir le mode de recueil des données). C'est une des règles de la démarche scientifique : le même ensemble de données ne peut servir à la fois pour construire des hypothèses et les vérifier.

4.2 Principe de l'inférence statistique classique

4.2.1 Calcul d'une probabilité associée

Etant donné un indicateur statistique, on calcule une probabilité associée selon un principe analogue à celui utilisé dans le cadre du modèle intérieur aux données. La seule différence (mais elle est de taille) est que le modèle utilisé fait intervenir la population et le mode de tirage dont est issu l'échantillon (ce modèle n'est donc pas intérieur aux données).

Nous décrivons cette démarche pas à pas dans le contexte d'un sondage. Chaque **pas** est illustré par le cas de la comparaison entre deux moyennes (*cas 1*) et celui de l'examen d'un coefficient de corrélation (*cas 2*).

Pas 1. Problématique et description du mode d'obtention des données

Cas 1 : on étudie la différence entre les résultats trimestriels en mathématiques des filles et des garçons. Pour cela, on a tiré au hasard un échantillon comportant n_1 filles et un échantillon comportant n_2 garçons. Pour chaque élève, on relève sa note en mathématiques.

Cas 2 : on étudie la liaison entre la note en mathématiques (X) et la note en physique (Y) ; pour cela, on a tiré au hasard un échantillon de N élèves. Pour chacun, on relève sa note en mathématiques et sa note en physique.

Pas 2. Formalisation du mode d'obtention des données par un modèle théorique

Cas 1 : on considère l'ensemble des filles et celui des garçons comme deux populations comportant un nombre infini d'individus. A chaque individu correspond la valeur d'une variable. Cette variable a pour moyenne μ_1 pour la première population et μ_2 pour la seconde (traditionnellement, on utilise les lettres grecques pour les paramètres des populations théoriques). Le tirage au hasard est celui de n_1 individus d'une part et de n_2 individus d'autre part.

Cas 2 : on considère l'ensemble des élèves comme une population comportant un nombre infini d'individus. A chaque individu, on associe les valeurs de deux variables X et Y. Pour l'ensemble de la population, le coeffîcient de corrélation entre X et Y vaut ρ_{xy}. Le tirage au hasard est celui de N individus.

Pas 3. Définition de l'hypothèse de référence associée à la question posée

Principe : on traduit la question posée dans les termes des paramètres du modèle ; la situation neutre (absence de différence, de liaison, etc.) constitue une hypothèse de référence (notée H_0 qui se lit H zéro) par rapport à laquelle on examine les données.

Cas 1 : existe-t-il une différence entre les deux moyennes des populations ? L'hypothèse de référence correspondant à cette question spécifie l'absence d'une différence entre les moyennes des deux populations soit : $\mu_1=\mu_2$.

Cas 2 : les deux variables sont-elles liées linéairement au niveau de la population ? L'hypothèse de référence correspondant à cette question spécifie l'absence d'une liaison linéaire entre les deux variables étudiées soit : $\rho_{xy}=0$

Pas 4. Distribution de l'indicateur statistique dans le cadre de l'hypothèse de référence

Principe : du fait du tirage au hasard, l'indicateur statistique est une variable aléatoire. Il faut en connaître la distribution dans le cadre de l'hypothèse de référence, ce qui nécessite des hypothèses techniques (et donc simplement utilitaires, par opposition à l'hypothèse de référence intrinsèquement liée à la question posée ; on distingue quelquefois ces hypothèses techniques sous le nom de suppositions) sur le modèle. Si ces hypothèses ne sont pas vérifiées, la distribution obtenue n'est qu'une approximation (souvent suffisante en pratique) de la vraie.

Cas 1 : dans chacune des deux populations, la variable étudiée est supposée distribuée selon une loi normale. Alors, dans le cadre de l'hypothèse de référence, la différence entre les deux moyennes des échantillons est distribuée selon une loi normale de moyenne nulle (le calcul pratique de la variance de cette loi pose un problème technique qui peut être résolu de différentes façons ; cf. Fiche 7).

Cas 2 : on suppose que, dans la population, le couple de variable X,Y est distribué selon une loi normale à 2 dimensions. Alors, dans le cadre de l'hypothèse de référence, la distribution du coefficient de corrélation calculé à partir d'un échantillon a une moyenne nulle et peut être déduite d'une loi de Student (cf. Fiche 5).

Pas 5. Calcul d'une probabilité associée

A partir de la distribution de l'indicateur statistique, on calcule la probabilité que, dans le cadre de l'hypothèse de référence, le tirage au hasard conduise à une valeur de l'indicateur statistique au moins aussi éloignée de celle de l'hypothèse de référence que ne l'est la valeur effectivement observée.

Cas 1 : l'hypothèse de référence spécifie une différence nulle entre les deux moyennes ; la probabilité cherchée est celle d'obtenir une différence au moins aussi grande (en valeur absolue) que celle effectivement observée, alors qu'il n'y a pas de différence au niveau des populations.

Cas 2 : l'hypothèse de référence spécifie un coefficient de corrélation nul ; la probabilité cherchée est celle d'obtenir un coefficient de corrélation au moins aussi grand (en valeur absolue) que celui effectivement observé, alors que la corrélation est nulle au niveau de la population.

4.2.2 Utilisation de la probabilité associée lorsque les données dérivent effectivement d'un tirage au hasard

Dans le cas où les données sont issues d'un tirage au hasard, la probabilité associée à un indicateur statistique s'interprète comme la probabilité d'observer tel phénomène dans l'échantillon alors que dans la population ce phénomène n'existe pas (par exemple, probabilité d'observer dans l'échantillon au moins un point d'écart entre les notes en mathématiques au bac des filles et des garçons, alors qu'en fait, sur l'ensemble des populations, il n'y a pas de différence).

Cette probabilité constitue un lien entre l'échantillon et la population ; elle aide à décider ou non de généraliser à l'ensemble de la population ce que l'on observe sur l'échantillon.

Exemple (fictif) illustrant le principe de l'inférence classique dans un cas simple

Sur un échantillon d'élèves tirés au hasard, on observe pour la note en mathématiques au bac une différence de 1 point entre les filles et les garçons. Le calcul de la probabilité associée à cette différence donne : .00001 ; on effectue une sorte de raisonnement par l'absurde :

1) si, au niveau de la population totale, il n'existe aucune différence entre les notes en mathématiques des filles et des garçons, alors il est quasiment impossible d'observer une telle différence dans un échantillon tiré au hasard.
2) il est donc raisonnable de considérer que dans la population totale il existe une différence entre les notes en mathématiques des filles et des garçons.

Ce faisant, on a généralisé (ou extrapolé) les résultats observés sur l'échantillon à un ensemble plus vaste.

4.3 Seuils, erreurs et risques

Définitions et problèmes

La démarche de l'inférence statistique classique consiste à généraliser un résultat observé lorsque sa probabilité associée est inférieure ou égale à un seuil fixé a priori, habituellement 5% (ce choix peut être considéré comme une norme). Un tel résultat est dit significatif, terme dont le sens en statistique ne correspond pas au sens usuel ; un résultat associé à une probabilité inférieure à 1% est quelquefois dit « hautement significatif ».

Cette valeur de 5% s'interprète comme la probabilité de commettre un certain type d'erreur : généraliser à tort un résultat. Cette erreur est appelée *erreur de première espèce* et la probabilité correspondante, notée toujours α, *risque de première espèce*.

Il existe un autre type d'erreur, dite *erreur de seconde espèce* : ne pas généraliser à tort un résultat. La probabilité associée, notée toujours β, est dite *risque de seconde espèce*.

L'existence de ces deux types d'erreur met en évidence les risques inhérents à toute décision dans l'incertain ; le caractère incertain provient du fait que l'on ne connaît qu'une partie (l'échantillon) des données sur lesquelles on conclut (la population). Or ces risques de première et de seconde espèce sont antagonistes : diminuer l'un engendre automatiquement une augmentation de l'autre. Ceci peut être illustré par le cas limite où l'on annule l'un des deux risques : si, quelles que soient les données, on décide de ne pas généraliser, le risque de première espèce est nul (on ne généralisera jamais à tort) mais le risque de seconde espèce est maximum (même s'il faut le faire, on ne généralise pas). Remarquons au passage que ce type de règle de décision, qui consiste à ne pas tenir compte des données, n'est pas si rare qu'il le paraît.

Attention ! Observer un résultat significatif ne prouve pas que le résultat est vrai au niveau de la population ; ce n'est qu'un élément qui incite à considérer comme raisonnable l'attitude qui consiste à faire comme si c'était vrai.

En pratique

La démarche inférentielle classique est la plus répandue dans les ouvrages de statistique. A tel point que la plupart des utilisateurs la considère comme un "théorème" auquel on ne peut que se plier, alors qu'il ne s'agit que d'une formalisation du problème de décision dans l'incertain. Etre la plus répandue dans les ouvrages n'implique absolument pas qu'elle soit la seule possible, ni qu'elle soit adaptée à la plupart des situations concrètes, ni même enfin qu'elle soit la plus utilisée en pratique.

Deux points importants ne sont pas pris en compte dans l'inférence classique :

- les idées a priori (on dit aussi *degré de croyance*) ; or, le simple bon sens implique de ne pas avoir la même attitude dans l'examen d'un résultat lorsque l'on vérifie dans un cas particulier un résultat largement acquis par ailleurs ou lorsque l'on explore un domaine nouveau ;
- la possibilité de différer la décision en attendant de disposer d'informations ou de données supplémentaires ; les situations où l'on est acculé à prendre une décision sur la vue d'un seul résultat sont en fait assez rares.

En pratique, confronté à des probabilités voisines du seuil qu'il s'est pourtant lui-même choisi, l'utilisateur généralement réserve sa décision. Ainsi, sa démarche s'apparente à la démarche classique, mais en est bien distincte puisque la probabilité associée ne s'interprète plus comme une probabilité d'erreur.

D'un point de vue pratique, la démarche classique est et doit être plutôt considérée comme une aide à la décision que comme une règle de décision (même si cela est rarement dit explicitement).

4.4 A quelle population généraliser les résultats observés ?

En toute rigueur, l'inférence classique ne permet de généraliser les résultats observés qu'à la population dont les données ont été tirées au hasard explicitement. Or, l'utilisateur a très souvent besoin d'aller au-delà. Nous en donnons deux exemples.

Lorsque l'on réalise un sondage, on considère toujours que les résultats valent pour une période plus grande que celle de l'enquête proprement dite (cela est typique des enquêtes sur les intentions de vote). Or, en toute rigueur, la démarche classique ne considère que l'extrapolation au moment de l'enquête : toute autre extrapolation, indépendamment du fait qu'elle soit raisonnable ou non, relève d'une inférence non formalisée en dehors du champ de la démarche classique.

Lorsque l'on réalise une expérimentation, les individus (ou unités expérimentales) sont très rarement tirés au hasard : en sciences humaines on fait appel à des volontaires, en agronomie on utilise les parcelles expérimentales disponibles, etc. L'objectif est généralement l'étude de la réaction des individus en fonction d'un ensemble de conditions expérimentales : ainsi on étudie le rendement d'une culture en fonction de différentes pratiques culturales, etc. Dans ces cas, le tirage au hasard n'intervient que dans l'affectation des conditions expérimentales aux individus : la population à laquelle l'inférence classique permet d'accéder n'excède pas l'ensemble des unités expérimentales. Dans de tels cas, l'inférence statistique permet simplement d'émettre des propositions du type : si l'on avait affecté différemment les conditions expérimentales aux individus, on aurait très vraisemblablement observé les mêmes résultats. Autrement dit, elle ne permet pas de généraliser au-delà des unités expérimentales lorsque ces dernières ne sont pas tirées au hasard.

4.5 En conclusion

L'inférence est la généralisation d'un résultat observé sur un ensemble de données à un ensemble plus vaste. Il s'agit d'une problématique très générale. L'inférence statistique classique est l'une des formalisations de cette problématique générale qui n'est pleinement réaliste que dans des cas très particuliers.

L'inférence repose sur le fait que l'ensemble des données sur lequel on effectue les calculs (appelé en général l'échantillon) est représentatif de l'ensemble plus vaste auquel on généralise (appelé en général population). Dans l'inférence statistique, cette représentativité est assurée par le tirage au hasard. Si les données ne dérivent pas d'un tirage au hasard, il ne reste que l'inférence non formalisée (ou non statistique), dans laquelle la représentativité ne peut être que supposée, cette supposition pouvant toutefois être argumentée par la vérification que, pour certains critères a priori liés au problème étudié, l'échantillon ressemble à la population (par exemple la fixation de quotas dans les sondages).

Distribution d'une moyenne

1. Moyenne d'un échantillon

On s'intéresse au "comportement" de la moyenne d'un échantillon tiré au hasard dans une population. Pour cela, on étudie l'ensemble des valeurs possibles que peut prendre cette moyenne.

1.1 Illustration à partir d'un exemple

Population

Soit un ensemble de I individus appelé ici population. A chaque individu est associée la valeur d'une variable x. Ainsi à l'individu i on associe : $x_i=x(i)$. Etudier la population revient à étudier l'ensemble de I valeurs : $E_x=\{x_1, \ldots, x_i, \ldots, x_I\}$. Cet ensemble a pour moyenne $\bar{x}$ et pour écart-type s_x.

Ci-après nous étudions, pour une population de 4 individus, la variable x telle que :

$E_x=\{-2, -1, 1, 2\}$; $\bar{x}=0$; $s_x^2=2.5$; $s_x\approx 1.58$

*Ensemble des échantillons possibles (cf. **Tab. I**)*

On considère le tirage au hasard d'un échantillon de taille $n=2$. ou, ce qui revient au même, l'ensemble de tous les échantillons possibles de taille 2 que l'on peut extraire de E_x.

L'échantillon numéro j est à la fois :

- le couple d'individus $\{j_1, j_2\}$ en appelant j_1 et j_2 les numéros des individus extraits ; par exemple on a extrait les individus 1 et 3 ;
- le couple des valeurs de x prises par j_1 et j_2 ; soit : $\{x(j_1), x(j_2)\}$; par exemple, on a extrait les valeurs -2 et 1.

j_1, j_2	1, 2	1, 3	1, 4	2, 3	2, 4	3, 4
$x(j_1), x(j_2)$	-2, -1	-2, 1	-2, 2	-1, 1	-1, 2	1, 2
y_j	-1.5	-.5	0	0	.5	1.5

Tableau 1. *Ensemble des échantillons de taille 2 qu'il est possible d'extraire d'une population de 4 individus.*
y_j : moyenne de l'échantillon j

*Ensemble des moyennes possibles (cf. **Tab. 1**)*

A chaque échantillon j, on associe la moyenne y_j de ses valeurs, soit : $y_j=(1/2)[x(j_1)+x(j_2)]$

*Distribution des moyennes (cf. **Fig. 1**)*

On considère indifféremment :

- l'ensemble E_y des 6 valeurs de y_j ;
- l'ensemble des valeurs distinctes affectées de leur fréquence (distribution des moyennes ou encore de la variable y).

Ces ensembles ont pour moyenne et variance :

$\bar{y}=0$; $s_y^2=5/6=.833$

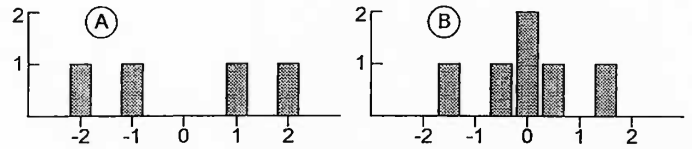


Figure 1. *Distribution d'une population de 4 individus (A) et distribution des moyennes des échantillons de taille 2 (B).*

Comparaison entre les distributions de x et de y

Elle tient en trois points principaux :

- les deux distributions ont la même moyenne ;
- la dispersion de y est plus petite que celle de x ;
- la distribution de y est plus dense autour de la moyenne qu'ailleurs.

1.2 Résultats généraux

Les propriétés mises en évidence dans ce petit exemple sont générales. En effet, lorsque l'on compare la distribution des moyennes des échantillons possibles et celle de la population dont ils sont issus, on peut montrer les trois résultats suivants.

Moyennes des deux distributions

La distribution des moyennes d'échantillons et celle de la population parente ont la même moyenne.

Variances des deux distributions

La variance de la distribution des moyennes d'échantillons est toujours plus petite que celle de la distribution de la population parente. La variance des moyennes se déduit de la variance de la population parente par la relation :

$$s_y^2=\left(\frac{I-n}{I-1}\right)\frac{1}{n}s_x^2$$

Dès l'instant que la taille de l'échantillon (n) est très petite devant celle de la population (I), le rapport $(I-n)/(I-1)$ est voisin de 1 ; dans ce cas, s_y^2 est voisin de s_x^2/n, valeur exacte obtenue lorsque l'on effectue un tirage avec remise (i.e. qui autorise qu'un même individu soit tiré plusieurs fois et donc apparaisse plusieurs fois dans un même échantillon).

La variance des moyennes est donc :

- d'autant plus grande que la population possède une grande variance ;
- d'autant plus petite que la taille de l'échantillon est grande ;
- d'autant plus petite que la différence entre la taille de l'échantillon et celle de la population est petite.

Remarque. Cette relation illustre l'intérêt d'augmenter la taille de l'échantillon dans un sondage : une telle augmentation diminue la variabilité des moyennes qu'il est possible d'obtenir et « assure » que la moyenne de l'échantillon n'est pas trop éloignée de celle de la population.

Forme des deux distributions

Sauf lorsque $n = I$, la distribution des moyennes d'échantillons ressemble toujours plus à une distribution normale que celle de la population. Cette ressemblance avec la loi normale augmente avec n (à condition que $n < I/2$).

2. Moyenne de deux ou de plusieurs variables

Cette situation est très différente de la précédente.

2.1 Exemples

Exemple extrait du fichier des notes

On a calculé, pour chaque matière, la moyenne annuelle (moyenne équipondérée des trois notes trimestrielles). Les écarts-types de notes sont regroupés dans le tableau 2.

	1er tr.	2ème tr.	3ème tr.	moy. an.
mathématiques	2.85	2.77	2.98	2.63
physique	2.97	2.79	3.16	2.70
sciences nat.	2.41	2.43	2.89	2.16
histoire-géo.	2.13	2.17	2.39	1.93
philosophie	2.00	2.12	2.33	1.88

Tableau 2. *Ecarts-types des notes trimestrielles et des moyennes entre les trois trimestres.*

Dans cet exemple où les écarts-types des notes dont on calcule la moyenne sont proches, l'écart-type de la moyenne est toujours plus petit que celui des notes dont on calcule la moyenne. Les paragraphes suivants font apparaître le caractère systématique de ce type de résultat dès lors que l'on calcule des moyennes entre plusieurs variables.

Exemple de données choisies

Nous détaillons quelques calculs de moyennes sur un petit jeu de données qui met en évidence différents cas de figure (cf. **Tab. 3**). La moyenne est calculée entre une variable A

et successivement trois autres variables qui sont, soit liée linéairement et positivement à A ($r(A,B)=1$), soit liée linéairement et négativement à A ($r(A,C)=-1$), soit indépendante linéairement de A ($r(A,D)=0$)

	A	B	C	D	$(A+B)/2$	$(A+C)/2$	$(A+D)/2$
1	6	8	10	10	7	8	8
2	8	10	8	14	9	8	11
3	10	12	6	8	11	8	9
4	12	14	4	12	13	8	12
moyenne	9	11	7	11	10	8	10
ec.-type	5	5	5	5	5	0	2.5

Tableau 3. *Exemples de calculs de moyennes sur des données choisies : 4 variables A, B, C et D mesurées sur 4 individus 1, 2, 3 et 4.*

Bien que calculées à partir de notes ayant chacune le même écart-type, les trois moyennes ont des écarts-types très différents. Ceci est dû au fait que l'écart-type de la moyenne dépend de la liaison entre les variables dont on calcule la moyenne.

2.2 Résultats généraux

Soit la moyenne entre les variables X et Y. On démontre les relations suivantes, en notant s_x (resp. s_y) l'écart-type de X (resp. Y), $s(X+Y)$ l'écart-type de $X+Y$ et $r(X,Y)$ le coefficient de corrélation entre X et Y :

pour la somme : $s^2(X+Y)=s_x^2+s_y^2+2s_xs_yr(X,Y)$

pour la moyenne : $s^2[(X+Y)/2]=(1/4)[s_x^2+s_y^2+2s_xs_yr(X,Y)]$

L'écart-type de la moyenne $(X+Y)/2$ est d'autant plus grand que les écarts-types s_x et s_y des variables dont on fait la moyenne sont grands.

L'écart-type de la moyenne $(X+Y)/2$ est toujours inférieur ou égal à la moyenne des écarts-types s_x et s_y des variables dont on fait la moyenne :

- il est égal à cette moyenne si $r(X,Y)=1$;
- si $r(X,Y)>0$, il est d'autant plus grand que les variables sont liées linéairement ;
- si $r(X,Y)<0$, il est d'autant plus petit que les variables sont liées linéairement.

Remarque. On déduit immédiatement des relations précédentes que la variance d'une somme de variables non corrélées est égale à la somme des variances des variables sommées. Ce résultat, qui formalise un aspect de l'addition des erreurs de mesure, est un des arguments en faveur de la variance pour appréhender la variabilité.

Cas de l'exemple de notes

Dans chaque matière, les notes trimestrielles sont corrélées fortement et positivement (cf. Tab. 9.1). Aussi les moyennes annuelles ont-elles un écart-type seulement légèrement inférieur à la moyenne des écarts-types des notes trimestrielles.

Index

(§ 14.3 : paragraphe 3 du chapitre 14 ; Tab. F4.2 : tableau 2 de la fiche 4)

T

V

Bibliographie

Pour le lecteur désireux d'approfondir et/ou de s'ouvrir à d'autres méthodes, nous donnons une liste d'ouvrages en langue française orientés vers les applications, centrés sur l'approche classique de données présentant une ou deux variables ([3], [4], [5], [7], [13], [16]), centrés sur l'analyse des données multidimensionnelles ([1], [2], [8], [9], [11], [12]) ou généraux ([6], [14], [15]).

[1] **Benzécri J.-P. et coll.** (1973) - *L'analyse des données*. Tome 1 : *la taxinomie*. 615p. Tome 2 : *correspondances*. 619p. Dunod, Paris.

[2] **Benzécri J.-P., Benzécri F.** (1980) - *Analyse des correspondances : exposé élémentaire*. 424p. Dunod, Paris.

[3] **Ceresta** (1995) - *Aide-mémoire statistique*. 285p. Cisia, Saint-Mandé, France.

[4] **Dagnélie P.** (1992) - *Statistique théorique et appliquée*. Tome 1 : 492p. Presses agronomiques, Gembloux, Belgique.

[5] **Dagnélie P.** (1986) - *Théorie et méthodes statistiques*. Vol. 2 461p. Presses agronomiques, Gembloux, Belgique.

[6] **Dodge Y.** (1993) - *Statistique : dictionnaire encyclopédique*. 409p. Dunod, Paris.

[7] **Droesbeke J.-J.** (1992) - *Eléments de statistique*. 527p. Ellipses, Paris.

[8] **Escofier B., Pagès J.** (1993) - *Analyses factorielles simples et multiples : objectifs, méthodes, interprétation*. 274p. Dunod, Paris.

[9] **Fénelon J.-P.** (1981) - *Qu'est-ce que l'analyse des données ?* 311p. Lefonen, Paris.

[10] **Grangé D. Lebart L.** (1993) - *Traitements statistiques des enquêtes*. 255p. Dunod, Paris.

[11] **Jambu M.** (1989) - *Exploration informatique et statistique des données*. 505p. Dunod, Paris.

[12] **Lebart L., Morineau A., Piron M.** (1995) - *Statistique exploratoire multidimensionnelle*. 439p. Dunod, Paris.

[13] **Rouanet H., Bernard J.-M., Le Roux B.** (1990) - *Statistique en sciences humaines : analyse inductive des données*. 190p. Dunod, Paris.

[14] **Saporta G.** (1990) - *Probabilités, analyse des données et statistiques*. 493p. Technip, Paris.

[15] **Tenenhaus M.** (1994) - *Méthodes statistiques en gestion*. 373p. Dunod, Paris.

[16] **Wonnacott T., Wonnacott R.** (1991) - *Statistique*. 920p. Economica, Paris.

Logiciels. La plupart des graphiques présentés dans cet ouvrage ont été obtenus à l'aide d'un tableur de grande diffusion. Les traitements statistiques, en particulier les tableaux triés de valeurs-tests, ont été obtenus à partir du logiciel SPADN diffusé par CISIA 1, av. Herbillon F-94160 Saint-Mandé. Beaucoup de probabilités ont été obtenues à l'aide du logiciel *LeProbabiliste*, également diffusé par le CISIA.

DIDACT STATISTIQUE

Initiation aux Traitements statistiques

Méthodes, méthodologie

Prenons un fichier de donnés, posons-nous quelques questions et introduisons graduellement les outils statistiques nécessaires pour y répondre.

À partir de cette idée simple, les auteurs proposent une présentation originale de la statistique, véritablement ancrée dans la pratique. S'appuyant sur leur riche expérience tant en formation initiale qu'en formation continue, ils ont bâti **le cours idéal de statistique pour non-mathématiciens**.

Ce manuel s'adresse à un large public.

• **L'étudiant** y trouvera l'approche intuitive mais rigoureuse des méthodes usuelles, absente des manuels courants dont le formalisme mathématique constitue souvent un obstacle à la lecture.

• **L'enseignant** y puisera bon nombre d'exemples et de présentations pédagogiques.

Tous les utilisateurs seront particulièrement intéressés par des problèmes méthodologiques importants en pratique (mais rarement abordés dans les livres) comme l'étude et la prise en compte de données manquantes et de données aberrantes.

BRIGITTE ESCOFIER est l'une des fondatrices de l'école française d'analyse des données. Docteur d'état ès sciences mathématiques, professeur à l'université de Rennes, elle a enseigné à l'Institut National des Sciences Appliquées (INSA) de Rennes et à l'IUT de Vannes. Brigitte Escofier s'est éteinte en juillet 1994.

JÉRÔME PAGÈS est ingénieur agronome, docteur en économie et habilité à diriger des recherches en statistique. Professeur de l'enseignement supérieur agronomique, il anime actuellement l'unité de statistique appliquée de l'Institut National Supérieur de Formation Agro-alimentaire (INSFA) situé sur le pôle agronomique de Rennes.

Les deux auteurs ont travaillé en équipe pendant douze années. On leur doit de nombreux travaux de recherches et un ouvrage de référence sur l'analyse factorielle tradut en plusieurs langues.

EN COUVERTURE :
CALLIGRAPHIE DE RICHARD DELÉCOLLE.

9 782868 472311

ISBN 2-86847-231-1
PRIX : 95 F

les PUR Presses Universitaires Rennes
Université de Rennes 2 Haute Bretagne