

Chapitre 7 : L'analyse des données

Il existe trop souvent un partage des tâches entre ceux qui mènent concrètement la collecte des données et ceux qui font l'analyse. Cet état des faits correspond à une difficulté réelle : sans supposer nécessairement des connaissances théoriques très approfondies — souvent, le bon sens suffit —, l'analyse suppose de pouvoir définir et interpréter les outils statistiques dont on a besoin. Il est souhaitable que le chercheur soit capable de manipuler certains outils informatiques (logiciels d'analyse statistique), ou qu'il soit, à défaut, en mesure de passer une commande précise à un statisticien informaticien. Ce chapitre aura atteint son objectif s'il convainc le lecteur non initié qu'il peut parvenir à des résultats concrets assez vite et que l'essentiel, ici comme souvent, est avant tout de se lancer.

L'analyse des données peut avoir des objectifs aussi divers que :

- Faire un bilan du niveau moyen des élèves, ce qui peut s'accompagner de comparaisons avec d'autres pays ou avec la situation des années précédentes. De tels bilans peuvent être un moyen de fortement mobiliser les acteurs du système scolaire ;
- Faire une analyse détaillée des acquis des élèves, par types de connaissances et de savoir-faire. Ce genre d'analyses peut permettre aussi bien de penser une refonte des programmes que de modifier les accents de la formation continue à donner aux maîtres en désignant les lacunes à combler chez les élèves.
- Faire un bilan des inégalités régionales ou sociales en termes d'acquis et de moyens scolaires. Cela peut déboucher sur la mise en place de politiques ciblées sur certaines catégories de la population scolaire.
- Évaluer l'efficacité des différentes conditions d'enseignement et des différentes pratiques pédagogiques. L'objectif est alors d'aider le décideur politique à choisir la meilleure combinaison de moyens pour rendre l'école aussi efficace que possible.

Une analyse complète des données aborde tous ces points – et d'autres encore peut-être.

Dans le cadre d'un guide pratique d'évaluation, une sélection est inévitable. En particulier, il ne sera que brièvement fait mention de l'analyse des résultats par type de savoir. La présentation retenue a trois soucis principaux que la typographie met en valeur :

1. Bien marquer les différentes **étapes nécessaires** d'une analyse, en mettant l'accent sur les enjeux et les difficultés (corps du texte) ;
2. Donner un **exemple concret suivi** (passages en italiques). *L'exemple choisi est celui de l'effet de la taille des classes en Côte d'Ivoire en CMI sur les progrès des élèves, tel que le PASEC permet de le mesurer pour l'année scolaire 1995-96.*
3. Donner quelques précisions techniques sous la forme d'encadrés simples qui permettent de manipuler les **outils statistiques usuels** en comprenant leur sens général.

Pour commencer, voici comment s'enchaînent les huit étapes de l'analyse que l'on peut recenser, avec en vis-à-vis les différents outils statistiques utilisables :

ETAPES	OUTILS
Préparation des données	
1. Construction de variables pour l'analyse	Qu'est-ce qu'une bonne variable statistique ?
2. Fusion des différents fichiers de données en un unique fichier d'analyse	
3. Vérification des données	Le résumé des données
Analyse descriptive	
4. Analyse des scores	Outils d'analyse d'une distribution
5. Analyse des variables explicatives	
6. Descriptions bivariées	Description statistique d'une corrélation
Analyse causale	
7. Sélections des variables du modèle explicatif	L'identification des effets causaux dans la régression multivariée
8. Lecture et amélioration du modèle statistique	La significativité en statistiques

A. La préparation des données

1. Etape 1 : la construction des variables :

Revenons un peu en arrière par rapport à la fin du chapitre 6, au moment où les données de niveau élève, maître et directeur sont encore rangées dans trois fichiers distincts (respectivement CI_EL.dta, CI_QM.dta et CI_QD.dta). C'est à l'intérieur de chacun de ces fichiers de données que s'effectue la construction des variables. Si nous la présentons dans ce chapitre, c'est que nous voulons insister sur son importance qui en fait, vraiment, la première étape dans l'analyse des données.

On peut légitimement être embarrassé, dans un premier temps, devant la masse considérable de données dont on dispose. Il est important de se ramener assez vite à un nombre raisonnable de variables, si on ne veut pas se disperser. Pourtant, le choix de ces variables n'est pas sans conséquence : il peut conduire à perdre une partie de l'information, au point de faire porter un regard biaisé sur les données. Or le risque existe que, faute de temps ou d'énergie, on ne revienne jamais aux données élémentaires¹. La sélection et la construction d'un petit nombre de variables que l'on va analyser est donc une étape relativement décisive.

La construction des variables s'inscrit logiquement dans la poursuite des objectifs de l'étude. En général, les tests et questionnaires ont été conçus avec une idée a priori des variables qu'on entendait en tirer.

- Parfois, la variable correspond directement à l'information recueillie.

On veut par exemple en Côte d'Ivoire évaluer les performances pédagogiques selon le genre du maître. On demande son sexe au maître dans le questionnaire maître, et on peut créer une variable MAITRFEM codée 1 quand le maître est une femme, 0 sinon.

¹ C'est la raison pour laquelle il est de bonne pratique, au sein de la communauté scientifique, de permettre à d'autres chercheurs d'accéder aux données élémentaires collectées, en amont de tout travail de construction de variables.

- D'autres fois, l'information ne peut être recueillie si simplement : les questions ne permettent que d'approcher le concept auquel on s'intéresse.

*En Côte d'Ivoire, on voulait introduire une variable qui donne le niveau de vie de la famille de l'élève. Demander à l'élève le revenu de ses parents n'aurait guère eu de sens. On a choisi de poser une série de questions sur les objets présents dans la maison de l'élève (télévision, réfrigérateur, lits,...). On va alors créer une variable qui approche le niveau de vie par le nombre d'objets présents à la maison : 0 si aucun objet n'est possédé par la famille, 1 s'il y en a un seul, 2 s'il y en a deux, 3 s'il y en a trois, etc. Dans le cas de la Côte d'Ivoire, on a ainsi créé la variable **RICHESSSE** qui donne le nombre d'objets possédés parmi 13 (fauteuil, réfrigérateur, robinet, lit, électricité, voiture, vélo, motocyclette, vidéo, télévision, radio, cuisinière à gaz, toilettes avec eau courante). On a aussi créé la variable **NIVEAUVI** qui donne le nombre d'objets parmi 3 seulement (réfrigérateur, vidéo, voiture). **Il y a une part d'arbitraire dans la création de telles variables.** Ce qui importe cependant, c'est qu'elles correspondent à un **concept clair** : ici, le niveau de vie est mesuré par les biens possédés ; tout ce qu'on demande, c'est que s'il existe de fortes différences de niveau de vie, on puisse le voir en constatant que **RICHESSSE** prend des valeurs élevées pour certaines familles, faibles pour d'autres. Si le niveau de vie intervient dans les résultats des élèves, la variable **NIVEAUVI** devrait permettre de capter cet effet : c'est cela qu'on veut. On peut vérifier que **NIVEAUVI** ou **RICHESSSE** mesurent bien approximativement la même chose en calculant la corrélation entre ces deux variables¹ : elle est ici de 0.77, ce qui est relativement élevé et indique que ces deux variables sont des mesures globalement concordantes du niveau de vie.*

Qu'est-ce qu'une 'bonne' variable explicative ?

Une variable est toujours construite en vue de rendre compte d'une réalité dont on a une idée a priori. En ce sens, il ne s'agit pas de d'introduire dans un modèle une variable par 'curiosité statistique', pour 'voir si ça marche'.

Bien entendu, la qualité d'une variable dépend de ce qu'on veut étudier. Pour nous, une variable est **pertinente** lorsque la réalité qu'elle représente

¹ Voir l'encadré technique sur la corrélation de deux variables (étape 6).

est liée, au moins potentiellement, au processus d'apprentissage scolaire, et elle est **opératoire** si elle permet concrètement de tester et de quantifier ce lien.

On est sûr qu'une variable est pertinente lorsqu'on est capable de détailler la façon dont elle peut jouer sur les apprentissages scolaires. Ce ne sont que des hypothèses, mais elles doivent être le plus précises possible. Par exemple, on suppose que les pères d'élèves plus éduqués ont les connaissances requises pour aider leurs enfants et que cela permet aux enfants de progresser plus vite. C'est parce que cette hypothèse est intéressante qu'il est légitime de chercher à introduire une variable de niveau d'éducation du père. Attention : le résultat peut confirmer ou non cette hypothèse, mais **l'absence d'effet est aussi un résultat intéressant en soi**.

Pour être opératoire, une variable doit permettre de tester et de quantifier la relation dont on a fait l'hypothèse. Il y a pour cela deux conditions. Il faut tout d'abord que la variable puisse être mesurée de façon fiable. Ensuite, lorsque la variable définit des catégories, il faut que chacune d'entre elles comporte assez d'individus pour qu'on puisse tirer des conclusions statistiques solides. *Par exemple, pas question de mesurer l'effet des classes de plus de 90 élèves s'il n'y en a qu'une dans l'échantillon. On pourrait créer la variable correspondante, on obtiendrait toujours un résultat, mais comment le généraliser à partir d'une classe ?*

Il peut y avoir tension entre ces deux objectifs : souvent, la variable la plus pertinente que l'on puisse imaginer n'est pas opératoire, soit parce qu'elle n'est pas mesurable, soit parce que sa mesure comporterait une trop grosse part d'erreur, soit parce que les données ne permettent pas de distinguer des catégories aussi fines. Il y a donc un **arbitrage** à faire dans certains cas entre **ce qu'on voudrait mesurer** et **ce qu'il est possible de mesurer** de façon fiable. Cet arbitrage se fait au cas par cas...

*Si on reprend l'exemple de la Côte d'Ivoire, une variable particulièrement pertinente est le **nombre d'années d'études effectuées par le père**. On peut le demander dans le questionnaire aux élèves. Mais est-ce opératoire ? Beaucoup d'élèves risquent de ne pas répondre, ou de répondre de façon très imprécise. On envisage alors une autre variable : **le père est-il***

ou non alphabétisé ? Cette variable est moins pertinente dans la mesure où deux pères alphabétisés mais ayant fait respectivement 5 et 12 années d'études n'ont certainement pas les mêmes compétences pour aider leur enfant. Mais elle est plus opératoire car plus facile à mesurer de façon fiable, dans la mesure où l'enfant peut dire si son père sait lire et écrire. Quelle variable est alors préférable ? C'est sans doute la variable d'alphabétisation du père. Elle constitue un bon compromis entre ce qu'on aimerait mesurer (la capacité du père à aider l'enfant dans son travail) et ce qu'on peut espérer mesurer de façon fiable.

2. Etape 2 : fusion des différents fichiers en un fichier d'analyse unique :

On dispose à ce stade de variables à trois niveaux : élève, classe et école. La plus grosse part des analyses sera menée au niveau des élèves puisque l'objectif central est d'expliquer la progression des élèves. La fusion des différents fichiers de variables permet de rassembler toute l'information dans un tableau de données où à chaque élève (en lignes) est associé une série de variables (en colonnes) qui donnent ses caractéristiques individuelles, mais aussi celles de sa classe et de son école.

Le logiciel d'analyse utilisé comporte en général une commande pour exécuter cette fusion. En particulier, le chapitre 6 a présenté la procédure utilisée dans STATA. Celle-ci suppose une bonne identification de chaque élève, au moyen de trois variables qui indiquent son école, sa classe et son numéro d'identification au sein de la classe. Il est conseillé de bien vérifier le fichier d'analyse obtenu. C'est ce que permet en particulier l'étape 3 :

3. Etape 3 : vérification de la cohérence des données :

Cette étape n'est ni longue ni difficile ; mais l'oublier peut s'avérer particulièrement coûteux ! Il s'agit, sur la trentaine de variables qu'on a construites, de s'assurer qu'il n'y a pas d'incohérences, d'observations perdues, d'erreurs de codage,...

Pour avoir un coup d'œil global sur les données, on utilise le résumé des données :

Le résumé des données

Le résumé des données comprend en général les informations suivantes : nombre d'observations disponibles, valeurs moyenne, minimale et maximale prises par la variable.

Soit l'exemple du résumé d'une partie des données en Côte d'Ivoire au CM1, tel que le logiciel STATA permet de l'obtenir :

Variable	Obs	Mean	Std. Dev.	Min	Max
NUMECOLE	2295	59.86187	34.58081	1	120
NUMCLASS	2295	5	0	5	5
NUMELEVE	2295	10.22571	5.714933	1	20
(...)					
MAITRFEM	2266	.0264784	.1605885	0	1
AGEMAITR	2266	37.03266	5.30357	23	52

Lecture : pour chaque variable, la première colonne donne le nombre d'élèves pour lesquels l'information est disponible (Obs) ; la seconde donne la valeur moyenne que prend la variable (Mean) ; la troisième donne l'écart type¹ (Std. Dev.), la quatrième et la cinquième donnent les valeurs minimale et maximale prises dans l'échantillon (Min et Max). On lit par exemple que 2,6% des élèves ont une femme pour maître (variable MAITRFEM), et que l'âge moyen de leurs maîtres est de 37 ans, avec un minimum de 23 et un maximum de 52.

Même si cela ne suffit pas toujours, un simple coup d'œil sur ces grandeurs permet souvent de dépister les erreurs les plus grossières : par exemple, si l'âge minimal des maîtres que l'on obtient est de 9 ans, cela indique vraisemblablement une faute de saisie. Ou encore, si la moyenne de la variable RURAL est de 0.30, ce qui voudrait dire que 30% des élèves sont dans des écoles en milieu rural, alors que l'échantillon est à peu près représentatif d'un pays à 75% rural, on doit se demander si on n'a pas interverti les caractéristiques 'rural' et 'urbain'.

Des vérifications plus poussées sont possibles et souvent souhaitables : par exemple, il faut s'interroger si on découvre que tous les élèves d'une

¹ Voir l'encadré technique sur les outils d'analyse d'une distribution.

classe ont un score nul au test (ont-ils vraiment été testés ? était-ce un problème de langue ?). C'est en regardant les données par classe, par administrateur de tests ou de saisie, et en faisant de fréquents aller et retour des documents papier aux fichiers informatiques qu'on peut rectifier des erreurs et, parfois, prendre conscience des limites des données.

Le résumé des données permet aussi de voir le nombre d'**observations manquantes** pour chaque variable. Il faut s'interroger sur ces observations manquantes : elles posent deux problèmes :

- **Biais de sélection** : En général, les non-réponses ne sont pas dues au hasard. Correspondent-elles à une partie spécifique de l'échantillon ? (par exemple, manque-t-il des questionnaires maîtres et directeurs dans les écoles plus reculées ? les maîtres qui ne répondent pas aux questions sont-ils ceux qui ont le plus de critiques mais qui n'osent les formuler ?) Cela risque-t-il de biaiser notre perception de la réalité représentée par cette variable ?
- **Réduction de la taille de l'échantillon** : Lorsqu'on veut utiliser conjointement plusieurs variables pour lesquelles il y a des valeurs manquantes¹, on est obligé de laisser de côté toute observation pour laquelle l'une au moins des variables est manquante. Même si un faible pourcentage d'observations manque pour chaque variable, cela peut conduire au total à abandonner un grand nombre d'observations. Il peut être alors souhaitable d'**imputer les valeurs manquantes**. Les logiciels statistiques peuvent présenter une commande pour cela. Elle consiste à attribuer une valeur probable pour la variable qui manque, en fonction des valeurs prises par les autres variables. Par exemple, si on ignore le genre d'un maître mais qu'on sait que ce maître est âgé de plus de cinquante ans et que la plupart des maîtres âgés de plus de cinquante ans dans l'échantillon sont des hommes, et si d'autres variables conduisent de la même façon à penser qu'il s'agit d'un homme, l'imputation va considérer que le maître en question est un homme. Attention : il convient d'utiliser ces fonctions d'imputation avec prudence. En particulier, on évitera systématiquement d'utiliser le score des élèves dans les imputations. On risquerait en effet de créer artificiellement de fausses relations entre la variable qu'on veut expliquer et les variables explicatives, et de fausser ainsi l'analyse causale ultérieure.

¹ C'est en particulier le cas dans la régression multiple, que nous examinons plus bas.

B. L'analyse descriptive des données

La description des données constitue une étape à part entière de l'analyse. C'est le moment de faire un diagnostic d'ensemble du système éducatif tel qu'il est, avant d'envisager, avec l'analyse causale, les moyens de l'améliorer. L'analyse descriptive peut procéder en trois étapes qui correspondent à autant de questions :

- quels sont les résultats obtenus par les élèves (analyse des scores) ?
- quels sont les conditions d'apprentissage et les moyens mis en œuvre dans les écoles (analyse des variables contextuelles et politiques) ?
- comment résultats et moyens se répartissent-ils sur la population scolaire (analyse bivariée) ?

1. Etape 4 : L'analyse des scores des élèves :

Les erreurs d'analyse à éviter :

Lorsqu'on dispose de données de tests, le réflexe spontané serait de procéder comme pour l'analyse des résultats d'un examen scolaire ordinaire : regarder le pourcentage d'élèves qui ont la moyenne, donner le pourcentage de bonnes ou de très mauvaises notes,...

Mais le test n'a pas été conçu comme un examen qui est 'réussi' lorsque l'élève a la moyenne. Il a surtout été construit de façon à différencier les élèves, à mettre en lumière le mieux possible la diversité des apprentissages réalisés. Un bon test comprend donc à la fois des items 'trop faciles' et des items 'trop difficiles' qui permettent l'obtention de ce dégradé de résultats nécessaire à l'analyse causale. **Ce qu'il faut donc garder à l'esprit, c'est que le test est une construction, et que les résultats au test dépendent à la fois des connaissances des élèves et de la construction du test.**

Au Cameroun en CM1, le score moyen en mathématiques est de 44/100 en début d'année, et de 41/100 en fin d'année. Si on ne prend pas garde au fait que les tests étaient différents et qu'il est tout à fait possible que le test de fin d'année soit de difficulté sensiblement supérieure, on risque de commenter hâtivement que les élèves ont régressé en mathématiques. Ce n'est manifestement pas le bon usage à faire de ces données.

Les analyses possibles : approche critériée et analyse comparative :

- Même si cela ne peut se faire naïvement par simple référence à la moyenne (50/100), il demeure intéressant d'évaluer la 'réussite' à un test par rapport à certains objectifs pédagogiques. Cela peut se faire moyennant certaines précautions. Le principe est de réunir à l'avance, au moment de l'élaboration des tests, des critères de réussite par domaine. On pourra ainsi évaluer pour chaque domaine le pourcentage des enfants dont on juge qu'ils sont parvenus à une maîtrise satisfaisante.

Soit au CM1 le domaine du calcul, et l'objectif de maîtrise de la multiplication. Le test de début d'année propose trois exercices comportant respectivement quatre items, 1 item et 1 item qui portent sur la multiplication. Au moment de l'élaboration des tests, on s'est mis d'accord sur un critère à partir duquel on estime qu'un élève maîtrise la multiplication : par exemple, à partir de quatre réponses exactes sur six. Une fois que le test a été administré, il est possible de calculer le pourcentage d'élèves qui a atteint l'objectif.

- Outre cette approche critériée, les résultats aux tests sont particulièrement intéressants à regarder dans une **perspective comparative** :

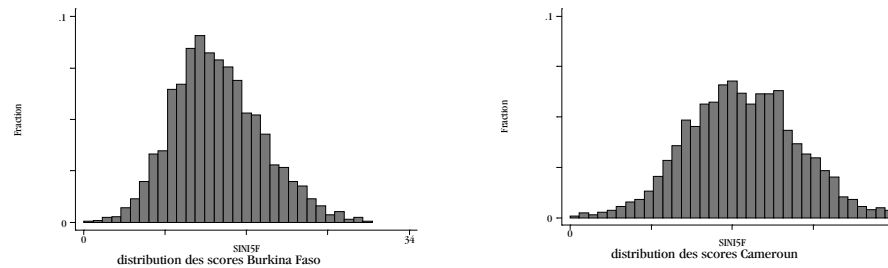
La comparaison peut se faire dans le temps ou dans l'espace. Dans le temps, on comparera les résultats au même test¹ deux années différentes (ce qui suppose la mise en place d'un suivi périodique des apprentissages). Dans l'espace, la comparaison est possible si tout ou partie d'un même test a été dispensé dans deux pays différents. Mais il est également possible de comparer les résultats dans deux régions différentes, entre filles et garçons, etc².

Que peut-on dire des résultats des élèves camerounais en début de CM1 en français, par comparaison avec ceux obtenus par les élèves du Burkina Faso ? La première chose est de regarder la distribution des scores des élèves dans ces deux pays :

¹ Il n'est pas nécessaire que les tests soient identiques dans leur entier ; ils peuvent ne comprendre qu'un certain nombre d'items communs, dits items d'ancrage. La comparaison porte alors sur les résultats à cette partie de test.

² De telles comparaisons sont déjà des analyses bivariées ; la procédure est détaillée ci-dessous à l'étape 6.

Distributions comparées des scores au Burkina Faso et au Cameroun (début de CM1)

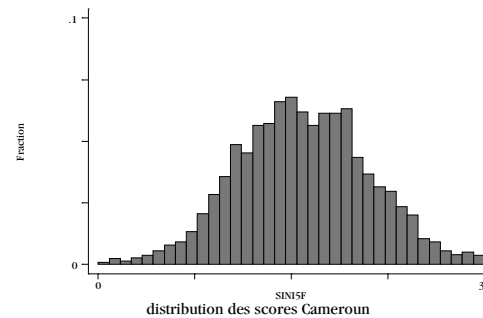


On voit facilement un certain nombre d'éléments intéressants : en moyenne, les élèves au Cameroun sont plus nombreux dans la partie droite du graphique, autrement dit, le test est en moyenne mieux réussi. Mais les scores au Burkina Faso sont davantage concentrés, ce qui indique un niveau plus homogène.

Comment quantifier ces observations ? L'encadré technique sur l'analyse d'une distribution propose quelques outils statistiques descriptifs souvent utilisés.

Outils d'analyse d'une distribution :

Comme souvent en statistique, il est bon de commencer par une représentation graphique. La représentation habituelle de la distribution est donnée par un histogramme comme celui-ci :



En abscisses sont portées les valeurs prises par la variable (par exemple, les différents scores obtenus par les élèves) ; la fréquence de ces valeurs (le pourcentage d'élèves obtenant ces scores) est représentée par la hauteur des rectangles.

En elle même, cette représentation comporte toute l'information sur la distribution. Certaines statistiques usuelles servent ensuite à décrire l'un ou l'autre aspect de la distribution :

- la **moyenne** ;
- la **médiane** (valeur autour de laquelle se répartissent les deux moitiés de la classe, dans le cas de la distribution des scores au sein d'une classe) et tous les indicateurs semblables : le **1^{er} quartile**, en deçà duquel se situent les 25% d'élèves les plus faibles, le **1^{er} décile**, en deçà duquel se situent les 10% d'élèves les plus faibles, le **9^{ème} décile**, au delà duquel se situent les 10% des scores les plus élevés,...
- **l'écart type** : c'est une mesure de la **dispersion des scores**. Plus l'écart type est élevé, plus les scores prennent des valeurs extrêmes autour de la moyenne. En fait, l'écart type est une sorte de distance moyenne de tous les individus à la valeur moyenne de la distribution. Par exemple, si on n'a que 4 élèves dans une classe, avec pour scores 0, 0, 2 et 2, la moyenne est de 1 et chaque individu se situe à 1 point de cette moyenne : l'écart type est de 1. Mais si les scores sont 1, 1, 1 et 1, la moyenne reste 1 mais la dispersion est nulle : l'écart type vaut 0.

L'écart type permet de formaliser notre comparaisons des scores au Burkina Faso et au Cameroun : au Burkina Faso, l'écart type est de 4,8 ; au Cameroun, il est de 5,9 points. On mesure ainsi que, pour cet indicateur, la dispersion au Cameroun dépasse de 23% (c'est-à-dire $(5,9-4,8)/4,8$) celle observée au Burkina Faso.

¹ La formule exacte de l'écart type est la suivante : $\sigma = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}}$ où le score de l'élève

i est noté y_i , la moyenne des scores est notée \bar{y} et n est le nombre d'élèves de l'échantillon.

2. Etape 5 : L'analyse des variables explicatives :

L'analyse descriptive des variables explicatives va nous permettre de décrire les moyens mis en œuvre et les marges de manœuvre qui existent.

On peut distinguer trois types de variables explicatives :

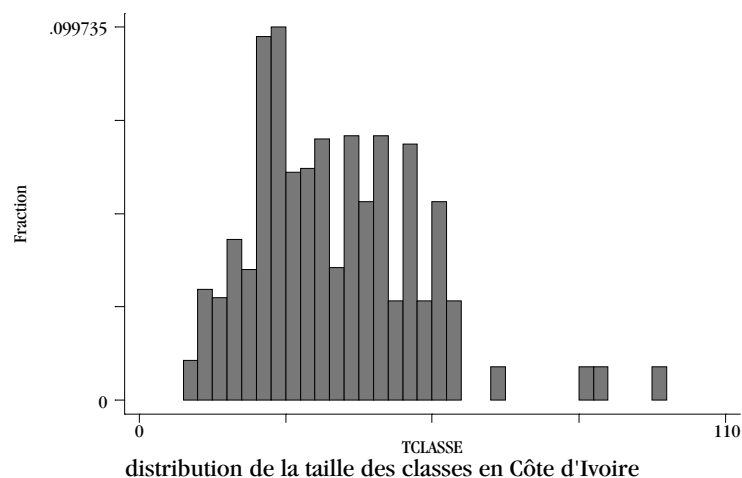
1. Les **variables quantitatives** (la valeur prise mesure une quantité : nombre d'années d'expérience du maître, nombre d'enfants dans la classe) ;
2. Les **variables qualitatives dichotomiques**, qui peuvent prendre 0 ou 1 comme valeur. Par exemple : le maître est-il une femme (la variable MAITRFEM prend la valeur 1) ou un homme (la variable MAITRFEM prend la valeur 0) ? ;
3. Les **variables qualitatives catégorielles**, par exemple lorsqu'on code par 1, 2, 3 ou 4 l'appartenance à 4 catégories (catégories d'âge, type de formation,...)

Les outils descriptifs utilisés varient selon que la variable est qualitative ou quantitative :

Type de variable	Variables quantitatives	Variables qualitatives
Exemples	<ul style="list-style-type: none"> - Scores des élèves - Expérience des maîtres (en années) - Taille des classes (nombre d'élèves) 	<ul style="list-style-type: none"> - Genre du maître - Lieu de la formation initiale
Outils descriptifs	<ul style="list-style-type: none"> - Représentation graphique de la distribution - Moyenne - Médiane, déciles et quartiles - Ecart type 	<ul style="list-style-type: none"> - Représentation graphique de la distribution - Tableau des effectifs des différentes catégories

Pour bien illustrer les différences entre les différents types de variables, prenons l'exemple de la taille des classes en Côte d'Ivoire en CM1.

- **La taille des classes comme variable quantitative** : Dans un premier temps, une variable quantitative discrète a été construite, faisant correspondre à chaque classe le nombre d'élèves dans cette classe. La distribution des tailles de classe est donnée dans le graphique de la page suivante.



- **La taille des classes comme variable dichotomique :** La partition la plus simple consiste à prendre d'un côté les classes pour lesquelles on a moins de 35 élèves, de l'autre celles de plus de 35 élèves. On crée ainsi la variable T35PL qui prend la valeur 1 pour une classe de plus de 35 élèves, et 0 pour une classe de 35 élèves ou moins.

La description univariée de cette variable est très simple :

	Nombre d'élèves	Pourcentages
Classes de moins de 35 élèves	1131	49,28%
Classes de plus de 35 élèves	1164	50,72%
Total (toutes classes confondues)	2295	100%

(Remarque : Comme la variable prend des valeurs 0 et 1 seulement, le pourcentage d'élèves étudiant dans des classes de plus de 35 élèves est aussi la valeur moyenne prise par la variable T35PL, soit 0,5072).

- **La taille des classes comme variable catégorielle :** Il peut cependant paraître utile de distinguer plus de deux catégories. On choisit d'en définir cinq : classes de moins de 25 élèves (CATEG=1), de 26 à 35 (CATEG=2), de 36 à 45 (CATEG=3), de 46 à 55 (CATEG=4) et de plus de 55 élèves

(CATEG=5). Les effectifs de ces différentes catégories sont donnés dans le tableau suivant :

Catégories (selon le nombre d'élèves dans la classe) :	Nombre d'élèves dans les classes considérées	Pourcentages
1. Moins de 25	675	29,4%
2. de 26 à 35	536	23,4%
3. de 36 à 45	460	20,0%
4. de 46 à 55	375	16,3%
5. 56 et plus	249	10,9%
Toutes classes confondues	2295	100%

3. Etape 6 : L'analyse descriptive bivariée :

Après avoir étudié une par une les différentes variables que nous avons créées, il convient de les étudier conjointement. Nombre de questions qu'on se pose fréquemment font intervenir des descriptions bivariées. Par exemple : les écoles rurales sont-elles aussi bien dotées en livres que les écoles urbaines ? Les filles ont-elles le même niveau que les garçons ? Les élèves progressent-ils plus vite dans des classes moins nombreuses ? Mais arrêtons-nous justement sur cette dernière question pour bien mesurer son ambiguïté :

La description bivariée ne prouve jamais une relation causale :

Le langage présente parfois des pièges redoutables. Soit la question : *les élèves progressent-ils plus vite dans les classes moins nombreuses ?* Cette phrase peut signifier deux choses sensiblement différentes :

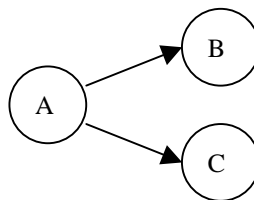
1. *des classes moins nombreuses font-elles progresser plus vite les élèves ?*
2. *certaines élèves sont dans des classes peu nombreuses. Se trouve-t-il que ce sont aussi des élèves qui progressent plus vite ?*

La seconde formulation reste complètement silencieuse sur les mécanismes causaux qui sont à l'œuvre : si on trouve que les élèves des classes moins nombreuses progressent plus vite, ce peut être pour une multitude de raisons qui n'ont rien à voir avec la taille de la classe.

La première formulation, elle, est clairement causale : elle dit que c'est parce qu'ils sont dans des classes moins nombreuses que ces élèves progressent plus vite.

La description bivariée et la mise en évidence d'une corrélation ne suffisent jamais pour montrer qu'il y a une relation causale entre deux variables. En effet, la corrélation entre deux variables, aussi intense soit-elle, peut toujours être due à une troisième variable cachée.

On rencontre souvent la configuration du graphique ci-dessous : A cause B et A cause C, et du coup B et C sont corrélés sans qu'il n'y ait de relation causale entre eux. L'erreur à ne pas commettre est donc d'étudier B et C isolément (sans tenir compte de A) et de conclure à une relation causale de B vers C (ou de C vers B).



En Côte d'Ivoire en CM1, la corrélation entre la progression des élèves et la taille de la classe est positive. En particulier, les élèves des classes de plus de 35 élèves obtiennent de meilleurs résultats que ceux des classes de moins de 35 élèves. Mais le rôle de la variable cachée A est peut-être joué par le milieu géographique : on constate en effet que 78% des classes de moins de 35 élèves sont en milieu rural, contre 25% des classes plus nombreuses. Si certaines classes sont moins nombreuses, n'est-ce pas parce qu'elles sont en milieu rural? Et si les résultats y sont moins bons, n'est-ce pas dû aussi au milieu rural? N'est-ce pas cela, en somme, qui explique que les classes moins nombreuses présentent de moins bons résultats?

On ne pourra répondre à ces questions tant qu'on en restera à l'analyse bivariée. L'analyse bivariée, à la différence de la régression multivariée, ne permet pas de répondre à cette objection de la 'variable cachée'. C'est pour cela qu'elle ne prouve jamais l'existence d'une relation causale.

Quel est alors le bon usage de la description bivariée ?

La description bivariée permet d'identifier des publics cibles :

La description bivariée est complémentaire de l'analyse causale. Connaître les mécanismes des apprentissages plus ou moins efficaces (grâce à l'analyse

causale) donne des leviers d'action pour agir sur certains publics cibles défavorisés (identifiés par l'analyse descriptive).

En particulier, des analyses bivariées des résultats des élèves permettent de mettre en évidence les inégalités d'apprentissage entre régions, entre garçons et filles, entre catégories sociales,... **C'est donc l'analyse bivariée qui permet d'identifier les groupes qui ont besoin d'une attention spécifique parce qu'ils cumulent des handicaps que l'analyse causale considère, elle, isolément.**

Les outils des analyses bivariées : la mise évidence d'une corrélation :

Jusqu'ici, on a utilisé le terme de corrélation sans en préciser le contenu. Pour montrer que la taille de la classe était positivement corrélée aux progrès des élèves, on s'est contenté de comparer les progrès moyens dans des classes plus de 35 et de moins de 35 élèves. L'encadré qui suit présente de façon plus systématique les outils qui permettent la mise en évidence de la corrélation de deux variables.

Description statistique d'une corrélation :

- L'approche la relation entre deux variables diffère selon qu'il s'agit de variables quantitatives ou qualitatives (variables dichotomiques et catégorielles).

		Nature de la variable 1 :	
		Qualitative	Quantitative
Nature de la variable 2 :	Qualitative	<ul style="list-style-type: none"> - Tableau croisé des effectifs dans les différentes catégories des variables 1 et 2. - Test d'indépendance du Chi-2 	<ul style="list-style-type: none"> - Moyennes de la variable 1 selon les catégories de la variable 2 - Test de comparaison de moyennes de Student
	Quantitative	<ul style="list-style-type: none"> - Moyennes de la variable 2 selon les catégories de la variable 1 - Test de comparaison des moyennes de Student 	<ul style="list-style-type: none"> - Coefficient de corrélation ; graphique croisé - Test de Student à partir de la régression simple

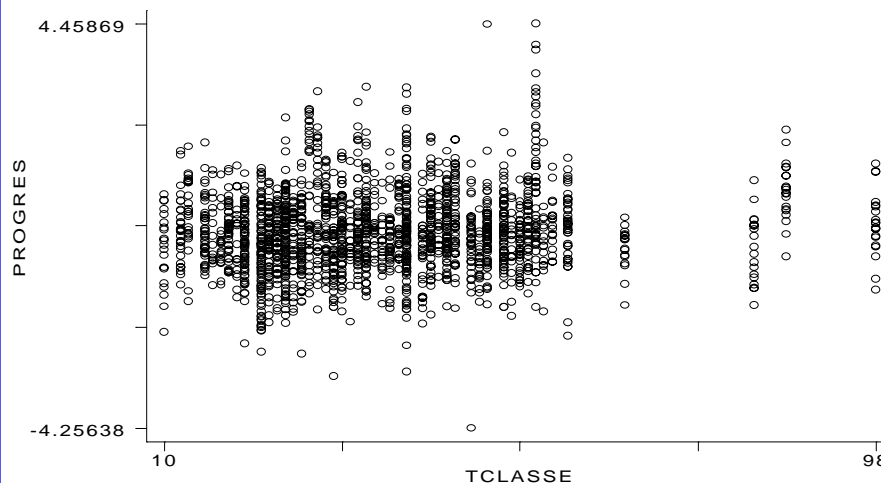
Envisageons successivement les trois cas ainsi définis et détaillons les outils utilisés :



• Deux variables quantitatives :

Le premier réflexe est sans doute de représenter graphiquement la relation entre les deux variables.

Prenons l'exemple des progrès des élèves en fonction de la taille de la classe en Côte d'Ivoire en 5^{ème} année :



Le graphique donne un nuage de points indiquant pour chaque élève sa progression en fonction de la taille de sa classe. On constate d'abord que, pour une même taille de classe, les progressions des élèves varient fortement, ce qui montre que la relation n'est pas très étroite. On peut mentalement essayer, cependant, de tracer une droite qui résume le nuage de points. Elle semble avoir une pente légèrement positive.

Lorsque la relation entre les deux variables est linéaire (c'est-à-dire lorsqu'on pense que la résumer graphiquement par une droite est adéquat), un indicateur de l'**intensité de la relation linéaire** est le **coefficient de corrélation linéaire**. Cet indicateur d'intensité de la relation prend des valeurs entre -1 et 1 ; lorsque l'intensité de la relation est élevée, le coefficient de corrélation linéaire est élevé en valeur absolue (proche de -1 ou de $+1$) ; plus l'intensité de la relation est faible, plus il est proche de 0 . Le signe du coefficient de corrélation indique simplement si les variables varient en même sens ou en sens contraire.

Par exemple, le coefficient de corrélation des progrès des élèves et de la taille de la classe est de 0,1 ; indiquant que les variables sont faiblement mais positivement corrélées.

Attention, le coefficient de corrélation n'est pas un outil de prédiction, c'est-à-dire qu'il ne permet pas de savoir si une des deux variables "cause" la variation de l'autre. Dans notre exemple, il permet juste de dire : *Les élèves des classes plus nombreuses se trouvent aussi être en général des élèves qui progressent plus vite, sans que la relation ne soit très prononcée.*

• **Une variable qualitative et une variable quantitative :**

Un résumé utile est donné par un tableau qui compare les moyennes de la variable quantitative en fonction de la variable qualitative.

Par exemple, on crée la variable qualitative correspondant aux catégories suivantes en Côte d'Ivoire : la classe a plus de 35 élèves (T35PL=1) et la classe a moins de 35 élèves (T35PL=0). Pour étudier sa relation avec la variable de progression des élèves, on peut donner le tableau suivant :

	<i>Nombre d'élèves</i>	<i>Progression moyenne</i>
<i>Classe de moins de 35 élèves</i>	1081	-0,08
<i>Classe de plus de 35 élèves</i>	938	0,04

*En moyenne, les élèves des classes de moins de 35 élèves progressent moins que les élèves des classes de plus de 35 élèves. (Bien sûr, cela ne veut pas dire qu'ils progressent moins **parce qu'ils** sont dans des classes moins nombreuses !)*

Cette différence de 0,12 points est-elle significative ? Autrement dit, avec quel degré de certitude peut-on, à partir de notre échantillon, affirmer que les élèves de classes moins nombreuses progressent moins vite¹ ?

Le test utilisé est le test de comparaison de deux moyennes, qui recourt à la statistique de Student. Voici la façon dont ce test est fourni par Stata :

¹ Cette question de la significativité est reprise plus en détail dans un encadré à l'étape 8.

Two-sample t test with equal variances					0: Number of obs =	1081
					1: Number of obs =	938
Variable	Mean	Std. Err.	t	P> t	[95% Conf. Interval]	
0	-.082058	.0257154	-3.19101	0.0015	-.1325158	-.0316002
1	.0373272	.029003	1.28701	0.1984	-.0195911	.0942456
diff	-.1193852	.0386263	-3.09078	0.0020	-.1951368	-.0436337
Degrees of freedom: 2017						
Ho: mean(0) - mean(1) = diff = 0						
Ha: diff < 0		Ha: diff ~= 0		Ha: diff > 0		
t = -3.0908		t = -3.0908		t = -3.0908		
P < t = 0.0010		P > t = 0.0020		P > t = 0.9990		

Le résultat se lit de la façon suivante : au vu des données de notre échantillon, la probabilité de se tromper en affirmant que les moyennes des deux groupes sont différentes est de deux pour mille (0,0020), et on peut affirmer avec 95% de chances de ne pas se tromper que la différence se situe entre -0,195 et -0,044 au désavantage des élèves des classes de moins de 35 élèves.

Deux variables qualitatives :

Lorsque les deux variables sont qualitatives, il est possible de décrire leur distribution jointe dans un seul tableau. On utilise alors un test du chi2 pour vérifier si les deux variables sont ou non indépendantes.

Prenons par exemple la question suivante : la taille des classes est elle indépendante de l'environnement géographique, rural ou urbain ? L'unité d'observation, cette fois, est la classe. On peut résumer ainsi la distribution croisée :

RURAL	T35PL		Total
	0	1	
0	14	39	53
	26.42	73.58	100.00
	21.54	75.00	45.30
1	51	13	64
	79.69	20.31	100.00
	78.46	25.00	54.70
Total	65	52	117
	55.56	44.44	100.00
	100.00	100.00	100.00

Pearson chi2(1)=33.3219 **Pr=0.00**

Les caractères en gras donnent le nombre de classes dans chacune des quatre catégories. Par exemple, sur 117 classes de l'échantillon, 51 sont des classes rurales avec moins de 35 élèves. Les deux lignes suivantes donnent dans chaque case les pourcentages en ligne et en colonne. Par exemple, en lisant la troisième ligne de chaque case, on constate que les classes de moins de 35 élèves sont à 21,54% en milieu urbain (RURAL=0), et à 78,46% en milieu rural (RURAL=1) ; alors que les classes de plus de 35 élèves sont à 75% en milieu en milieu urbain et à 25 % en milieu rural.

Si les deux distributions étaient indépendantes, on devrait avoir le même pourcentage de classes urbaines et rurales dans les deux colonnes, que la classe soit nombreuse ou non, et ce devraient être les pourcentages moyens (respectivement 45,30% et 54,70%).

C'est cela qu'utilise le test du Chi2 : avec quel degré de certitude la distribution observée nous permet-elle d'affirmer que les deux variables sont ou non indépendantes ?

La réponse est donnée par la probabilité limite : on a moins d'une chance sur mille ($Pr=0,000$) de se tromper en affirmant que les deux variables ne sont pas indépendantes.

Remarque : Une condition de validité du test du Chi2 est que les effectifs dans chaque case soient au moins égaux à 5 individus. Cette condition est ici bien vérifiée.

C. L'analyse causale

La question de la causalité est aussi redoutable qu'incontournable.

On peut penser la causalité par rapport l'expérimentation telle que la pratiquent les sciences expérimentales : on répète la même expérience plusieurs fois, mais en modifiant seulement une des conditions de l'expérience. Si on constate que les résultats varient systématiquement en fonction de cette condition, on infère que cette condition a un effet causal donné sur les résultats.

En sciences sociales, l'expérimentation n'est guère possible. C'est par l'analyse comparative de situations existantes qu'on essaie de mettre en évidence des relations causales.

Même si expérimentation et analyse comparative apparaissent fondamentalement différentes, il reste intéressant de penser l'analyse comparative par

rapport à l'expérimentation idéale. En effet, l'analyse comparative fait un double effort pour se rapprocher des conditions de l'expérimentation : le choix d'un **échantillon** comprenant une multiplicité d'individus fait écho à la **répétition** de l'expérience ; et la reconstitution de conditions «**toutes choses égales par ailleurs**» est l'analogue du **contrôle des conditions** de l'expérimentation.

Prenons l'exemple de la taille de la classe. L'expérimentation idéale, ce serait le clonage ! On prendrait une classe de 30 élèves. On dupliquerait deux fois ces trente élèves de façon à former deux classes, l'une de 30 et l'autre de 60 élèves, comportant strictement les mêmes individus. On dupliquerait également le maître et toutes les conditions de scolarisation. Il ne resterait plus qu'à laisser travailler les deux classes pendant un an, et de comparer les résultats finaux pour mesurer l'effet du doublement de la taille d'une classe.

*A l'autre extrême, la **comparaison naïve** consisterait à trouver dans la réalité des classes de 30 et des classes de 60 élèves, et à comparer les progrès des élèves dans ces classes. Mais on se heurterait à une cascade d'objections qui tournent toutes autour de la même idée : ces classes de 30 et de 60 élèves sont-elles comparables ? En particulier, sont-elles également dotées en maîtres expérimentés et bien formés, leurs élèves bénéficient-ils du même environnement socio-culturel ? — toutes choses qui, on le pressent, ont des effets sur la progression des élèves. En d'autres termes, la comparaison simple (dont nous avons vu que la description bivariée est un exemple) se heurte à l'objection des «effets de contexte»¹ ou de «l'erreur de la variable cachée»².*

*La troisième solution consiste à mener la **comparaison tout en essayant de mesurer les effets de contexte, afin de contrôler leurs effets et d'isoler ce qui correspond à la relation directe entre les deux variables**. Par exemple, dans la comparaison des classes nombreuses et peu nombreuses on va constater que ces classes diffèrent par leur localisation, par leur public d'élèves, par les moyens éducatifs mis en œuvre. On va mesurer les effets de ces différences. On va ôter ces effets de la différence de résultats mesurée entre classes nombreuses et classes peu nombreuses. La différence restante sera attribuée à la taille de classe.*

¹ Pour reprendre la terminologie de Durkheim dans *Les règles de la méthode sociologique*.

² Pour reprendre la terminologie économétrique qui parle de «biais de la variable manquante» ou «biais de la variable cachée».

C'est ce troisième raisonnement qui sous-tend la régression multiple : « Une fois que j'ai tenu compte des effets de localisation, de publics d'élèves et de moyens éducatifs, je constate une différence de résultats entre classes nombreuses et peu nombreuses. Cette différence semble donc directement liée à la différence de taille de classe. C'est en ce sens que je parle d'effet net, et que je le considère comme une approche de l'effet causal des différences de taille de classe. »

Ce type de raisonnement acquis, il faut voir comment la régression multiple le met en œuvre. Il y a en effet une difficulté pratique : pour mesurer l'effet causal de la différence de tailles de classe, je dois connaître l'effet causal des autres variables. Mais la réciproque est vraie : pour connaître l'effet causal des autres variables, je dois connaître l'effet causal des différences de taille de classe. Prenons par exemple l'effet causal du milieu rural : je ne peux le déduire de la comparaison simple des résultats en milieu urbain et en milieu rural. Car on pourrait objecter que les classes de milieu rural sont aussi en général moins nombreuses, et que cela trouble la comparaison. Il y a donc un problème de circularité si on veut travailler de façon récursive. C'est pourquoi la régression multivariée identifie de façon simultanée les différents effets :

L'identification des effets causaux dans la régression multivariée

Il existe de multiples façons de présenter la régression multivariée. Mais pour comprendre ce qui se passe, le mieux est de partir d'un modèle très simple :

On veut mesurer l'effet des classes de plus de 35 élèves sur la progression des élèves. On sait cependant que les classes de plus de 35 élèves sont concentrées en milieu urbain, car l'analyse bivariée l'a montré. On ne veut pas que l'effet du milieu environnant interfère avec notre mesure de l'effet des classes nombreuses.

La régression multivariée consiste à effectuer la modélisation suivante : le progrès des élèves est expliqué par la somme des effets de deux facteurs, celui du milieu rural et celui des classes de plus de 35 élèves. Soit sous forme mathématique :